Sigrid Leithe

Statistical Methods for the Analysis of Data with a Lower Limit of Detection

A Simulation Study with Application to two Datasets from Medicine

June 2019

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Master's thesis

2019

Sigrid Leithe

Master's thesis

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of
Science and Technology

Statistical Methods for the Analysis of Data with a Lower Limit of Detection

A Simulation Study with Application to two Datasets from Medicine

Sigrid Leithe

Master of Science in Physics and Mathematics
Submission date: June 2019
Supervisor: Turid Follestad, Unit of Applied Clinical Research
Co-supervisor: Sara Martino, Department of Mathematical Sciences

Norwegian University of Science and Technology
Department of Mathematical Sciences

Summary

In this thesis, we address the problem of analyzing data measured on a continuous scale with a lower limit of detection and zero inflation, for univariate, bivariate, and longitudinal data. We specify a censored two-part mixture model, as proposed by Moulton and Halsey (1995). The model consists of one discrete part representing the proportion of the sample with zero values, and one continuous part for the magnitude of the response. This thesis provides a detailed evaluation using simulations of the two-part model with interval censoring compared to its simpler variants, the Tobit model and the uncensored two-part model, as well as naive substitution of the censored observations with half the detection limit. We simulate data scenarios with varying detection limits, parameter values, and proportions of zeroes. The three simpler models resulted in misleading parameter estimates as their assumptions were violated, but also the censored two-part model was inappropriate in some scenarios due to over-parameterization.

The four candidate models are applied to two datasets: (1) Borrelia antibody concentrations in Sør-Trøndelag, and (2) data on cytokine concentrations in pregnant women with different autoimmune rheumatic diseases. The cluster structure of the data due to repeated measurements in the latter application is accounted for by including random effects in both parts of the model.

In the former application, the two-part model with interval censoring is demonstrated to work well for estimating the prevalence of borrelia infections, but the high amount of uncertainty due to a low number of uncensored observations makes the simpler logistic regression more feasible in this particular case. In the second application, three cytokines with different proportions of censored samples are analyzed. For all three, the binary mixture models are found to be superior to the one-part models. With the two-part models, significant differences in the time profiles between the diagnostic groups were found in two of the cytokines.

In a search for multivariate methods for analysis of the cytokine data, we specified three bivariate models; A bivariate Tobit model, a two-part mixture model, and a four-part mixture model. The two first-mentioned have shown promise in other applications, but none of the models were suitable for the problem at hand.

Sammendrag

I denne oppgaven tar vi opp problemet med å analysere data målt på en kontinuerlig skala med en nedre deteksjonsgrense og null-inflasjon, for univariate, bivariate og longitudinelle data. Vi spesifiserer en sensurert binær blandingsmodell, som foreslått av Moulton og Halsey (1995). Modellen består av en diskret del som representerer andelen nuller i utvalget, og en kontinuerlig del for størrelsen på de positive responsene. Denne oppgaven gir en detaljert evaluering av den binære blandingsmodellen med intervallsensurering, sammenlignet med dens enklere varianter, Tobitmodellen og den usensurerte binære blandingsmodellen, samt å naivt bytte ut de sensurerte observasjonene med halvparten av deteksjonsgrensen. Vi simulerer data med varierende deteksjonsgrenser, parameterverdier og mengde nuller. De tre enklere modellene gir misvisende resultater ettersom de underliggende antakelsene brytes. Også den binære blandingsmodellen med intervallsensurering er upassende i enkelte scenario grunnet overparametrisering.

De fire kandidatmodellene blir anvendt på to datasett: (1) Konsentrasjoner av borrelioseantistoff i Sør-Trøndelag, og (2) data med cytokinkonsentrasjoner hos gravide kvinner med ulike autoimmune revmatiske sykdommer. Klyngestrukturen i de sistnevnte dataene grunnet repeterte målinger blir tatt hånd om ved å inkludere tilfeldige effekter i begge delene av blandingsmodellen.

I den første anvendelsen viser den binære blandingsmodellen med intervallsensurering seg å fungere godt til å estimere prevalensen av borrelioseinfeksjoner, men et lavt antall usensurerte observasjoner gjør at estimeringene blir svært usikre. Derfor kan enkel logistisk regresjon sies å være mer praktisk. I den andre anvendelsen ble tre cytokiner med ulik andel sensurerte observasjoner analysert. For alle tre viste de binære blandingsmodellene seg å være overlegne endelsmodellene. Signifikante forskjeller i tidsprofilene mellom diagnosene ble funnet i to av cytokinene.

I søket etter multivariate metoder for å analysere cytokindataene, ble tre bivariate modeller undersøkt; En bivariat Tobitmodell, en binær blandingsmodell og en firedels blandingsmodell. De to førstnevnte har vist lovende resultater i andre anvendelser, men ingen av modellene var adekvate for de aktuelle dataene.

Preface

This thesis is written as a Masters degree in Industrial Mathematics at the Norwegian University of Science and Technology, Department of Mathematical Sciences, during the spring semester of 2019. Some previous work was conducted in a project thesis during the autumn semester of 2018, thus some parts are based on or inspired by work from this project. These are Section 2.2, Section 3.1.1 - 3.1.3, Section 3.2, and Section 3.4 in Chapter 3, and Section 7.1 and some of the analysis in Section 7.2.1 in Chapter 7.

I would like to thank my supervisors, Associate Professor Turid Follestad at the Unit of Applied Clinical Research, NTNU, and Associate Professor Sara Martino at the Department of Mathematical Sciences, NTNU, for excellent guidance. I would also like to thank Mona Høysæter Fenstad at the Department of Immunology and Transfusion Medicine, St. Olavs Hospital, for the great help with understanding multiplex assays and for giving me the opportunity to work with the cytokine data. For the chance to work on the borrelia data, I would like to thank Jan Egil Afset at the Department of Medical Microbiology, St. Olavs Hospital. The data was collected as part of a medical undergraduate thesis by Marie M. Holt and Rino Eriksen in 2018.

Trondheim, June 2019
Sigrid Leithe

Table of Contents

Summary i

Sammendrag ii

Preface iii

Table of Contents v

Abbreviations vii

1 Introduction 1

2 Background 3
2.1 Borrelia Antibody Concentration across Regions . . . 3
2.1.1 Lyme Borreliosis and its Prevalence . . . 3
2.1.2 Enzyme-Linked Immunosorbent Assays (ELISA) . . . 4
2.2 Cytokine Concentrations during Pregnancies . . . 5
2.2.1 Autoimmune Rheumatic Diseases and Pregnancy . . . 5
2.2.2 Multiplex Assays . . . 5

3 Statistical Methods 7
3.1 Statistical Models for Data Subject to a Lower Limit of Detection . 7
3.1.1 Tobit Models . . . 7
3.1.2 Two-Part Models . . . 9
3.1.3 Two-Part Models with Interval Censoring . . . 10
3.1.4 Substitute Model . . . 11
3.1.5 Marginalized Parameterization . . . 12
3.1.6 Choice of Continuous Distribution and Link Function . . . 13
3.2 Expansions to Longitudinal Data . . . 15
3.2.1 Including Time in the Models . . . 17
3.2.2 Prediction . . . 18
3.3 Expansions to Bivariate Data . . . 19
3.3.1 Tobit Models . . . 19
3.3.2 Two-Part Models . . . 21

3.3.3 Four-Part Models . . . 23
3.4 Model Estimation . . . 25
3.4.1 Numerical Methods . . . 25
3.4.2 Estimation of Standard Errors . . . 27
3.4.3 Likelihood Ratio Tests . . . 28
3.5 Model Evaluation . . . 30
3.5.1 Model Selection Criteria . . . 30
3.5.2 Prediction and Scoring Rules . . . 32
3.6 Monte Carlo Simulation Studies . . . 33

4 Design of Simulation Study **41**
4.1 Models . . . 41
4.2 Simulated Datasets . . . 42

5 Results from Simulation Study **49**
5.1 Exploration and Visualization of Results . . . 49
5.2 Inference . . . 53
5.2.1 Two-Part Model w/ Interval Censoring . . . 55
5.2.2 Two-Part Model . . . 57
5.2.3 Tobit Model . . . 58
5.2.4 Substitute Model . . . 59
5.2.5 Concluding Remarks . . . 59
5.3 Prediction . . . 60
5.4 Discrete and Continuous Effect in Opposite Directions . . . 60
5.5 Multiple Covariates . . . 63

6 Application to Borrelia Data **67**
6.1 Description of the Data . . . 67
6.2 Statistical Analysis . . . 68

7 Application to Cytokine Data **75**
7.1 Description of the Data . . . 75
7.2 Longitudinal Statistical Analysis . . . 82
7.2.1 Cytokine TNF-α . . . 84
7.2.2 Cytokine MCP1 . . . 92
7.2.3 Cytokine IL8 . . . 95
7.3 Bivariate Statistical Analysis . . . 98

8 Conclusion and Further Work **103**

References **105**

Appendix **111**
A Simulation Study - R Code . . . 111
B Longitudinal Analysis - SAS code . . . 125
C Bivariate Analysis - R code . . . 131
D Application to Cytokine Data - Results . . . 136

Abbreviations

CRPS	=	Continuous ranked probability score
ELISA	=	Enzyme-linked immunosorbant assay
LOD	=	Limit of detection
LRT	=	Likelihood ratio test
MC	=	Monte Carlo
MCSE	=	Monte Carlo standard error
MTP	=	Marginalized two-part
TP	=	Two-part
TPIC	=	Two-part with interval censoring

1 Introduction

It is a common phenomenon in medical research to come across data characterized by a point mass at zero and a continuous distribution of positive responses. The concentration of observations at zero could arise from several circumstances. Excluding the possibility that the point mass is caused by some technical error, the observations in the point mass are typically either true zeroes or low values indistinguishable from zero. A classic example of the former is data on the occurrence of some phenomenon over a specific period of time, such as alcohol consumption (Liu et al., 2016), symptom severity (Mahmud et al., 2010; Xing et al., 2017), and medical expenditures (Smith et al., 2014). In these cases, the phenomenon in question truly has not occurred for parts of the population, while the rest has a positive score. Alternatively, the measurements are subject to a lower limit of detection (LOD), and parts or all of the observations recorded as zero would be positive if the measurements were more sensitive. This is typical for measurements of concentration, such as viral loads (Dagne and Huang, 2015; Su and Luo, 2017), cytokine levels (Bernhardt, 2018), and antibody concentrations (Moulton and Halsey, 1995).

In this thesis, the focus will primarily lie on the second category, where the true values of the observations recorded as zero are not known, but they are determined to be below a certain limit. This is motivated by two applications with data on this form:

(1) Estimation of the prevalence of Lyme disease from borrelia antibody concentrations in blood sera.

(2) Comparison of cytokine levels in women with autoimmune rheumatic diseases and healthy controls, during pregnancies.

The first application is conducted on a dataset with 981 measurements of antibody concentrations in blood sera from former Sør-Trøndelag county that was collected as part of a medical undergraduate research thesis by Holt and Eriksen (2018). The concentrations were measured with Enzyme-Linked Immunosorbent Assays (ELISA) and subject to a lower detection limit. The prevalence of Lyme disease is increasing in many countries, including Norway (Jore et al., 2011). In order to measure possible changes in the rate of borreliosis infections, it is of importance to determine the present prevalence in the population, which is currently not known.

The second application is concerned with autoimmune rheumatic diseases. It is well established that these diseases are affected differently by pregnancies. The

reasons for this are not fully understood, but some of the findings have been linked to changes in cytokine levels (Swain and Jena, 2016). It is therefore hypothesized that cytokines may be used in targeted treatment of autoimmune rheumatic diseases. For the purpose of examining this link, cytokine levels are measured in 75 pregnant women, 56 of which have an autoimmune rheumatic diagnosis, at different time points before, during, and after pregnancy in a study conducted at NTNU. The cytokine concentrations are measured with multiplex assays and subject to lower detection limits.

There is a considerable amount of literature proposing different methods for dealing with the problem of detection limits. Replacing the recorded zeroes with a substitute value, such as half the detection limit, may work well when the proportion of censored observations is low. For larger proportions, it is known to cause bias in estimates and predictions (Hornung and Reed, 1990).

Tobin (1958) suggested assuming an underlying continuous distribution whose values below the detection limit are considered unobserved. This model is shown to work well when the proportion of censored observations is close to what is expected from the distribution of the non-censored observations. This constraint was relaxed by Cragg (1971), which proposed a two-part mixture model where the probability of falling below the limit of detection and the magnitude of the non-censored response are determined by two separate processes. Moulton and Halsey (1995) expanded the model further by explicitly allowing for the probability that some of the censored observations are the results of interval censoring from the continuous distribution. Mixed effects were included in both model parts by Berk and Lachenbruch (2002) to account for the correlation between measurements from the same individual commonly present in longitudinal studies.

In the bivariate setting, Lyles et al. (2001) introduced a censored model, analogous to the model by Tobin (1958) in the univariate setting, based on the assumption that all the data has originated from a latent bivariate continuous distribution. This model was generalized by Chu et al. (2005), which proposed adding a sub-LOD component to relax the restrictions on the proportion of both variables falling below the LOD.

We start this thesis by introducing some relevant background in Chapter 2 on the prevalence of Lyme disease and the link between cytokines and autoimmune rheumatic diseases in pregnancies. In this chapter, we also describe the technologies used to conduct the measurements giving rise to the detection limits. Then, in Chapter 3, the statistical methods used in the thesis are presented. First, we specify the statistical models with expansions to longitudinal and bivariate data. This is followed by methods for estimating the model parameters and evaluating the model performance. Lastly, Monte Carlo simulation studies are introduced.

Chapter 4 and 5 are dedicated to the design and results of a simulation study that examines the performance of a set of candidate models commonly used to handle lower detection limits on data with various properties. The findings are utilized in the analysis of the borrelia data in Chapter 6, and both longitudinal and bivariate analysis of the cytokine data in Chapter 7. Finally, in Chapter 8 we offer a conclusion and point out possible directions for further work on the topic.

2 Background

In this Chapter, we present some relevant background on the two applications motivating this thesis, namely estimating the prevalence of Lyme borreliosis in Sør-Trøndelag and the time profiles of cytokine levels in pregnant women with autoimmune rheumatic diseases. We also described the technologies used to conduct the measurements to gain an understanding of why the data is subject to lower detection limits.

2.1 Borrelia Antibody Concentration across Regions

2.1.1 Lyme Borreliosis and its Prevalence

Lyme borreliosis, also known as Lyme disease, is the most common tick-borne disease in Europe and North-Amerika. It is caused by an infection of bacteria from the group *Borrelia burdorferi* sensu lato, which consists of multiple species. Hereafter this group will be referred to as *B. burdorferi*. The disease transmits to humans or other animal hosts when unfed flat ticks attach to the skin of the host and inject saliva during feeding. Usually, a feeding period of at least 36 hours is needed for transmission to occur, but it can be more rapid (Stanek et al., 2012).

The most common manifestation of Lyme disease is *erythema migrans*, a red rash at the site of the tick bite that eventually resolves, even without treatment. The infection can, however, spread to other tissues and organs, and may cause more severe reactions involving the skin, nervous system, joints, or heart. Since there are currently no vaccines against *B. burdorferi*, the best way to prevent infection is to avoid walking through vegetation with exposed skin and checking the skin for ticks regularly.

The prevalence of Lyme disease is increasing in many countries, including Norway. Jore et al. (2011) compared the then prevalence of ticks in Norway with historical data from 1983 and 1943, and found clear evidence that ticks are now present at higher altitudes and latitudes than before. The reasons for this are debated, but many ascribe the shifts to climate changes. In order to measure possible changes in the rate of borreliosis infections, it is of importance to determine the current prevalence in the population. Since only cases of disseminated Lyme disease are reported, the exact prevalence in Norway is not currently known. Some studies have estimated the prevalence in certain regions of Norway, but only one

study known to us have included Sør-Trøndelag county. Vestrheim et al. (2016) measured the concentration of Borrelia IgG antibodies in blood serum using two different types of immuno assays, and found the prevalence in Sør Trøndelag to be $3.9\%(95\ \%\ \text{CI: } 2.3\% - 6.4\%)$ and $3.7\%(95\ \%\ \text{CI: } 2.4\% - 5.7\%)$. In this study, young people were over-represented. Thus the results may be unreliable.

2.1.2 Enzyme-Linked Immunosorbent Assays (ELISA)

ELISA measures the concentration of substances, such as antibodies and proteins. The analyte is added to a plastic well which is covered with a substance that the analyte of interest binds to, such that the analyte is immobilized. This technology makes it possible to wash the wells to remove unbound material without removing the substance of interest (Crowther, 1995).

The sample is added to a dry well such that the analyte binds to the substance covering the bottom. The unbound materials are removed through washing before detection-antibodies bound to enzymes are added. They attach to the bounded analyte, and again the residue is removed. Lastly, a substrate that reacts with the enzymes in a color reaction is added. The color reaction may be chromogenic, fluorescent, or chemiluminescent. The first mentioned results in visual color, while the last two must be measured with specific instruments. Because the amount of enzymes is directly linked to the quantity of the analyte, the resulting color intensity corresponds to the initial concentration in the sample. A schematic representation of the process is shown in Figure 2.1.

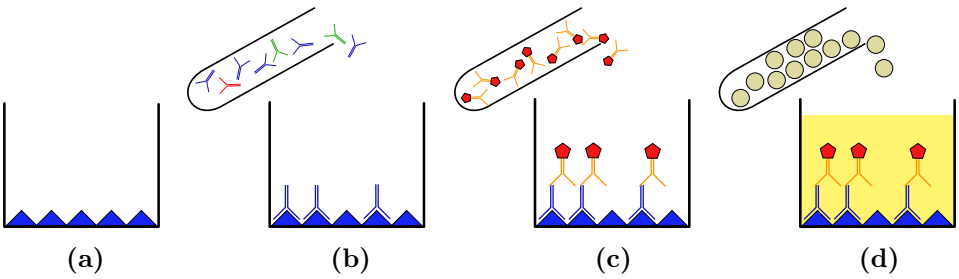

Figure 2.1: Schematic representation of ELISA for measuring the concentration of an antibody. **(a)** Plastic wells covered with analyte-specific antigens. **(b)** The sample is added and the antibodies attach to the antigens. **(c)** ELISA-antibodies with enzymes are added and attach to the antibodies of interest. **(d)** A substrate is added that reacts with the enzymes and gives a color reaction. The resulting color intensity corresponds with the concentration in the sample.

The limit of detection (LOD) is the lowest concentration that can be determined to be significantly different from a blank, i.e. the lowest quantity of the substance that can be distinguished from no substance (Fortunato, 2016). Typically, the LOD is set to two standard deviations above the mean of the blanks, based on multiple measurements of the color intensity of blanks.

2.2 Cytokine Concentrations during Pregnancies

2.2.1 Autoimmune Rheumatic Diseases and Pregnancy

Some autoimmune rheumatic diseases are known to be affected by pregnancies. Among these are systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA), including seronegative rheumatoid arthritis (SN-RA).

SLE is characterized by reoccurring inflammations in connective tissue. In particular, patients with SLE are prone to kidney failure. The disease is predominantly seen in females, which typically have a first onset between puberty and menopause (Mok and Lau, 2003). Rheumatoid arthritis is characterized by inflammation and swelling in the joints of fingers and toes. If left untreated, the disease can cause bone deformation and destruction (McInnes and Schett, 2011). Patients with seropositive RA has a presence of autoantibodies in the blood serum, that is not found in patients with SN-RA. The two groups are often treated as one, but recent studies have suggested that the seronegative status is associated with a different prognosis of the disease (Ajeganova and Huizinga, 2015).

Numerous studies have shown a higher risk of flare in SLE patients during pregnancy, compared to non-pregnant SLE-patients. This coincides with an increased rate of premature delivery and fetal loss (Gordon, 2004). On the other hand, approximately 75% of patients with RA experience some degree of improvement during pregnancy. Over 50% improve as early as the first trimester. Some studies have indicated that the probability of improvement is even greater among patients with SN-RA. Only a quarter show no improvement or worsening during the course of the pregnancy. However, most patients who improve, relapse *postpartum* (after birth) (Swain and Jena, 2016).

Cytokines are small proteins that have specific effects on the interaction and communication between cells (Zhang and An, 2007). There are many types of cytokines, and they play a vital role in the outcome of a pregnancy. Studies of pregnant women with RA and SLE have shown some significant differences in cytokine expressions compared to healthy women. This has been linked to activity in the underlying disease. Because of this, it is hypothesized that cytokines can be used in targeted treatment of autoimmune rheumatic diseases (Østensen et al., 2006).

2.2.2 Multiplex Assays

Multiplex assays make use of magnetic beads to measure multiple analytes simultaneously (Gupta et al., 2014). Each bead is given a unique color through mixtures of red and infrared color and is coated with antibodies to target a specific cytokine. When a sample is added to the mixture of color-coded beads, the cytokines are captured by the antibodies. After a series of washes in order to remove unbound materials, detection antibodies are added. These antibodies attach to the bounded cytokines and form an antibody sandwich around the analytes. Lastly, streptavidin-phycoerythrin (SA-PE) is added and binds to the detection antibodies. A schematic representation is given in Figure 2.2.

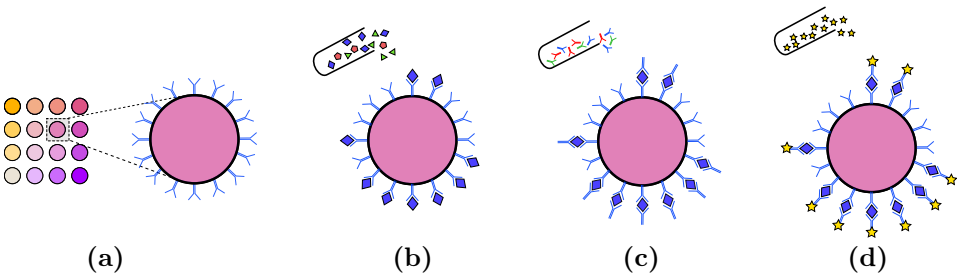

Figure 2.2: Schematic representation of a multiplex assay. **(a)** Uniquely colored beads coated with different analyte-specific antibodies. **(b)** The sample is added and the cytokines attach to the antibodies. **(c)** Detection antibodies are added and form a sandwich of antibodies around the cytokines. **(d)** SA-PE is added and binds to the detection antibodies. The figure is inspired by Bio-Rad Laboratories, Inc. (nd).

The mixture is analyzed in an instrument with two lasers. One detects the color of the bead, and thus which cytokine is analyzed. The other laser measures the fluorescence of the SA-PE, which in turn is used to determine the concentration of the cytokine. The concentration is determined from the fluorescence intensity through standard curves, which are calculated for each cytokine by a set of standard samples with known concentrations. An example of a standard curve is provided in Figure 2.3.

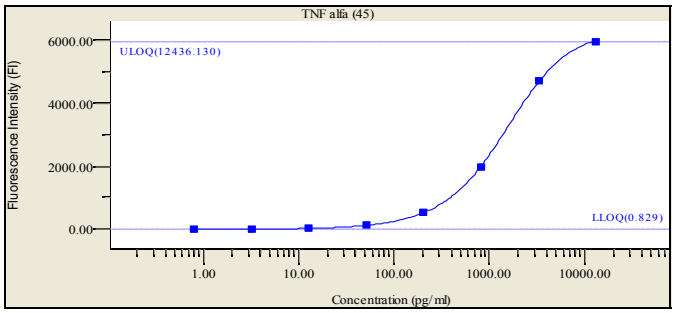

Figure 2.3: Example of standard curve for TNF-α. Each point represents a measurement whit known concentration, and the curve is obtained with logistic interpolation.

The measurements are subject to a lower limit of detection. The limit of detection (LOD) is determined by adding two standard deviations to the average of the median fluorescence intensity for ten replicates of the standard curve blank (Gupta et al., 2010). This means that observations that are recorded as below the LOD are considered to not be significantly different from zero. In addition, some of the lower concentrations are extrapolated from the standard curve, and thus are subject to greater uncertainty.

3 Statistical Methods

In this Chapter, statistical methods for modeling and model evaluation of data subject to a lower limit of detection. First, statistical models for data subject to a detection limit are specified in Section 3.1, followed by expansions to longitudinal data in Section 3.2 and to bivariate data in Section 3.3. In Section 3.4, methods for estimating the model parameters in a frequentist framework are presented, and in Section 3.5 methods for evaluating the performance of the models are described. Lastly, Monte Carlo simulation studies are introduced in Section 3.6.

3.1 Statistical Models for Data Subject to a Lower Limit of Detection

The semicontinuous data considered in this thesis is characterized by a point mass at zero and a continuous distribution above a detection limit. There exists a wide variety of methods for modeling data on this form, some of which will be presented in this section. Two of the specified models are binary mixture models with one continuous and one discrete part. We finish off with presenting an alternative way of parameterizing these two-part models and a discussion of some common choices of link functions for the discrete part and continuous distributions for the continuous part.

3.1.1 Tobit Models

The Tobit model (Tobin, 1958) treats the observations below the LOD as latent continuous observations that have been left-censored. This can be formulated as

$$y_i = \begin{cases} 0, & y_i^* \leq T \\ y_i^*, & y_i^* > T \end{cases},$$

where T is the LOD, and y_i is the observed concentration. The latent value y_i^* comes from a continuous parametric distribution $f(y^*)$ with positive support. This model assumes that all observations come from the same underlying distribution.

Let I_i indicate whether observation y_i is censored, i.e.

$$I_i(y_i) = \begin{cases} 1, & y_i = 0 \\ 0, & y_i > T \end{cases}. \tag{3.1}$$

The density of the observed value y_i can be written as

$$g(y_i) = F(T)^{I_i} f(y_i)^{(I_i - 1)}, \tag{3.2}$$

where $f(\cdot)$ is the continuous distribution of the latent variable with corresponding cumulative density function $F(\cdot)$. Thus, the probability of observing a zero is given by

$$P(Y_i = 0) = F(T).$$

A conceptual illustration of the latent and observed distribution is provided in Figure 3.1.

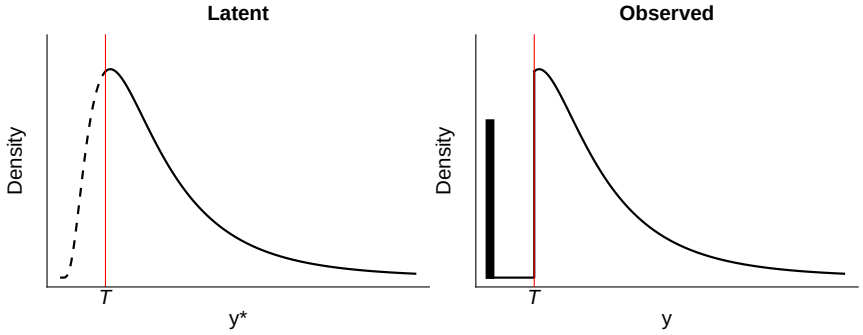

Figure 3.1: Conceptual illustration of the assumed latent distribution of y_i^* and the resulting observed values y_i under the Tobit model. The latent distribution is continuous, whereas the observed values follow a left truncated continuous distribution and a point mass at zero with weight $F(T)$.

The Lognormal Tobit Model

A typical choice for the latent distribution is a the lognormal distribution,

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}} \tag{3.3}$$

Covariates can be introduced to the mean of the distribution by letting

$$\mu_i = \mathbf{z}_i' \boldsymbol{\gamma}, \tag{3.4}$$

where $\mathbf{z}_i$ is a vector of covariates and $\boldsymbol{\gamma}$ is a vector of the corresponding fixed effects.

In this case the mean of the response Y_i is given by

$$\begin{aligned} E(Y_i) &= P(Y_i > 0)E[Y_i|Y_i > 0] = P(Y_i > 0)E[Y_i|Y_i > T] \\ &= (1 - F(T; \mu_i, \sigma))E[Y_i|Y_i > T] \\ &= \left[1 - \mathbf{\Phi}\left(\frac{\ln(T) - \mathbf{z}_i'\boldsymbol{\gamma}}{\sigma} - \sigma\right)\right] e^{\mathbf{z}_i'\boldsymbol{\gamma} + \sigma^2/2}. \end{aligned}$$

3.1.2 Two-Part Models

The Tobit model is shown to work well when the proportion of censored observations is close to what is expected from the continuous distribution. This constraint can be relaxed by assuming that the discrete and continuous data arises from two distinct stochastic processes, which is commonly done by introducing a point mass distribution below the LOD (Cragg, 1971). The resulting model is a binary mixture model with one continuous part representing the positive responses, and one discrete part representing the zeroes. The density can be expressed as

$$g(y_i) = \pi_i^{I_i}[(1 - \pi_i)f(y_i)]^{(1-I_i)} \tag{3.5}$$

where

$$\pi_i = P(Y_i = 0)$$

denotes the probability of observing $Y_i = 0$, $f(\cdot)$ is the probability density function of the continuous data, and I_i is the indicator variable defined in (3.1). A conceptual illustration is provided in Figure 3.2.

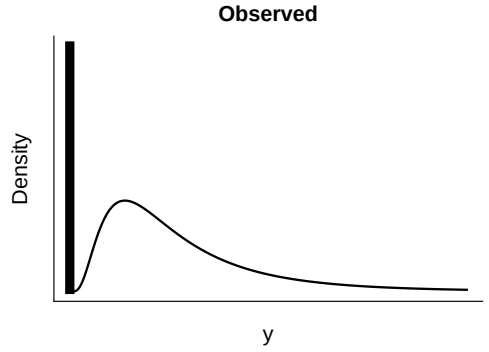

Figure 3.2: Conceptual illustration of the assumed distribution of y in the two-part model.

The Probit/Lognormal Mixture Model

A typical choice for the continuous part $f(y_i)$ is the lognormal distribution (3.3). In order to introduce covariates to the probability π_i a link function must be utilized. Here we will use the probit-link such that

$$\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}).$$

The continuous distribution is assumed to be lognormal with

$$\mu_i = \mathbf{z}_i'\boldsymbol{\gamma}.$$

Here $\mathbf{x}_i$ and $\mathbf{z}_i$ are sets of (possibly distinct) covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of fixed effects. In this parametrization γ_j is the effect of z_{ij} conditional on having a positive response whereas β_j is the effect on the probability of being positive. Thus, the marginal mean of Y_i is

$$E(Y_i) = (1-\pi_i)E(Y_i|Y_i > 0) = (1 - \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}))e^{(\mathbf{z}_i'\boldsymbol{\gamma}+\sigma^2/2)}. \tag{3.6}$$

3.1.3 Two-Part Models with Interval Censoring

In the previously specified two-part model there is no latent distribution, hence $f(\cdot)$ models observed distribution of the positive responses. When the measurements are subject to a detection limit, the distribution of the positive responses becomes truncated. Thus, fitting a log-normal distribution to the positive responses will not be appropriate. Moulton and Halsey (1995) explicitly allowed for interval censoring of the continuous distribution on the interval $[0, T]$ by expanding the model to

$$g(y_i) = [\pi_i + (1-\pi_i)F(T)]^{I_i}[(1-\pi_i)f(y_i)]^{(1-I_i)}, \tag{3.7}$$

where $F(\cdot)$ is the cumulative density function corresponding to $f(\cdot)$. This model distinguishes between a population whose responses follows the distribution $f(\cdot)$, which might result in responses below the LOD, and a separate sub-LOD population regarded as true zeroes. The interpretation of π_i changes from the probability of falling below the LOD to the probability of belonging to the sub-LOD population. Now there are two possible reasons for observing $y_i = 0$. Either y_i belongs to the sub-LOD population such that $y_i^* = 0$, which has probability π_i, or it is a censored realization from the continuous distribution $y_i^* < T$, which has probability $(1-\pi_i)F(T)$. An illustration of the assumed latent distribution of y^* and the resulting observed distribution of y is provided in Figure 3.3.

The two-part model with interval censoring has the two-part model (without interval censoring) defined in (3.5) as its latent distribution. Thus, they are equivalent when there is no detection limit ($T = 0$) and are expected to behave similarly when the LOD is small. Furthermore, it contains the Tobit model as a special case when $\pi_i = 0$.

The Probit/Lognormal Mixture Model

Covariates may be introduced in the same manner as in the two-part model:

$$\begin{aligned} \pi_i &= \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}), \\ \mu_i &= \mathbf{z}_i'\boldsymbol{\gamma}, \end{aligned} \tag{3.8}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are sets of (possibly distinct) covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of fixed effects.

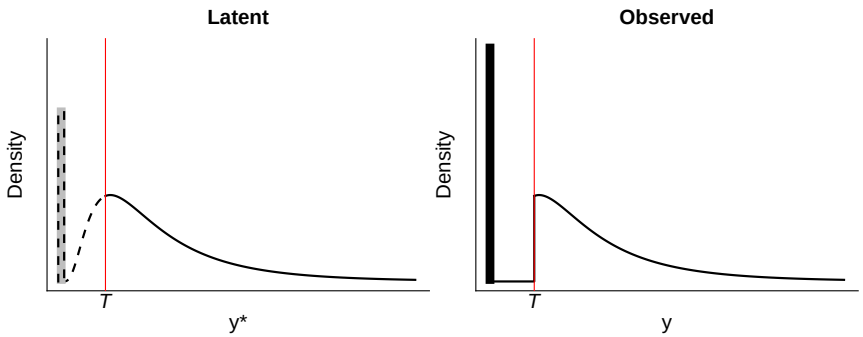

Figure 3.3: Conceptual illustration of the assumed latent distribution of y_i^* and the resulting observed values y_i under the two-part model with interval censoring. The latent distribution is a mixture of a point mass distribution at zero with weight π_i and a continuous distribution $f(y^*)$. The observed distribution is a mixture of point mass distribution with weight $\pi_i + (1-\pi_i)F(T)$ and a truncated continuous distribution $(1-\pi_i)f(y^*)$ on $y > T$.

Calculating the marginal mean of the observed response Y_i is similar as for the Tobit model. In this case the probability of a positive response is $P(Y_i > 0) = (1-\pi_i)(1-F(T))$, which gives

$$\begin{aligned} E(Y_i) &= P(Y_i = 0)E(Y_i|Y_i = 0) + P(Y_i > 0)E(Y_i|Y_i > 0) \\ &= (1-\pi_i)(1-F(T))E(Y_i|Y_i > T) \\ &= (1-\boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}))\left[1-\boldsymbol{\Phi}\left(\frac{\ln(T)-\mathbf{z}_i'\boldsymbol{\gamma}}{\sigma}-\sigma\right)\right]e^{\mathbf{z}_i'\boldsymbol{\gamma}+\sigma^2/2}. \end{aligned} \tag{3.9}$$

3.1.4 Substitute Model

The most naive approach to modeling left-censored data due to a limit of detection, is to replace the censored observations with a substitute value. Common choices are $T/2$ and $T/\sqrt{2}$. The former implicitly assumes that the data below the LOD follows a uniform distribution, while the latter implicitly assumes a triangular shape. Hornung and Reed (1990) showed that $T/\sqrt{2}$ is superior to $T/2$ for estimating the geometric mean and standard deviation of a lognormal distribution, unless the LOD greatly surpasses the mode of the distribution. For larger proportions of zeroes, both of these methods are known to produce bias in estimates and predictions.

Let y_i be the observed data and $S \in (0, T]$ be the chosen substitute value, such that the transformed data $\tilde{y}_i$ is obtained by letting

$$\tilde{y}_i = \begin{cases} S, y_i = 0 \\ y_i, y_i > T \end{cases},$$

Now, the transformed data $\tilde{y}_i$ is assumed to follow a continuous distribution $f(\cdot)$ with positive support. Alternatively, the probability density can be formulated in

terms of the original data, in which case

$$g(y_i) = f(S)^{I_i} f(y_i)^{(1-I_i)} = f(\tilde{y}_i), \tag{3.10}$$

where I_i is the indicator function defined in (3.1).

Lognormal Distribution

After the substitution is performed a linear model is fitted to the log-transformed data without distinguishing between the censored and non-censored observations, such that

$$\ln(Y_i) = \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma).$$

The marginal mean is given by

$$E(Y_i) = \exp(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma^2/2). \tag{3.11}$$

This model is equivalent to the Tobit model when there is no LOD ($T = 0$). Thus, they are expected to behave similarly when the LOD is low.

3.1.5 Marginalized Parameterization

The covariates $\boldsymbol{\gamma}$ in the continuous part of previously specified two-part models must be interpreted conditionally on having observed a positive response. This makes it complicated to interpret the overall effect of a covariate on the marginal mean. For instance, one covariate might increase the probability of observing a positive response, while simultaneously decreasing the expected positive response. In such cases it is unclear how the marginal mean is affected by said covariate. As demonstrated, it is possible to find an expression for the marginal mean $E(Y_i)$ by removing the conditioning on $Y_i > 0$ and transforming back from $\ln(y)$-space to y-space. The resulting marginal effects are however dependent on the values of the remaining covariates and confidence intervals are not easily obtained. For instance, assuming $\mathbf{z}_i' = \mathbf{x}_i'$ the multiplicative marginal effect of the jth covariate on the mean with the standard two-part model (3.6) is given by

$$\frac{E(Y_i|x_{ij} = x+1, \mathbf{x}_{i,-j})}{E(Y_i|x_{ij} = x, \mathbf{x}_{i,-j})} = \frac{1 - \boldsymbol{\Phi}(\mathbf{x}_{i,-j}\boldsymbol{\beta}_{-j} + \beta_j \cdot (x+1))}{1 - \boldsymbol{\Phi}(\mathbf{x}_{i,-j}\boldsymbol{\beta}_{-j} + \beta_j \cdot x)} \exp(x_{ij}\gamma_j). \tag{3.12}$$

Unless $\beta_j = 0$, all fixed effects $\mathbf{x}_i$ must be specified in order to determine the effect of x_{ij}. This is also the case for the standard two-part model with interval censoring (3.9). In applications where the marginal effects are of interest, this might be a substantial drawback.

Smith et al. (2014) proposed a "marginalized" two-part (MTP) model which is parameterized in terms of the marginal mean, thus giving more interpretable effect estimates. The model is written on the same form as the standard two part model,

$$g(y_i) = \pi_i^{I_i}[(1-\pi_i)f(y_i|y_i > 0)]^{(I_i-1)}$$

where $\pi_i = P(Y_i = 0)$ denotes the probability of y_i belonging to the point mass and $f(\cdot)$ is the continuous distribution of the positive responses.

The MTP model differs from the conventional two-part model in how covariates are introduced to the continuous part. In the MTP model they are parametrized in terms of the marginal mean, i.e.

$$E(Y_i) = \exp(\mathbf{z}_i'\boldsymbol{\gamma}). \tag{3.13}$$

The Probit/Lognormal Mixture

As in the standard two part model, covariates are included to the point mass by

$$\pi_i = \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}). \tag{3.14}$$

When using the lognormal distribution for the continuous part we know that the marginal mean is given by

$$E(Y_i) = (1 - \pi_i)\exp(\mu_i + \sigma^2/2).$$

Equating this with (3.13) and solving for μ_i gives

$$\mu_i = \mathbf{z}_i'\boldsymbol{\gamma} - \sigma^2/2 - \ln(1 - \pi_i)$$

as the resulting parameter of the lognormal distribution.

For simplicity, assume $\mathbf{z}_i' = \mathbf{x}_i'$. From (3.13), the marginal effect of x_{ij} in the MTP model is simply given by

$$\frac{E(Y_i|x_{ij} = x + 1, \mathbf{x}_{i,-j})}{E(Y_i|x_{ij} = x, \mathbf{x}_{i,-j})} = \exp(\gamma_j).$$

This is the same as in the substitute model (3.11), which makes the parameters in $\boldsymbol{\gamma}$ directly comparable across the two models. Unlike for the standard two part models, a confidence interval for the marginal effect can easily be obtained by plugging in the confidence limits of γ_j.

3.1.6 Choice of Continuous Distribution and Link Function

So far the models have been presented using a lognormal distribution for the continuous distributions, and a probit-link for the discrete parts of the two-part models. There are however countless other options, a few of which will be discussed here.

Link Functions

A link function defines the relationship between the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ and the mean of the distribution. The discrete part of the two-part models represents a Bernoulli process where each observation y_i has a probability π_i of being a true zero. In order to introduce covariates to this probability, we need a link function that maps to $[0, 1]$. Two common choices are the logit link

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}},$$

and the probit link

$$\pi_i = \mathbf{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}).$$

Chambers and Cox (1967) studied the differences between these two link functions. The functions are close to indistinguishable, but the logit link has slightly heavier tails. They find that it is only possible to discriminate between the link functions with large sample sizes and certain patterns in the data. Therefore, it is commonly advised that the choice of link function is largely a matter of taste (Hahn and Soyer, 2005).

Continuous Distributions

The lognormal distribution (3.3) is a popular choice for the continuous part of the distribution, but it may not be suitable in all situations. For instance, it assumes a symmetric distribution of the log-transformed data, which is not always the case. Such cases raise the need for a more flexible alternative for the continuous part. This was the motivation behind the probit/log-skew-normal mixture introduced by Chai and Bailey (2008).

The skew-normal distribution is a class of distributions that contains the normal distribution as a special case. Using the same parametrization as Chai and Bailey (2008) a variable X is said to have a skew-normal distribution if its probability density function is

$$f(x|\mu,\sigma,\delta) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}\right). \tag{3.15}$$

The corresponding cumulative density function can be expressed as

$$F(x|\mu,\sigma,\delta) = \Phi\left(\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}\right) - 2T\left(\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}, \frac{\delta}{\sigma}\right), \tag{3.16}$$

where $T(\cdot,\cdot)$ is the Owen's T-function

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}dx.$$

The parameter δ is called the skew-parameter, as it defines the degree of skewness in the distribution. Note that $\delta = 0$ gives $\Phi\big(\frac{\delta}{\sigma}\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}\big) = \frac{1}{2}$ and $T\big(\frac{x-\mu}{\sqrt{\sigma^2+\delta^2}}, \frac{\delta}{\sigma}\big) = 0$, which is equivalent to the normal distribution with mean μ and variance σ^2. Furthermore, $\delta > 0$ gives a distribution that is skewed to the left, and $\delta < 0$ gives a right-skewed distribution. The effects of the skewness parameter are demonstrated in Figure 3.4.

The mean and variance of X is given by

$$\begin{aligned} E(X) &= \mu + \delta\sqrt{\frac{2}{\pi}}, \\ \mathrm{Var}(X) &= \sigma^2 + \delta^2\Big(1 - \frac{2}{\pi}\Big). \end{aligned} \tag{3.17}$$

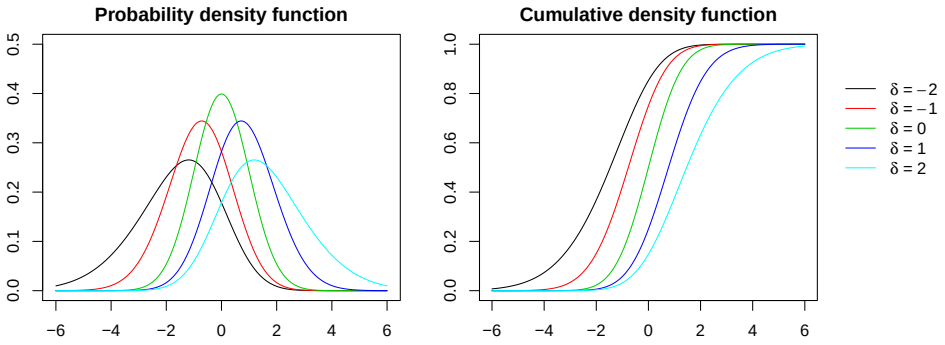

Figure 3.4: The probability density function $f(x|\mu, \sigma, \delta)$ and cumulative density function $F(x|\mu, \sigma, \delta)$ of the skew-normal distribution for $\sigma = 1$, $\mu = 0$ and different values of δ.

From these expressions it becomes clear that the mean increases and the variance grows as the magnitude of the skew-parameter δ increases. For future reference we write $X \sim \text{SN}(\mu, \sigma, \delta)$.

A variable X is said to be log-skew-normal distributed if $\ln(X) \sim \text{SN}(\mu, \sigma, \delta)$ follows a skew-normal distribution. The moment generating function of the skew-normal distribution is given by $M_X(t) = E(e^{tX}) = 2e^{\mu t + \sqrt{\sigma^2+\delta^2}t^2/2}\mathbf{\Phi}(\delta t)$. This can be used to derive the mean and variance of a log-skew-normal distributed variable, by

$$\begin{aligned} E(X) &= E(e^{\ln X}) = M_{\ln X}(1) = 2e^{\mu+(\sigma^2+\delta^2)/2}\mathbf{\Phi}(\delta), \\ \text{Var}(X) &= E(X^2) - E(X)^2 = M_{\ln X}(2) - M_{\ln X}(1)^2 \\ &= 2e^{2\mu+\sigma^2+\delta^2}[e^{\sigma^2+\delta^2}\mathbf{\Phi}(2\delta) - 2\mathbf{\Phi}(\delta)^2]. \end{aligned}$$

There are countless other possible methods for dealing with non-symmetric log-transformed data, such as using skewed distributions like the gamma distribution, and *ad hoc* transformations of the data to obtain symmetry. The advantages of the log-skew-normal distribution include that it eliminates the need for transforming the data and that the interpretations of the regression coefficients (except the intercept) are the same as for the lognormal distribution (Chai and Bailey, 2008). Other studied alternatives to the lognormal distribution in the specified two-part models include the log-gamma (Moulton and Halsey, 1996), log-skew-T (Dagne, 2017; Xing et al., 2017), and generalized gamma (Liu et al., 2016; Jaffa et al., 2018) distributions.

3.2 Expansions to Longitudinal Data

Let $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ be the vector consisting of the n_i observations of individual i, such that y_{ij} is obtained at time point t_{ij} with associated indicator variable I_{ij} as defined in (3.1). When there are multiple observations from the same individual

measured at different time points it is no longer reasonable to assume independence. The correlation between measurements from the same individual can be accounted for by introducing random effects to the model. In order to calculate the liklelihood of $\mathbf{y}_i$ the random effects must be integrated out, giving marginalized likelihoods. In general the marginal likelihood of the observations $\mathbf{y}_i$ from the ith individual is on the form

$$\mathcal{L}_i = \int \left(\prod_{i=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, \tau_i) \right) f_\tau(\boldsymbol{\tau}_i|\boldsymbol{\Sigma}) \; d\tau_i, \tag{3.18}$$

where $g(\cdot)$ is the density of y_{ij}, n_i is the number of observations from the individual, $\boldsymbol{\theta}$ is the parameters of $g(\cdot)$, and $f_\tau(\boldsymbol{\tau}_i|\boldsymbol{\Sigma})$ is the probability density function of the random effects $\boldsymbol{\tau}_i$. The parameterizations and resulting marginal likelihoods are presented below.

One-Part Models

In the Tobit-models and the substitute models a single random effect can be introduced to (3.4) by letting

$$\mu_{ij} = \mathbf{z}'_{ij}\boldsymbol{\gamma} + \tau_i,$$

where $\tau_i \sim \mathcal{N}(0, s)$ is normally distributed with variance s.

For the Tobit model the marginal likelihood can be found by plugging the probability density function (3.2) into the general form (3.18), which gives

$$\mathcal{L}_i(= \int \left(\prod_{i=1}^{n_i} F(T|\boldsymbol{\theta}, \tau_i)^{I_{ij}} f(y_{ij}|\boldsymbol{\theta}, \tau_i)^{(1-I_{ij})} \right) f_\tau(\tau_i|s) \; d\tau_i, \tag{3.19}$$

where $\boldsymbol{\theta}$ is the parameters of $f(\cdot)$, including $\boldsymbol{\gamma}$. The probability density of the substitute model is defined in equation (3.10), and gives the marginal density

$$\mathcal{L}_i = \int \left(\prod_{i=1}^{n_i} f(S|\boldsymbol{\theta}, \tau_i)^{I_{ij}} f(y_{ij}|\boldsymbol{\theta}, \tau_i)^{(1-I_{ij})} \right) f_\tau(\tau_i|s) \; d\tau_i, \tag{3.20}$$

where $S \in (0, T]$ is the chosen substitute value.

Standard Two-Part Models

In the standard two-part models, both with and without interval censoring, random effects can be introduced to both parts of the model. This is demonstrated by Mahmud et al. (2010) for the standard two-part model and Berk and Lachenbruch (2002) for the standard two-part model with interval censoring. The random effects are introduced to (3.8) by letting

$$\begin{aligned} \mu_{ij} &= \mathbf{z}'_{ij}\boldsymbol{\gamma} + \tau_{1i}, \\ \pi_{ij} &= \boldsymbol{\Phi}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \tau_{2i}), \end{aligned} \tag{3.21}$$

where τ_{1i} and τ_{2i} are random effects. They are assumed to be bivariate normally distributed with mean zero, i.e. $\boldsymbol{\tau}_i = (\tau_{1i}, \tau_{2i}) \sim \mathcal{N}_2(0, \boldsymbol{\Sigma})$. Let the (2×2) covariance matrix be

$$\boldsymbol{\Sigma} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}. \tag{3.22}$$

Here s_{11} and s_{22} are the variances of the random effects, and s_{12} is their covariance.

As before, the marginal likelihoods are found by inserting the probability density functions into the general marginal likelihood in (3.18). The standard two-part model without interval censoring defined by the probability density (3.5) gives the marginal likelihood

$$\mathcal{L}_i = \int \left(\prod_{j=1}^{n_i} \pi_{ij}^{I_{ij}} \{(1-\pi_{ij}) f(y_{ij}|\boldsymbol{\theta}, \tau_i)\}^{(1-I_{ij})} \right) f_\tau(\boldsymbol{\tau}_i|\boldsymbol{\Sigma})\ d\boldsymbol{\tau}_i, \tag{3.23}$$

where π_{ij} is defined in (3.21), $f(\cdot)$ is the assumed continuous density of the positive observations, and $f_\tau(\cdot)$ is the bivariate normal denstiy of the random effects $\boldsymbol{\tau}$ with mean zero and covariance $\boldsymbol{\Sigma}$.

Likewise, the marginal likelihood of $\mathbf{y}_i$ under the standard two-part model with interval censoring defined in (3.7) is given by

$$\mathcal{L}_i = \int \left(\prod_{j=1}^{n_i} \{\pi_{ij} + (1-\pi_{ij}) F(T|\boldsymbol{\theta}_{ij})\}^{I_{ij}} \{(1-\pi_{ij}) f_(y_{ij}|\boldsymbol{\theta}, \boldsymbol{\tau}_i)\}^{(1-I_{ij})} \right) f_\tau(\boldsymbol{\tau}|\boldsymbol{\Sigma})\ d\boldsymbol{\tau}, \tag{3.24}$$

where π_{ij} is defined in (3.21), $f(\cdot)$ is the assumed distribution of the continuous part of the latent distribution with corresponding cumulative density function $F(\cdot)$, and $f_\tau(\cdot)$ is the bivariate normal density of the random effects $\boldsymbol{\tau}$ with mean zero and covariance $\boldsymbol{\Sigma}$.

Marginalized Two-Part Models

The marginalized two-part model is expanded to lognitudinal data by Jaffa et al. (2018). Random effects are introduced to both model parts (3.13) and (3.14) by letting

$$\begin{aligned} E(Y_i) &= \exp(\mathbf{z}'_{ij}\boldsymbol{\gamma} + \tau_{1i}), \\ \pi_{ij} &= \boldsymbol{\Phi}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \tau_{2i}), \end{aligned}$$

where τ_{1i} and τ_{2i} are bivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}$ as in (3.22). As the marginalized two-part model assumes the same distribution as the standard two-part model, the marginal likelihood is as in (3.23).

3.2.1 Including Time in the Models

With longitudinal data it is often of interest to include time in the model in order to study how the passage of time affects the response. The time parameter t_{ij} can be included as a covariate in both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in countless ways. Among the common

alternatives are including t_{ij} as a categorical covariate (Berk and Lachenbruch, 2002), a linear covariate (Su and Luo, 2017) and a quadratic covariate (Mahmud et al., 2010; Dagne, 2016). The first mentioned has the advantage of making no assumptions about the relationship between the response and the passage of time. This comes at the cost of many additional parameters if the number of time points is high. Assuming a linear or quadratic relationship gives fewer additional parameters, but also less flexibility.

Another important consideration is whether the mixing probability π should depend only on the individual (Dagne, 2016), or if it may vary with time as well (Dagne, 2017; Berk and Lachenbruch, 2002; Mahmud et al., 2010). In the model without interval censoring described by likelihood (3.5), π_{ij} is the probability of falling below the LOD. As the mean μ_{ij} of the continuous distribution is assumed to depend on time, the proportion falling below the LOD will also change. Hence, the only reasonable choice for this model is to let π_{ij} depend on time. This is not as straightforward for the model with interval censoring described by likelihood function (3.7). Here, π_{ij} is the probability of belonging to a separate sub-LOD population. This probability might be constant over time, even if the mean μ_{ij} of the continuous part changes.

3.2.2 Prediction

There are two approaches to prediction for mixed effect models. Say we want to estimate the probability π_{ij} of an observation belonging to the discrete part. The *population averaged* probability can be found by integrating out the random effects. It is given by

$$\hat{\pi}_{ij} = \int \mathbf{\Phi}(\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \tau_{2i}) \frac{1}{\sqrt{s_{22}}} \phi\Big(\frac{\tau_{2i}}{\sqrt{s_{22}}}\Big)\, d\tau_{2i},$$

where $\frac{1}{\sqrt{s_{22}}}\phi(\frac{\tau_{2i}}{\sqrt{s_{22}}})$ is the assumed marginal distribution of τ_{2i}. This is an estimate of the overall probability in the population, and can be used to estimate the outcome for new unobserved individuals.

Subject specific prediction refers to estimating π_{ij} conditioned on the random effect, i.e.

$$\hat{\pi}_{ij}|\tau_{2i} = \mathbf{\Phi}(\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \tau_{2i}).$$

This can be used to predict the outcome for a previously measured individual with estimated random effect τ_{2i}. For example, if an individual has missing data for one or more time points the outcome at these time points can be predicted conditioned on the individuals random effect. The random effects $\boldsymbol{\tau}_i$ are unobserved latent variables that can be estimated by the empirical Bayes estimator (Pinheiro and Bates, 1995). The rationale is to use the parameter estimates $\hat{\boldsymbol{\theta}}$ and the observed data to calculate the posterior mode of random effects, i.e.

$$\begin{aligned}\hat{\boldsymbol{\tau}}_i &= \arg\max_{\boldsymbol{\tau}_i} \Big\{ f(\boldsymbol{\tau}_i|\mathbf{y}_i, \mathbf{x}_i, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}) \Big\} \\ &= \arg\max_{\boldsymbol{\tau}_i} \Big\{ f(\mathbf{y}_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}, \boldsymbol{\tau}_i) f(\boldsymbol{\tau}_i|\hat{\boldsymbol{\Sigma}}) \Big\}.\end{aligned}$$

Here $f(\mathbf{y}_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}, \boldsymbol{\tau}_i)$ is the subject specific likelihood of the observations $\mathbf{y}_i$ conditioned on the random effects, and $f(\boldsymbol{\tau}_i|\hat{\boldsymbol{\Sigma}})$ is the bivariate normal prior distribution of the random effects.

If the predictor is a linear function of the random effects, the population averaged probability is equivalent to plugging in the expected value of the random effect. For the continuous part of the model we have

$$\hat{\mu}_{ij} = \int (\mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}} + \tau_{1i}) \frac{1}{\sqrt{s_{11}}} \phi\Big(\frac{\tau_{1i}}{\sqrt{s_{11}}}\Big)\, d\tau_{1i} = \mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}},$$

which is equivalent to plugging in $\tau_{1i} = 0$. Thus, the marginal means can be read directly from the estimated parameters.

3.3 Expansions to Bivariate Data

Let $\mathbf{y}_i = (y_{1i}, y_{2i})'$ be a pair of two observations from individual i that both are subject to a lower limit of detection, denoted as T_1 and T_2, respectively. The two observations may for instance represent measurements conducted with two different methods, measurements at two different time points, or measurements of two different phenomenons assumed to be related. If a non-negligible amount of observations falls below the detection limit naive methods, such as replacing the censored observations with a fraction of the detection limit, will in general produce bias (Chu et al., 2008). This has motivated the use of censored bivariate models to estimate the correlation. The bivariate models also have other applications, such as estimating the ratio between the means of the two measurements (Andersen et al., 2013).

As in the univariate case, we define a latent variable $\mathbf{y}^*_i = (y^*_{1i}, y^*_{2i})'$ such that

$$y_{1i} = \begin{cases} 0, & y^*_{1i} \leq T_1 \\ y^*_{1i}, & y^*_{1i} > T_1 \end{cases},$$

and equivalently for y_{2i}.

3.3.1 Tobit Models

Lyles et al. (2001) proposed a bivariate model based on the assumption that all the observed pairs arises from the same underlying distribution $f(\mathbf{y}^*_i)$ with support on $\mathbb{R}^2_+ = (0, \infty) \times (0, \infty)$. One or both of the variables may be censored due to lower limits of detection, and are therefore recorded as zero. This is analogous to the Tobit model in the univariate setting. A conceptual illustration is provided in Figure 3.5.

There are four possible types of observed pairs, which corresponds to the four regions separated by the LODs in Figure 3.5: (1) Both y^*_{1i} and y^*_{2i} are observed, (2) y^*_{1i} is observed and $y^*_{2i} \leq T_2$, (3) $y^*_{1i} \leq T_1$ and and y^*_{2i} is observed, and (4) both

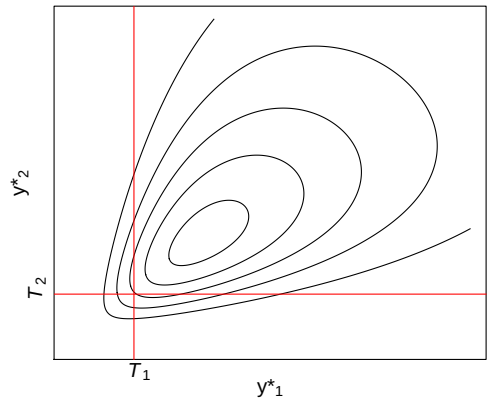

Figure 3.5: Conceptual illustration of the assumed distribution of $\mathbf{y}_i^*$ with the bivariate Tobit model.

$y_{1i}^* \leq T_1$ and $y_{2i}^* \leq T_2$. The four types of pairs has the following contributions to the likelihood:

$$\begin{aligned}
(1): &\quad \mathcal{L}_i = f(y_{1i}^*, y_{2i}^*) \\
(2): &\quad \mathcal{L}_i = P(Y_{2i}^* \leq T_2 | Y_{1i}^* = y_{1i}^*) \cdot f(y_{1i}) \\
(3): &\quad \mathcal{L}_i = P(Y_{1i}^* \leq T_1 | Y_{2i}^* = y_{2i}^*) \cdot f(y_{2i}^*) \\
(4): &\quad \mathcal{L}_i = P(Y_{1i}^* \leq T_1, Y_{2i}^* \leq T_2)
\end{aligned} \tag{3.25}$$

In pairs of type (1) the true value of both y_{1i}^* and y_{2i}^* is observed, so the contribution is simply their joint probability density $f(y_{1i}^*, y_{2i}^*)$. When one of the observations is censored, as in pair (2) and (3), the contribution can be expressed as the probability of falling below the LOD for the censored observation conditioned on the non-censored observation, times the marginal probability density of the non-censored observation. In the last case, when both observations are censored, the contribution is the joint probability of falling below the LOD.

The Lognormal Bivariate Tobit Model

The bivariate lognormal model has five parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. $\mathbf{Y} = (Y_1, Y_2)'$ is said to follow a bivariate lognormal distribution if

$$\ln(\mathbf{Y}) \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

With this distribution the likelihoods in (3.25) can be expressed as follows:

$$\begin{aligned}
(1): \quad \mathcal{L}_i &= \frac{1}{\sigma_2\sigma_{1|2}}\phi\left(\frac{\ln(y_{1i})-\mu_{1|y_{2i}}}{\sigma_{1|2}}\right)\phi\left(\frac{\ln(y_{2i})-\mu_2}{\sigma_2}\right)\\
(2): \quad \mathcal{L}_i &= \frac{1}{\sigma_1}\phi\left(\frac{\ln(y_{1i})-\mu_1}{\sigma_1}\right)\mathbf{\Phi}\left(\frac{\ln(T_2)-\mu_{2|y_{1i}}}{\sigma_{2|1}}\right)\\
(3): \quad \mathcal{L}_i &= \frac{1}{\sigma_2}\mathbf{\Phi}\left(\frac{\ln(T_1)-\mu_{1|y_{2i}}}{\sigma_{1|2}}\right)\phi\left(\frac{\ln(y_{2i})-\mu_2}{\sigma_2}\right)\\
(4): \quad \mathcal{L}_i &= \int_0^{T_1}\frac{1}{\sigma_1}\phi\left(\frac{\ln(z)-\mu_1}{\sigma_1}\right)\mathbf{\Phi}\left(\frac{\ln(T_2)-\mu_{2|z}}{\sigma_{2|1}}\right)\,dz
\end{aligned} \tag{3.26}$$

Here $\sigma^2_{1|2} = \sigma^2_1(1-\rho^2)$ and $\mu_{1|y_{2i}} = \mu_1+(\rho\sigma_1/\sigma_2)(\ln(y_{2i})-\mu_2)$, and vice versa. The contribution from pairs of type (4) is found by calculating the joint probability by conditioning on the observation of y_{1i} and integrating over all the possible values of y_{1i}, i.e. $P(Y^*_{1i} \le T_1, Y^*_{2i} \le T_2) = \int_0^{T_1} P(Y^*_{2i} \le T_2 | Y^*_{1i} = z) \cdot f_{Y^*_{1i}}(z)\,dz$.

As demonstrated by e.g Andersen et al. (2013), covariates may be introduced to both μ_1 and μ_2 by

$$\begin{aligned}
\mu_{1i} &= \mathbf{x}_i'\boldsymbol{\gamma}_1,\\
\mu_{2i} &= \mathbf{x}_i'\boldsymbol{\gamma}_2,
\end{aligned}$$

where $\mathbf{x}_i$ is a set of covariates with corresponding fixed effects in $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ for the two model parameters.

3.3.2 Two-Part Models

The bivariate Tobit model has the same restriction as in the univariate case, namely that the number of censored observations must correspond to the distribution of the non-censored observations. This restriction was relaxed by Chu et al. (2005), which proposed two-part mixture model for data where the south-west tail of the observed distribution is incompatible with the high amount of censored observations. The mixture consists of two components, one lower component denoted $f_L(\mathbf{y})$ corresponding to the low-rosponders and one higher component $f_H(\mathbf{y})$ corresponding to individuals with higher responses. Let π_i be the probability that $\mathbf{y}_i$ is from the lower component. Then the latent mixture density is given by

$$f(\mathbf{y}) = \pi_i f_L(\mathbf{y}) + (1-\pi_i) f_H(\mathbf{y}).$$

Furthermore, we assume that the lower component $f_L(\mathbf{y})$ is entirely located on the domain $[0, T_1] \times [0, T_2]$ below the detection limits. Thus, its shape is irrelevant, and it can be thought of as a point mass at $(0, 0)$. This is analogous to the two-part model with interval censoring in the univariate setting. A conceptual illustration is provided in Figure 3.6. Again, there are four possible types of observed pairs, which corresponds to the four sections in Figure 3.6 separated by the LODs. Their contributions to the likelihood are described in (3.25).

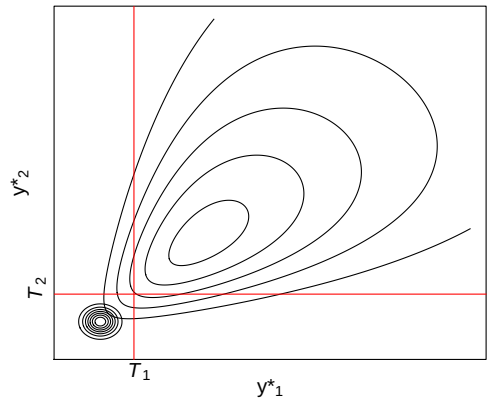

Figure 3.6: Conceptual illustration of the assumed distribution of $\mathbf{y}_i^*$ with the bivariate two-part model.

In this model it is assumed that both measurements are low responses simultaneously. Thus, observed pairs where one of the measurements are censored have to come from the higher distribution, meaning that the amount of half-observed pairs must correspond to the shape of the observed higher distribution. This is a logical assumption in many cases, for instance if y_{1i} and y_{2i} are two measurements of the same phenomenon, either with two different methods or at two different time points. Then a low response of y_{1i} can be assumed to be accompanied by a low response of y_{2i}, and vice versa. If on the other hand y_{1i} and y_{2i} are measurements of two distinct phenomenons, it may not be reasonable to assume that both are always low-responders simultaneously.

The Lognormal Bivariate Two-Part Model

If the bivariate lognormal distribution is used for the higher component $f_H(\mathbf{y})$ and the lower component $f_L(\mathbf{y})$ is assumed to fall within the domain $[0, T_1] \times [0, T_2]$, the likelihoods in (3.25) can be expressed as follows:

$$
\begin{aligned}
(1): \quad & \mathcal{L}_i = \frac{1-\pi_i}{\sigma_2 \sigma_{1|2}} \phi\left(\frac{\ln(y_{1i}) - \mu_{1|y_{2i}}}{\sigma_{1|2}}\right) \phi\left(\frac{\ln(y_{2i}) - \mu_2}{\sigma_2}\right) \\
(2): \quad & \mathcal{L}_i = \frac{1-\pi_i}{\sigma_1} \phi\left(\frac{\ln(y_{1i}) - \mu_1}{\sigma_1}\right) \mathbf{\Phi}\left(\frac{\ln(T_2) - \mu_{2|y_{1i}}}{\sigma_{2|1}}\right) \\
(3): \quad & \mathcal{L}_i = \frac{1-\pi_i}{\sigma_2} \mathbf{\Phi}\left(\frac{\ln(T_1) - \mu_{1|y_{2i}}}{\sigma_{1|2}}\right) \phi\left(\frac{\ln(y_{2i}) - \mu_2}{\sigma_2}\right) \\
(4): \quad & \mathcal{L}_i = \pi_i + (1-\pi_i) \int_0^{T_1} \frac{1}{\sigma_1} \phi\left(\frac{\ln(z) - \mu_1}{\sigma_1}\right) \mathbf{\Phi}\left(\frac{\ln(T_2) - \mu_{2|z}}{\sigma_2}\right) dz
\end{aligned}
\tag{3.27}
$$

Here $\sigma^2_{1|2} = \sigma^2_1(1-\rho^2)$ and $\mu_{1|y_{2i}} = \mu_1 + (\rho\sigma_1/\sigma_2)(\ln(y_{2i}) - \mu_2)$, and vice versa. Thus, there are five parameters toestimate, $\boldsymbol{\theta} = (\pi, \mu_1, \mu_2, \sigma^2_1, \sigma^2_2, \rho)$. There are no parameters for the lower component, apart from the mixing probability π, since the only observable information about this component is its relative weight. As in the univariate setting, covariates can be introduced to the π_i by utilizing e.g. the probit link function. Thus, covariates can be introduced to the model by letting

$$\begin{aligned}\mu_{1i} &= \mathbf{x}_i'\boldsymbol{\gamma}_1,\\ \mu_{2i} &= \mathbf{x}_i'\boldsymbol{\gamma}_2,\\ \pi_i &= \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}).\end{aligned}$$

3.3.3 Four-Part Models

The described two-part model is only suitable when the two measurements y_{1i} and y_{2i} are always low-responders simultaneously. This is not always the case. For instance the amount of excess zeroes in y_{1i} may be different from the amount of excess zeroes in y_{2i}. We propose a four-part mixture model where the amount of excess zeroes in each variable is determined by distinct processes. The mixture consists of a higher component $f_H(\mathbf{y})$, and three lower components denoted $f_L(\mathbf{y})$, $f_{L_1}(\mathbf{y})$ and $f_{L_2}(\mathbf{y})$. As in the two-part model, $f_L(\mathbf{y})$ represents the low-responders in both variables. The lower component $f_{L_1}(\mathbf{y})$ corresponds to the low-responders in y_{2i} and high-responders in y_{1i}, and $f_{L_2}(\mathbf{y})$ corresponds to the low-responders in y_{1i} and high-responders in y_{2i}. Let π, π_1 and π_2 be the mixing probabilities of $f_L(\mathbf{y})$, $f_{L_1}(\mathbf{y})$ and $f_{L_2}(\mathbf{y})$, respectively. Then the latent mixture density is given by

$$f(\mathbf{y}) = \pi_{1i} f_{L_1}(\mathbf{y}) + \pi_{2i} f_{L_2}(\mathbf{y}) + \pi_i f_L(\mathbf{y}) + (1 - \pi_{1i} - \pi_{2i} - \pi_{3i}) f_H(\mathbf{y}).$$

A conceptual illustration of the model is provided in Figure 3.7. Furthermore, we assume that the lower components are located entirely below the LOD of the variable it is a low-responder in. In other words, the domain of $f_{L_1}(\mathbf{y})$ is assumed to be within $[0,\infty) \times [0, T_2]$, the domain of $f_{L_2}(\mathbf{y})$ is assumed to be within $[0, T_1] \times [0,\infty)$, and the domain of $f_L(\mathbf{y})$ is assumed to be within $[0, T_1] \times [0, T_2]$.

If the two variables are independent, the censored variables in y_{2i} have the same marginal distribution in y_{1i} as the uncensored variables in y_{2i}. Thus, $f_{L_1}(\mathbf{y})$ and $f_{L_2}(\mathbf{y})$ have the same marginal distributions as $f_H(\mathbf{y})$. In this case, the marginal distributions of each of the variables are two-part models with interval censoring.

The Lognormal Bivariate Four-Part Model

Only the mixing probability π is observable for the component $f_L(\mathbf{y})$ contained below both the LODs, thus it contributes with only one parameter. For the two half-censored components we observe the marginal distribution of the non-censored component as well as their relative weight. If the marginal distributions are assumed to be lognormal they contribute with two parameters each. We denote the parameters of $f_{L_1}(\mathbf{y})$ as μ_{L_1} and σ_{L_1}, and likewise the parameters of $f_{L_2}(\mathbf{y})$ as μ_{L_2} and σ_{L_2}. This gives a total of twelve parameters, $\boldsymbol{\theta} =$

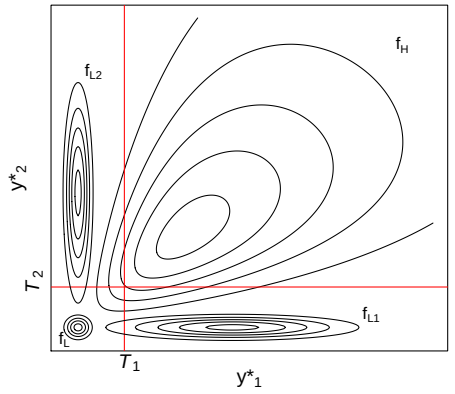

Figure 3.7: Conceptual illustration of the assumed distribution of $\mathbf{y}_i^*$ with the bivariate four-part model.

$(\pi_1, \pi_2, \pi, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho, \mu_{L_1}, \sigma_{L_1}, \mu_{L_2}, \sigma_{L_2})$. The contributions to the likelihood for each of the four possible types of observed pairs is given by

$$
\begin{aligned}
(1): \quad & \mathcal{L}_i = \frac{1-\pi_{1i}-\pi_{2i}-\pi_i}{\sigma_2\sigma_{1|2}}\phi\left(\frac{\ln(y_{1i})-\mu_{1|y_{2i}}}{\sigma_{1|2}}\right)\phi\left(\frac{\ln(y_{2i})-\mu_2}{\sigma_2}\right) \\
(2): \quad & \mathcal{L}_i = \frac{\pi_{1i}}{\sigma_{L_1}}\phi\left(\frac{\ln(y_{1i})-\mu_{L_1}}{\sigma_{L_1}}\right) \\
& \quad + \frac{1-\pi_{1i}-\pi_{2i}-\pi_i}{\sigma_1}\phi\left(\frac{\ln(y_{1i})-\mu_1}{\sigma_1}\right)\mathbf{\Phi}\left(\frac{\ln(T_2)-\mu_{2|y_{1i}}}{\sigma_{2|1}}\right) \\
(3): \quad & \mathcal{L}_i = \frac{\pi_{2i}}{\sigma_{L_2}}\phi\left(\frac{\ln(y_{2i})-\mu_{L_2}}{\sigma_{L_2}}\right) \\
& \quad + \frac{1-\pi_{1i}-\pi_{2i}-\pi_i}{\sigma_2}\mathbf{\Phi}\left(\frac{\ln(T_1)-\mu_{1|y_{2i}}}{\sigma_{1|2}}\right)\phi\left(\frac{\ln(y_{2i})-\mu_2}{\sigma_2}\right) \\
(4): \quad & \mathcal{L}_i = \pi_i + \pi_{1i}\mathbf{\Phi}\left(\frac{\ln(T_1)-\mu_{L_1}}{\sigma_{L_1}}\right) + \pi_{2i}\mathbf{\Phi}\left(\frac{\ln(T_2)-\mu_{L_2}}{\sigma_{L_2}}\right) \\
& \quad + (1-\pi_{1i}-\pi_{2i}-\pi_i)\int_0^{T_1}\frac{1}{\sigma_1}\phi\left(\frac{\ln(z)-\mu_1}{\sigma_1}\right)\mathbf{\Phi}\left(\frac{\ln(T_2)-\mu_{2|z}}{\sigma_2}\right)\,dz
\end{aligned}
$$

As before, the fully observed pairs of type (1) comes from the high component $f_H(\mathbf{y})$. The half-observed pairs of type (2) and (3) may come from the high component or the respective low-component representing low-responders in one of the variables. The fully censored pairs of type (4) may come from any of the four model parts.

There are many possible ways of simplifying this model. For instance if the low-responders in y_{2i} is assumed to follow the same marginal distribution in y_{1i} as

the high-responders one can set $\mu_{L_1} = \mu_1$ and $\sigma_{L_1} = \sigma_1$. The same can be done to eliminate μ_{L_2} and σ_{L_2} from the model. Under these assumptions there are only eight parameters to estimate, $\boldsymbol{\theta} = (\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

3.4 Model Estimation

In this section we present methods for estimating the parameters of the models with random effects specified in Section 3.2. The same methods can however be utilized when random effects are not included in the model and for the bivariate models in Section 3.3. In these cases the numerical integration is not needed. The methods presented are based on a frequentist approach.

Let $\boldsymbol{\theta}$ be the parameters of the model used, and $\mathbf{x}$ be all the observed data. In the frequentist approach the likelihood to be maximized is

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \prod_i \mathcal{L}_i(\boldsymbol{\theta}|\mathbf{x}), \tag{3.28}$$

with $\mathcal{L}_i$ as in defined in Section 3.2 for the different models. The general form of the likeihoods is provided in (3.18),

$$\mathcal{L}_i = \int \left(\prod_{i=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, \tau_i) \right) f_\tau(\boldsymbol{\tau}_i|\boldsymbol{\Sigma}) \; d\tau_i,$$

where $g(\cdot)$ is the density of y_{ij}, n_i is the number of observations from the individual, $\boldsymbol{\theta}$ is the parameters of $g(\cdot)$, and $f_\tau(\boldsymbol{\tau}_i|\boldsymbol{\Sigma})$ is the probability density function of the random effects $\boldsymbol{\tau}_i$.

The theory in this section is mainly obtained from the book by Casella and Berger (2002).

3.4.1 Numerical Methods

There are two numerical challenges with the frequentist approach. Firstly, the integral in the marginal likelihoods (3.18) can not be solved analytically. Secondly, the likelihood function must be maximized. We will continue to describe methods for solving these problems.

Numerical Integration by Adaptive Gaussian Quadrature

A numerical integration method that is shown to work well with moderate cluster sizes is Gaussian quadrature (Rabe-Hesketh et al., 2005). In this method the area under the curve is approximated by summing over split areas. The areas are represented by quadrature points with corresponding weights, and the accuracy depends on the number of quadrature points.

First, we substitute the integration variable $\boldsymbol{\tau}_i$ with $\boldsymbol{v}_i$, which is obtained by

$$\boldsymbol{\tau}_i = \mathbf{Q}\boldsymbol{v}_i,$$

where $\mathbf{Q}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$. Since the covariance matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite we have $\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{Q}^*$, where $\mathbf{Q}$ is lower triangular. The resulting integration variables $\boldsymbol{v}_i = (v_{i1}, v_{i2})$ are independent standard normal variables. Thus, the integral can be expressed as

$$\begin{aligned}\mathcal{L}_i &= \int \phi(v_{i2}) \int \phi(v_{i1}) \prod_{j=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, \boldsymbol{v}_i)\ dv_{i1} dv_{i2} \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v_{i2}^2} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v_{i1}^2} \prod_{j=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, \boldsymbol{v}_i)\ dv_{i1} dv_{i2},\end{aligned}$$

where $\phi(\cdot)$ is the standard normal probability density. We now have two nested integrals on the form $\int e^{-x^2} f(x)\ dx$. The Gauss-Hermite quadrature for integrals on this form is given by

$$\begin{aligned}\mathcal{L}_i &\approx \sum_{r_2=1}^{R_2} p_{r_2} \sum_{r_1=1}^{R_1} p_{r_1} \prod_{j=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, a_{r_1}, a_{r_2}) \\ &= \sum_{r=1}^{R} w(\mathbf{a}_r) \prod_{j=1}^{n_i} g(y_{ij}|\boldsymbol{\theta}, \mathbf{a}_r),\end{aligned}$$

where $\mathbf{a}_r$ are locations centered around zero and $w(\mathbf{a}_r)$ are the associated weights of R-point Gaussian quadrature. The total number of quadrature points is $R = R_1 \cdot R_2$. Thus, the number of quadrature points increases exponentially with the number of random effects.

This method can be substantially improved by using adaptive quadrature. The rationale behind this method is to take the form of the integrand into account. Instead of using preset quadrature points $\mathbf{a}_r$ centered around zero the algorithm finds quadrature points centered at the approximate mode of the integrand. Consequently, the adaptive method can use fewer quadrature points to achieve the same level of precision.

Maximum Likelihood Estimation by Quasi-Newton Optimization

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}),$$

where $\boldsymbol{\Theta}$ is the parameter space of $\boldsymbol{\theta}$. This is equivalent to finding the roots of the gradient vector of the log-likelihood, i.e. solving

$$g(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \nabla \ln \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \mathbf{0}.$$

A popular class of methods for numerical root-finding is the Quasi-Newton algorithms (Gould et al., 2006). These are based on the Newton-Raphson algorithm.

Starting from an initial position $\boldsymbol{\theta}_0$ the Newton-Raphson algorithm calculates new positions as

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + s[-H(\boldsymbol{\theta}_i)]^{-1} g(\boldsymbol{\theta}_i|\mathbf{y}),$$

where s is the step size and $H(\boldsymbol{\theta}_i)$ is the matrix of second derivatives, known as the Hessian. This is repeated until a convergence criterion is reached. The default step size is $s = 1$, but there exists methods for adaptive step size calculation which gives higher efficiency.

For larger problems it can be very computationally expensive to calculate the Hessian. This is avoided in Quasi-Newton methods by iteratively estimating $H(\boldsymbol{\theta}_i)$ at every step. There are many methods for doing this. Popular choices are the Davidon–Fletcher–Powell (DFP) forumla and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula. Let A_i be the estimate of the Hessian at step i such that

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + [-A_i]^{-1} g(\boldsymbol{\theta}_i|\mathbf{y}).$$

Both DFP and BFGS calculate A_{i+1} such that

$$(\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i) = A_{i+1}(g(\boldsymbol{\theta}_{i+1}|\mathbf{y}) - g(\boldsymbol{\theta}_i|\mathbf{y})).$$

When convergence is reached, the resulting matrix A_n can be used as an estimate of the Hessian at the maximum.

3.4.2 Estimation of Standard Errors

The Hessian of the log-likelihood function is the matrix consisting of the partial second derivatives, defined as

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}).$$

The Fisher information matrix is defined as

$$\mathcal{I}(\boldsymbol{\theta}) = -E[H(\boldsymbol{\theta})|\boldsymbol{\theta}]$$

It can be estimated by inserting the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, which gives

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = -H(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y}).$$

In Quasi-Newton optimization, $H(\hat{\boldsymbol{\theta}})$ is estimated numerically along with maximizing the likelihood, as described above. By the Cramer-Rao bound we have

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \geq [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}.$$

Thus, the inverse Fisher information matrix $C(\hat{\boldsymbol{\theta}}) = [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}$ can be used as an estimate of the covariance matrix. Consequently, the estimated standard error of each estimated parameter $\hat{\theta}_k$ is the square root of the diagnoal elements of the inverse Fisher information matrix,

$$\widehat{\mathrm{SE}}(\hat{\theta}_k) = \sqrt{C(\hat{\boldsymbol{\theta}})_{kk}},$$

where $C_{kk}(\hat{\boldsymbol{\theta}})$ denotes the kth diagnoal element of $C(\hat{\boldsymbol{\theta}}) = [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}$.

Test for Fixed Effects

A common way to test the significance of the estimated parameters is to utilize the property that the test statistic

$$T_k = \frac{\hat{\theta}_k - \theta_k}{\widehat{\mathrm{SE}}(\hat{\theta}_k)}$$

is approximately Student's t-distributed with ν degrees of freedom. It is however not straight forward to calculate the degrees of freedom when random effects are present. Satterthwaite (1946) developed a popular method for approximating ν, but it is not well defined for non-linear models (Molenberghs and Verbeke, 2004). One possible work around is to resort to a Wald-test on the basis that the Student's t-distribution is asymptotically normal as $\nu \to \infty$, i.e.

$$\frac{\hat{\theta}_k - \theta_k}{\widehat{\mathrm{SE}}(\hat{\theta}_k)} \approx \mathcal{N}(0, 1).$$

In this case $(1 - \alpha)100\%$ confidence intervals are given by

$$\theta_k = \hat{\theta}_k \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\theta}_k).$$

Note that this will give anti-conservative results, as the Student's t-distribution have heavier tails than the normal distribution.

Another possibility is to test the significance of θ_k with a likelihood ratio test. This method is described in the following section.

3.4.3 Likelihood Ratio Tests

Consider the hypotheses

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0, \qquad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c.$$

A likelihood ratio test is a test on the statistic

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\Theta}_0} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}{\sup_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{x})}$$

with rejection region on the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$. Here $\boldsymbol{\Theta}_0$ is a subset of the parameter space $\boldsymbol{\Theta}$, and $\boldsymbol{\Theta}_0^c = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$ is its complement. Thus, the numerator of $\lambda(\mathbf{x})$ is the maximum likelihood under the restriction $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and the denominator is the maximum likelihood without the restriction. Because $\boldsymbol{\Theta}_0$ is a subset of $\boldsymbol{\Theta}$ we have $\lambda(\mathbf{x}) \in [0, 1]$. If the ratio of the likelihoods is sufficiently small we reject H_0 in favor of H_1, and conclude that removing the restriction $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ gives a significant improvement in the likelihood.

A level α test is obtained by choosing c such that the probability of observing $\lambda(\mathbf{X}) \leq c$ equals α under H_0. This probability can be estimated by utilizing the asymptotic property

$$-2\log(\lambda(\mathbf{X})) \sim \chi^2_\nu$$

when the sample size $n \to \infty$. Assuming that $\boldsymbol{\Theta}_0$ is in the interior of $\boldsymbol{\Theta}$, the degrees of freedom ν is the difference in the number of free parameters when $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and under the restriction $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. This test statistic can be computed directly from the log-likelihoods of the models, as

$$-2\log(\lambda(\mathbf{X})) = -2(\log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\mathbf{x}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{x})).$$

Test for Fixed Effects

The significance of fixed effects in the model can be tested by considering the hypotheses

$$H_0 : \boldsymbol{\theta}_k = \mathbf{0}, \qquad H_1 : \boldsymbol{\theta}_k \neq \mathbf{0},$$

where $\boldsymbol{\theta}_k$ is a vector of parameters corresponding to the fixed effects, possibly of length one. The test statistic will be asymptotically chi-squared distributed with degrees of freedom equal to the number of parameters in $\boldsymbol{\theta}_k$.

Test for Random Effects

Recall that the two-part models include two random effects $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}) \sim \mathcal{N}_2(0, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}.$$

Testing the significance of random effects can be done by performing likelihood ratio tests on the elements of the covariance matrix $\boldsymbol{\Sigma}$ (Baey et al., 2017). Testing for the presence of the random effect τ_{i1} can be done by letting

$$\begin{aligned} \boldsymbol{\Theta}_0 &= \{\boldsymbol{\theta}|\boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\gamma} \in \mathbb{R}^p, \sigma \in \mathbb{R}^+, \delta \in \mathbb{R}, s_{11} = 0, s_{12} = 0, s_{22} \in \mathbb{R}^+\}, \\ \boldsymbol{\Theta} &= \{\boldsymbol{\theta}|\boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\gamma} \in \mathbb{R}^p, \sigma \in \mathbb{R}^+, \delta \in \mathbb{R}, \boldsymbol{\Sigma} \in \mathbb{S}_+^2\}, \end{aligned}$$

where $\mathbb{S}_+^2$ is the set of all symmetric, positive semi-definite (2×2) matrices. The same setup can be used for testing the significance of τ_{i2} by switching s_{11} with s_{22} in the definition of $\boldsymbol{\Theta}_0$.

Since $\boldsymbol{\Theta}_0$ lies on the boundary of $\boldsymbol{\Theta}$, the asymptotic property used in the previous section does not hold. It can however be proved that the limiting distribution of $-2\log(\lambda(\mathbf{X}))$ is the distribution $\frac{1}{2}(\chi_0^2 + \chi_1^2)$ when the random effects are independent (i.e. $s_{12} = 0$), and the distribution $\frac{1}{2}(\chi_1^2 + \chi_2^2)$ when the random effects are correlated (i.e. $s_{12} \neq 0$) (Baey et al., 2017). Testing for a random effect in the one-part models is equivalent to the case where the random effects are uncorrelated. The cumulative density function of a distribution on the form $\frac{1}{2}(\chi_{k-1}^2 + \chi_k^2)$ can be expressed as

$$P(\tfrac{1}{2}(\chi_{k-1}^2 + \chi_k^2) \leq x) = \tfrac{1}{2}(P(\chi_{k-1}^2) \leq x) + P(\chi_k^2) \leq x)).$$

Thus, the p-value of the likelihood ratio test can by found by taking the average of the p-values provided by the two distributions in the mixture (Goldman and Whelan, 2000).

Test for the Discrete Part

The Tobit model (3.2) can be viewed as a special case of the two-part model with interval censoring (3.7) where the probability π_i of true zeroes is set to zero. Therefore, the significance of the discrete part can be tested with a likelihood ratio test (Berk and Lachenbruch, 2002). The difference in the number of parameters is the number of parameters q in $\boldsymbol{\beta}$ plus the variance s_{22} and covariance s_{12} of the random effect τ_{i2}. This can be formulated as

$$\begin{aligned}\boldsymbol{\Theta}_0 &= \{\boldsymbol{\theta} | \beta_0 = -\infty, \beta_{i>0} = 0, \boldsymbol{\gamma} \in \mathbb{R}^p, \sigma \in \mathbb{R}^+, s_{11} \in \mathbb{R}^+, s_{12} = 0, s_{22} = 0\},\\ \boldsymbol{\Theta} &= \{\boldsymbol{\theta} | \boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\gamma} \in \mathbb{R}^p, \sigma \in \mathbb{R}^+, \boldsymbol{\Sigma} \in \mathbb{S}_+^2\},\end{aligned}$$

where $\mathbb{S}_+^2$ is the set of all symmetric, positive semi-definite (2×2) matrices. Two of the restrictions in $\boldsymbol{\Theta}_0$ are on the boundary of the parameter space, $\beta_0 = -\infty$ and $s_{22} = 0$. This scenario is studied by Self and Liang (1987). The resulting asymptotic distribution is a mixture of χ^2_{q+2}, χ^2_{q+1} and χ^2_q, where q is the number of parameters in $\boldsymbol{\beta}$. More degrees of freedom gives more conservative p-values. Therefore, the resulting p-value from assuming a χ^2_{q+2} distribution can be used as an upper bound for the true p-value, and likewise the p-value based on q degrees of freedom gives a lower bound. Alternatively, a model selection criterion could be used. This option will be discussed in Section 3.5.1.

3.5 Model Evaluation

Here we present two frameworks for evaluating the performance of statistical models. The model selection criteria evaluate the model fit based on the resulting likelihood of the data. This is akin to likelihood ratio tests, but does not require the candidate models to be nested. Scoring rules evaluate the predictive power of the models by assigning a numerical score to predictive distributions based on the true outcome.

3.5.1 Model Selection Criteria

The problem of model selection can be seen as a trade-off between complexity and model fit (Vrieze, 2012). Increased complexity will lead to a better fit to the data, reflected in a higher likelihood. But, it also increases the chance of overfitting. Two common model selection criteria that address this trade-off are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Both can be expressed in terms of the log-likelihood and a penalty for increased complexity,

$$\begin{aligned}\text{AIC} &= -2\ell(\hat{\boldsymbol{\theta}}) + 2k,\\ \text{BIC} &= -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)k,\end{aligned} \tag{3.29}$$

where $\ell(\cdot)$ is the log-likelihood, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the model parameters, k is the number of model parameters and n is the number of observations in the dataset.

Even though the formulas for the AIC and BIC are strikingly similar, they are derived on very different grounds. The AIC is founded on information theory. The idea is to estimate how much information is lost by representing the true underlying process with the candidate model. Of course, the true distribution is not known, so we can not know for sure which candidate model has the lowest information loss, but the difference in the AIC can be used as an estimate of the relative information loss. The candidate model with the lowest AIC has the lowest expected information loss when used to represent the true process (Vrieze, 2012).

The BIC is founded on Bayesian statistical analysis. From Bayes' theorem the probability of model $\mathcal{M}$ being true after having observed the data $\mathbf{x}$ is

$$P(\mathcal{M}|\mathbf{x}) \propto P(\mathbf{x}|\mathcal{M})P(\mathcal{M}),$$

where

$$P(\mathbf{x}|\mathcal{M}) = \int_{\boldsymbol{\Theta}_{\mathcal{M}}} P(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}$$

is often called the evidence for model $\mathcal{M}$ provided by the data $\mathbf{x}$. By Laplace approximation of the integral around $\hat{\boldsymbol{\theta}}$ and by omitting the terms that do not depend on n, it can be shown that $P(\mathbf{x}|\mathcal{M}) \approx e^{\ell(\hat{\boldsymbol{\theta}})}n^{-k/2}$ for large n. Thus, when using a flat prior for $\mathcal{M}$ the probability becomes

$$P(\mathcal{M}|\mathbf{x}) \approx C \cdot e^{\ell(\hat{\boldsymbol{\theta}})}n^{-k/2} \propto e^{-\mathrm{BIC}/2},$$

where C is a proportionality constant (Wit et al., 2012). Thus, the model with the lowest BIC is the model with the highest posterior probability of being true.

Consistency vs. Efficiency

The greatest advantage of the BIC is that it is *consistent*, meaning that it will always select the true model when the sample size grows to infinity. This does not hold for the AIC, which has a much smaller penalty for increased complexity, and therefore has a non-zero probability of choosing an overparameterized variant of the true model. The consistency property is however only beneficial if the true model is known to be among the candidate models. In practice, this is rarely the case, and it is impossible to select the true model. In these cases, minimization of a loss function, such as mean squared error (MSE) of predictions, is of interest. The AIC is asymptotically *efficient* in prediction MSE when the true model is not in the candidate set, meaning that the prediction MSE is minimized given the candidate models. This is not a property of the BIC. Therefore, the AIC is preferable when the candidate models are assumed to be over-simplifications of the true model (Vrieze, 2012).

Comparison to Likelihood Ratio Tests

When two candidate models are nested, a likelihood ratio test (LRT) can be used to choose between the two. H_0 is that the least complex model is true, and H_1 is that the more generalized variant is true. H_0 is rejected in favour of the more

complex model if the difference in log-likelihoods is larger than some critical value depending on the difference in complexity. This is analogous to using the AIC or the BIC.

Assume we compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$, where $\mathcal{M}_2$ has one more parameter than $\mathcal{M}_1$. Using a significance level of $\alpha = 0.05$, $\mathcal{M}_2$ will be preferred by the LRT if

$$-2(\ell(\hat{\theta})_{\mathcal{M}_1} - \ell(\hat{\theta})_{\mathcal{M}_2}) > 3.84.$$

If the AIC was used, $\mathcal{M}_2$ would be preferred if $\text{AIC}_{\mathcal{M}_2} < \text{AIC}_{\mathcal{M}_1}$, i.e.

$$-2(\ell(\hat{\theta})_{\mathcal{M}_1} - \ell(\hat{\theta})_{\mathcal{M}_2}) > 2.$$

This corresponds to a significance level of 0.16 in the likelihood ratio test. Thus, the AIC is far less conservative than the LRT.

Now, suppose the sample size of the study was $n = 100$. The BIC would favour $\mathcal{M}_2$ if

$$-2(\ell(\hat{\theta})_{\mathcal{M}_1} - \ell(\hat{\theta})_{\mathcal{M}_2}) > 4.61,$$

corresponding to a significance level of 0.03. Furthermore, $n = 1000$ would correspond to a significance level of 0.009 and $n = 10000$ would correspond to a significance level of 0.002. The BIC becomes more conservative then the LRT already at $n = 47$.

This example illustrates the importance of being aware of the different behaviour of the three methods. Each of them are suitable for different goals. The LRT answers whether the more complex model gives a significantly better model fit. The AIC attempts to answer if higher complexity is beneficial for prediction purposes. The BIC attempts to find the true model among the candidates.

3.5.2 Prediction and Scoring Rules

Assume we have fitted a model with distribution $g(y_i|\mathbf{x}_i)$ to a dataset. Let $\mathbf{x}_0$ be a new sample. In many cases, it can be of interest to predict its corresponding response y_0. One possibility is to report a point estimate $\hat{y}_0$, which typically is the mean, mode or median of $g(y|\mathbf{x}_0)$. The performance of the prediction can be reported as the absolute or squared value of the difference $\hat{y}_0 - y_0$. This makes it possible to compare how close $\hat{y}_0$ falls to y_0 for different models, but it does not say anything about how well the models reflect the truth.

Another possible approach is probabilistic prediction, where the entire distribution $g(y|\mathbf{x}_0)$ represents our *predictive distribution* of y_0. This means that we expect y_0 to follow this distribution. A good predictive distribution is both *calibrated* and *sharp* (Gneiting and Raftery, 2007). Calibration refers to the consistency between the predictive distribution and the observed value. A predictor is said to be calibrated if the observed values follow the predictive distribution. Sharpness refers to the concentration of the predictive distributions and reflects the confidence of the predictor. Calibration is obtained by choosing the right model for the data, while sharpness can be improved by gathering more information (e.g. including more relevant covariates). Gneiting and Raftery (2007) argue that the goal of prediction is

to maximize the sharpness subject to calibration, i.e. having the sharpest possible predictive distributions while being calibrated.

Let $G(y|\mathbf{x})$ be the cumulative density function corresponding to $g(y|\mathbf{x})$. Scoring rules assign numerical scores to the predictive cumulative distribution $G(y|\mathbf{x}_0)$ based on the true outcome y_0. This makes it possible to compare how well different predictive distributions perform with regard to both sharpness and calibration. We use the notation $s(G, y)$ for the score that is assigned to the predictive distribution G when y is the true outcome. The scores are negatively oriented, meaning that smaller is better. A scoring rule is said to be proper when the expected value of $s(G, y)$ for an observation $y \sim F$ is minimized when $G = F$.

A popular choice in cases where the response y is continuous is the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007). It is defined in terms of the cumulative density G and observation y as

$$\text{CRPS}(G, y) = \int_{-\infty}^{\infty} (G(z) - 1(z \geq y))^2 \, dz, \tag{3.30}$$

where $1(\cdot)$ is the indicator function. An illustration of how the score is calculated is provided in Figure 3.8. This scoring rule is proper and provides a direct way to compare different probabilistic predictions using only a single metric. It reaches its theoretical minimum of zero if $G(z) = 1(z \geq y)$, meaning that the response is predicted to equal y with no uncertainty. Illustrations of the resulting shape of the CRPS for different predictive distributions is shown in Figure 3.9.

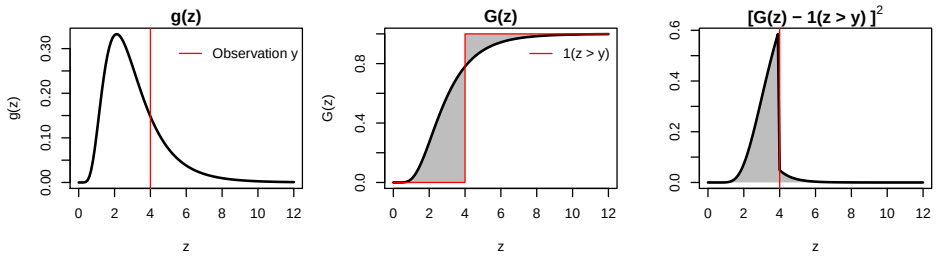

Figure 3.8: Illustration of how the CRPS is calculated based on the predictive distribution $g(z)$ and the observed value y. (a) The predictive distribution g(z) and the observed value y. (b) The cumulative predictive distribution $G(z)$ and the indicator function $1(z \geq y)$. The difference between the two functions is illustrated by the shaded area. (c) The integrand of the CRPS, $(G(z) - 1(z \geq y))^2$. The shaded area is the resulting CRPS.

3.6 Monte Carlo Simulation Studies

The term *Monte Carlo* is used to denote methods that rely on repeated random sampling to generate numerical results (Thomopoulos, 2012). There is a wide variety of Monte Carlo methods with countless important applications, such as estimating the posterior distribution in Bayesian analysis trough Markov Chain Monte

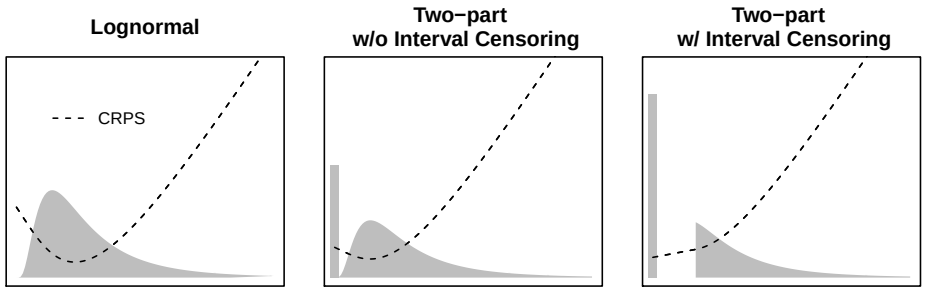

Figure 3.9: Illustration of the shape of the CRPS(G,y) for different predictive distributions.

Carlo (MCMC) methods and bootstrap resampling inferential methods (Boos and Stefanski, 2013).

Monte Carlo simulation studies, often referred to as Monte Carlo studies or just simulation studies, is an important application of Monte Carlo techniques for the purpose of studying how certain statistics depend on different factors. We will continue to use the term 'simulation studies'. The rationale behind simulation studies is to generate a large number of datasets mimicking a true population in order to perform empirical estimation of the sample distributions of various estimators of interest. Since the true distribution of the generated datasets is known, it is possible to evaluate the accuracy of the estimators. This is particularly useful in cases where it is unfeasible to derive the distribution of the estimators analytically (Burton et al., 2006). There are many ways of using simulation studies, including checking that the code operates as expected, evaluating new statistical methods to see if it works for the scenarios it is designed for, and comparative evaluation of statistical methods (Morris et al., 2017).

Simulation studies rely on large sample theorems (Boos and Stefanski, 2013), which in short state that the sample mean and variance of independent identically distributed variables converge to the true mean and variance as the number of samples goes to infinity. Thus, it is possible to estimate the average performance of an inference method by simulating independently generated datasets and taking the average of the performances to estimate the true mean. Likewise, the sample variance can be used to estimate the true variance of the performance.

We will continue to cover some fundamental large sample theory before we go into how to perform simulation studies.

Large Sample Theory

This section is based on the book by Boos and Stefanski (2013).

Let $X_1, ..., X_n$ be a sequence of independent identically distributed random variables with mean $\mu = E(X_i)$ and variance $\sigma^2 = \text{Var}(X_i)$. The Laws of Large Numbers guarantee that the sample mean $\bar{X}$ is close to the mean μ when n is large.

There are two forms of the law. The Strong Law of Large Numbers states that

$$P\Big(\lim_{n\to\infty} |\bar{X} - \mu| < \epsilon, \text{ for every } \epsilon > 0\Big) = 1,$$

which means that the events in which $\bar{X}$ does not approach μ has probability zero. This is known as *almost sure convergence*.

The Weak Law of Large Numbers states that

$$\lim_{n\to\infty} P\big(|\bar{X} - \mu| < \epsilon\big) = 1, \text{ for every } \epsilon > 0.$$

This means that the probability of the sample mean $\bar{X}$ being far away from the expected value μ becomes smaller and smaller as n increases, also known as *convergence in probability*. Clearly, the strong law implies the weak law, but not the converse.

Furthermore, the Central Limit Theorem describes the limiting behaviour of $\bar{X}$ when n tends to infinity. The theorem states that

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq t\right) = \mathbf{\Phi}(t),$$

where $\mathbf{\Phi}(t)$ is the standard normal cumulative distribution function. In other words, the asymptotic distribution of $\bar{X}$ is a normal distribution with the true average μ as mean and variance σ^2/n.

The Laws of Large Numbers and the Central Limit Theorem also applies to statistics that are asymptotically equivalent to averages. This includes functions of averages and statistics that are implicitly defined by averages.

Pseudo-Random Numbers

It is stated above that simulation studies are based on random sampling, but in practice we only have access to pseudo-random number generators. The term pseudo-random is used because the number generating mechanisms are fully deterministic, and can only generate numbers with similar behavior as truly random numbers (Gamerman and Lopes, 2006). Large sample theorems, such as the weak and strong laws of large numbers and the central limit theorem, in general, applies to pseudo-random numbers. Therefore, pseudo-random numbers are adequate for simulation studies (Boos and Stefanski, 2013).

While the applicability of pseudo-random numbers may be questioned in certain cases, Morris et al. (2017) argues that there are several advantages to using deterministic number generators in simulation studies. Most importantly, it is possible to pick a starting state for the random number generator such that the simulation can be re-run under that state. In many cases, this makes debugging and analysis easier, and it becomes possible for other researchers to get an exact reproduction of the results.

Study Design

Just like any other experiment, simulation studies require careful planning and analysis. Boos and Stefanski (2013) argue that simulation studies should be held to the same standard as other studies regarding, among other things, reproducibility and transparency. The tutorial on simulation studies by Morris et al. (2017) is the basis for this section. Some relevant notation is described in Table 3.1.

Table 3.1: Description of notation.

θ	The true value of the estimand
$\hat{\theta}_i$	The estimate of θ from the ith simulation
n_{obs}	Sample size of a simulated dataset
n_{sim}	Number of simulated datasets
$\text{Var}(\hat{\theta})$	The empirical long-run variance of $\hat{\theta}$
$\widehat{\text{Var}}(\hat{\theta}_i)$	The estimate of $\text{Var}(\hat{\theta})$ from the ith simulation

In simulation studies a large number n_{sim} of pseudo-random datasets $\{X_{i,1}, ..., X_{i,n_{\text{obs}}}\}$ are generated in order to calculate a set of statistics $\hat{\theta}_1, ..., \hat{\theta}_{n_{\text{sim}}}$. This is used to investigate the properties of the sampling distribution of the estimator $\hat{\theta}$. Morris et al. (2017) lists five important parts of planning a simulation study; Aims, data-generating mechanisms, methods, estimands, and performance measures. They are briefly described below:

(1) **Aim:** The aim of the study must be formulated based on what one wants to learn. Typical properties to investigate are bias, coverage, power, and variance.

(2) **Data-generating mechanism:** How the random datasets are going to be generated. In most cases this is done by random sampling from parametric distributions, but there are other possibilities such as repeated sampling from a specific dataset. The choice will depend on the aims of the study. It might produce realistic data for the sake of being relevant, or completely unrealistic data in order to stretch to a breaking point. Several methods may be used in order to cover different scenarios.

(3) **Methods:** Decide which methods/models for analysis to investigate. This requires knowledge of previous work on the area in order to include serious contenders.

(4) **Estimands:** Decide which estimands θ to investigate. This is chosen based on what the aims of the analysis are. When fitting a model this may be a specific regression parameter β if the aim is inference, or the fitted values $E(Y)$ if the aim is prediction.

(5) **Performance measures:** The performance measures are numerical quantities used to assess the performance of the methods. The choice depends on the target of the study. If the target is estimation, the performance measure

may be bias, mean squared error or coverage. If the target is testing, a typical performance measure is power.

Performance Measures and the MCSE

The Monte Carlo standard error (MCSE) denotes the uncertainty in the estimates due to using a finite number of simulations n_{sim}. Morris et al. (2017) stresses the importance of calculating and reporting this error, as no estimate should be reported without its corresponding uncertainty. Some examples of performance measures and their corresponding Monte Carlo standard errors are provided in this Table 3.2.

Table 3.2: Common performance measures and their corresponding Monte Carlo standard errors.

Performance Measure	Formula	Monte Carlo SE
Bias	$\frac{1}{n_{\mathsf{sim}}}\sum_{i=1}^{n_{\mathsf{sim}}}(\hat{\theta}_i - \theta)$	$\sqrt{\frac{1}{n_{\mathsf{sim}}(n_{\mathsf{sim}}-1)}\sum_{i=1}^{n_{\mathsf{sim}}}(\hat{\theta}_i - \theta)^2}$
Model SE	$\sqrt{\frac{1}{n_{\mathsf{sim}}}\sum_{i=1}^{n_{\mathsf{sim}}}\widehat{\mathsf{Var}}(\hat{\theta}_i)}$	$\sqrt{\frac{\mathsf{Var}[\widehat{\mathsf{Var}}(\hat{\theta}_i)]}{4n_{\mathsf{sim}}(\widehat{\mathsf{Model\ SE}})^2}}$
Empirical SE	$\sqrt{\frac{1}{n_{\mathsf{sim}}-1}\sum_{i=1}^{n_{\mathsf{sim}}}(\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\mathsf{Empirical\ SE}}{\sqrt{2(n_{\mathsf{sim}}-1)}}$
RelErrorSE	$100\left(\frac{\mathsf{ModSE}}{\mathsf{EmpSE}} - 1\right)$	$100\left(\frac{\mathsf{ModSE}}{\mathsf{EmpSE}}\right)\sqrt{\frac{\mathsf{Var}[\widehat{\mathsf{Var}}(\hat{\theta}_i)]}{4n_{\mathsf{sim}}(\widehat{\mathsf{ModSE}})^2} + \frac{1}{2(n_{\mathsf{sim}}-1)}}$
Coverage	$\frac{1}{n_{\mathsf{sim}}}\sum_{i=1}^{n_{\mathsf{sim}}}1(\hat{\theta}_{\mathsf{low},i} \leq \theta \leq \hat{\theta}_{\mathsf{upp},i})$	$\sqrt{\frac{\mathsf{Coverage}\times(1-\mathsf{Coverage})}{n_{\mathsf{sim}}}}$
Power	$\frac{1}{n_{\mathsf{sim}}}\sum_{i=1}^{n_{\mathsf{sim}}}1(p_i \leq \alpha)$	$\sqrt{\frac{\mathsf{Power}\times(1-\mathsf{Power})}{n_{\mathsf{sim}}}}$

The bias is the average deviation of the estimators from the true value. The optimal bias is zero, which means that the expected value of the estimator $\hat{\theta}_i$ equals the true parameter θ. However, one can often tolerate slight biases because of other desirable qualities, such as better error prediction. The model SE denotes the mean of the estimated standard errors, and quantifies the level of confidence in the parameter estimate $\hat{\theta}_i$. The true observed standard error of $\hat{\theta}_i$ is called the empirical SE, which is the observed variation in $\hat{\theta}_i$. In order to assess whether the model SE correctly estimates the variance in $\hat{\theta}_i$, the relative percentage error can be computed, here denoted as RelErrorSE.

A useful way to assess the joint performance of the estimator $\hat{\theta}$ and its estimated

standard error $\widehat{\text{SE}}(\hat{\theta})$, is the proportion of confidence intervals that contains the true value, called the coverage. A Wald-type confidence interval is given by

$$[\hat{\theta}_{\text{low},i}, \hat{\theta}_{\text{upp},i}] = [\hat{\theta}_i - z_{\alpha/2}\widehat{\text{SE}}(\hat{\theta}_i), \hat{\theta}_i + z_{\alpha/2}\widehat{\text{SE}}(\hat{\theta}_i)],$$

where α is the level of significance and $z_{\alpha/2}$ is the critical value of the standard normal distribution. The coverage is an estimate of the probability that this interval contains the true value. If all assumptions hold this probability is $100(1-\alpha)\%$. Under-coverage may happen due to bias in the estimates and/or under-estimation of the standard error. Over-coverage is a result of over-estimation of the standard error.

The best models in terms of coverage are models that give unbiased estimates with correctly estimated standard errors. The size of the standard error is irrelevant, as long as it correctly reflects the confidence in the corresponding estimate. A small standard error is however beneficial if one wants to target a hypothesis test. A Wald-type p-value from testing the significance of $\hat{\theta}_i$ is given by

$$p_i = 2\left(1 - \mathbf{\Phi}\left(\frac{|\hat{\theta}_i|}{\widehat{\text{SE}}(\hat{\theta}_i)}\right)\right).$$

The power of a model is the proportion of hypothesis tests that concludes that $\hat{\theta}_i$ is significantly different from zero. All other things being equal, a greater power is better. However, the power must be seen in context with the other performance measures. If a higher power is the result of biased estimates or under-estimation of the standard error, it does not make the model better.

Simulation Sample Size

Morris et al. (2017) state that the simulation sample size n_{sim} should be chosen based on the resulting Monte Carlo SE. This can be done by performing a preliminary simulation study with low n_{sim} in order to get an estimate of the performance measure, and then solving the MCSE for n_{sim}. For instance, if our key performance measure is bias, and our preliminary analysis shows that $\text{Var}(\hat{\theta}) = \frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}}(\hat{\theta}_i - \theta)^2 < 0.04$, the number of simulations needed to get an MCSE of at most 0.005 is

$$n_{\text{sim}} = \frac{\text{Var}(\hat{\theta})}{\text{MCSE}^2} > \frac{0.04}{0.005^2} = 1600.$$

Similar computations can be done for any performance measure of interest.

Analyzing the Results

There are several possible choices of software packages for performing simulation studies. According to a review of simulation studies in *Statistics in Medicine* performed by Morris et al. (2017) the most common choices are R and SAS. In this thesis R will be used. Guides for performing simulation studies in R are provided

by e.g. Hallgren (2013) and Abonazel (2018). There are multiple add-on packages in R for performing different types of simulation studies. In particular the package `rsimsum` (Gasparini, 2018), which is modeled upon the command `simsum` (White, 2010) in Stata, is useful for summarizing the results of a simulation study and calculating performance measures and their corresponding Monte Carlo standard errors.

4 Design of Simulation Study

4.1 Models

In this simulation study, we will compare the performance of four different models. An overview of the models and their parameters is provided in Table 4.1. The binary mixture model defined in (3.5) is included, as well as the variant with interval censoring defined in (3.7). They are hereafter denoted as respectively the TP model and the TPIC model for brevity. The TP model models the observed data directly, while the TPIC model models a latent variable that is assumed to be left-censored on $[0, T]$. Two one-part models are also included in the study, the Tobit model described in equation (3.2) and the substitute value approach described in Section 3.1.4 with substitute value $T/2$. Both of these models assume that all the observations arise from a single process, but they handle the zeroes very differently. The Tobit model handles the zeroes by assuming that they are left-censored observations from a latent distribution, and can therefore be viewed as a one-part variant of the TPIC model. The substitute model sets all the zeroes to the value $T/2$ and fits a continuous distribution to the observed data.

Table 4.1: An overview of the methods considered in the simulation study.

		Model Parameters	
Abbrivation	Full Name	Discrete Part	Continuous Part
TP	Two-Part Model	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$, σ (conditional)
TPIC	Two-Part Model w/ Interval Censoring	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$, σ (conditional)
Tobit	Tobit Model	—	$\boldsymbol{\gamma}$, σ (marginal)
Substitute	Substitute Model ($S = T/2$)	—	$\boldsymbol{\gamma}$, σ (marginal)

Because the models are based on different assumptions, their parameters have quite different interpretations. One important distinction is that the parameters in the continuous part the two-part models must be interpreted conditional on having observed a positive response, whereas the one-part models are parameterized in terms of the marginal means. Therefore, the parameters are not directly comparable across the models. In order to determine whether the estimated parameters of the one-part models are correct, they must be compared to the true marginal effects of the covariates in the data.

4.2 Simulated Datasets

All simulated data is generated from a standard two-part model with interval censoring. The continuous part is set to be lognormal, such that

$$\ln(Y_i^*)|Y_i^* > 0 = \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma).$$

A limit of detection is introduced such that the observed data is

$$Y_i = \begin{cases} 0, & Y_i^* \leq T \\ Y_i^*, & Y_i^* > T \end{cases}.$$

The discrete part is modelled with the same covariates,

$$\pi_i = P(Y_i^* = 0) = \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}).$$

The total amount of zeroes in the data is determined by both T and π_i:

$$\begin{aligned} P(Y_i = 0) &= P(Y_i^* = 0) + P(Y_i^* > 0)P(Y_i^* < T|Y_i^* > 0) \\ &= \pi_i + (1 - \pi_i)\boldsymbol{\Phi}\Big(\frac{\ln(T) - \mathbf{x}_i'\boldsymbol{\gamma}}{\sigma}\Big). \end{aligned}$$

In this model the expected value of the observation is

$$\begin{aligned} E(Y_i|\mathbf{x}_i) &= P(Y_i > T|\mathbf{x}_i)E(Y_i|Y_i > T, \mathbf{x}_i) \\ &= (1 - \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}))\left[1 - \boldsymbol{\Phi}\left(\frac{\ln(T) - \mathbf{x}_i'\boldsymbol{\gamma}}{\sigma} - \sigma\right)\right]e^{\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2/2} \\ &= (1 - \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}))(1 - \boldsymbol{\Phi}(Z_{\mathbf{x}} - \sigma))e^{\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2/2}, \end{aligned}$$

where $Z_{\mathbf{x}} = \frac{\ln(T) - \mathbf{x}_i'\boldsymbol{\gamma}}{\sigma}$. In general, it is not possible to calculate the marginal effect of one covariate on the expected value, as it depends on the values of the other covariates.

The main part of the analysis will be conducted on simulated datasets with only one binary covariate, where the covariate shifts the expected value in the same direction in both model parts. In practice, this means that the signs of the associated parameters in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are opposites. This is arguably the most intuitive, as an increased mean in the continuous part coincides with a reduced probability of being a true zero from the discrete component, and vice versa. This is a simple setup which makes it easy to interpret the result, and since it has only one covariate it is possible to compute the marginal effect of the covariate in order to accurately assess the performance of the one-part models. As a supplement, we will also simulate datasets where the effect of the binary covariate is in the opposite direction in each model part, in order to see how this affects the conclusions. Lastly, we will simulate datasets with multiple covariates to assess whether the results hold when more covariates are introduced. An overview of the three setups is provided in Table 4.2.

Table 4.2: Overview of simulated datasets. Four different values are used for T and β_0, labeled 1 - 4 in the table. The values are varied in a full factorial experiment, giving a total of 16 scenarios for each setup.

	Parameter																
	T				β_0				β_1	β_2	β_3	γ_0	γ_1	γ_2	γ_3	σ	n_{obs}
Setup	1	2	3	4	1	2	3	4									
1	0.5	1	2	3	-1.2	-0.7	-0.5	-0.2	-0.2	—	—	1	0.2	—	—	0.7	300
2	0.5	1	2	3	-1.4	-0.9	-0.7	-0.4	0.2	—	—	1	0.2	—	—	0.7	300
3	0.5	1	2	3	-1.2	-0.7	-0.4	-0.1	-0.2	0.4	-0.1	0.8	0.2	-0.2	0.1	0.7	300

Setup 1: One covariate, discrete and continuous effect in same direction.
Setup 2: One covariate, discrete and continuous effect in opposite directions.
Setup 3: Three covariates.

In all simulations the number of observations is set to $n_{\text{obs}} = 300$ and the variance of the error term ε_i is set to $\sigma = 0.7$, which mimics the behaviour of the data on the cytokine TNF-α. Two parameters, the LOD and the intercept β_0, are varied in a full factorial experiment with four values each, giving a total of 16 scenarios. The detection limit is set to $T = 0.5, 1, 2, 3$, and the parameters of $\boldsymbol{\gamma}$ are chosen to give approximately 0.5%, 5%, 30% and 50% of the continuous distribution being censored. The parameters of $\boldsymbol{\beta}$ are chosen such that approximately 10%, 20%, 30% and 40% of the observations arises from the discrete part. The effect sizes are chosen in order to give a power that is not too close to one or zero, such that it is possible to assess how the factors affect the power.

Setup 1: One Covariate

In the main part of the simulation study we set

$$\begin{aligned} \ln(Y_i^*)|Y_i^* > 0 &= \gamma_0 + \gamma_1 x_{1,i} + \varepsilon_i \\ &= 1 + 0.2x_{1,i} + \varepsilon_i, \end{aligned} \tag{4.1}$$

where $x_{1,i} \sim$ Bernoulli(0.5) is a binary covariate and $\varepsilon_i \sim \mathcal{N}(0, \sigma = 0.7)$ is a normally distributed error term. This results in approximately approximately 0.5%, 5%, 30% and 50% of the continuous distribution being censored, when the LOD is set to $T = 0.5$, 1, 2, and 3, respectively.

The discrete part is modelled with the same covariate, i.e

$$\begin{aligned} \pi_i = P(Y_i^* = 0) &= \boldsymbol{\Phi}(\beta_0 + \beta_1 x_{1,i}), \\ &= \boldsymbol{\Phi}(\beta_0 - 0.2x_{1,i}). \end{aligned} \tag{4.2}$$

For all datastes, β_1 is set to -0.2, while the intercept β_0 is set to -1.2, -0.7, -0.5 and -0.2 in order to give about 10, 20, 30 and 40 %, respectively, of the data arising from the discrete part of the distribution. A description of the 16 scenarios is provided in Table 4.3.

Naturally, the amount of zeroes increases with increasing LOD and with increasing discrete proportion, giving expected proportions of observed zeroes ranging from 10.3% to 68.9%. The marginal effect of the covariate increases with increasing LOD, because the amount of censored data increases most in the group with the

Table 4.3: Description of the 16 scenarios with only one covariate. The censored proportions are the approximate proportions of observations from the continuous part that falls below the LOD, and the discrete proportion is the overall approximate proportion of data that arises from the discrete part. The expected percent of observed zeroes from the underlying distribution is calculated from (4.2) and the overall marginal multiplicative effects of $x_{i,1}$ are calculated from the marginal effects in (3.9).

		β_0 = -1.2 10 % discrete		β_0 = -0.7 20 % discrete		β_0 = -0.5 30 % discrete		β_0 = -0.2 40 % discrete	
LOD	Censored proportion	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect
0.5	0.5 %	10.3 %	1.27	21.7 %	1.32	27.9 %	1.34	38.6 %	1.38
1	5 %	15.2 %	1.28	26.0 %	1.33	31.8 %	1.35	41.9 %	1.39
2	30 %	35.2 %	1.34	43.4 %	1.39	47.8 %	1.42	55.5 %	1.46
3	50 %	54.7 %	1.43	60.4 %	1.48	63.5 %	1.51	68.9 %	1.55

lowest mean, $x_{1,i} = 0$, which increases the difference between the two groups. Furthermore, the marginal effect increases with increasing discrete proportion, because the group difference is greater in the discrete part.

A preliminary simulation study with 100 simulations is performed in order to estimate the variance of the estimated parameters. The parameter with the largest estimated variance is found to be $\hat{\beta}_1$ in the TPIC model when $T = 3$ and $\beta_0 = -1.2$. In this scenario, the variance of $\hat{\beta}_1$ is found to be significantly smaller than 3.31. In order to achieve a Monte Carlo SE of at most 0.05 in the estimates of the bias in $\hat{\beta}_1$ we will need

$$n_{\text{sim}} = \frac{\text{Var}(\hat{\beta}_1)}{\text{MCSE}^2} > \frac{3.31}{0.05^2} = 1324.$$

Most of the estimates will get a considerably lower MCSE. For instance, the variance in $\hat{\gamma}_1$ in the same scenario is estimated to be significantly smaller than 0.034, which with $n_{\text{sim}} = 1324$ simulations will give a MCSE of at most

$$\text{MCSE} = \sqrt{\frac{\text{Var}(\hat{\gamma}_1)}{n_{\text{sim}}}} < \sqrt{\frac{0.034}{1324}} = 0.005$$

in the estimate of the bias in $\hat{\gamma}_1$. In order to ensure a satisfactory confidence in the results with a comfortable margin we will perform the simulation study with $n_{\text{sim}} = 2000$.

Setup 2: Discrete and Continuous Effect in Opposite Direction

In the data described in the previous section, the covariate $x_{1,i}$ shifted the mean in the same direction in both model parts. γ_1 was positive, such that $x_i = 1$ increased the mean of the continuous part, and β_1 was negative such that $x_i = 1$ decreased the probability of belonging to the discrete part. It is however entirely possible that the effect is opposite in the two model parts. Imagine that x_i represents sex and that the response y_i is a symptom of some disease. It is clearly possible that females have a higher chance of catching the disease, while males have a greater

symptom severity given that they have the disease. In such cases, sex might not affect the marginal mean, even if it affects each model part.

We will now take a look into what happens in such scenarios by changing the sign of β_1, such that $\beta_1 = 0.2$. In order to keep the proportion of data from the discrete part similar the values of β_0 are shifted in the opposite direction and are now set to -1.4, -0.9, -0.7 and -0.4. Everything else is kept equal. The scenarios are described in Table 4.4.

Table 4.4: Description of the 16 scenarios with one covariate and effects in opposite directions. The censored proportions are the approximate proportions of observations from the continuous part that falls below the LOD, and the discrete proportion is the overall approximate proportion of data that arises from the discrete part. The expected percent of observed zeroes from the underlying distribution is calculated as $P(Y_i^* = 0) = \boldsymbol{\Phi}(\beta_0 + 0.2x_{1,i})$, and the overall marginal multiplicative effects of $x_{i,1}$ are calculated from the marginal effects in (3.9).

		β_0 = -1.4 10 % discrete		β_0 = -0.9 20 % discrete		β_0 = -0.7 30 % discrete		β_0 = -0.4 40 % discrete	
LOD	Censored proportion	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect	Percent zeroes	Marginal effect
0.5	0.5 %	10.3 %	1.18	21.7 %	1.14	27.9 %	1.11	38.6 %	1.08
1	5 %	15.2 %	1.19	26.1 %	1.14	31.9 %	1.12	42.0 %	1.09
2	30 %	35.4 %	1.24	43.7 %	1.20	48.2 %	1.18	55.9 %	1.14
3	50 %	54.9 %	1.32	60.7 %	1.28	63.9 %	1.25	69.3 %	1.21

The observed proportions of zeroes are almost identical to the previous setting, which is expected since the discrete proportion and the proportion of censored observations from the continuous part is kept equal. As before, the multiplicative marginal effect increases with increasing LOD. This setup differs from the previous one in that the marginal effect decreases with increasing discrete proportion, because the group with the highest mean in the continuous part has a greater chance of belonging to the discrete part. Thus, the groups have an almost identical marginal mean when $T = 0.5$ and $\beta_0 = -0.4$.

We perform a preliminary simulation study with $n_{\text{sim}} = 100$ simulations. As before, the parameter with the highest estimated variance is $\hat{\beta}_1$ in the TPIC model when $T = 0.5$ and $\beta_1 = -1.2$. It is found to be significantly lower than 2.46. Thus, we conclude that $n_{\text{sim}} = 2000$ will give satisfactory confidence in the results in this setup as well.

Setup 3: Multiple covariates

In the data described so far only one covariate has been included. This made it possible to calculate the marginal effects for all methods and directly compare the results across all models. However, an analysis will often contain more than one covariate. Therefore, we will investigate what happens when multiple covariates are included. The data is generated by letting

$$\begin{aligned}\ln(Y_i^*)|Y_i^* > 0 &= \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i} + \varepsilon_i \\ &= 0.8 + 0.2x_{1,i} - 0.2x_{2,i} + 0.1x_{3,i} + \varepsilon_i\end{aligned}$$

where $x_1 \sim$ Bernoulli(0.5), $x_2 \sim \mathcal{N}(0,1)$, $x_3 \sim$ Pois(1) and $\varepsilon_i \sim N(0, 0.7)$. As before, a limit of detection is introduced at $T = 0.5, 1, 2, 3$, such that the below observations are observed as zero. This gives approximately the same proportions of observations under the LOD in each scenario. The same covariates are included in the discrete part of the model, such that the probability of belonging to the discrete part is

$$\begin{aligned} \pi_i = P(Y_i^* = 0) &= \mathbf{\Phi}(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}) \\ &= \mathbf{\Phi}(\beta_0 - 0.2x_{1,i} + 0.4x_{2,i} - 0.1x_{3,i}), \end{aligned} \tag{4.3}$$

with intercepts $\beta_0 = -1.2, -0.7, -0.4, -0.1$ such that the proportion of observations from the discrete part is kept to approximately 10, 20, 30 and 40% as in the previous analysis. The scenarios are described in Table 4.5.

With this data it is not possible to calculate the effect of each covariate on the marginal mean, as the marginal effect of one covariate depends on the value of the other two covariates.

Table 4.5: Description of the 16 scenarios with multiple covariates. The censored proportion is the approximate proportion of observations from the continuous part that falls below the LOD, and the discrete proportion is the overall approximate proportion of data that arises from the discrete part. The expected percent of observed zeroes from the mixture distribution is calculated from (4.3) for each scenario.

		Discrete proportion			
LOD	Censored proportion	10 % (β_0 = -1.2)	20 % (β_0 = -0.7)	30 % (β_0 = -0.4)	40 % (β_0 = -0.1)
0.5	0.5 %	10.8 %	21.1 %	29.7 %	39.7 %
1	5 %	17.5 %	26.8 %	34.6 %	43.8 %
2	30 %	39.7 %	46.1 %	51.6 %	58.1 %
3	50 %	58.9 %	63.0 %	66.6 %	70.9 %

Since more covariates are included, the estimates are subject to more variance. A preliminary simulation study with $n_{\text{sim}} = 100$ simulations showed that the parameter with the largest estimated variance was $\hat{\beta}_1$ in the TPIC model when $T = 2$ and $\beta_0 = -1.2$. Its variance was estimated to be significantly lower than 123.2. In order to achieve a Monte Carlo SE of at most 0.05 we will need

$$n_{\text{sim}} = \frac{\text{Var}(\hat{\beta}_1)}{\text{MCSE}^2} > \frac{123.2}{0.05^2} = 49280.$$

A variance of 132.2 on a parameter with value -0.2 is however a clear sign that the model does not behave well in this scenario. Therefore, aiming to achieve a low MCSE in this case is arguably an unnecessarily strict requirement. If we consider the TP model instead, the parameter with the highest estimated variance was $\hat{\beta}_1$ when $T = 0.5$ and $\beta_0 = -1.2$. Its variance was estimated to be significantly lower than 0.07, thus requiring only

$$n_{\text{sim}} = \frac{\text{Var}(\hat{\beta}_1)}{\text{MCSE}^2} > \frac{0.07}{0.05^2} = 28$$

in order to achieve a MCSE of at most 0.05. Furthermore, a MCSE of at most 0.005 is achieved by using $n_{\text{sim}} > 2800$. The Tobit and Substitute models have even lower estimated variances in the parameters. We therefore conclude that $n_{\text{sim}} = 3000$ will give satisfactory confidence in the results in all relevant cases.

5 Results from Simulation Study

The main focus of this simulation study is data with one covariate, as described in Table 4.2 as setup 1. First we will explore and visualize the raw results in Section 5.1, before we analyze the results for each model in terms of inference in Section 5.2 and prediction in Section 5.3. Furthermore, a briefer analysis of data with discrete effect in the opposite direction and data with multiple covariates is included in Section 5.4 and Section 5.5, respectively. The R code used to perform the simulation study is included in Appendix A.

5.1 Exploration and Visualization of Results

Before computing the performance measures, we explore the raw results. Recall that γ_1 is the effect of the covariate $x_{1,i}$ in the continuous part. In the two-part models, it represents the effect conditioned on being a non-zero response, while in the one-part models it represents the marginal effect of the covariate. Scatter plots of the estimated parameters $\hat{\gamma_1}$ versus the estimated standard errors $\widehat{\mathrm{SE}}(\hat{\gamma_1})$ are displayed in Figure 5.1, together with density plots of $\hat{\gamma_1}$.

There are several clear patterns in Figure 5.1. When the LOD is small, the two-part models are seemingly indistinguishable. The differences between the two-part models increase as the LOD increases. In short, the estimates from the TP model becomes more and more biased, and the TPIC gets a larger and larger variance in the results, which is also reflected in increased estimated standard errors. The one-part models are also close to equivalent when the LOD is low, and when the discrete proportion is low.

Scatter plots and marginal density plots of the estimates of β_1 are displayed in Figure 5.2. It is the effect of the covariate $x_{1,i}$ in the discrete part, and is therefore only present in the two-part models. As for γ_1, the results are indistinguishable for the smallest LOD. As the LOD increases, the TP model seems to get a slightly negative bias, and the TPIC model provides very dispersed results, accompanied by extremely large standard errors. There is a bump in the marginal density of $\hat{\beta}_1$ around -3 and 3 in the results for the largest LOD. Thus, the assumption of normally distributed estimates is clearly violated.

The estimates of the effect β_1 must be seen in context with the estimates of the intercept β_0. A scatterplot of the estimates of two parameters with marginal

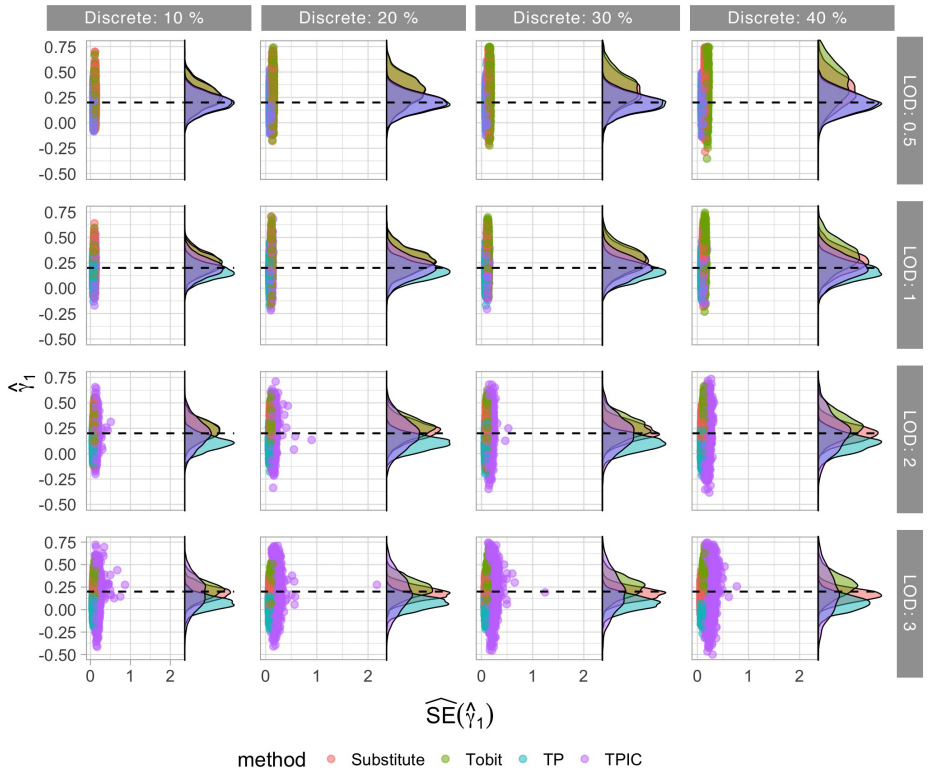

Figure 5.1: Scatter plot of of the estimated parameters $\hat{\gamma}_1$ and corresponding estimated standard errors $\widehat{\text{SE}}(\hat{\gamma}_1)$ with marginal densities of $\hat{\gamma}_1$. The true value of γ_1 is marked with the horizontal dashed line. The parameter γ_1 is the covariate effect in the continuous part, as specified in (4.1).

densities is provided in Figure 5.3. This figure shows that the highest values of $\hat{\beta}_1$, $\hat{\beta}_1 \approx 2$, are accompanied by very low estimates of the intercept, $\hat{\beta}_0 \approx -3.5$. In these cases the probability of belonging to the discrete part of the distribution for the group $x_i = 0$ is estimated to $\mathbf{\Phi}(-3.5) \approx 0$, while the probability for the group $x_i = 1$ is estimated to $\mathbf{\Phi}(-4 + 3) \approx 0.07$, which is close to the truth. The estimated effect $\hat{\beta}_1 \approx 3$ is however very misleading, as the true value is $\beta_1 = -0.2$. The lowest estimates of $\hat{\beta}_1 \approx -3$ are accompanied by relatively correct estimates of the intercept, $\hat{\beta}_0 \approx -1$. In these cases the estimated probability of belonging to the discrete part for the group $x_i = 0$ is close to the truth, while the estimated proportion when $x_i = 1$ is $\mathbf{\Phi}(-1 - 3) \approx 0$.

Failed solutions

The TPIC model fails to provide solutions for some datasets, either due to non-convergence or producing a Hessian matrix that is not negative-definite. The latter

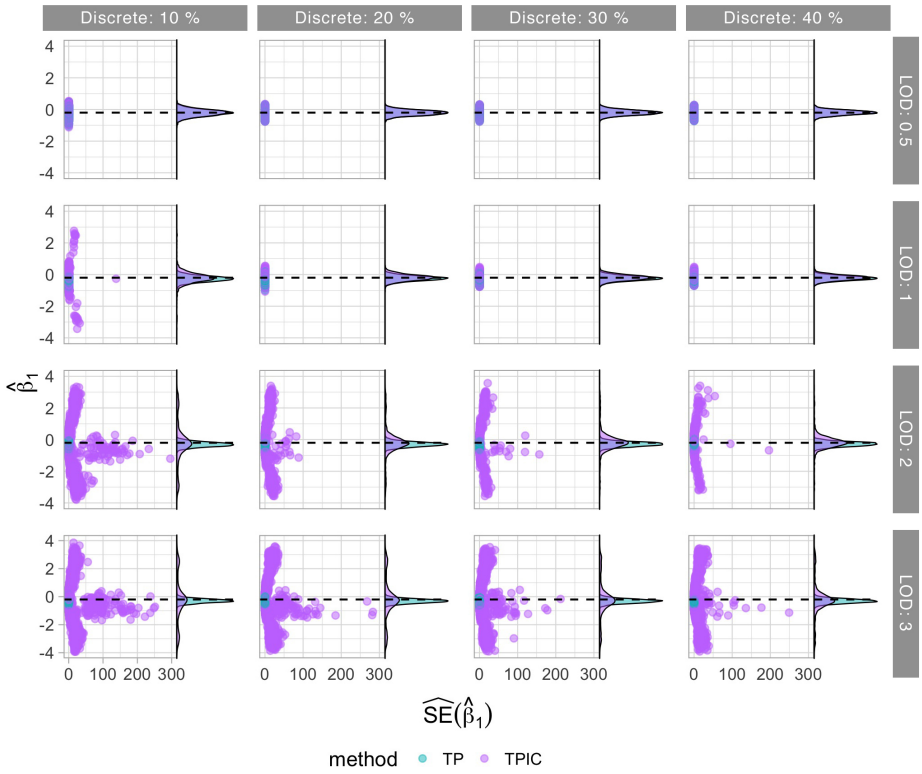

Figure 5.2: Scatter plot of of the estimated parameters $\hat{\beta}_1$ and corresponding estimated standard errors $\widehat{\text{SE}}(\hat{\beta}_1)$ with marginal densities of $\hat{\beta}_1$. The true value of β_1 is marked with the horizontal dashed lines.

is the result of not reaching a maximum of the likelihood function, and leads to negative estimates of the variance in one or more of the parameters. This clearly indicates that the results are not reliable. We therefore regard these solutions as missing, even if all convergence criteria are met. The amount of missing solutions in each scenario is listed in Table 5.1. The problem of non-convergence is clearly linked to the LOD, as there are no missing solutions for the two lowest LODs and missing solutions in all scenarios with the two highest LODs. The proportion also highest for the lowest discrete proportions.

A deeper look into the simulated data that resulted in failed solutions for the TPIC model reveals that in most cases the expected amount of zeroes based on the observed part of the continuous distribution is higher than or close to the observed amount of zeroes in the data for both groups. Thus, all the observed zeroes can be explained by left censoring of the continuous part, making the discrete part superfluous. This is in most cases caused by unexpectedly many observations close to the LOD, which makes the left tail of the continuous part appear heavier than it truly is. An example is provided in Figure 5.4. The figure illustrates how the

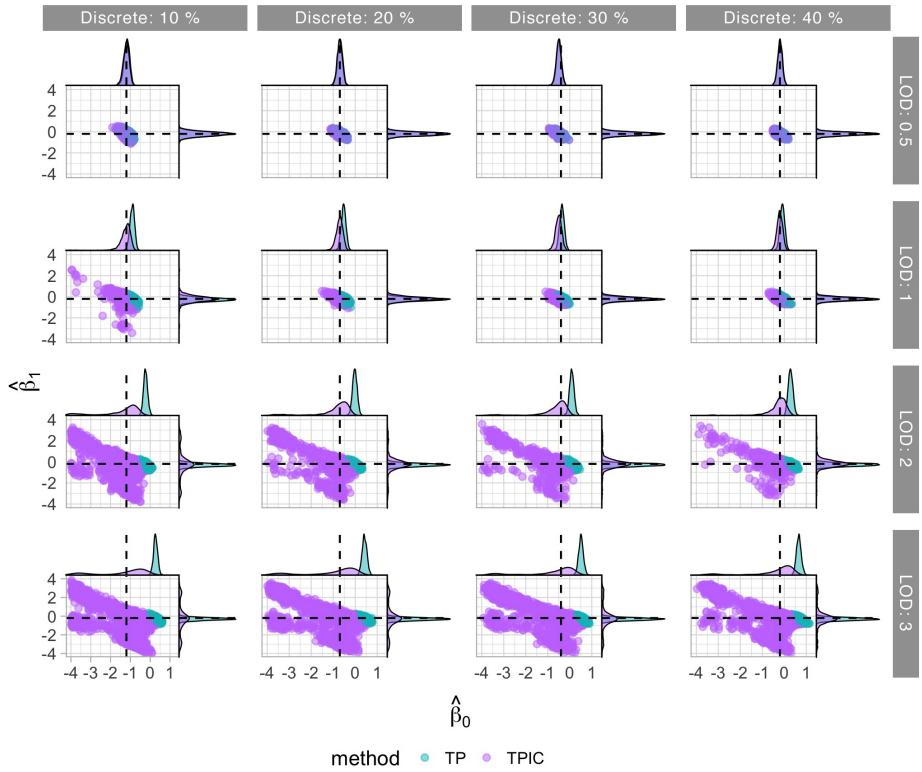

Figure 5.3: Scatter plot of of the estimated parameters $\hat{\beta}_1$ and $\hat{\beta}_0$ with marginal densities. The true values are marked with dashed lines.

Table 5.1: Amount of missing solutions with setup 1. Only the TPIC model leads to missing soultions with this setup.

TPIC				
	Discrete			
LOD	10 %	20 %	30 %	40 %
0.5	—	—	—	—
1	—	—	—	—
2	1.4 %	0.3 %	0.3 %	0.1 %
3	1.3 %	1.3 %	1.1 %	0.5 %

observed positive data has a distribution that leads to overestimation of the left tail of the continuous part of the distribution for both groups in the data.

In the scatter plots of $\hat{\beta}_1$ versus $\widehat{\text{SE}}(\hat{\beta}_1)$ in Figure 5.2 it is clear that there are some outliers in the data with extremely large estimated standard errors by the TPIC model. The highest estimated standard error is $\widehat{\text{SE}}(\hat{\beta}_1) = 295.9$. A standard error this high is a clear sign that the method has failed to converge properly. This small number of outliers shifts the mean estimated standard error substantially in

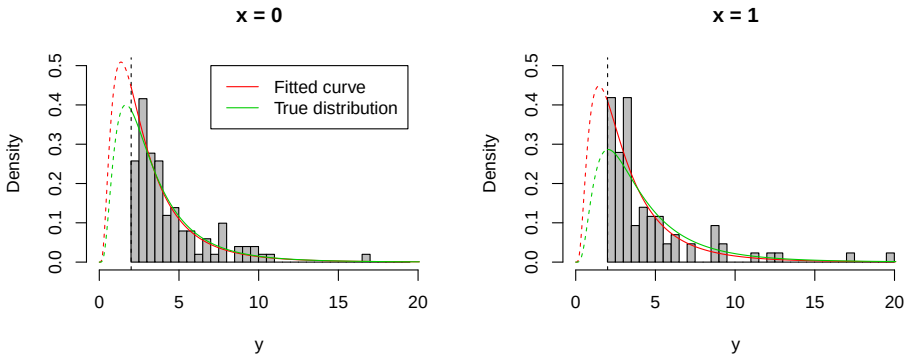

Figure 5.4: Illustration of a dataset that failed to provide a solution with the TPIC model. It shows the result of fitting a truncated lognormal distribution to the observed positive data (red curve), compared to the true underlying distribution of the continuous data (green curve). The predicted censored left tailes are shown with dashed lines. In this scenario $T = 2$ and $\beta_0 = -1.2$.

certain scenarios. Therefore, we choose to regard solutions with an estimated standard error larger than 10 standard deviations from the mean estimated standard error as failed in the further analysis. For instance, in the scenario with $T = 3$ and $\beta_0 = -0.2$ (40% discrete proportion) the estimated standard errors $\widehat{\text{SE}}(\hat{\beta}_1)$ from the TPIC model has mean of 4.64 and standard deviation of 11.6, which gives a cutoff at $\widehat{\text{SE}}(\hat{\beta}_1)_{\max} = 4.63 + 10 \cdot 11.6 = 120.6$. Three observations lie above this limit and are excluded in further analysis. This procedure is performed on all parameters across all methods and scenarios, resulting in a total of 46 additional missing solutions, all of which are from the TPIC model. In relative terms, this makes up 0.14% of all the simulated results.

5.2 Inference

In this section, we will investigate how the different models perform when the goal is to make inferences about the underlying process. In particular, it is of interest to determine how well the models are able to distinguish between the two groups in the data. In the two-part models, the group effect is quantified by γ_1 in the continuous part, conditioned on having observed a non-zero response, and by β_1 in the discrete part. In the one-part models, the group effect is quantified by only γ_1, in terms of how the covariate affects the marginal mean. The performance measures related to γ_1 are displayed in Figure 5.5 and the performance measures related to β_1 are displayed in Figure 5.6.

The parameter γ_1 have dissimilar interpretations in the one- and two-part models, and is therefore not expected to be the same across all models. In order to assess the performance of the one-part models, we must assess whether they cor-

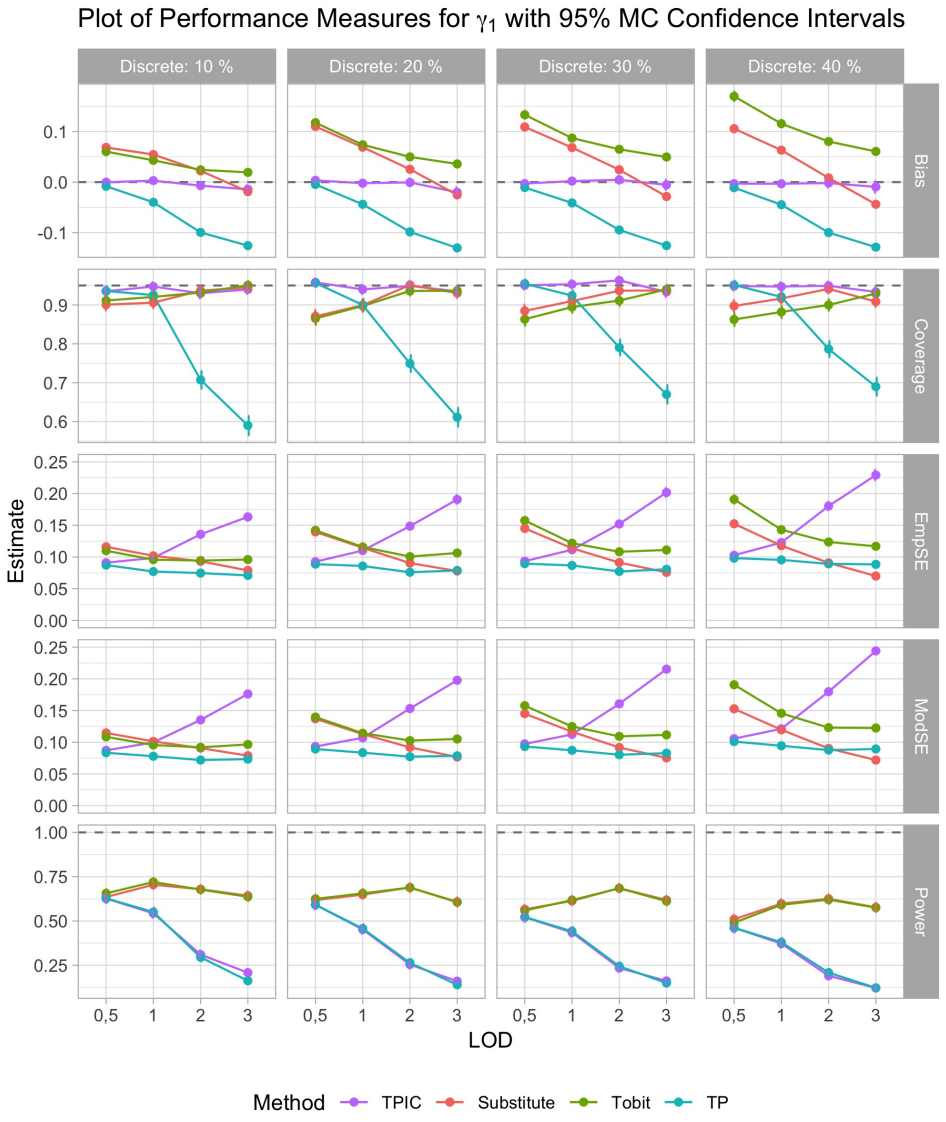

Figure 5.5: Performance measures related to γ_1 with 95 % Monte Carlo confidence intervals.

rectly estimate the marginal effect. The marginal expected values for each group and the multiplicative marginal effect of the group parameter is calculated for each model in every scenario. The resulting bias in the estimates are displayed in Figure 5.7. Fitted densities for each model is shown for selected scenarios in Figure 5.8.

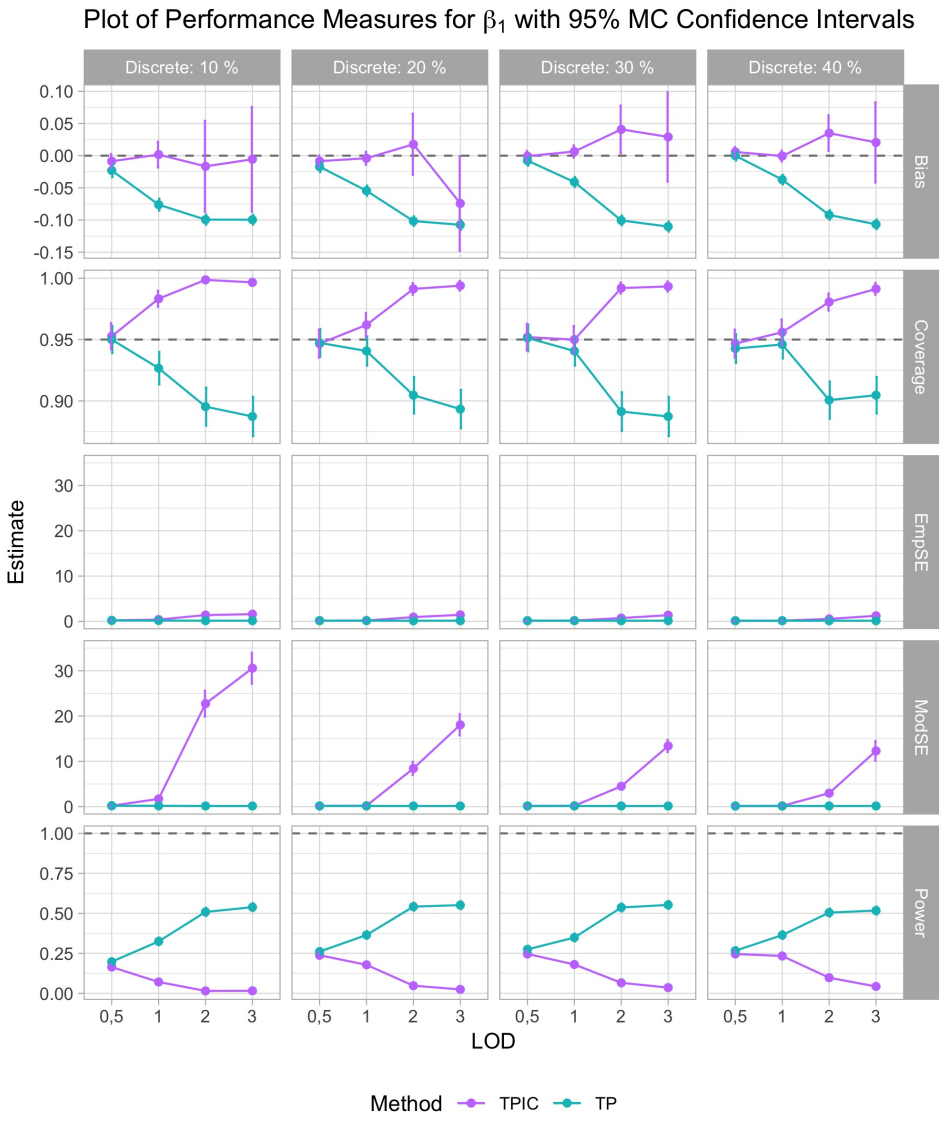

Figure 5.6: Performance measures related to β_1 with 95 % Monte Carlo confidence intervals.

5.2.1 Two-Part Model w/ Interval Censoring

The two-part model with interval censoring (TPIC) is the model used to generate the data for this simulation study. Therefore, it is expected that it gives unbiased estimates of γ_1 in all scenarios. The estimated standard errors $\widehat{\text{SE}}(\hat{\gamma}_1)$, called the model SE, increases with increasing LOD and with increasing discrete proportion.

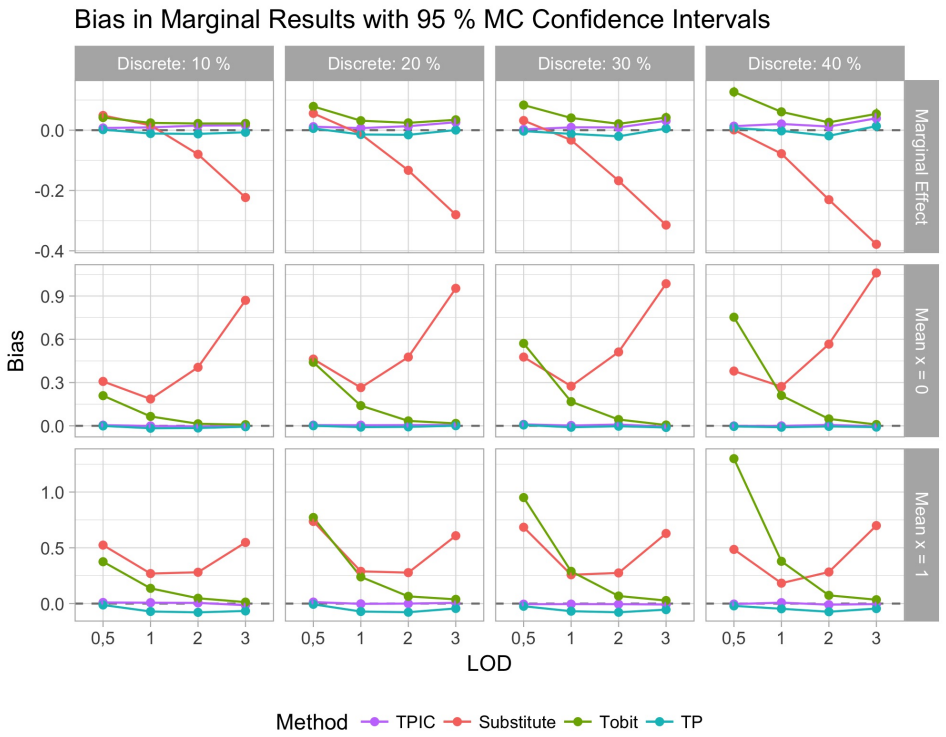

Figure 5.7: Bias in the estimated marginal means for each group and the marginal multiplicative group effect.

This is to be expected, as the amount of observed data from the continuous distribution decreases. As a result, the power decreases. The model SE is close to the empirical SE of $\hat{\gamma}_1$ in all scenarios, but the standard error is somewhat over-estimated for the largest LOD. This gives a slight decrease in the coverage, which otherwise is close to the expected 95%.

When it comes to the group effect in the discrete part of the distribution, β_1, the scatter plot in Figure 5.2 showed that the results for the TPIC model became very unstable for large LODs, especially when the discrete proportion is low. The performance measures displayed in Figure 5.6 gives a better insight into this behavior. The model SE is greatly over-estimated, with an average relative error of at worst 2000% when LOD = 3 and the discrete proportion is 10%. As a consequence, the coverage increases to almost a 100% and the power decreases to nearly zero for the highest LODs. Due to the high variance in the estimates, the MCSE in the estimated bias is also very large.

Given the unbiased estimates of β_1 and γ_1, it comes to no surprise that the model provides close to unbiased estimates of marginal means and the multiplicative marginal effect, as shown in Figure 5.7. The fitted densities in Figure 5.8 also

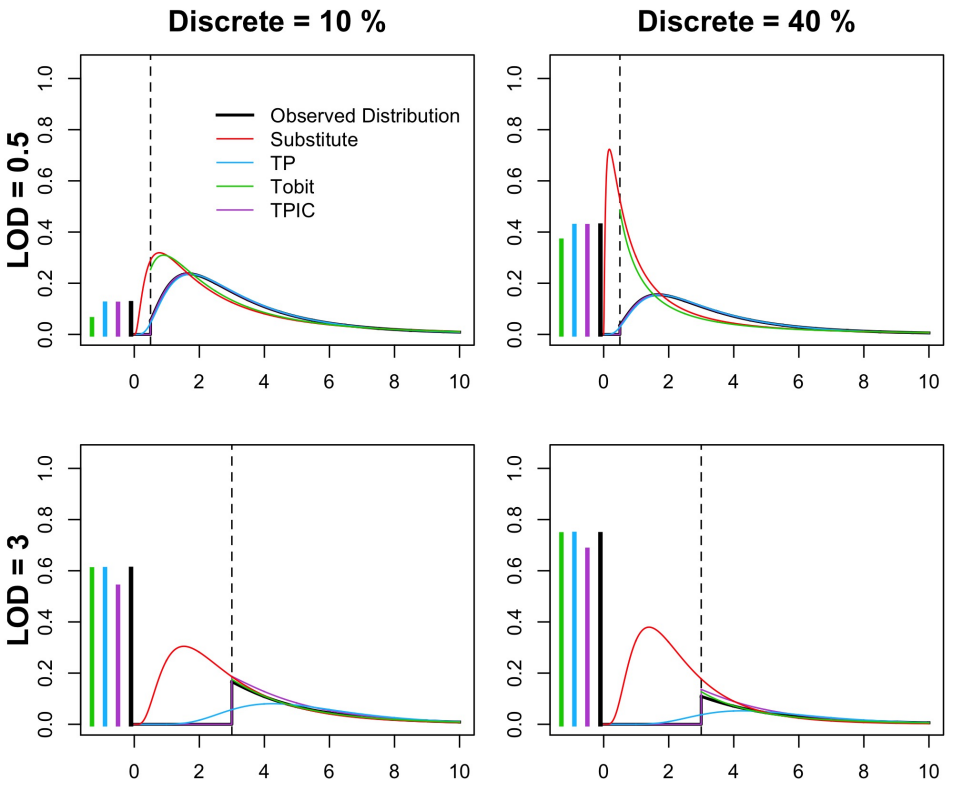

Figure 5.8: The fitted densities based on the mean of the parameter estimates in four of the scenarios when $x_{1,i} = 0$. The results when $x_{1,i} = 1$ look similar.

reveals a relatively good fit to the underlying curve. The discrete proportion is however on average somewhat under-estimated, due to a negative bias in β_0, which is evident in Figure 5.3.

The instability in the results when a large proportion of the censored observations come from the continuous distribution, can be taken as a clear sign that the model is over-parameterized in these scenarios. The model struggles to distinguish between the true zeroes from the discrete part of the distribution and the "false" zeroes that are censored observations from the continuous part. Therefore, the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are highly unstable.

5.2.2 Two-Part Model

The underlying assumptions of the two-part (TP) model differ from the distribution of generated datasets in that the entire continuous part is assumed to be observed. The consequences of this misassumption are expected to amplify as the LOD increases.

From Figure 5.5 it is evident that γ_1 becomes increasingly under-estimated as the LOD increases. As a consequence, the coverage decreases. This is a consequence of attempting to fit a lognormal distribution to a truncated lognormal distribution. It is an interesting result that the TP model has equally good power as the TPIC model, even if the effect is underestimated. This is because the power is a function of both the estimate $\hat{\gamma}_1$ and its corresponding estimated standard error $\widehat{\text{SE}}(\hat{\gamma}_1)$. The model SE of the TP model does not increase with increasing LOD, and consequently, it suffers no more loss in power than the TPIC model. The model SE of the TP model correctly estimates the empirical SE in all scenarios.

The effect β_1 in the discrete part of the model is increasingly over-estimated as the LOD increases. In this model, the discrete part represents all the observed zeroes, including the observations that are truly censored observations from the continuous part. Therefore, β_1 represents the combined group effect on the probability of being observed as zero both from the discrete and continuous part. This leads to biased results and decreased coverage. In terms of power, the TP model performs a lot better than the TPIC model for high LODs, and will much more frequently find a significant group difference in the discrete part. This will however not be particularly useful, as it is not possible to determine whether the group difference is caused by the effect γ_1.

It is interesting to note that the estimated marginal means and the multiplicative marginal group effect is close to unbiased in all scenarios, despite the bias in both $\hat{\beta}_1$ and $\hat{\gamma}_1$. Clearly, the biases cancel out such that the resulting estimated means are unbiased.

Figure 5.8 illustrates how the TP model behaves in selected scenarios. For the lowest LOD it is indistinguishable from the true distribution, but for the largest LOD its disadvantages are clear. The lognormal distribution does not fit well to the truncated-lognormal distribution, and therefore gives misleading results. However, the results from the TP model is shown to provide much more stable results than the TPIC model for the high LODs, which nevertheless can make the TP model a better choice.

5.2.3 Tobit Model

The Tobit model is a special case of the TPIC model where the discrete part is not present, such that the latent continuous distribution accounts for all the observed zeroes. This means that the entire group difference is quantified by γ_1, which is the marginal effect of the $x_{1,i}$. Therefore, this model is expected to give biased estimates for γ_1, which is confirmed in Figure 5.5. The parameter is over-estimated in all scenarios, as it also includes the effect in the discrete part of the model. The estimated effect decreases with increasing LOD and increases with increasing discrete proportion. The former can be explained by that discrete part of the data constitutes a smaller portion of the observed zeroes as the LOD increases, and therefore the effect in the discrete part becomes less prominent. Likewise, $\hat{\gamma}_1$ increases with increasing discrete proportion, because this makes the discrete part more prominent. The model SE, $\widehat{\text{SE}}(\hat{\gamma}_1)$, decreases with increasing LOD, which

is in agreement with the empirical SE. This leads to the power being relatively similar in all scenarios, between 50% and 75%.

The behavior of the Tobit model is nicely illustrated by the fitted curves in Figure 5.8. For the lowest LOD, the curve is considerably left-skewed, because the proportion of zeroes is way higher than what is expected from the continuous part alone. This becomes worse as the discrete proportion increases. For the highest LOD, the Tobit model provides a good fit to the true distribution, as the proportion of zeroes is much closer to what is expected from the non-censored observations. This behavior is reflected in the bias in the marginal means in Figure 5.7. The estimated marginal means are close to unbiased for the highest LOD.

5.2.4 Substitute Model

In the substitute model, all zeroes are set to half the detection limit before a lognormal distribution is fitted. This is a common technique when the proportion of censored data is low. The model consists of one continuous part, where the group effect is quantified by γ_1. Thus, γ_1 represents the effect on the marginal mean, as opposed to the conditional effect which γ_1 underlying distribution of the data. Therefore, we expect the estimated parameters of γ_1 by the substitute model to be biased estimates of the underlying parameter γ_1 used to generate the data. In Figure 5.5, we see that the estimated effect decreases as the LOD increases, which is expected since the distributions of the two groups become more similar when a larger proportion falls below the LOD and is set to $T/2$.

When the LOD and the discrete proportion is low, all the performance measures are very similar as for the Tobit model. The differences between the one-part models become greater as the LOD and discrete proportion increases. In particular, the substitute model estimates a smaller group effect γ_1 than the Tobit model, as well as smaller corresponding estimated standard errors. In the substitute model, all the zeroes are fixed to $T/2$ for both groups in the data, while the distribution below the LOD is flexible in the Tobit model. Therefore, it is justifiable that the Tobit model detects a greater difference between the groups. In terms of power, the one-part models have almost identical performance.

The substitute model over-estimates the marginal means for both groups (see Figure 5.7). This is expected as all the zeroes are taken to be $T/2$. For the lowest LODs, the marginal effect is however close to correctly estimated, but as the LOD increases, the multiplicative marginal effect is greatly underestimated, due to decreasing estimates $\hat{\gamma}_1$.

5.2.5 Concluding Remarks

As expected, the full TPIC model has the overall least biased estimates, as this is the model used to generate the data. The model does however not behave well when the discrete part of the model constitutes a smaller proportion of the total amount of zeroes. These problems seem to kick in when less than 70% of the total amount of zeroes arises from the discrete part of the model, and the analysis becomes very problematic when less than half of the zeroes arises from the discrete

part. In such scenarios, the TPIC model is clearly over-parameterized, and one should either gather more data or use another model.

The Tobit model gives good results for larger LODs, and can therefore be a good alternative when the TPIC model is over-parameterized. It provides nearly unbiased estimates of the marginal means, and it achieves a much better power than the TPIC model. A drawback with this approach is that the Tobit model consists of only one part, such that both the effect in the discrete part and the continuous part is contained in one parameter. In situations where it is desirable to isolate these effects, the TP model can be a better option. It is however important to note that the TP model gives biased results for higher LODs. Already when $T = 1$ and 5% of the observations from the continuous part of the data is censored, the TP model has a substantial drop in performance in terms of bias and coverage.

The substitute model is inferior in all scenarios, but when the LOD is low it provides good estimates of the marginal effect. It is by far the easiest model to apply since all familiar regression techniques can be used on the data after substituting the zeroes with $T/2$. Here we have shown that it can be a valid choice when the proportion of censored observations is low.

5.3 Prediction

So far, we have compared the models in terms of inference. There are however many situations where the goal is not to correctly identify the underlying properties of the distribution, but rather to be able to predict the outcome of a new observation. In order to compare how well the models are able to do this, we simulate new data from the same underlying distribution and compare their mean continuous rank probability score (CRPS) (3.30) for each scenario. The results are displayed in Figure 5.9. Recall that a lower score is better.

Figure 5.9 shows that the two-part models achieve a close to optimal score in all scenarios, with slightly better scores for the TPIC model than the TP model when the LOD increases. The Tobit model performs poorly for the lowest LOD, but achieves the lowest score for the highest LODs, i.e. when the proportion of censored data form the continuous distribution exceeds 30 %. The substitute model is inferior in all scenarios, but performs considerably better for low LODs than high LODs.

5.4 Discrete and Continuous Effect in Opposite Directions

The setup is displayed in Table 4.2 as Setup 2, and the data in the 16 scenarios is described in Table 4.4. It differs from the previously analyzed data in that the sign of β_1 is reversed such that $\beta_1 = 0.2$, and the values of β_0 are shifted such that the weight of the point mass at zero is the same. As a consequence, the marginal effect of $x_{1,i}$ now decreases with increasing discrete proportion, thus giving a close to zero marginal effect in some scenarios. The bias in the estimates of γ_1 and β_1

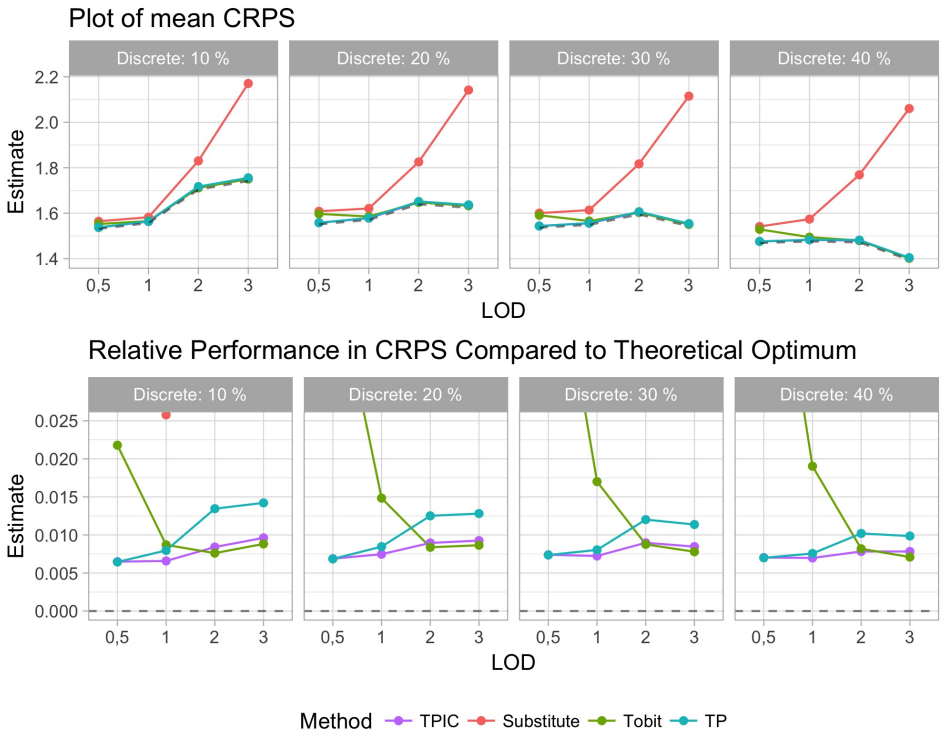

Figure 5.9: The top panel shows the mean CRPS for each method and scenario, as well as the theoretical optimal mean CRPS achieved by the true distribution represented by the dashed lines. In the bottom panel, the theoretical optimal mean CRPS is subtracted, so that the loss in performance compared to the true model is shown. The plot is zoomed in so that the differences between the methods are visible.

is shown in Figure 5.10 and the power in the same parameters is shown in Figure 5.11.

The performance of the TPIC model is overall similar as in the previous analysis. It provides close to unbiased estimates of γ_1, but suffers great instability for the larger LODs, especially in $\hat{\beta}_1$. The TP model performs equally good in estimating γ_1 as in the previous setup, but the results for β_1 are highly misleading for the higher LODs. Recall that the true value is $\beta_1 = 0.2$. When $T = 1$ the TP model estimates β_1 to be close to zero, as the increased probability of belonging to the discrete part associated with $x_{1,i} = 1$ is canceled out by the decreased probability of being censored from the continuous part. For the highest LODs β_1 is estimated to be negative, because the increased probability of belonging to the discrete part is dominated by the decreased probability of being censored from the discrete proportion.

In the previously studied scenarios, the one-part models achieved better power

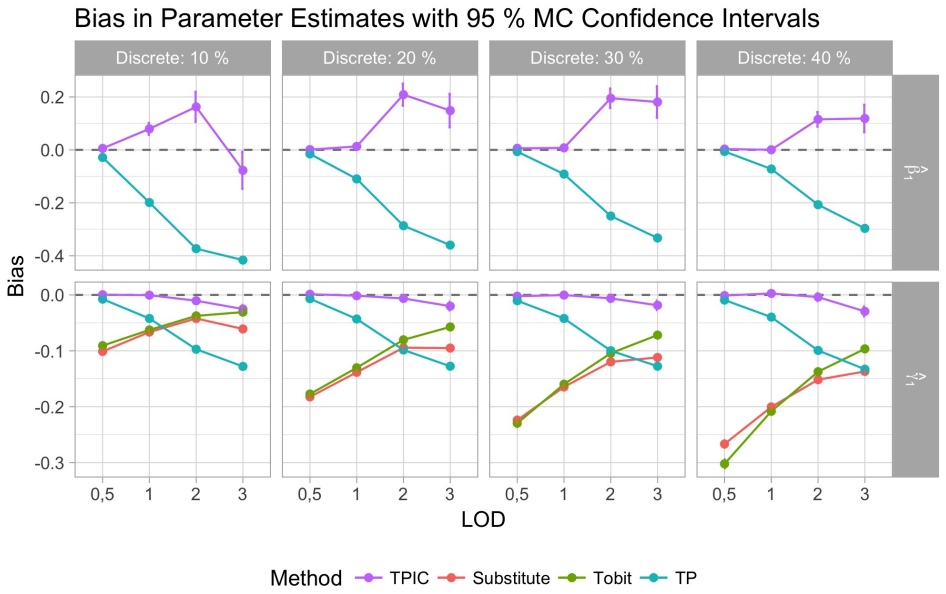

Figure 5.10: Bias in estimates of β_1 and γ_1 in setup 2. The true values are $\beta_1 = 0.2$ and $\gamma_1 = 0.2$

than the two-part models and would therefore be preferable if power was the only concern. This was because combining both the discrete and the continuous effect into one parameter resulted in parameter estimates with greater magnitude. When the discrete effect and the continuous effect affect the mean in opposite directions, this result is reversed. The marginal effect on the mean is low because the effect of γ_1 is canceled by the effect of β_1, therefore the one-part models give small, or even negative, estimates for γ_1, resulting in low power in many scenarios, as shown in Figure 5.11. As a result, the two-part models performs best in terms of power in most scenarios, except for when $T = 3$ where the marginal effect is greatest.

This example illustrates why it might be important to isolate the effect in the discrete and continuous part of the underlying model. The Tobit or the substitute model is clearly not a good alternative to the TPIC model for in many scenarios, as they only detect the marginal effect of $x_{1,i}$. The TP model gives misleading estimates of β_1, but it achieves a just as good power in γ_1 as the TPIC model, without the issues related to convergence. The performance in prediction is similar as before, with the TPIC model performing best for the lowest LODs and the Tobit model being the best when $T = 3$.

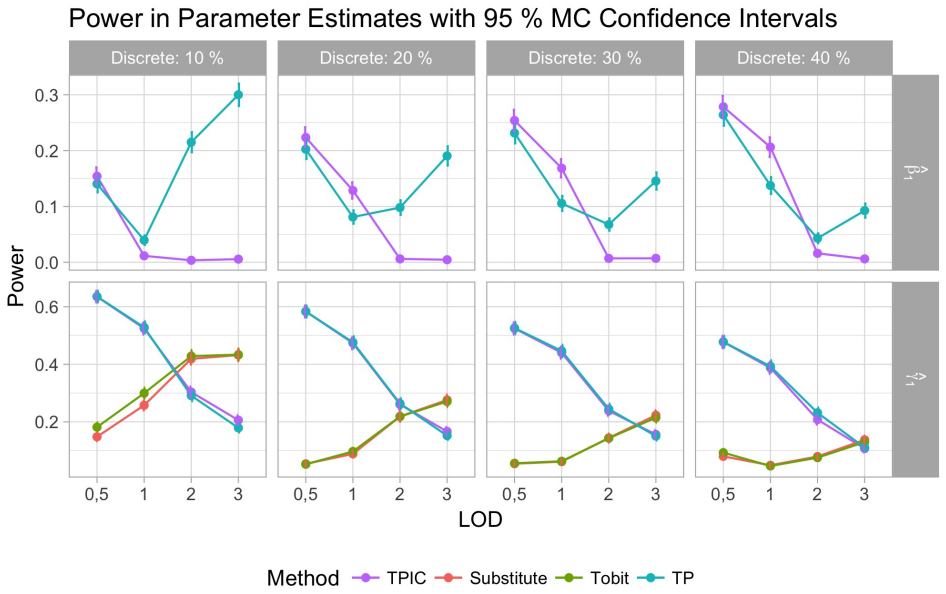

Figure 5.11: Power of estimates of β_1 and γ_1 in setup 2.

5.5 Multiple Covariates

The setup is displayed in Table 4.2 as Setup 3, and the data in the 16 scenarios is described in Table 4.5. Compared to the previous analysis, two additional covariates are included in both parts of the underlying model. This gives four additional parameters in the two-part models, and two additional parameters in the one-part models.

The overall results with multiple covariates are similar as with one covariate. In general, there is more variance in the results, as the models contain substantially more parameters. Most notably, the instability in the TPIC model performs considerably worse and there is a much larger proportion of failed solutions. The proportions of failed solutions are displayed in Table 5.2. Overall, the proportions are about ten times as large as when only one covariate was present. There is also a greater proportion of outliers with extremely large estimated standard errors. In total 171 solutions, making up 0.36% of the simulations, from the TPIC model has estimated standard errors greater than ten standard deviations from the mean estimated standard error. This makes it even more important to look for alternatives to the TPIC model when the discrete part of the model makes out a small proportion of the total amount of observed zeroes.

Figure 5.12 shows the bias in $\hat{\beta}_1$ and $\hat{\gamma}_1$ for all methods and scenarios when multiple covariates are present. The biases in $\hat{\gamma}_1$ are very similar to the biases when only one covariate was included, as shown in Figure 5.5. The bias in $\hat{\beta}_1$ in

Table 5.2: Amount of missing solutions with multiple covariates.

	TPIC			
	Discrete			
LOD	10 %	20 %	30 %	40 %
0.5	—	—	—	—
1	1.1 %	—	—	—
2	5.6 %	3.1 %	2.2 %	1.5 %
3	8.7 %	6.3 %	5.5 %	5.1%

the TP model is also similar to before, with an increasingly negative bias as the LOD increases. With only one covariate the estimates of β_1 by the TPIC model were close to unbiased. When more covariates are introduced $\hat{\beta}_1$ gets a substantial negative bias for the largest LODs.

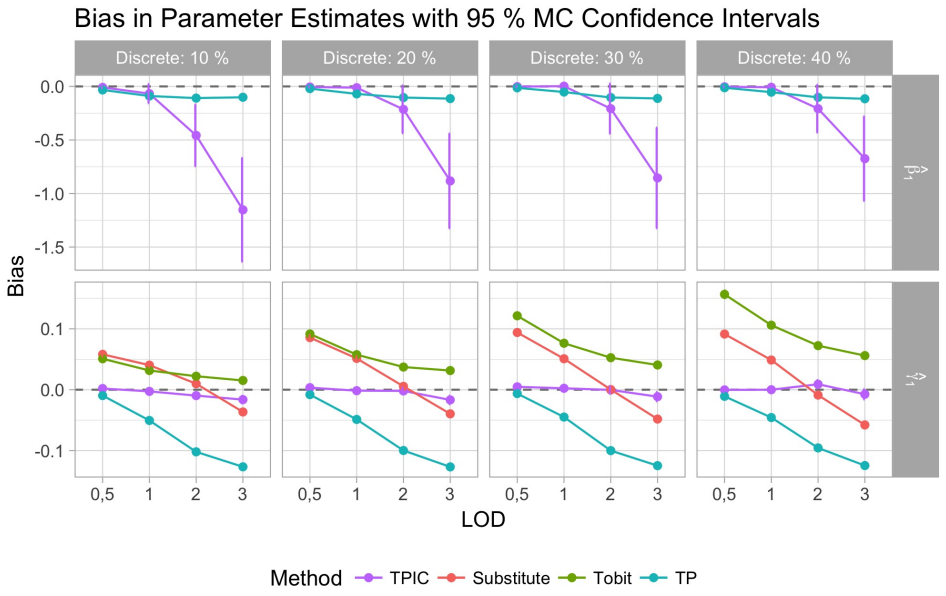

Figure 5.12: Bias in estimates of β_1 and γ_1 when multiple covariates are included.

The long-run empirical SE in $\hat{\beta}_1$ and $\hat{\gamma}_1$ is shown in Figure 5.13. The estimator $\hat{\beta}_1$ is subject to extreme variation for the largest LODs when the TPIC model is used, with an empirical SE around 12 in all the scenarios with $T = 3$. In practice, this means that the estimates range from about -300 to 100. This is also the case for the other parameters in $\boldsymbol{\beta}$. The empirical SE of $\hat{\beta}_1$ in the TPIC model with one covariate was only about 1.5 in the same scenarios, with estimates ranging from about -4 to 4. This is a drastic increase in variance that is not seen in any of the other models.

The results of including more covariates can be summarized in that the estimators as expected has a higher variance, and that this has a particularly great impact

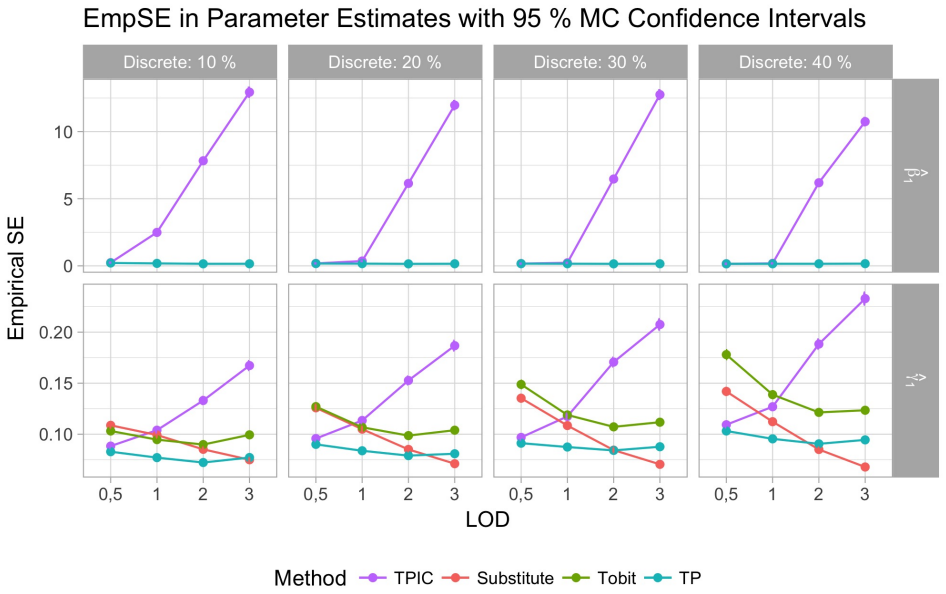

Figure 5.13: The long-run empirical SE of $\hat{\beta}_1$ and $\hat{\gamma}_1$ when multiple covariates are included.

on the performance of the TPIC model. As with one covariate, both the TP model and the Tobit model provide viable alternatives to the TPIC model for the highest LODs. In terms of prediction, the Tobit model is the best alternative, while the TP model might provide better results in terms of inference as it separates the two model parts. This is the same conclusion as was drawn with one covariate, and can therefore be taken as evidence that the results found for one covariate generalizes to multiple covariates.

6 Application to Borrelia Data

In this Chapter the different models presented in Chapter 3.1 will be demonstrated on the borrelia antibody data introduced in Section 2.1, with the primary goal of estimating the prevalence of Lyme borreliosis.

6.1 Description of the Data

The data consists of 981 serum samples from medical offices in Sør-Trøndelag from the period 2013 - 2017. It was collected as part of a medical undergraduate research thesis by Holt and Eriksen (2018). The samples were analyzed for borrelia antibody type IgG using the ELISA method LIAISON®and reported in AU/ml. This method is chemiluminescent, and the light signal is measured with a photomultiplier. Measurements lower than 5 AU/ml are indistinguishable from zero with this method. Thus, the observations are subject to a lower limit of detection. A presence of this antibody in the blood serum indicates that the individual has at some point been infected by *B. burdorferi*.

A database of 12318 samples that satisfied the selection criteria was available. The available samples were stratified based on region and age, which resulted in 36 strata. The number of samples to be selected from each stratum was decided based on the population and age distribution in each region, such that the resulting samples were representative for the population in Sør-Trøndelag. Trondheim was not included in the study because the population of Trondheim consists of many students from other from regions and countries, which may affect the results.

The county of Sør-Trøndelag was categorized into four regions; Inland, Mid-East, Mid-West, and Coastal. The goal is to investigate whether the coastal region has more occurrences of Lyme borreliosis than the other regions, therefore the three first regions are combined to a region called non-coastal. The number of participants in each region and the number of non-censored observations is presented in Table 6.1. Across all regions, the number of non-censored observations is low. In total 96.8% of the measurements in the datasets falls below the LOD, with 93.3% in the coastal regions and 98.5% in non-coastal regions. A plot of the non-censored measurements is provided in Figure 6.1.

Figure 6.1 shows that the non-censored concentrations have a higher mean in the coastal region, and from Table 6.1 the proportion of censored observations is

Table 6.1: The number of participants and the population in each region, and the proportion of the measurements in each region that are measured as above the LOD. The population figures are from SSB (Statistics Norway) dated 01.01.2018.

	Participants	Non-censored observations	Population
Non-Coastal	**668**	**10 *(1.5 %)***	**88828**
Inland	210	4 *(1.9 %)*	28352
Mid-West	259	3 *(1.2 %)*	34029
Mid-East	199	3 *(1.5 %)*	26446
Coastal	**313**	**21 *(6.7 %)***	**41998**
Total	**981**	**31 *(3.2 %)***	**130825**

highest in the non-coastal region. Both points toward a higher occurrence of Lyme borreliosis in the coastal region.

6.2 Statistical Analysis

The dataset consists of one binary covariate x_{reg}, which indicates whether the observation is from a coastal ($x_{\text{reg}} = 1$) or non-coastal ($x_{\text{reg}} = 0$) region. We will fit all the four models presented in Section 3.1 to the presented cytokine data. A brief overview of the models and their parameters is provided in Table 6.2.

Table 6.2: An overview of the models used in the statistical analysis. The parameter δ is only present when the log-skew-normal distribution is used for the continuous part.

		Model Parameters	
Abbrivation	Full Name	Discrete Part	Continuous Part
TP	Two-Part Model w/o Interval Censoring	$\boldsymbol{\beta}$	γ, σ, δ (conditional)
TPIC	Two-Part Model w/ Interval Censoring	$\boldsymbol{\beta}$	γ, σ, δ (conditional)
Tobit	Tobit Model	—	γ, σ, δ (marginal)
Substitute	Substitute ($S = T/2$)	—	γ, σ, δ (marginal)

In the Tobit model (3.2) all the observations are assumed to arise from the same latent continuous distribution $f(\cdot)$ with all responses below the LOD being observed as zero. In the simpler substitute models (3.10) all the censored observations are set to equal $S = T/2$.

The two-part models have an additional discrete point mass at zero. The standard two-part (TP) model (3.5) can be expressed as binary mixture with a point mass at zero with weight $\pi_i = P(Y_i = 0)$ and a continuous part $f(\cdot)$ with weight $(1 - \pi_i)$. Left-censoring of the continuous part below the LOD is taken into account in the two-part with interval censoring (TPIC) model (3.7), such that $\pi_i = P(Y_i^* = 0)$ is the probability of belonging to a separate sub-LOD population and the total probability of being censored is $P(Y_i = 0) = \pi_i + (1 - \pi_i)F(T; \mu_i)$.

In all of these models the covariate is introduced to the continuous distribution

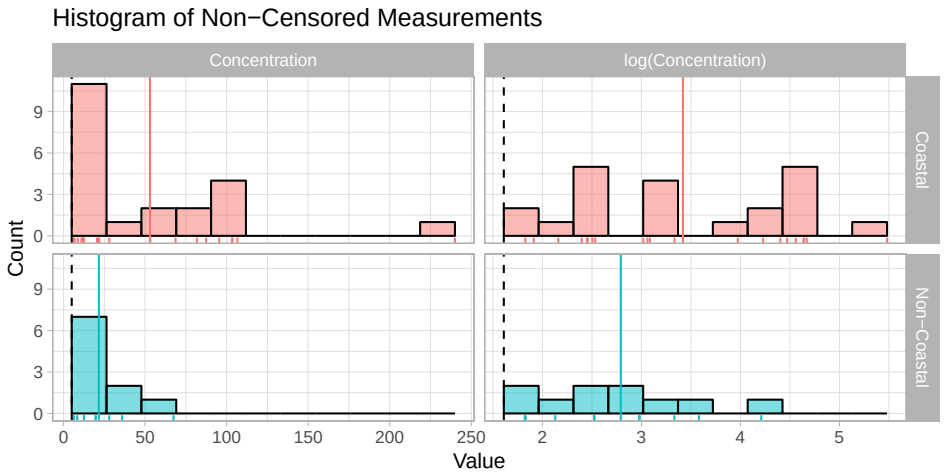

Figure 6.1: Histograms of the non-censored measurements divided by group on the measured scaled (left) and on log-scale (right). The black dashed vertical lines represents the LOD, while the colored solid vertical lines represents the mean of the non-censored measurements. A rug plot is included in the bottom margins to show the individual measurements.

$f(\cdot)$ by letting

$$\mu_i = \gamma_0 + \gamma_{\text{reg}} x_{\text{reg}}.$$

In the two-part models the probit link is utilized to introduce the covariate to π_i, such that

$$\pi_i = \mathbf{\Phi}(\beta_0 + \beta_{\text{reg}} x_{\text{reg}}).$$

In the dataset of borrelia antibody concentrations, it is expected that a large proportion does not have the antibody, therefore a two-part model is assumed to best represent the underlying process in the data. The histograms in Figure 6.1 shows many observations close to the LOD, which indicates a presence of censored positive observations below the LOD. This points towards the TPIC model providing the best fit. The simulation study in Chapter 5 did, however, show that the TPIC model becomes very unreliable when the majority of the censored observations comes from the continuous part. This is however unlikely to be a problem here, as the prevalence of Lyme borreliosis is previously found to be only a few percents (Vestrheim et al., 2016). Therefore, the large majority of the censored observations is expected to be true zeroes. The prevalence of ticks has previously been shown to be highest in the coastal areas of Norway (Jore et al., 2011). Consequently, Lyme borreliosis is expected to be more prevalent in the coastal region of Sør-Trøndelag.

The statistical analysis is performed in R using the command `optim()` with `method = "BFGS"` for numerical optimization of the likelihood functions. This is a Quasi-Newton optimization method that is described in Section 3.4.1. The models are fitted using both a probit/lognormal mixture and a probit/log-skew-normal

mixture. The difference between these two continuous distributions is a skew-parameter δ, which allows for more flexibility. The lognormal distribution is a special case of the log-skew-normal distribution with $\delta = 0$. The R functions used to fit the models are provided in Appendix A.1.

Table 6.3: Resulting maximum likelihood parameter estimates from fitting the models with a lognormal continuous distribution with 95 % Wald-type confidence intervals.

Parameter	TPIC	TP	Tobit	Substitute
γ_0	2.07 (0.59, 3.55)	2.79 (2.18, 3.40)	-6.32 (-9.04, -3.60)	0.94 (0.91, 0.98)
γ_{reg}	1.09 (-0.31, 2.49)	0.63 (-0.11, 1.37)	2.52 (1.15, 3.89)	0.14 (0.08, 0.20)
σ	1.25 (0.71, 1.78)	0.98 (0.74, 1.23)	3.65 (2.49, 4.80)	0.44 (0.42, 0.46)
β_0	1.99 (1.58, 2.40)	2.17 (1.93, 2.41)	—	—
β_{reg}	-0.55 (-0.98, -0.12)	-0.67 (-1.00, -0.35)	—	—
$\ell(\hat{\boldsymbol{\theta}})$	-169.3	-172.4	-172.8	-579.6
AIC	348.5	354.8	351.6	1165.2

The results from using the lognormal distribution as the continuous part are shown in Table 6.3. Firstly, all fitted models find at least one of the group parameters γ_{reg} and β_{reg} to be significantly different from zero, which is strong evidence for the presence of an underlying difference between the two groups. In both two-part models $\hat{\beta}_{\text{reg}}$ is found to be negative and $\hat{\gamma}_{\text{reg}}$ is found to be positive, which means that the coastal region has a lower estimated probability of belonging to the discrete part and a higher estimated expected concentration for the positive responses. In the one-part models, $\hat{\gamma}_{\text{reg}}$ is also positive, which again means that the coastal region has a higher estimated concentration of the borrelia antibody. This is consistent with the earlier studies on the prevalence of ticks and Lyme borreliosis (Vestrheim et al., 2016; Jore et al., 2011).

Test for Skew

The results from using the log-skew-normal distribution for the continuous distribution is shown in Table 6.4. Using a log-skew-normal distribution for the continuous part gives a slight improvement in the AIC for the TPIC model. However, a likelihood ratio test for the presence of the skew-parameter δ based on the hypotheses finds it not to be significant ($p = 0.07$), as shown in Table 6.5. The test is based on the hypotheses

$$H_0 : \ \delta = 0, \qquad\qquad H_1 : \ \delta \neq 0,$$

and the test statistic follows a χ^2_1-distribution.

It is interesting to note that the substitute model provides a very good fit when the log-skew-normal distribution is used. It estimates an extremely low $\hat{\sigma} = 3{\cdot}10^{-4}$. Recall that the skew-normal distribution (3.15) can be expressed as

$$f(y|\mu,\sigma,\delta) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{y-\mu}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{y-\mu}{\sqrt{\sigma^2+\delta^2}}\right).$$

When $\sigma \to 0$ the last multiplicative term becomes zero for $y < \mu$ and one for $y > \mu$, such that the result is a normal distribution with mean μ and variance δ^2

that is left truncated at $y = \mu$. In this case $\mu \approx T/2$ for both groups, such that the likelihood at $T/2$ is very high, $f(T/2; \mu, \sigma, \delta) \approx 1.8$. As 96.8% of the observations are censored, and therefore set to $y = T/2$ in the substitute model, this gives a remarkably high likelihood for the model. The resulting model is however not useful for inference purposes, as it does not provide any insight to the underlying process.

Table 6.4: Resulting maximum likelihood parameter estimates from fitting the models with a log-skew-normal continuous distribution with 95 % Wald-type confidence intervals.

Parameter	TPIC	TP	Tobit	Substitute
γ_0	3.96 (3.12, 4.80)	2.81 (-5.03, 10.7)	2.37 (0.67, 1.32)	0.91 (0.87, 0.94)
γ_{reg}	1.18 (0.24, 2.12)	0.63 (-0.11, 1.37)	2.49 (1.32, 3.66)	0.004 (-0.03, 0.04)
σ	0.40 (0.01, 0.79)	0.98 (0.72, 1.24)	1.14 (0.17, 2.11)	$3{\cdot}10^{-4}(2{\cdot}10^{-5}, 6{\cdot}10^{-4})$
β_0	1.38 (-2.88, 5.63)	2.17 (1.93, 2.41)	—	—
β_{reg}	-0.71 (-2.59, 1.16)	-0.67 (-1.00, -0.35)	—	—
δ	-10.4 (-93.6, 71.7)	-0.02 (-9.83, 9.78)	-42.0 (-72.2, -11.76)	0.45 (0.43, 0.47)
$\ell(\hat{\boldsymbol{\theta}})$	-167.7	-172.4	-170.3	75.5
AIC	347.5	356.8	348.6	-143.1

The Presence of Interval Censoring

The TP model achieves a lower likelihood than the TPIC model with the same number of parameters. Thus the interval censoring improves the fit of the model. The fitted continuous part of the two-part models is shown in Figure 6.2. The TPIC model estimates a heavier left tail than the TP model, due to assuming that a proportion of the censored observations arises from the continuous part. Chai and Bailey (2008) argues that a well-specified model should include interval censoring if it is known to be present in the data. In this dataset, there are many observations close to the LOD. Thus, it seems reasonable to believe that some of the censored observations are in fact non-zero concentrations that could have been detected with a more sensitive measurement technique. All of this indicates that the interval censoring should be included in the model.

Test for the Discrete Part

The TPIC model is found to be superior in terms of AIC, with the Tobit model performing nearly as well. The AIC is known to give somewhat anti-conservative results, therefore a likelihood ratio test (LRT) is performed for the presence of the discrete part. The Tobit model is a special case of the TPIC model with $\beta_0 = -\infty$ and $\beta_{\text{reg}} = 0$. Therefore, a LRT can be performed based on the hypotheses

$$H_0 : \ (\beta_0, \beta_{\text{reg}}) = (-\infty, 0), \qquad H_1 : \ (\beta_0, \beta_{\text{reg}}) \neq (-\infty, 0).$$

Since the restriction on β_0 lies on the boundary of the parameter space $\mathbb{R}^2$, the resulting test statistic follows a mixture of two chi-square distributions, $\frac{1}{2}(\chi_1^2 + \chi_2^2)$ (Self and Liang, 1987). The discrete part is found to be significant with a p-value of 0.02, as shown in Table 6.5. This confirms that the TPIC model provides the best

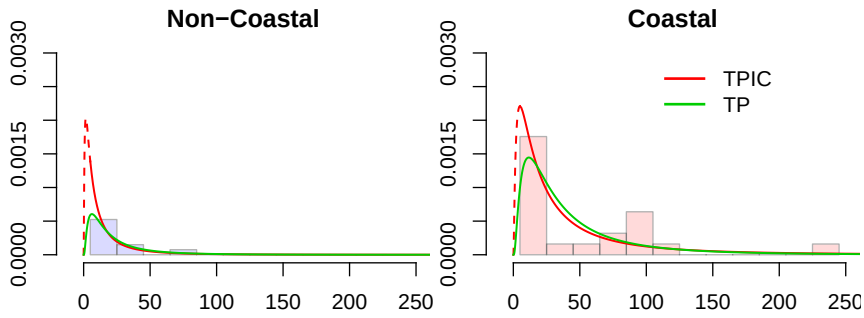

Figure 6.2: Plot of the fitted continuous parts of the two-part models with lognormal continuous distribution along with a histogram of the positive observations for both groups.

Table 6.5: Results of likelihood ratio tests for the presence of skew, discrete part and fixed effects.

Model	-2*loglikelihood	-2*Difference in loglikelihoods		Distribution	*p*-value
(a) TPIC with skew	335.4				
Skewness:					
(b) TPIC	338.6	3.2	(b-a)	χ^2_1	0.07
Discrete part:					
(c) Tobit	345.6	7.0	(c-b)	$\frac{1}{2}(\chi^2_1 + \chi^2_2)$	0.02
Fixed effects:					
(d) TPIC without γ_{reg}	341.3	2.7	(d-b)	χ^2_1	0.10
(e) TPIC without β_{reg}	340.9	2.3	(e-b)	χ^2_1	0.13
(f) TPIC without β_{reg} and γ_{reg}	358.6	23.2	(f-b)	χ^2_2	$9{\cdot}10^{-6}$

fit, and is in agreement with the assumed structure of the underlying process, which is expected to have a prominent discrete part at zero representing the individuals without the borrelia antibody.

Test for Fixed Effects

As the lognormal TPIC model is shown to provide the best fit for the data, the reminder of the analysis is focused on this model. Likelihood ratio tests are performed for the fixed effects in the model. Both γ_{reg} and β_{reg} do not significantly increase the likelihood of the model when tested separately. When the discrete effect β_{reg} is restricted to zero the continuous effect γ_{reg} increases in magnitude and captures a lot of the group difference previously explained by β_{reg}, and vice versa. However, when both parameters are tested simultaneously the drop in likelihood is highly significant ($p = 9 \cdot 10^{-6}$), which is strong evidence for including the covariate x_{reg} in the model.

Estimated Prevalence

If we define an occurrence of Lyme borreliosis as having a presence of borrelia antibody in the blood serum, the estimated prevalence by the TPIC model without skew is

$$\hat{P}(Y_i^* > 0 | x_{\text{reg}} = 0) = 1 - \mathbf{\Phi}(\hat{\beta}_0) = 1 - \mathbf{\Phi}(1.99) = 2.3\%$$

for the non-coastal region and

$$\hat{P}(Y_i^* > 0 | x_{\text{reg}} = 1) = 1 - \mathbf{\Phi}(\hat{\beta}_0 + \hat{\beta}_{\text{reg}}) = 1 - \mathbf{\Phi}(1.99 - 0.55) = 7.5\%$$

in the coastal region. The difference between the groups is significant, as $\hat{\beta}_{\text{reg}}$ has a p-value of 0.01.

Usually, a more strict definition of prevalence is used. For the given data, a sample is regarded as definitely positive if the concentration is above 15.6 AU/ml. If this definition is used, the estimated prevalence is

$$\hat{P}(Y_i^* > 15.6 | x_{\text{reg}} = 0) = (1 - \mathbf{\Phi}(\hat{\beta}_0))(1 - F(15.6 | \hat{\mu}_i = \hat{\gamma}_0, \hat{\sigma})) = 0.7\%$$

in the non-coastal region and

$$\hat{P}(Y_i^* > 15.6 | x_{\text{reg}} = 1) = (1 - \mathbf{\Phi}(\hat{\beta}_0 + \hat{\beta}_{\text{reg}}))(1 - F(15.6 | \hat{\mu}_i = \hat{\gamma}_0 + \hat{\gamma}_{\text{reg}}, \hat{\sigma})) = 4.7\%$$

in the coastal region.

The overall prevalence in Sør-Trøndelag is estimated by fitting a model without the region covariate. This gives

$$\begin{aligned} \hat{P}(Y_i^* > 15.6) &= (1 - \mathbf{\Phi}(\hat{\beta}_0))(1 - F(15.6 | \hat{\gamma}_0, \hat{\sigma})) \\ &= (1 - \mathbf{\Phi}(1.76))(1 - F(15.6 | 2.76, 1.34)) = 2.0\% \end{aligned}$$

The estimated prevalence is a function of the three estimates $\hat{\boldsymbol{\theta}} = (\hat{\gamma}_0, \hat{\sigma}, \hat{\beta}_0)$. It is not straight forward to calculate its variance, as it is the function is nonlinear. The variance can however be estimated by the delta method based on error propagation (Powell, 2007). Let $\hat{P}(Y_i^* > 15.6) = h(\hat{\boldsymbol{\theta}})$. The delta method can be formulated as

$$\text{Var}(h(\hat{\boldsymbol{\theta}})) \approx \nabla h(\hat{\boldsymbol{\theta}})' \text{Cov}(\hat{\boldsymbol{\theta}}) \nabla h(\hat{\boldsymbol{\theta}})$$

This approximation gives $\text{SE}(h(\hat{\boldsymbol{\theta}})) \approx 0.023$. This is a relatively large standard error compared to the estimated value of $h(\hat{\boldsymbol{\theta}}) = 0.020$. This is mainly caused by the great variance in $\hat{\gamma}_0$ and $\hat{\sigma}$, due to few observations above the detection limit. The Wald-type confidence interval truncated at zero is $(0\% - 6.5\%)$, but this is not reliable as the assumption of normality is clearly violated.

The estimated prevalence of 2.0% is somewhat lower than the previously estimated prevalence in Sør-Trøndelag by Vestrheim et al. (2016), which estimated the prevalence to be 3.9%(95 % CI: 2.3% − 6.4%) and 3.7%(95 % CI: 2.4% − 5.7%) with two different assays and sample size $n = 301$. They used the same definition of a positive sample. Unlike for the data studied here, Vestrheim et al. (2016) did include Trondheim in the study. Since Trondheim constitutes 60 % of the population in Sør-Trøndelag, this may have contributed greatly to the difference in the estimates.

Comparison to Logistic Regression

A common approach to estimating the prevalence is categorizing the data in positive and negative samples and using logistic regression to estimate the probability of the sample being positive. Let

$$\tilde{Y}_i = \begin{cases} 1, Y_i > 15.6 \\ 0, Y_i \leq 15.6 \end{cases},$$

such that $\tilde{Y}_i = 1$ indicates a positive sample. We want to estimate $p_i = P(\tilde{Y}_i = 1|\mathbf{x}_i)$. This is done by using the logit link, such that

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_{\text{reg}} x_{\text{reg},i}.$$

Solving for p_i gives the estimated probability,

$$p_i = \frac{\exp(\beta_0 + \beta_{\text{reg}} x_{\text{reg},i})}{1 + \exp(\beta_0 + \beta_{\text{reg}} x_{\text{reg},i})}.$$

The parameters can be estimated using the function `glm` in R, which gives $\hat{\beta}_0 = -4.88(-5.76, -4.00)$ and $\hat{\beta}_{\text{reg}} = 1.75(0.71, 2.79)$ with 95 % confidence intervals. Thus, the estimated prevalences are

$$\hat{P}(\tilde{Y}_i = 1|x_{\text{reg}} = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = 0.7\%$$

in the non-coastal region and

$$\hat{P}(\tilde{Y}_i = 1|x_{\text{reg}} = 1) = \frac{\exp(\beta_0 + \beta_{\text{reg}})}{1 + \exp(\beta_0 + \beta_{\text{reg}})} = 4.2\%$$

in the coastal region. The estimated prevalence in the non-coastal region is similar as when the TPIC model was used, but the estimated prevalence in the coastal region is lower. This indicates that the amount of observations above the limit 15.6 is lower than what is expected from the probit/lognormal mixture fitted to all the data. Fitting the model without the covariate gives

$$\hat{P}(\tilde{Y}_i = 1) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = \frac{\exp(-3.98)}{1 + \exp(-3.98)} = 1.8\%,$$

with 95 % confidence interval $(1.2\% - 2.9\%)$. The estimated prevalence with the TPIC model was $2.0\%(\text{SE} \approx 2.3\%)$. Thus, the logistic regression gives a slightly lower estimate of the prevalence with a much smaller uncertainty. The TPIC model takes more information into account, as it is based on the entire distribution of the data, whereas the in the logistic regression the data is collapsed into two categories. One can however argue that in this case with very few observations above the detection limit, logistic regression is a better option because the TPIC model provides results with much larger uncertainty.

7 | Application to Cytokine Data

In this chapter, the models presented in Chapter 3 will be demonstrated on the cytokine data presented in Section 2.2. First, a longitudinal analysis will be conducted in Section 7.2 with the methods presented in Section 3.2 on three individual cytokines with different proportions of censored observations. A comprehensive analysis of the cytokine TNF-α with 32.5% censored observations is followed by briefer analyses of the cytokines MCP1 with 3.6% censored observations and IL8 with 76.9% censored observations. The primary goal is to identify differences in the time profiles across the diagnostic groups. Then, a bivariate analysis is conducted on two of the cytokines, TNF-α and IL8, in Section 7.3 in order to estimate their correlation.

7.1 Description of the Data

The data consists the concentrations of 34 cytokines measured simultaneously with multiplex assays, as well as the time period of the measurement, the diagnosis of the individual, the age of the individual and a code indicating which individual the measurement is from. Measurements indistinguishable from zero are recorded as below the LOD.

There is a total of 308 measurements from 75 patients across seven time-points. Five patients have measurements from more than two pregnancies. In these cases, we regard the distinct pregnancies as two separate individuals, which gives a total of 80 patients. The patients have one of four diagnoses; 19 patients are healthy controls, 28 are diagnosed with SLE, 23 are diagnosed with RA, and ten are diagnosed with SN-RA. The number of measurements per patient ranges from only one to seven. Some patients have more than one recording for certain time points.

The possible time points, labeled 0 - 6, corresponds to different time periods before, during, and after the pregnancy. Time point 0 is before the pregnancy, time points 1 - 3 are respectively the first, second and third trimester of the pregnancy, and time points 4 - 6 are six weeks, six months and twelve months postpartum. They are illustrated on a timeline in Figure 7.1.

The age is recorded at least once per individual. Naturally, their ages increase through the pregnancies. For simplicity, we regard the first recorded age during pregnancy as the patients' age, unless the only available measurement is before

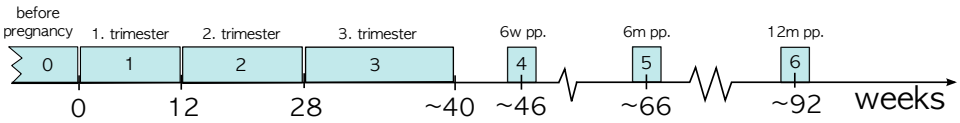

Figure 7.1: Illustration showing when the measurements are obtained on an axis denoting weeks. Note that the time of birth will vary, so there is some uncertainty regarding how much time has passed from the *prepartum* (during pregnancy) measurements to the *postpartum* (after birth) measurements.

pregnancy. Table 7.1 displays the number of patients for each diagnosis, along with the mean age of the group. We see that there are considerably fewer patients with the SN-RA diagnosis. It might be desirable to merge this group with seropositive RA to not over-parameterize the models.

Table 7.1: Patient characteristics.

	Healthy	SLE	RA	SN-RA	Total
Patients *n* (%)	19 (24)	28 (35)	23 (29)	10 (12)	80
Age *mean* (*sd*)	30 (2.5)	30 (5.3)	33 (4.4)	31 (3.1)	31 (4.5)

The number of measurements taken at each time point per diagnosis is displayed in Table 7.2. We see that there are some variations across the time points. Perhaps most noteworthy, there are no measurements of healthy controls before pregnancy (0) and twelve months postpartum (6). In general, there are most measurements during the first two trimesters of the pregnancy (1 - 2), and gradually fewer after that.

Table 7.2: The number of measurements per time period for each diagnosis.

	Time Period							
	0	1	2	3	4	5	6	Total
Healthy	0	19	20	19	20	16	0	94
SLE	9	19	20	18	15	13	13	107
RA	15	13	13	10	6	9	8	74
SN-RA	5	6	6	3	5	3	5	33
Total	**29**	**57**	**59**	**50**	**46**	**41**	**26**	**308**

The further analysis focuses on the log-transformed concentrations. The histograms of the log-transformed concentrations are displayed in Figure 7.2, along with the percentage of left-censored observations. The proportion of left-censored observations ranges from zero to 99% (TNF-β). Four cytokines have no censored observations, while for seven cytokines more than 90% of the observations are below the LOD.

The Spearman's correlations between the log-transformed concentrations are depicted in Figure 7.3. The Spearman's correlation denotes the correlation between the ranks of the observations, which makes it more suitable for data known

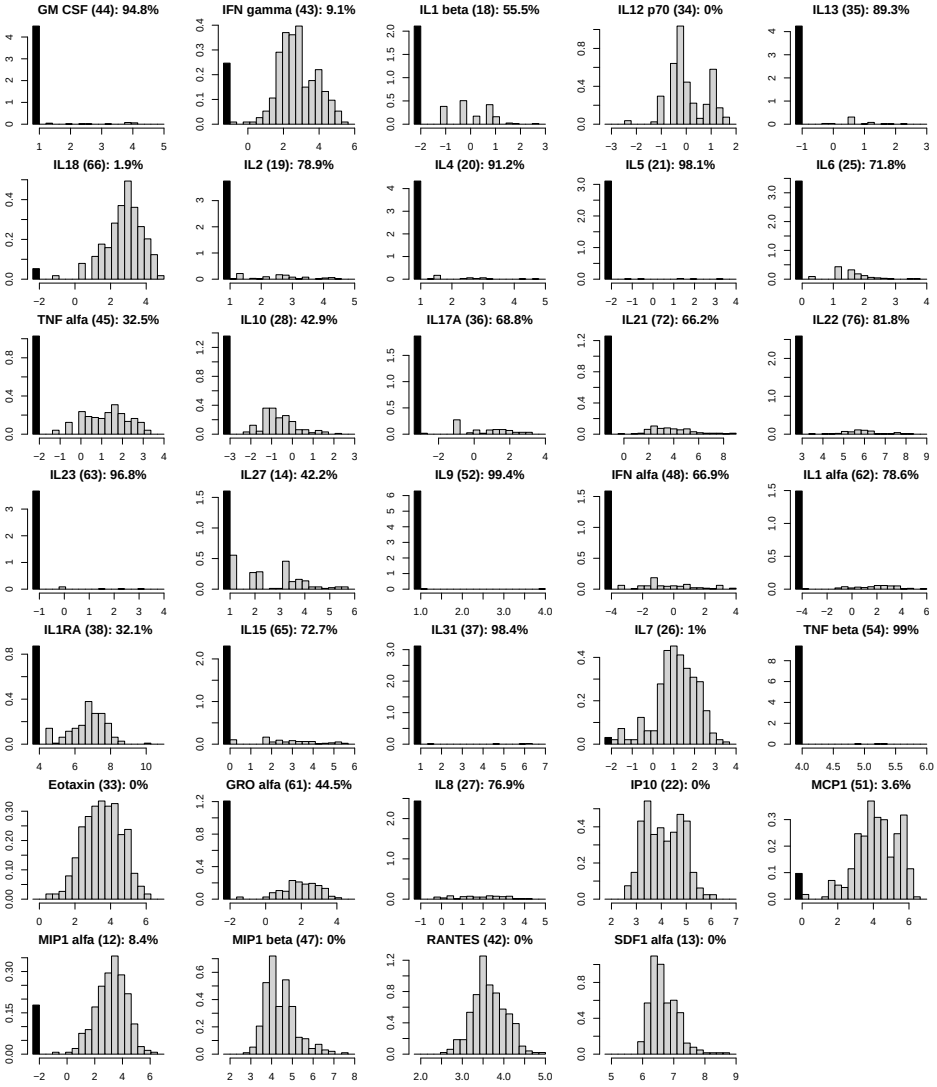

Figure 7.2: Density histograms of the log-transformed concentrations for each cytokine across all patients and time points. The black bins represent the left censored observations. The proportion of censored observations is specified in the headings.

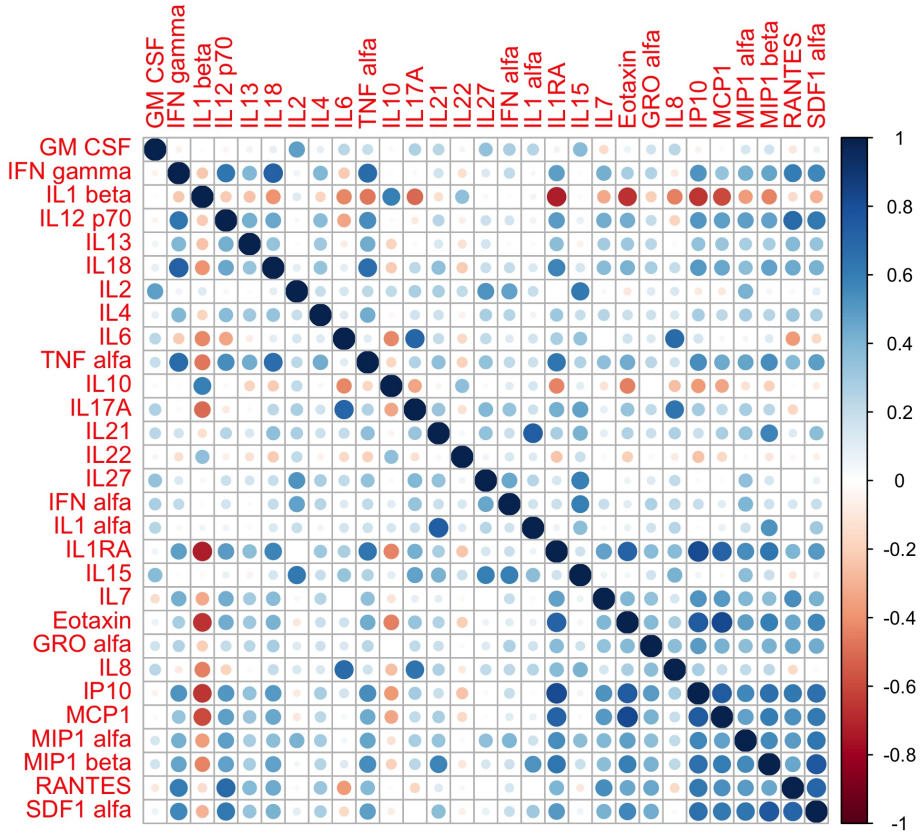

Figure 7.3: Illustration of the Spearman's correlation matrix between the log-transformed concentrations of the cytokines. The color indicates the sign of the correlation and the size of the circle increases with the magnitude. Cytokines with less than 10 non-censored observations are excluded from the plot.

to not be normally distributed, such as data subject to detection limits (Kaune and Kettrup, 1994). The censored observations are set to LOD/2, however their values are arbitrary, as only the ranks are used to calculate the correlations. Cytokines with ten or fewer non-censored measurements are excluded from the figure. Figure 7.3 shows a preponderance of positive correlations, meaning that a high concentration of one cytokine tends to be accompanied by high concentrations of the other cytokines.

We will continue to focus on the cytokines TNF-α with 32.5% censored observations, MCP1 with 3.6% censored observations, and IL8 with 76.9% censored observations as illustrative examples. The first-mentioned is a pro-inflammatory cytokine that has been linked to the improvement of RA during pregnancy (Swain and Jena, 2016). The second and third are chemoattractant cytokines. MCP1 has been associated with both RA and SLE (Deshmane et al., 2009), and IL8 is linked to bone-erosion and pain in RA (Ridgley et al., 2018).

Cytokine TNF-α

More insight into the data on the cytokine TNF-α is provided in Figure 7.4. There are several noteworthy takeaways from the figure. Firstly, the proportion of censored observations in each group varies from 21.6% in the RA group to almost the double, 40.4%, in the healthy group. This is indicative of a group effect on the probability of being censored. There is however no very apparent difference in the distributions of the non-censored observations.

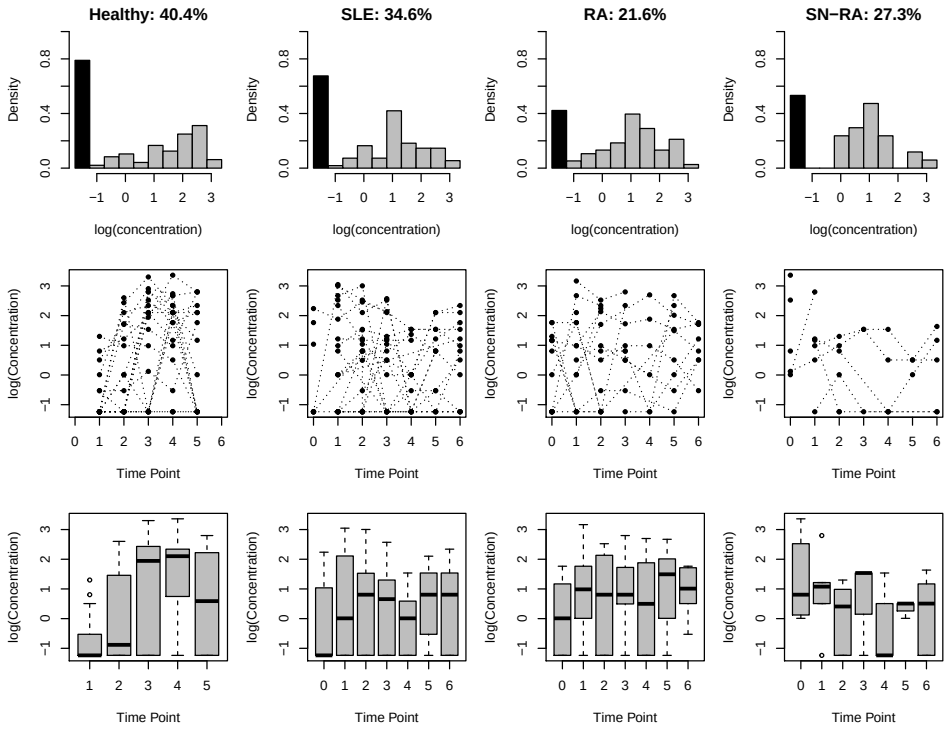

Figure 7.4: Density histograms, spaghetti plots and box plots of the log-transformed concentrations of TNF-α. In the histograms the censored observations are represented by a black bin below the LOD. In the spaghetti plots and box plots the censored observations are set to the detection limit.

From the spaghetti plots and box plots in Figure 7.4, the concentration of TNF-α seems to increase throughout the pregnancy for the healthy patients with a peak at time point 4 (six weeks postpartum) and a decline to time point 5 (six months postpartum). This trend is not seen in any of the other groups, which display no apparent temporal trends. The three diagnosed groups appear to have close to equal mean concentrations, that remains stable across all the time points.

Cytokine MCP1

A more detailed look into the data on the cytokine MCP1 is provided in Figure 7.5. It shows separate histograms for each diagnosis, and spaghetti plots and box plots that illustrate the development over time. From Figure 7.5 we see that the concentration of MCP1 increases throughout the pregnancy for the healthy patients, with a peak at time point 4 (six weeks postpartum). No clear trend is visible in any of the diagnosed groups.

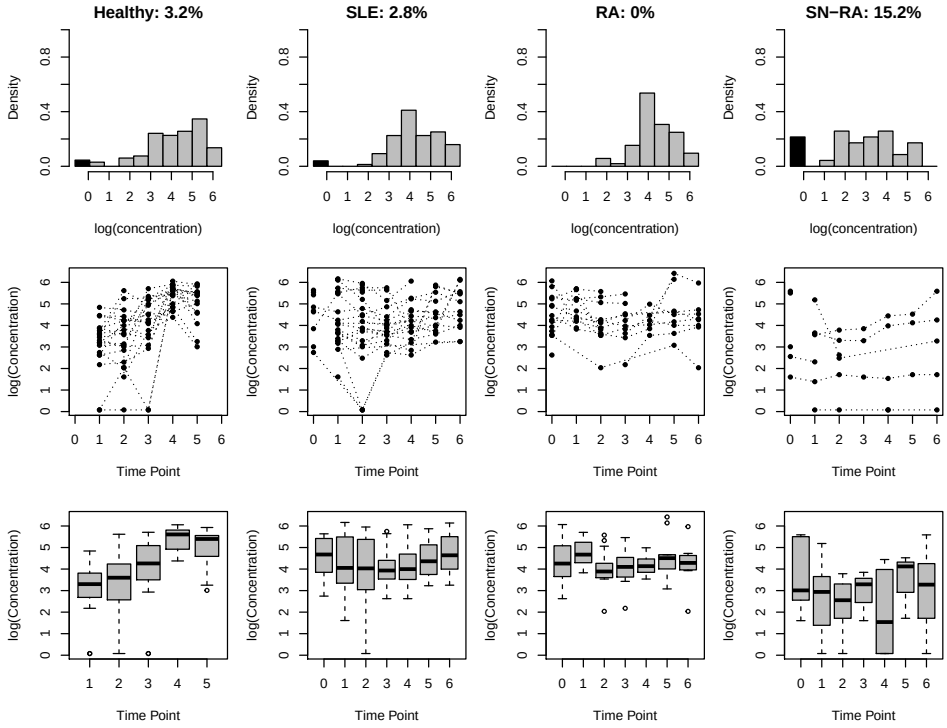

Figure 7.5: Density histograms, spaghetti plots and box plots of the log-transformed concentrations of MCP1. In the histograms the censored observations are represented by a black bin below the LOD. In the spaghetti plots and box plots the censored observations are set to the detection limit.

The amount of censored observations in each group ranges from 15.2% for the SN-RA group and zero for the RA group. Note however that all the censored observations in the SN-RA group come from the same individual. Based on the spaghetti plots there seems to be a considerable correlation between the measurements from the same individual, which indicates that the measurements cannot be regarded as independent.

Cytokine IL8

A detailed overview of the data on the cytokine IL8 is provided in Figure 7.6. The proportion of censored observations varies from 70.1% in the SLE group to 86.2% in the healthy group. This is indicative of a group difference in the probability of being censored. There is however no clear trend over time in any of the groups.

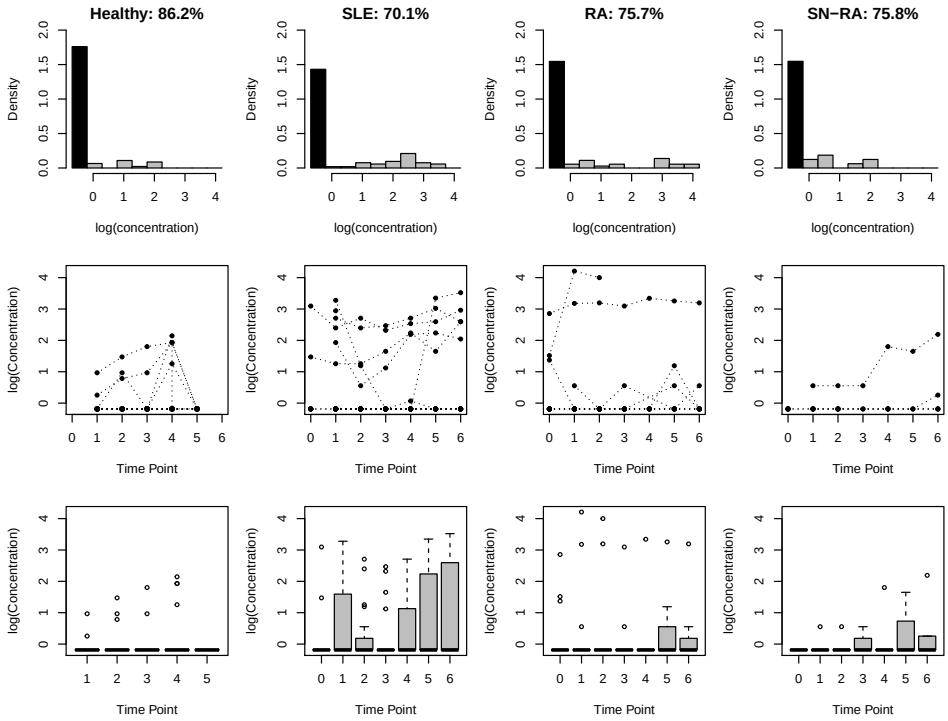

Figure 7.6: Density histograms, spaghetti plots and box plots of the log-transformed concentrations of IL8. In the histograms the censored observations are represented by a black bin below the LOD. In the spaghetti plots and box plots the censored observations are set to the detection limit.

7.2 Longitudinal Statistical Analysis

In this section, longitudinal statistical analysis is performed on data on the three cytokines TNF-α, MCP1, and IL8 with methods presented in Section 3.2. Methods for estimating the model parameters are described in 3.4.

There are many computational challenges with fitting the models. The integrals in the marginal likelihoods (3.18) must be estimated and the resulting likelihood function (3.28) must be maximized. The proposed mixture models are nonlinear, so it is not possible to use R-packages like `lme4` and `glmmTMB`. The latter does

have functionality for modeling zero-inflated data, but does not have options for including interval censoring or using the skew-normal distribution.

The SAS procedure `NLMIXED` handles non-linear mixed models with random effects and is very flexible. It offers a wide range of optimization techniques and allows the user to specify the likelihood function. The `NLMIXED` procedure is used by e.g. Berk and Lachenbruch (2002) and Mahmud et al. (2010) for similar problems. For the longitudinal analysis, we use the `NLMIXED` procedure with adaptive Gaussian quadrature for estimating the marginal likelihoods and quasi-newton optimization of the likelihood (`tech = QUANEW`) with the BFGS method for iteratively estimating the Hessian (`update = BFGS`). One drawback with SAS is that the skew-normal distribution is not implemented. However, the probability density function of the skew-normal distribution can easily be calculated from the normal probability density and cumulative density function, as expressed by

$$f(y_{ij}|\mu_{ij},\sigma,\delta) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{y_{ij}-\mu_{ij}}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{y_{ij}-\mu_{ij}}{\sqrt{\sigma^2+\delta^2}}\right).$$

In order to fit the Tobit model (3.2) and the TPIC model (3.7), the cumulative density function

$$F(T|\mu_{ij},\sigma,\delta) = \Phi\left(\frac{T-\mu_{ij}}{\sqrt{\sigma^2+\delta^2}}\right) - 2T\left(\frac{T-\mu_{ij}}{\sqrt{\sigma^2+\delta^2}},\frac{\delta}{\sigma}\right)$$

must be evaluated, which requires implementing Owen's T function

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}\,dx.$$

We implement a numerical approximation of Owen's T function in SAS based on the implementation in the R-package `sn` (Azzalini, 2017). It is based on work by Owen (1956), which showed that for small positive values of h and $a > 1$ the integral can be estimated by series expansion, and for larger positive values of h and $a > 1$ asymptotic approximation can be used. A number of reflection properties are utilized to cover the remaining cases. As Azzalini (2017), we set the cut-point for h at $h = 8$ and truncate the series expansion after 50 terms. The code used for this is provided in Appendix B.

We fit all the four models presented in Section 3.2 with random effects to account for the correlation between measurements from the same individual. Let y_{ij} be an observation from the ith individual taken at the jth time point. Among the considered models there are two one-part models, the Tobit model (3.19) and the Substitute model (3.20). The former is based on the assumption that all the observations come from the same latent continuous distribution, while the latter takes all the censored observations to be a substitute value S and fits a continuous distribution to the resulting data. As before, we will use the substitute value $S = T/2$. In both of these models covariates and random effects are introduced to the parameter μ_{ij} of the continuous distribution such that

$$\mu_{ij} = \mathbf{z}_{ij}'\boldsymbol{\gamma} + \tau_{1i},$$

where $\mathbf{z}'_{ij}$ is a set of covariates with corresponding fixed effects $\boldsymbol{\gamma}$, and τ_{1i} is a random effect for individual i. For the one-part models the random effect is assumed to be normally distributed with variance s_{11}^2, $\tau_{1i} \sim \mathcal{N}(0, s_{11})$.

In the two-part models, an additional discrete point mass at $y = 0$ is included. Let the weight of the point mass be π_{ij}. In the two-part (TP) model (3.23) the point mass represents the observed zeroes. Thus, $\pi_{ij} = P(Y_{ij} = 0)$ denotes the probability of falling below the LOD, and $f(\cdot)$ is the distribution of the non-censored responses. This model can be expanded to take left-censoring of the continuous part $f(\cdot)$ on the interval $[0, T]$ into account. This results in the TPIC model (3.24). In this setting π_{ij} represents the probability that the true latent concentration of y_{ij} comes from a separate sub-LOD population represented by a point mass at zero. In both two-part models covariates are introduced to π_{ij} by using the probit link,

$$\pi_{ij} = \boldsymbol{\Phi}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \tau_{i2}),$$

where $\mathbf{x}'_{ij}$ is a set of covariates with corresponding fixed effects $\boldsymbol{\beta}$, and τ_{2i} is a random effect for individual i. The two random effects are assumed to follow a bivariate normal distribution, $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}) \sim \mathcal{N}_2(0, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}.$$

Thus, s_{11} and s_{22} are the variances of respectively τ_{1i} and τ_{2i}, and s_{12} is their covariance.

The available covariates for this analysis are the patients diagnostic status, the time point of the measurement, and their age. Because there are relatively few patients with the diagnosis SN-RA, we choose to merge this diagnosis with the group of RA patients, as they are known to display similar behavior. There are also certain time points with few observations, in particular the first and last, which has no observations from healthy controls. In order to have observations from all groups in all time periods and reduce the number of parameters, we group the time points into three time periods denoted as t_1, t_2, and t_3. The first time period t_1 contains the measurements from before the pregnancy and during the first two trimesters, i.e. *before/early pregnancy*, the next period t_2 contains measurements from the third trimester and six weeks postpartum, i.e. *around birth*, and the last period t_3 contains measurements from 6 months and 12 months postpartum, i.e. *after birth*. With these simplifications, the data is categorized into three diagnostic groups and three time periods.

7.2.1 Cytokine TNF-α

With the specified simplifications of the dataset, there are three diagnostic groups and three time periods, as well as a continuous variable for the age of the patient. This gives a total of five fixed effects in each model part. Because it is of interest to determine whether the time period affects the diagnostic group differently, we also include interaction terms between time period and diagnosis, which gives four

additional parameters in each model part. Including the intercepts, this gives a total of ten parameters that may be present in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, such that

$$\begin{aligned}\mu_{ij} =&\gamma_0 + \gamma_{\text{age}}\ x_{\text{age},i} + \gamma_{\text{RA}}\ x_{\text{RA},i} + \gamma_{\text{SLE}}\ x_{\text{SLE},i} + \gamma_{t_2}\ x_{t_2,ij} + \gamma_{t_3}\ x_{t_3,ij} + \\ &\gamma_{t_2\times\text{RA}}\ x_{\text{RA},i}\ x_{t_2,ij} + \gamma_{t_3\times\text{RA}}\ x_{\text{RA},i}\ x_{t_3,ij} + \\ &\gamma_{t_2\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_2,ij} + \gamma_{t_3\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_3,ij} + \tau_{i1},\end{aligned}$$

$$\begin{aligned}\pi_{ij} =&\boldsymbol{\Phi}(\beta_0 + \beta_{\text{age}}\ x_{\text{age},i} + \beta_{\text{RA}}\ x_{\text{RA},i} + \beta_{\text{SLE}}\ x_{\text{SLE},i} + \beta_{t_2}\ x_{t_2,ij} + \beta_{t_3}\ x_{t_3,ij} + \\ &\beta_{t_2\times\text{RA}}\ x_{\text{RA},i}\ x_{t_2,ij} + \beta_{t_3\times\text{RA}}\ x_{\text{RA},i}\ x_{t_3,ij} + \\ &\beta_{t_2\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_2,ij} + \beta_{t_3\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_3,ij} + \tau_{i2}).\end{aligned}$$

Furthermore, the random effects contribute with one parameter, the variance s_{11}, in the one-part models and the three parameters of the covariance matrix $\boldsymbol{\Sigma}$ in the two-part models. Including the standard deviation σ of the normal distribution, this results in 12 parameters in the one-part models and 22 parameters in the two-part models. Using the skew-normal distribution gives the additional skew parameter δ.

Based on the histograms of the concentrations in Figure 7.4 it seems very unlikely that all the observations originate from the same continuous distribution, as the proportion of censored observations looks much larger than what could be expected from the distribution of the non-censored data. This indicates a presence of excess zeroes, therefore the two-part models are expected to provide a much better fit than the one-part models. Since there are very few observations close to the LOD, we expect the censored tail of the continuous distribution to be small. The simulation study in Chapter 5 showed that the two-part models are close to equivalent when that is the case, both giving close to unbiased parameter estimates.

The results of fitting the four models with a log-skew-normal continuous distribution are shown in Table 7.3. In all models, most of the interaction effects are found to be significant, which is strong evidence for an underlying difference in the time profiles across the diagnoses. The signs of the parameters in $\boldsymbol{\gamma}$ in the two-part models are in all cases except one opposite from the sign of the associated parameter in $\boldsymbol{\beta}$, which means that lower probabilities of belonging to the discrete part tend to be associated with a higher expected value in the continuous part, and vice versa.

In terms of AIC, the two-part models are superior, with the TPIC model achieving a slightly better score than the TP model. Furthermore, the Tobit model gets a much higher likelihood than the substitute model with the same number of parameters. Thus, the substitute model is clearly not suitable for this problem, which was expected as the proportion of censored data is 32.5%.

The Presence of Interval Censoring

From Table 7.3 we see that the results from the TPIC and TP model are very similar, which was expected based on the distribution of the data. The greatest differences lie in $\boldsymbol{\beta}$, which has a different interpretation in the two models. In the TP

Table 7.3: Maximum likelihood parameter estimates from fitting the four models with a log-skew-normal continuous part. Wald-type 95% confidence intervals are included in parenthesis for the fixed effects.

Parameter	TPIC	TP	Tobit	Substitute
γ_0	2.33 *(0.82, 3.83)*	2.30 *(0.84, 3.76)*	2.32 *(0.67, 3.97)*	2.12 *(0.41, 3.83)*
β_0	-0.66 *(-2.85, 1.52)*	-0.56 *(-2.65, 1.53)*	—	—
γ_{age}	-0.01 *(-0.06, 0.04)*	-0.01 *(-0.06, 0.03)*	-0.01 *(-0.06, 0.04)*	-0.02 *(-0.07, 0.03)*
β_{age}	0.03 *(-0.04, 0.10)*	0.03 *(-0.04, 0.09)*	—	—
γ_{RA}	0.23 *(-0.38, 0.83)*	0.22 *(-0.37, 0.80)*	0.79 *(0.04, 1.54)*	0.84 *(0.14, 1.54)*
β_{RA}	-1.24 *(-2.12, -0.36)*	-1.21 *(-2.06, -0.36)*	—	—
γ_{SLE}	0.21 *(-0.42, 0.85)*	0.22 *(-0.39, 0.84)*	0.57 *(-0.19, 1.33)*	0.58 *(-0.12, 1.29)*
β_{SLE}	-0.73 *(-1.59, 0.13)*	-0.70 *(-1.54, 0.14)*	—	—
γ_{t_2}	0.91 *(0.41, 1.41)*	0.90 *(0.42, 1.37)*	1.41 *(0.74, 2.07)*	1.51 *(0.90, 2.12)*
β_{t_2}	-1.44 *(-2.19, -0.70)*	-1.45 *(-2.18, -0.72)*	—	—
γ_{t_3}	1.10 *(0.47, 1.74)*	1.11 *(0.49, 1.72)*	1.21 *(0.46, 1.96)*	1.25 *(0.56, 1.93)*
β_{t_3}	-0.52 *(-1.37, 0.32)*	-0.54 *(-1.37, 0.30)*	—	—
$\gamma_{t_2 \times \text{RA}}$	-0.80 *(-1.47, -0.13)*	-0.76 *(-1.40, -0.13)*	-1.51 *(-2.35, -0.67)*	-1.58 *(-2.43, -0.72)*
$\beta_{t_2 \times \text{RA}}$	1.83 *(0.66, 3.00)*	1.81 *(0.67, 2.95)*	—	—
$\gamma_{t_3 \times \text{RA}}$	-1.14 *(-1.92, -0.35)*	-1.09 *(-1.84, -0.34)*	-1.20 *(-2.20, 0.20)*	-1.09 *(-2.04, 0.13)*
$\beta_{t_3 \times \text{RA}}$	-0.30 *(-1.76, 1.15)*	-0.12 *(-1.37, 1.12)*	—	—
$\gamma_{t_2 \times \text{SLE}}$	-1.29 *(-1.96, -0.63)*	-1.25 *(-1.87, -0.62)*	-1.70 *(-2.63, -0.78)*	-1.76 *(-2.55, -0.97)*
$\beta_{t_2 \times \text{SLE}}$	1.39 *(0.34, 2.43)*	1.44 *(0.44, 2.44)*	—	—
$\gamma_{t_3 \times \text{SLE}}$	-0.97 *(-1.77, -0.16)*	-0.98 *(-1.75, -0.21)*	-1.34 *(-2.44, -0.24)*	-1.16 *(-2.07, -0.25)*
$\beta_{t_3 \times \text{SLE}}$	0.01 *(-1.19, 1.21)*	-0.001 *(-1.16, 1.16)*	—	—
σ	0.26	0.29	0.20	0.34
δ	-1.31	-1.20	-3.35	-2.43
s_{11}	0.40	0.40	0.45	0.56
s_{12}	-0.03	-0.06	—	—
s_{22}	0.90	0.89	—	—
$\ell(\hat{\boldsymbol{\theta}})$	-437.0	-440.2	-484.2	-572.1
AIC	924.1	930.4	994.3	1170.2

model $\pi_{ij} = \mathbf{\Phi}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \tau_{2i})$ is the probability of falling below the LOD, while in the TPIC model π_{ij} is the probability of belonging to a separate sub-LOD population. Therefore, $\hat{\pi}_{ij}$ is expected to be larger in the TP model, which is reflected in a higher estimate of the intercept β_0. Furthermore, both models estimate a negative skew-parameter δ, which gives a heavier left tail. The magnitude of $\hat{\delta}$ is greater in the TPIC model, which takes left-censoring below the LOD into account.

The expected proportion of censored observations from the continuous distribution in the TPIC model is given by

$$P(Y^*_{ij} < T | Y^*_{ij} > 0) = F(T; \hat{\mu}_{ij}, \hat{\sigma}, \hat{\delta}),$$

where $F(\cdot)$ in this case is the log-skew-normal distribution. If we set the age to the average $\bar{x}_{\text{age},i} = 31$ the population averaged expected proportion varies from 0.2% in the healthy group in the last time period t_3, to 1.9% in the healthy group during the first time period t_1. The weighted average based on the number of individuals in each group is 1.0%. In the simulation study in Chapter 5 we showed that the TP model and TPIC model gave close to equivalent results when the proportion was around 0.5%, and that the TP model had a substantial bias and loss in coverage when the proportion was 5% and higher. Thus, the TP model may be slightly more biased than the TPIC model for this data. There are no major differences in the parameter estimates between the two models, so the difference in bias cannot be large, but we choose to continue with the TPIC model as this is likely to give better predictions. The results of likelihood ratio tests for the different parts of the model is shown in Table 7.4.

The Presence of Skew

In order to examine the effects of allowing for skewness in the log-transformed observations, we fit the TPIC model with a lognormal continuous part. The estimated skew-parameter of the TPIC model with log-skew-normal continuous part is $\hat{\delta} = -1.31$, which gives a heavier left tail. Since the lognormal distribution is a special case of the log-skew-normal distribution with $\delta = 0$, the significance of δ can be demonstrated with a likelihood ratio test, as shown in Table 7.4. The test statistic is 11.4 and follows a χ^2_1 distribution under the null hypothesis. This results in a p-value of $p = 7 \cdot 10^{-4}$, which is strong evidence for skewness in the log-transformed data.

The parameter estimates of the fitted TPIC model without skew are presented in Table 8.2 in Appendix D. The biggest differences lie in the regression intercept γ_0 and the parameter σ of the continuous part, which is justifiable since these parameters have different interpretations in the two models. In the skew-normal distribution, γ_0 does not represent the expected value when all covariates are zero, and σ shares the variability with δ, as stated in (3.17). Calculating the mean and variance of the skew-normal distribution gives

$$\begin{aligned} E(X) &= \hat{\gamma}_0 + \hat{\delta}\sqrt{\frac{2}{\pi}} = 2.33 - 1.31\sqrt{\frac{2}{\pi}} = 1.28, \\ \text{Var}(X) &= \hat{\sigma}^2 + \hat{\delta}^2\Big(1 - \frac{2}{\pi}\Big) = 0.26^2 + (-1.31)^2\Big(1 - \frac{2}{\pi}\Big) = 0.69, \end{aligned}$$

Table 7.4: Results of likelihood ratio tests for the presence of skew, discrete part and fixed effects. All models are with a log-skew-normal continuous part, except for model (c) used for testing the presence of skewness.

Model	-2*loglikelihood	-2*Difference in loglikelihoods		Distribution	*p*-value
(a) TPIC	874.0				
Discrete part:					
(b) Tobit	968.4	94.4	(b-a)	$\chi^2_{10}, \chi^2_{11}, \chi^2_{12}$	$<2.2\cdot10^{-16}$
Skewness:					
(c) TPIC without skew	885.5	11.5	(c-a)	χ^2_1	$7\cdot10^{-4}$
Correlation in Random Effects:					
(d) TPIC without s_{12}	874.1	0.1	(d-a)	χ^2_1	0.75
Random Effects:					
(e) TPIC without τ_{1i}	902.2	28.1	(e-d)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$6\cdot10^{-8}$
(f) TPIC without τ_{2i}	900.8	26.7	(f-d)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$1\cdot10^{-7}$
Fixed effects in $\boldsymbol{\gamma}$:					
(g) TPIC without age	874.4	0.3	(g-d)	χ^2_1	0.58
(h) TPIC without interactions	890.0	15.9	(h-d)	χ^2_4	0.003
(i) TPIC without time	892.0	2.0	(i-h)	χ^2_2	0.37
(j) TPIC without diagnosis	894.7	4.7	(j-h)	χ^2_2	0.10
Fixed effects in $\boldsymbol{\beta}$:					
(k) TPIC without age	874.1	0.03	(k-d)	χ^2_1	0.86
(l) TPIC without interactions	887.9	13.8	(l-d)	χ^2_4	0.008
(m) TPIC without time	896.4	9.5	(m-l)	χ^2_2	0.01
(n) TPIC without diagnosis	893.9	6.0	(n-l)	χ^2_2	0.05

which is similar to the intercept $\hat{\gamma}_0 = 1.25$ and variance $\hat{\sigma}^2 = 0.60$ of the TPIC model without skew. Thus, the estimated mean and variance is similar, even if the parameter estimates are different.

There is also a notable difference in the intercept β_0 of the discrete part. In the TPIC model with skew $\hat{\beta}_0 = -0.66$ while the model without skew gives a higher estimate $\hat{\beta}_0 = -0.62$. This is related to δ, since the skewed TPIC model estimates a heavier left tail, and therefore that a larger proportion of the censored observations comes from the continuous part. The result is lower estimates of the discrete weight π_{ij}, reflected in a lower estimate of β_0.

Since allowing for skewness in the log-transformed data gives a highly significant improvement of the model fit ($p = 7 \cdot 10^{-4}$), we choose to do the remainder of the analysis on the TPIC model with a log-skew-normal continuous part.

Test for the Discrete Part

Since the Tobit model is a special case of the TPIC model, it is possible to use a likelihood ratio test to assess the significance of the discrete part of the TPIC model. In total, the Tobit model has 12 fewer parameters than the TPIC model, namely the ten parameters in $\boldsymbol{\beta}$ and the variance and covariance of the random effect τ_{2i}. Two of the parameters, the intercept β_0 and the variance s_{22} in the random effect τ_{2i}, lies on the boundary of the parameter space. Therefore the test-statistic follows a mixture of χ^2_{10}, χ^2_{11}, and χ^2_{12} under the null hypothesis (Self

and Liang, 1987). Since more degrees of freedom gives higher p-values, we use the χ^2_{12} distribution to get an upper bound for the p-value. As expected, this provides solid evidence for the presence of the discrete part, which means that all of the data have not originated from a single log-skew-normal distribution.

Based on the fitted TPIC model, the left-censored tail of the continuous part is very small. Therefore, the amount of observations below the LOD is much larger than what is expected from the observed part of the continuous distribution. In the simulation study in Chapter 5 the Tobit gave very misleading results in this kind of scenarios. It suffered a substantial bias in the estimated marginal effect, and it achieved a much higher CRPS than the two-part models. Based on this, the Tobit model is not a suitable choice for these data.

Test for Random Effects

The random effect $\boldsymbol{\tau}_i = (\tau_{1i}, \tau_{2i})'$ contributes with the three parameters of their covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}.$$

The variances s_{11} and s_{22} quantifies to what degree the random effects varies between the individuals. A low variance indicates that the differences between individuals are small, while a higher variance signifies that which individual the measurement is taken from greatly affects the expected response. A variance of zero means that there is no individual effect on the response, and thus no individual random effect. The covariance s_{12} quantifies to which degree the two random effects correlate. For the TPIC model with skew the estimated covariance is $\hat{s}_{12} = -0.03$, which corresponds to a correlation of $\hat{\rho}_{\boldsymbol{\tau}} = -0.05$. This means that a higher random effect in the continuous part tends to be accompanied by a lower random effect in the discrete part. In other words, individuals with higher than average positive responses tend to have lower than average probabilities of belonging to the discrete component, and vice versa.

Likelihood ratio tests are performed for the significance of the three parameters. The presence of the covariance is assessed with the hypotheses

$$H_0 : \ s_{12} = 0, \qquad H_1 : \ s_{12} \neq 0.$$

The test statistic follows a χ^2_1 distribution, which gives a p-value of 0.75. Thus, the null hypothesis is not contradicted, and there is no evidence for the presence of correlation. The presence of the random effects are tested with the hypotheses

$$H_0 : \ s_{11} = 0, \qquad H_1 : \ s_{11} > 0,$$

for the presence of τ_{1i} and likewise with s_{22} for the presence of τ_{2i}. Since H_0 lies on the boundary of the parameter space, the test statistics follows the mixture $\frac{1}{2}(\chi^2_0 + \chi^2_1)$. The chi-squared distribution with zero degrees of freedom is a point mass at zero. Thus, the resulting p-values are half of the p-values obtained from using the χ^2_1 distribution (Goldman and Whelan, 2000). As shown in Table 7.4,

the respective resulting p-values are $p = 6 \cdot 10^{-8}$ and $p = 1 \cdot 10^{-7}$, which is strong evidence for the presence of both random effects.

Since the correlation s_{12} complicates computations and interpretations and the LRT did not provide evidence for its existence, we choose to remove it in the further calculations.

Tests for Fixed Effects

Likelihood ratio tests are performed for the presence of the fixed effects in both model parts. The results are presented in Table 7.4. In both model parts the age parameter is found to not be significantly different from zero with $p = 0.58$ in the continuous part and $p = 0.86$ in the discrete part. Furthermore, the interaction effects between time and diagnosis are highly significant in both parts, which is strong evidence for a difference in the time profiles between the diagnoses. The resulting p-values are $p = 0.003$ and $p = 0.008$ in the continuous part and the discrete part, respectively.

Based on the results of the likelihood ratio tests, we choose to remove the age parameters, as there is no evidence for an effect of age on the cytokine concentrations. This makes inference on the time profiles much more straightforward, as only categorical covariates are present.

Population Averaged Time Profiles

In order to estimate the time profiles we use the probit/log-skew-normal TPIC model without age and the correlation s_{12} between the random effects. The resulting parameter estimates are shown in Table 8.2 in Appendix D. The log-likelihood of the simplified model is -437.14, whereas the full model with γ_{age}, β_{age} and s_{12} has a log-likelihood of -437.03. Thus, a likelihood ratio test for the combined significance of the three parameters has a test statistic of 0.22 that follows a χ^2_3 distribution under the null hypothesis. This gives a p-value of $p = 0.97$.

The expected value of the latent concentration Y^*_{ij} given that the response comes from the continuous part can be expressed as

$$E(Y^*_{ij}|Y^*_{ij} > 0) = 2e^{\mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}}+\tau_{1i}+(\hat{\sigma}^2+\hat{\delta}^2)/2}\boldsymbol{\Phi}(\hat{\delta}),$$

where τ_{1i} is the unobserved random effect for the individual. The random effect is assumed to follow a normal distribution with mean zero and variance $\hat{s}_{11}$. In order to calculate the expected value for unobserved members of the population, we must integrate over the possible random effects. Thus, the population averaged expected value conditioned on a non-zero latent concentration is

$$\begin{aligned} E(Y^*_{ij}|Y^*_{ij} > 0) &= 2e^{\mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}}+(\hat{\sigma}^2+\hat{\delta}^2)/2}\boldsymbol{\Phi}(\hat{\delta}) \int e^{\tau_{1i}} \frac{1}{\sqrt{\hat{s}_{11}}}\phi\Big(\frac{\tau_{i1}}{\sqrt{\hat{s}_{11}}}\Big)\, d\tau_{1i} \\ &= 2e^{\mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}}+(\hat{s}_{11}+\hat{\sigma}^2+\hat{\delta}^2)/2}\boldsymbol{\Phi}(\hat{\delta}) \end{aligned}$$

Furthermore, the population averaged probability of belonging to the point mass

at zero is found by integrating over the random effect τ_{i2},

$$\hat{\pi}_{ij} = \int_{-\infty}^{\infty} \mathbf{\Phi}(\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \tau_{i2}) \frac{1}{\sqrt{\hat{s}_{22}}} \phi\Big(\frac{\tau_{i2}}{\sqrt{\hat{s}_{22}}}\Big)\, d\tau_{i2}.$$

The first term is the cumulative distribution of a normally distributed variable with mean $\mu = -\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}$ and variance $\sigma = 1$, i.e. if $Z \sim \mathcal{N}(-\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}, 1)$ then $P(Z \leq \tau_{i2}) = \mathbf{\Phi}(\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \tau_{i2})$. The second term is the probability density of a variable with mean zero and variance s_{22}. Thus, if $Q \sim N(0, s_{22})$, then $P(Q = \tau_{i2}) = \frac{1}{\sqrt{s_{22}}}\phi\Big(\frac{\tau_{i2}}{\sqrt{s_{22}}}\Big)$. Furthermore, if Q and Z are independent we can write the integral as $\hat{\pi}_{ij} = \int_{-\infty}^{\infty} P(Z \leq Q | Q = \tau_{i2}) P(Q = \tau_{i2})\, d\tau_{i2}$, which equals the unconditional probability $P(Z \leq Q) = P(Z - Q \leq 0)$. Since both Z and Q are normally distributed their difference is also normally distributed with mean $-\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}$ and variance $1 + \hat{s}_{22}$. Therefore, the desired integral can be written as

$$\hat{\pi}_{ij} = \mathbf{\Phi}\left(\frac{\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}}{\sqrt{1 + \hat{s}_{22}}}\right).$$

Finally, because τ_{i1} and τ_{i2} are assumed to be independent the expected value of the lantent variable Y^*_{ij} is given by

$$\begin{aligned} E(Y^*_{ij}) &= \hat{\pi}_{ij} E(Y^*_{ij} | Y^*_{ij} = 0) + (1 - \hat{\pi}_{ij}) E(Y^*_{ij} | Y^*_{ij} > 0) \\ &= (1 - \hat{\pi}_{ij}) E(Y^*_{ij} | Y^*_{ij} > 0) \\ &= \mathbf{\Phi}\left(\frac{-\mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}}{\sqrt{1 + \hat{s}_{22}}}\right) 2e^{\mathbf{z}'_{ij}\hat{\boldsymbol{\gamma}} + (\hat{s}_{11} + \hat{\sigma}^2 + \hat{\delta}^2)/2} \mathbf{\Phi}(\hat{\delta}). \end{aligned}$$

The resulting estimates of the three quantities are plotted for each diagnostic group and time period in Figure 7.7. In this calculation of the latent expected value it is assumed that sub-LOD population with weight π_{ij} has a point mass at zero, but since it is unobserved it can, in theory, be any distribution fully contained on the interval [0,T]. The true latent concentration might therefore be higher. However, it does not seem unreasonable to assume that the sub-LOD population contains responses with zero or negligible concentrations. In this case $T = 0.29$, so even if the sub-LOD population is assumed to be a point mass at T, the resulting expected latent concentrations are almost identical to the ones displayed.

Figure 7.7 shows that the expected outcomes for patients diagnosed with RA and SLE are relatively similar, while the healthy controls clearly differ from the diagnosed patients. The healthy controls experience increased levels of TNF-α around birth (second time period), this is reflected in both a lower probability of being censored and greater magnitude in the non-censored observations. The diagnosed patients have a more stable concentration throughout the course of the pregnancy, which is also reflected in both parts of the model.

It is interesting to observe that for the healthy controls the expected value of the positive responses increases from the second to the third time period, while the probability of falling below the LOD also increases, resulting in an overall reduced

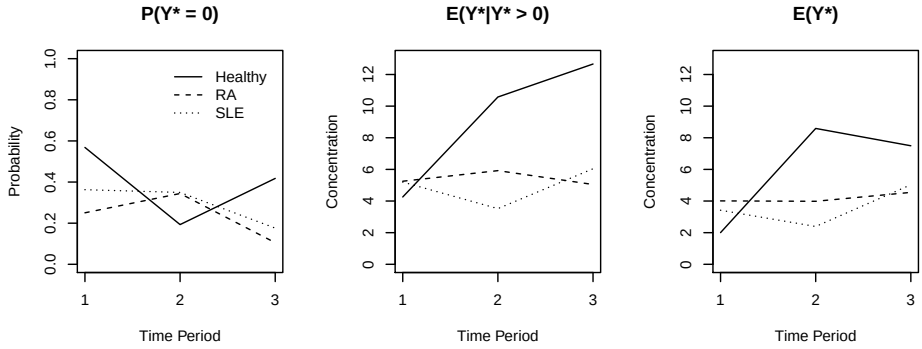

Figure 7.7: Probability of belonging to discrete part, expected value of the continuous part, and expected value of the latent concentration Y_{ij}^* of the cytokine TNF-α for each diagnosis and time period by the probit/log-skew-normal TPIC model. The first time period includes time point 0 - 2, the second time period is time point 3 - 4, and the last time period is time point 5 - 6.

expected observation from the second to the third time period. This means that more observations are censored, but those that are not censored tend to be higher. This example highlights that the mixing probability and the magnitude of the non-censored observations are the results of two independent processes. However, in general, an increase in the non-censored observations tend to coincide with a reduced probability of being censored.

7.2.2 Cytokine MCP1

For this cytokine, there are several time periods without any censored observations from several of the diagnostic groups. For instance, during the last time period t_3, there are no censored observations from Healthy controls or patients with SLE. In total, there are 11 censored observations, which are undoubtedly not enough to estimate all the ten parameters in $\boldsymbol{\beta}$ used in the analysis of TNF-α. Because the main focus of the analysis lies in detecting group differences, we choose to include only the diagnostic group effects in the discrete part. Thus,

$$\begin{aligned}\mu_{ij} =&\gamma_0 + \gamma_{\text{age}}\ x_{\text{age},i} + \gamma_{\text{RA}}\ x_{\text{RA},i} + \gamma_{\text{SLE}}\ x_{\text{SLE},i} + \gamma_{t_2}\ x_{t_2,ij} + \gamma_{t_3}\ x_{t_3,ij} + \\ &\gamma_{t_2\times\text{RA}}\ x_{\text{RA},i}\ x_{t_2,ij} + \gamma_{t_3\times\text{RA}}\ x_{\text{RA},i}\ x_{t_3,ij} + \\ &\gamma_{t_2\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_2,ij} + \gamma_{t_3\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_3,ij} + \tau_{i1},\end{aligned}$$

$$\pi_{ij} = \boldsymbol{\Phi}(\beta_0 + \beta_{\text{RA}}\ x_{\text{RA},i} + \beta_{\text{SLE}}\ x_{\text{SLE},i} + \tau_{i2}).$$

Including the parameter σ of the lognormal distribution and the variances of the random effects, this gives a total of 12 parameters in the one-part models and 17 parameters in the two-part models. Using the log-skew-normal distribution for the continuous part contributes with the additional skew parameter δ.

Based on the histograms in Figure 7.5 it is not entirely clear whether all the data may have arisen from a single continuous distribution, or if an additional part below the LOD is needed to explain the number of censored observations. Thus, the Tobit model may provide a good fit to the data. Since the amount of censored observations is so low, the substitute model may also be adequate.

The results of fitting the four models with a log-skew-normal continuous part are shown in Table 7.5.

Table 7.5: Resulting maximum likelihood parameter estimates from fitting the models with a log-skew-normal continuous distribution with 95 % Wald-type confidence intervals.

Parameter	TPIC	TP	Tobit	Substitute
γ_0	4.39 *(2.91, 5.88)*	4.39 *(2.90, 5.87)*	4.22 *(2.49, 5.96)*	4.23 *(2.51, 5.97)*
γ_{age}	-0.01 *(-0.06, 0.03)*	-0.01 *(-0.06, 0.03)*	0.01 *(-0.04, 0.06)*	0.01 *-0.04, 0.06)*
γ_{RA}	0.68 *(0.12, 1.24)*	0.68 *(0.12, 1.24)*	0.52 *(-0.14, 1.18)*	0.53 *-1.13, 1.18)*
γ_{SLE}	0.87 *0.30, 1.44)*	0.87 *(0.30, 1.44)*	0.69 *(0.03, 1.35)*	0.69 *(0.03, 1.35)*
γ_{t_2}	1.55 *(1.20, 1.90)*	1.55 *(1.20, 1.90)*	1.50 *(1.11, 1.88)*	1.50 *(1.20, 1.90)*
γ_{t_3}	1.64 *(1.18, 2.10)*	1.64 *(1.18, 2.10)*	1.65 *(1.14, 2.15)*	1.65 *(1.14, 2.16)*
$\gamma_{t_2 \times \text{RA}}$	-1.70 *(-2.26, -1.15)*	-1.70 *(-2.26, -1.15)*	-1.68 *(-2.27, -1.09)*	-1.69 *(-2.28, -1.09)*
$\gamma_{t_3 \times \text{RA}}$	-1.24 *(-1.87, -0.61)*	-1.24 *(-1.87, -0.61)*	-1.24 *(-1.95, -0.54)*	-1.25 *(-1.96, -0.54)*
$\gamma_{t_2 \times \text{SLE}}$	-1.67 *(-2.18, -1.17)*	-1.68 *(-2.18, -1.17)*	-1.50 *(-2.04,-0.96)*	-1.50 *(-2.05, -0.96)*
$\gamma_{t_3 \times \text{SLE}}$	-0.91 *(-1.54, -0.28)*	-0.91 *(-1.53, -0.28)*	-0.62 *(-1.34, 0.11)*	-0.61 *(-1.35, 0.12)*
β_0	-3.69 *(-6.27, -0.31)*	-3.67 *(-6.21, -1.13)*	—	—
β_{RA}	-1.38 *(-4.78, 2.01)*	-1.44 *(-4.90, 2.01)*	—	—
β_{SLE}	-0.50 *(-2.73, 1.73)*	-0.53 *(-2.77, 1.71)*	—	—
σ	0.60	0.60	0.39	0.39
δ	-0.82	-0.82	-1.39	-1.42
s_{11}	0.59	0.59	0.89	0.87
s_{12}	-1.19	-1.21	—	—
s_{22}	4.42	4.45	—	—
$\ell(\hat{\boldsymbol{\theta}})$	-432.7	-432.8	-461.4	-468.7
AIC	901.5	901.5	948.8	963.3

The Presence of Interval Censoring

The TP model and TPIC model achieve close to equal log-likelihoods, respectively -432.77 and -432.74, and the resulting parameter estimates from the two models are almost identical. When the age is set to the mean $\bar{x}_{age} = 31$, the population averaged proportion of the continuous distribution that falls below the LOD ranges from 0.0002% to 0.07% across the time periods and diagnoses.

The main argument for not including the interval censoring is that it makes the model more complex in several ways. The expression for the likelihood becomes more complex, since it includes the cumulative density function of the continuous part of the model. Furthermore, both parts of the mixture model play a role in the probability of falling below the LOD, which complicates computations and

interpretations. In the simulation study in Chapter 5, we showed that the interval censoring was superfluous when less than 0.5% of the continuous part was censored. Thus, it seems safe to conclude that the interval censoring is superfluous here.

In the TP model, π_{ij} represents the probability of falling below the LOD. Since there are sub-LOD responses in the data, the discrete parts have to be present. Alternatively, a Tobit or Substitute approach could be used to handle the sub-LOD observations. However, both of these models perform substantially worse in terms of AIC, so we conclude that the TP model is best suited for this data.

Table 7.6: Results of likelihood ratio tests for the presence of skew, discrete part and fixed effects. All models are fitted with a lognormal continuous part, except model (a) which is used to test the presence of skew.

Model	-2*loglikelihood	-2*Difference in loglikelihoods		Distribution	*p*-value
(a) TP with skew	865.5				
Skewness:					
(b) TP	866.8	1.3	(b-a)	χ^2_1	0.25
Correlation in Random Effects:					
(c) TP without s_{12}	873.5	6.7	(c-b)	χ^2_1	0.01
Random Effects:					
(d) TP without τ_{1i}	938.7	65.2	(d-c)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$3{\cdot}10^{-16}$
(e) TP without τ_{2i}	898.1	24.6	(e-c)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$4{\cdot}10^{-7}$
Fixed effects in $\boldsymbol{\gamma}$:					
(f) TP without age	867.0	0.2	(f-b)	χ^2_1	0.65
(g) TP without interactions	917.9	51.1	(g-b)	χ^2_4	$2{\cdot}10^{-10}$
(h) TP without time	956.1	38.2	(h-g)	χ^2_2	$5{\cdot}10^{-9}$
(i) TP without diagnosis	918.6	0.7	(i-g)	χ^2_2	0.70
Fixed effects in $\boldsymbol{\beta}$:					
(j) TP without diagnosis	868.2	1.4	(j-b)	χ^2_2	0.50

Tests for Model Parameters

The results of likelihood ratio tests for the presence of the different model parts of the TP model are displayed in Table 7.6. Based on these results, the skewness is insignificant ($p = 0.25$), and therefore omitted. As for the cytokine TNF-α, there is also no evidence for an effect of the patients' age on the response ($p = 0.65$). Unsurprisingly, due to the low number of observations below the LOD, there is no significant effect of the patients' diagnosis on the probability of falling below the LOD ($p = 0.50$). Omitting the diagnostic covariates in $\boldsymbol{\beta}$ leaves only the intercept β_0.

The results for this cytokine differs from the results on the cytokine TNF-α in that there is a significant covariance s_{12} at level of significance $\alpha = 0.01$ between the two random effects. Thus, the two random effects are not independent. This makes the computation of the population averaged time profiles more complicated. The simplified model upon removing the age covariate from $\boldsymbol{\gamma}$ and the diagnostic covariates from $\boldsymbol{\beta}$ is displayed in Table 8.1 in Appendix D.

Population Averaged Time Profiles

For the cytokine TNF-α we could calculate the population averaged $\hat{\pi}_{ij}$ and conditional expected value $E(Y_{ij}^*|Y_{ij}^* > 0)$ separately, and the resulting population averaged expected value was simply given by $(1-\hat{\pi}_{ij})E(Y_{ij}^*|Y_{ij}^* > 0)$. In the resulting model for the cytokine MCP1, the two random effects are correlated. Thus, the population averaged expected value is found by integrating over the two random effects simultaneously, i.e.

$$\begin{aligned} E(Y_{ij}) &= \iint (1 - \boldsymbol{\Phi}(\hat{\beta}_0 + \tau_{2i}))e^{\mathbf{x}'_{ij}\hat{\boldsymbol{\gamma}}+\tau_{1i}+\hat{\sigma}^2/2} f(\boldsymbol{\tau}|\hat{\boldsymbol{\Sigma}})\ d\boldsymbol{\tau} \\ &= e^{\mathbf{x}'_{ij}\hat{\boldsymbol{\gamma}}+\hat{\sigma}^2/2} \iint (1 - \boldsymbol{\Phi}(\hat{\beta}_0 + \tau_{2i}))e^{\tau_{1i}} f(\boldsymbol{\tau}|\hat{\boldsymbol{\Sigma}})\ d\boldsymbol{\tau} \\ &= 1.31\ e^{\mathbf{x}'_{ij}\hat{\boldsymbol{\gamma}}+\hat{\sigma}^2/2}, \end{aligned}$$

where $f(\boldsymbol{\tau}|\hat{\boldsymbol{\Sigma}})$ is the bivariate probability distribution of $\boldsymbol{\tau}$ with mean zero and covariance matrix $\hat{\boldsymbol{\Sigma}}$. As before, the population averaged probability of belonging to the discrete part is given by

$$\hat{\pi}_{ij} = \int \boldsymbol{\Phi}(\hat{\beta}_0 + \tau_{2i}) f(\tau_{2i}|\hat{s}_{22})\ d\tau_{2i} = \boldsymbol{\Phi}\left(\frac{\hat{\beta}_0}{\sqrt{1+\hat{s}_{22}}}\right) = \boldsymbol{\Phi}\left(\frac{-4.11}{\sqrt{1+3.85}}\right) = 3.1\%.$$

The conditional expected value of the continuous part is

$$E(Y_{ij}|Y_{ij} > 0) = \int e^{\mathbf{x}'_{ij}\hat{\boldsymbol{\gamma}}+\tau_{1i}+\hat{\sigma}^2/2} f(\tau_{1i}|\hat{s}_{11})\ d\tau_{1i} = 1.32\ e^{\mathbf{x}'_{ij}\hat{\boldsymbol{\gamma}}+\hat{\sigma}^2/2}.$$

Note that this is almost identical to the unconditional expected value $E(Y_{ij})$, due to the low amount of censored observations. Therefore, we plot only the unconditional expected value in Figure 7.8.

It is important to note that the TP model is used, thus the inference is done on the observed concentration Y_{ij}, and not the latent concentration Y_{ij}^*. This means that all the concentrations below the LOD is assumed to be zero. The TP model is however shown to give almost identical parameter estimates as the TPIC model for this data. Therefore, this distinction is virtually immaterial here.

7.2.3 Cytokine IL8

For this cytokine, there are 71 non-censored observations in total. The group with the highest proportion of censored observations is the healthy controls, which has no non-censored observations during the last time period t_3. Thus, there is not enough data to support three time periods with interactions in the continuous part, $\boldsymbol{\gamma}$. Therefore, we choose the divide the data into two time periods instead. The first time period, denoted t_1, contains the measurements conducted before birth (time point 0 - 3), and the last time period t_2 contains the postpartum measurements

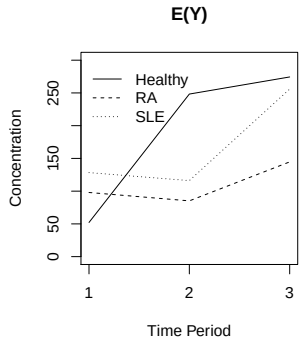

Figure 7.8: Expected value of the observed concentration Y_{ij} of the cytokine MCP1 for each diagnosis and time period by the probit/lognormal TP model. The first time period includes time point 0 - 2, the second time period is time point 3 - 4, and the last time period is time point 5 - 6.

(time point 4 - 6). The resulting parameterization is

$$\begin{aligned}\mu_{ij} =&\gamma_0 + \gamma_{\text{age}}\ x_{\text{age},i} + \gamma_{\text{RA}}\ x_{\text{RA},i} + \gamma_{\text{SLE}}\ x_{\text{SLE},i} + \gamma_{t_2}\ x_{t_2,ij} + \\ &\gamma_{t_2\times\text{RA}}\ x_{\text{RA},i}\ x_{t_2,ij} + \gamma_{t_2\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_2,ij} + \tau_{i1},\end{aligned}$$

$$\begin{aligned}\pi_{ij} =&\mathbf{\Phi}(\beta_0 + \beta_{\text{age}}\ x_{\text{age},i} + \beta_{\text{RA}}\ x_{\text{RA},i} + \beta_{\text{SLE}}\ x_{\text{SLE},i} + \beta_{t_2}\ x_{t_2,ij} + \\ &\beta_{t_2\times\text{RA}}\ x_{\text{RA},i}\ x_{t_2,ij} + \beta_{t_2\times\text{SLE}}\ x_{\text{SLE},i}\ x_{t_2,ij} + \tau_{i2}).\end{aligned}$$

Including the 2×2 covariance matrix $\mathbf{\Sigma}$ of the random effects and the parameter σ of the lognormal distribution, this gives a total of 9 parameters in the one-part models and 18 parameters in the two-part models. Using the log-skew-normal distribution for the continuous part gives one additional skew parameter δ.

Based on Figure 7.6, it seems highly unlikely that all the observations have originated from a single continuous distribution. Therefore, the two-part models are assumed to be superior. Furthermore, there are several observations close to the LOD, which is indicative of left-censoring from the continuous part. This points towards the TPIC model as the best choice.

The results of fitting the four models with a log-skew-normal continuous part are displayed in Table 7.7. As expected, the two-part models are superior to the one-part models in terms of AIC. In particular, the substitute model is clearly not suitable for this data. As with the borrelia antibody concentration data with 96.8% censored observations examined in Chapter 6, the substitute model estimates a very small $\hat{\sigma}$, such that the fitted distribution is in practice a lognormal distribution that is left-truncated at $y = \hat{\mu}_{ij}$.

Table 7.7: Maximum likelihood parameter estimates from fitting the four models with a log-skew-normal continuous part to the data on the cytokine IL8. Wald-type 95% confidence intervals are included in parenthesis for the fixed effects.

Parameter	TPIC	TP	Tobit	Substitute
γ_0	5.42 *(0.58, 10.3)*	3.72 *(-0.43, 7.87)*	-2.11 *(-6.90, 2.68)*	-0.34 *(-1.17, 0.49)*
β_0	-2.27 *(-12.2, 7.52)*	-5.76 *(-14.7, 3.17)*	—	—
γ_{age}	-0.21 *(-0.37, -0.04)*	-0.18 *(-0.32, -0.03)*	-0.04 *(-0.20, 0.13)*	-0.03 *(-0.05, 0.00)*
β_{age}	0.16 *(-0.18, 0.50)*	0.31 *(-0.02, 1.86)*	—	—
γ_{RA}	0.23 *(-0.38, 0.83)*	0.38 *(-0.83, 1.58)*	0.57 *(-0.98, 2.13)*	0.22 *(-0.14, 0.59)*
β_{RA}	-0.84 *(-3.84, 2.16)*	-0.85 *(-3.32, 1.61)*	—	—
γ_{SLE}	1.24 *(-0.31, 2.79)*	0.99 *(-0.22, 2.21)*	-0.08 *(-1.59, 1.42)*	0.37 *(-0.08, 0.84)*
β_{SLE}	0.29 *(-2.30, 2.89)*	-0.32 *(-2.82, 2.18)*	—	—
γ_{t_2}	1.20 *(0.65, 1.75)*	1.18 *(0.60, 1.76)*	0.75 *(-0.12, 1.62)*	-0.03 *(-0.12, 0.07)*
β_{t_2}	1.30 *(-0.15, 2.75)*	0.73 *(-0.39, 1.86)*	—	—
$\gamma_{t_2 \times \text{RA}}$	-0.81 *(-1.57, -0.05)*	-0.75 *(-1.49, -0.003)*	-0.36 *(-1.62, 0.90)*	0.05 *(-0.08, 0.18)*
$\beta_{t_2 \times \text{RA}}$	-1.55 *(-3.74, 0.64)*	-1.22 *(-2.70, 0.25)*	—	—
$\gamma_{t_2 \times \text{SLE}}$	-0.75 *(-1.42, -0.08)*	-0.69 *(-1.38, 0.01)*	0.19 *(-1.09, 1.48)*	0.07 *(-0.09, 0.23)*
$\beta_{t_2 \times \text{SLE}}$	-2.68 *(-4.96 -0.40)*	-2.03 *(-3.95, -0.12)*	—	—
σ	0.20	0.14	0.29	0.02
δ	-0.82	0.90	-2.61	0.88
s_{11}	3.35	1.61	27.8	0.67
s_{12}	-4.99	-3.67	—	—
s_{22}	10.4	10.0	—	—
$\ell(\hat{\boldsymbol{\theta}})$	-157.5	-164.4	-195.8	-305.4
AIC	353.0	366.9	411.5	610.8

The Presence of Interval Censoring

Based on the results in Table 7.7, we see that there are some substantial differences between the TP model and the TPIC model, and that the latter achieves a substantially better fit to the data. The population averaged proportion of censored observations from the continuous part, $F(T; \hat{\mu}_{ij}, \hat{\sigma}, \hat{\delta})$, for individuals with age $\bar{x}_{\text{age}} = 31$ ranges from 48.4% in the SLE group during the second time period to 67.7% in the healthy group during the first time period. In the simulation study in Chapter 5, the TP model was shown to give biased estimates in scenarios with a non-negligible censored left tail of the continuous distribution. Therefore, we continue with the TPIC model for this data.

Tests for Model Parameters

Likelihood ratio tests are performed for the presence of the various parameters of the TPIC model. The results are presented in Table 7.8. Neither the skew parameter δ, the correlation s_{12} between the random effects, nor the age parameters, are

Table 7.8: Results of likelihood ratio tests for the presence of skew, discrete part and fixed effects. All models are with a log-skew-normal continuous part, except for model (c) used for testing the presence of skewness.

Model	-2*loglikelihood	-2*Difference in loglikelihoods		Distribution	*p*-value
(a) TPIC with skew	315.0				
Discrete part:					
(b) Tobit with skew	391.5	76.5	(b-a)	$\chi^2_7, \chi^2_8, \chi^2_9$	$<7{\cdot}10^{-14}$
Skewness:					
(c) TPIC	319.0	4.0	(c-a)	χ^2_1	0.04
Correlation in Random Effects:					
(d) TPIC without s_{12}	324.1	5.1	(d-c)	χ^2_1	0.02
Random Effects:					
(e) TPIC without τ_{1i}	383.2	59.1	(e-d)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$7{\cdot}10^{-15}$
(f) TPIC without τ_{2i}	361.7	37.6	(f-d)	$\frac{1}{2}(\chi^2_0 + \chi^2_1)$	$4{\cdot}10^{-10}$
Fixed effects in $\boldsymbol{\gamma}$:					
(g) TPIC without age	325.9	1.8	(g-d)	χ^2_1	0.18
(h) TPIC without interactions	326.5	0.6	(h-d)	χ^2_2	0.74
(i) TPIC without time	340.2	13.7	(i-h)	χ^2_1	$2{\cdot}10^{-4}$
(j) TPIC without diagnosis	331.3	4.8	(j-h)	χ^2_2	0.09
Fixed effects in $\boldsymbol{\beta}$:					
(k) TPIC without age	324.3	0.2	(k-d)	χ^2_1	0.65
(l) TPIC without interactions	330.5	6.4	(l-d)	χ^2_2	0.04
(m) TPIC without time	330.5	0.02	(m-l)	χ^2_1	0.89
(n) TPIC without diagnosis	331.3	0.6	(n-l)	χ^2_2	0.44

significant at level of significance $\alpha = 0.01$. Therefore, they are omitted in order to simplify the inference.

The interaction terms, as well as the diagnosis covariates, are insignificant in both model parts, also when tested together ($p = 0.12$). This means that when splitting the time points into two periods, before birth and after birth, and regarding the RA and SN-RA patients as the same diagnosis, there are no significant differences in the measured concentrations across the diagnostic groups.

7.3 Bivariate Statistical Analysis

Methods for bivariate statistical analysis of variables with a lower limit of detection are presented in Section 3.3. In this section, these methods will be applied to data on two of the previously analyzed cytokines, TNF-α and IL8. As well as demonstrating the applicability of the methods, the main goal is to estimate the correlation between the two cytokines.

Let $\mathbf{y}_i = (y_{1i}, y_{2i}) = (y_{\text{TNF-}\alpha,i}, y_{\text{IL8},i})$ be a pair of observations from the same individual at the same time point, with detection limits T_1 and T_2. As both the measurements of TNF-α and IL8 are subject to a lower detection limit, there are four possible types of pairs; (1) Both y_{1i} and y_{2i} are observed, (2) y_{1i} is observed and $y_{2i} \leq T_2$, (3) $y_{1i} \leq T_1$ and and y_{2i} is observed, and (4) both $y_{1i} \leq T_1$ and $y_{2i} \leq T_2$.

The simplest model is a one-part bivariate Tobit model, which is based on the assumption that all the data have originated from a single bivariate lognormal distribution subject to censoring due to the detection limits. This model is described by the likelihoods in (3.26). Under this model, the amount of observed pairs of type (2) to (4) corresponds to the observed distribution of the non-censored pairs of type (1). The model consists of the five parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where (μ_1, σ_1^2) defines the marginal distribution of y_1, (μ_2, σ_2^2) defines the marginal distribution of y_2, and ρ is the correlation between the two variables.

The restriction on the number of observed pairs of type (4), where both variables are below the LOD, can be relaxed by introducing a lower part $f_L(\mathbf{y}_i)$ that is entirely contained on the domain $[0, T_1] \times [0, T_2]$. Let the high component $f_H(\mathbf{y}_i)$ be a bivariate lognormal distribution. The resulting distribution is a two-part mixture model

$$f(\mathbf{y}_i) = \pi_i f_L(\mathbf{y}_i) + (1 - \pi_i) f_H(\mathbf{y}_i),$$

where π_i denotes the probability that $\mathbf{y}_i$ belongs to the lower component. Under this model the amount of half-observed pairs of type (2) and (3) must correspond to the distribution of the fully observed pairs of type (1), while the amount of censored pairs of type (4) may exceed the expected proportion based on the rest of the data. This requires one additional parameter π, such that the total number of parameters is six, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho, \pi)$.

The restrictions can be relaxed further by introducing more parts to the mixture. As in the two-part mixture, let $f_H(\mathbf{y}_i)$ be a higher bivariate lognormal component, and $f_L(\mathbf{y}_i)$ be a lower component located entirely on the domain $[0, T_1] \times [0, T_2]$. In addition, let $f_{L_1}(\mathbf{y}_i)$ be a distribution on the domain $[0, \infty) \times [0, T_2]$, and $f_{L_2}(\mathbf{y}_i)$ be on the domain $[0, T_1] \times [0, \infty)$. In other words, $f_L(\mathbf{y}_i)$ represent low-repsonders in both cytokines, $f_{L_1}(\mathbf{y}_i)$ represents low-responders in y_2, and $f_{L_2}(\mathbf{y}_i)$ represents low-responders in y_1. We denote the mixing weights of $f_{L_1}(\mathbf{y}_i)$ and $f_{L_2}(\mathbf{y}_i)$ respectively π_{1i} and π_{2i}. The resulting four-part mixture distribution is

$$f(\mathbf{y}_i) = \pi_{1i} f_{L_1}(\mathbf{y}_i) + \pi_{2i} f_{L_2}(\mathbf{y}_i) + \pi_i f_L(\mathbf{y}_i) + (1 - \pi_{1i} - \pi_{2i} - \pi_i) f_H(\mathbf{y}_i).$$

For the two half-censored components $f_{L_1}(\mathbf{y}_i)$ and $f_{L_2}(\mathbf{y}_i)$ only the marginal distribution of the non-censored variable and their relative weight are observable. Thus, the two components contribute with three parameters each. This gives a total of twelve parameters, $\boldsymbol{\theta} = (\pi_1, \pi_2, \pi, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho, \mu_{L_1}, \sigma_{L_1}, \mu_{L_2}, \sigma_{L_2})$. Here μ_{L_1} and σ_{L_1} are the parameters of $f_{L_1}(\mathbf{y}_i)$, and μ_{L_2} and σ_{L_2} are the parameters of $f_{L_2}(\mathbf{y}_i)$. The model can be simplified by assuming that the half-observed lower components $f_{L_1}(\mathbf{y}_i)$ and $f_{L_2}(\mathbf{y}_i)$ has the same marginal distributions as the higher component $f_H(\mathbf{y}_i)$, such that $\mu_{L_1} = \mu_1$, $\sigma_{L_1} = \sigma_1$, $\mu_{L_2} = \mu_2$ and $\sigma_{L_2} = \sigma_2$.

In the previous longitudinal analysis, we showed that there is a significant correlation between measurements from the same individual, such that all the observed pairs cannot be regarded as independent. Therefore, we limit the analysis to one time-point, $t = 2$, which has the most observations. This time-point is the second trimester of the pregnancies, and includes a total of 59 observations. The number of observations of each type is shown in Table 7.9. Among the 59 pairs, only ten are non-censored in both variables.

Table 7.9: Number of observed pairs of each type at time point two, where y_{1i} is the cytokine TNF-α and y_{2i} is IL8.

	$y_{1i} > T_1$	$y_{1i} \leq T_1$	Total
$y_{2i} > T_2$	10 *(17 %)*	2 *(3 %)*	12 *(20 %)*
$y_{2i} \leq T_2$	27 *(46 %)*	20 *(34 %)*	47 *(80 %)*
Total	37 *(63 %)*	22 *(37 %)*	59

The results of fitting a bivariate lognormal distribution to only the ten fully observed pairs are shown in Table 7.10. The estimated correlation is $\hat{\rho} = 0.63$ (95% CI : $0.15, 0.89$). The estimate is subject to great uncertainty due to a low number of observations, but there is a significantly positive correlation between the measurements of type (1) of the two cytokines. This may, however, be a biased estimate of the underlying correlation, due to both variables being subject to detection limits. In the longitudinal analysis in the previous section, we showed that the data on both TNF-α and IL8 showed signs of non-negligible left-censored tails, based on the TPIC model providing a substantially better fit than the TP model. Therefore, one of the censored bivariate models might provide a better estimate of the correlation.

Table 7.10: Maximum likelihood parameter estimates from fitting a bivariate lognormal distribution to the pairs of type (1) that are non-censored in both variables. 95 % Wald-type confidence intervals are shown in parenthesis.

Parameter	Estimate
μ_1	1.39 *(0.73, 2.06)*
σ_1	1.07 *(0.60, 1.54)*
μ_2	1.67 *(0.87, 2.46)*
σ_2	1.28 *(0.72, 1.84)*
ρ	0.63 *(0.15, 0.89)*

Since both cytokines were shown to have a significant amount of excess observations below the LOD in the longitudinal analysis, the underlying assumptions of the bivariate Tobit model are not satisfied. The bivariate two-part model allows for excess observations below the LOD, but it is assumed that both variables are low responses simultaneously. This is not the case for this data, as there are a lot more excess zeroes in IL8 than in TNF-α. Based on this, it is expected that the bivariate four-part model provides the best fit.

The parameters of the three models, as well as the simplified version of the four-part model, are estimated using the command `optim()` in R with `method = "BFGS"` to maximize the likelihoods. The R code used to estimate the parameters is included in Appendix C. The results are presented in Table 7.11.

The behaviour of the models is investigated by plotting the resulting higher components $f_H(\mathbf{y}_i)$ in Figure 7.9. The upper left plot is the result of fitting a bivariate normal distribution to the log-transformed fully observed pairs of type (1). It does not take into account that there might be observations from the higher

Table 7.11: Resulting maximum likelihood parameter estimates from fitting the models with a log-skew-normal continuous distribution with 95 % Wald-type confidence intervals.

Parameter	Tobit	Two-Part	Four-Part	Simplified Four-Part
μ_1	0.19 *(-0.27, 0.66)*	0.81 *(0.09, 1.53)*	1.01 *(0.57, 1.46)*	1.11 *(0.65, 1.56)*
σ_1	1.67 *(1.26, 2.08)*	1.30 *(0.79, 1.82)*	0.97 *(0.65, 1.30)*	1.06 *(0.71, 1.41)*
μ_2	0.36 *(-0.27, 0.98)*	0.60 *(-0.01, 1.22)*	0.61 *(-0.08, 1.29)*	0.53 *(-0.09, 1.16)*
σ_2	1.18 *(0.72, 1.64)*	1.22 *(0.70, 1.55)*	1.24 *(0.74, 1.75)*	1.13 *(0.70, 1.55)*
ρ	0.51 *(0.12, 0.77)*	0.38 *(-0.05, 0.70)*	0.56 *(0.14, 0.83)*	0.42 *(0.02, 0.72)*
π	—	0.23	$2{\cdot}10^{-6}$	0.15
π_1	—	—	0.07	$2{\cdot}10^{-4}$
π_2	—	—	0.34	0.18
μ_{L_1}	—	—	2.48 *(2.07, 2.89)*	—
σ_{L_1}	—	—	0.29 *(0.01, 0.57)*	—
μ_{L_2}	—	—	1.11 *(0.91, 1.31)*	—
σ_{L_2}	—	—	0.14 *(0.01, 0.27)*	—
$\ell(\hat{\boldsymbol{\theta}})$	-120.2	-119.5	-110.6	-118.0
AIC	250.4	251.0	245.2	252.0

component that are partially or fully censored due to the detection limits. In the Tobit model shown in the upper right corner, one bivariate normal distribution is fitted to all the data. Therefore, a large portion of the distribution extends below the LODs to account for the censored observations. In the two-part model an additional part is introduced on the domain $[0, T_1] \times [0, T_2]$. Therefore, the higher component does not extend as far down in the lower left corner. In the four-part model two additional parts are introduced on the domains $[0, T_1] \times [0, \infty)$ and $[0, \infty) \times [0, T_2]$, resulting in a higher component that extends even less below the LODs. However, it still extends much farther below T_2 then the model based only on the observed pairs, so a substantial proportion of the censored observations of IL8 are expected to be censored observations from the higher component. This is in agreement with the results found in the longitudinal analysis, where the continuous part of the fitted TPIC model had a large left-censored tail.

The best model in terms of AIC is the full four-part model. It estimates the correlation to be $\hat{\rho} = 0.56$ (95% CI : $0.14, 0.83$). This model is, however, arguably overfitted. The estimated mean of y_{1i} in the higher component is $\hat{\mu}_1 = 1.01$. The lower component $f_{L_1}(\mathbf{y}_i)$, which represents the high-responders in y_{1i} and low-responders in y_{2i}, has a substantially higher estimated mean $\hat{\mu}_{L_1} = 2.48$. The observations of type (2) are assumed to have originated from either $f_H(\mathbf{y}_i)$ or $f_{L_1}(\mathbf{y}_i)$. Thus, the pairs of type (2) with higher observations of y_{1i} are expected to come from $f_{L_1}(\mathbf{y}_i)$ and be low-responders in y_{2i}, while the pairs of type (2) with lower observations of y_{1i} are expected to be censored observations from $f_H(\mathbf{y}_i)$. This does not seem reasonable, and the good fit is apparently a result of overfitting.

The room for overfitting is reduced in the simplified four-part model, where the marginal distributions of $f_{L_1}(\mathbf{y}_i)$ and $f_{L_2}(\mathbf{y}_i)$ are forced to be the same as the

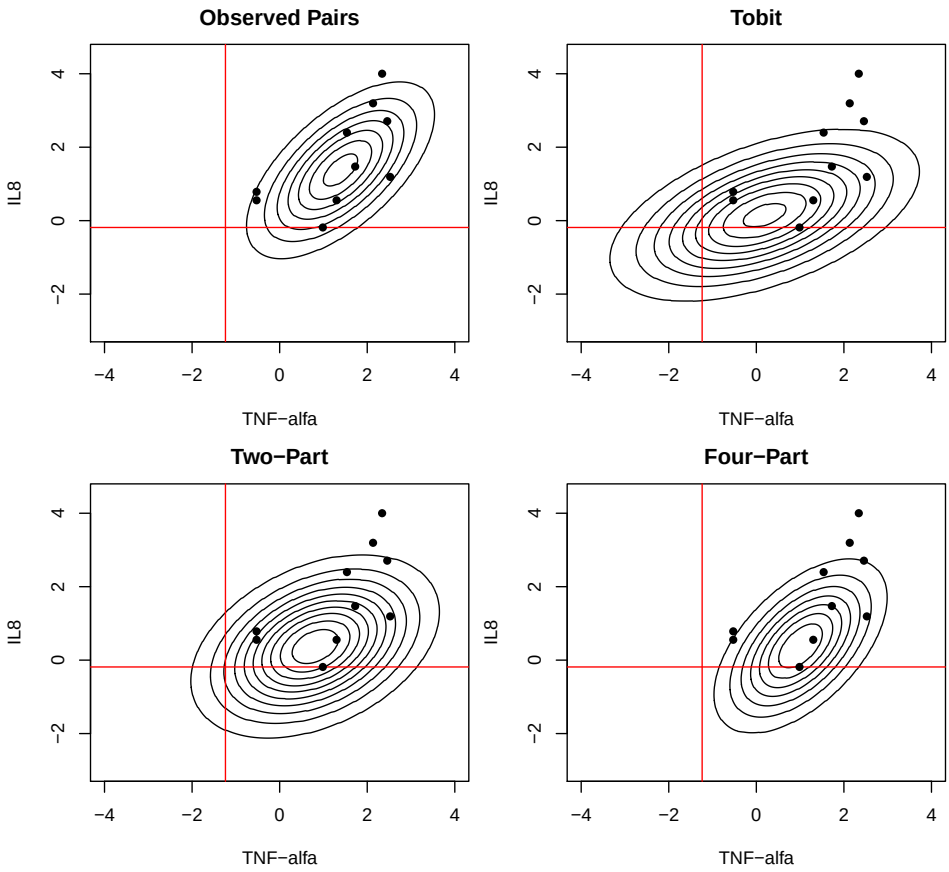

Figure 7.9: The fitted bivariate normal higher components $f_H(\mathbf{y}_i)$ to the log-transformed concentrations. The red lines are the detection limits, and the black points are the fully observed pairs of type (1). The upper left plot is the result of fitting a bivariate normal distribution to only the observed pairs and ignoring the rest of the data, with parameters shown in Table 7.10, while the other three are based on results in Table 7.11.

marginal distributions of $f_H(\mathbf{y}_i)$. This assumption is however problematic. Since the two cytokines apparently are not independent, it is not reasonable to assume that the measurements of TNF-α paired with a censored observation in IL8 follows the same distribution as the measurements of TNF-α paired with uncensored observations of IL8, and vice versa.

In conclusion, none of the models considered here are suitable for estimating the correlation in this particular data. The underlying assumptions of the Tobit and the two-part model are not met, and the attempt to generalize the two-part model by introducing two additional parts leads to overfitting. The overfitting is reduced by adding restrictions to the additional parts in the four-parts model, but still, the underlying assumptions are questionable at best.

8 | Conclusion and Further Work

In this thesis, we have studied methods for analyzing data subject to detection limits. Four models were specified; The Tobit model, a two-part mixture model, a censored two-part mixture model, and a naive substitute model. The performance of the four models was tested in numerous scenarios with data generated from a two-part model with interval censoring in a simulation study. This provided insight into the behavior of the models under different circumstances. In particular, it was demonstrated that even if the underlying process was a censored two-part model, there are certain scenarios where other models are better choices for analysis, both in terms of inference and prediction. When the true zeroes constituted less than half of the censored observations, the two-part model with interval censoring was over-parameterized, as there were not enough data to distinguish between the true and false zeroes. Both the Tobit model and the uncensored two-part model provided viable alternatives in these scenarios, depending on the problem at hand. In terms of prediction, the Tobit model achieved the best CRPS for the highest detection limits, while the two-part model was shown to be better for inference purposes with certain structures in the data.

The application on the borrelia data showed that the binary mixture model with interval censoring can be a useful alternative to logistic regression when estimating prevalence in a population. The two approaches differ in that the mixture model takes more information into account as a model is fitted to directly to the observed concentrations. However, for the borrelia data, the number of non-censored observations was so low that the continuous part of the mixture was subject to great uncertainty. Thus, logistic regression was arguably more practical.

The applicability of the four candidate models in a longitudinal setting was demonstrated on three cytokines in the cytokine data. For the cytokine TNF-α with 32.5% censored observations, there was a significant skew in the log-transformed data. Thus the probit/log-skew-normal two-part model with interval censoring provided the best fit. For the cytokine MCP1 with 3.6% censored observations, the estimated left censored tail of the continuous distribution was negligible, making the interval censoring superfluous. The probit/lognormal two-part model was best suited for the data. The cytokine IL8 with 76.9% censored observations had substantial estimated left-censoring of the continuous part, making the interval censoring crucial. Significant differences in the time profiles between the diagnostic groups were detected in the two first-mentioned cytokines.

There are methods for handling left-censored data that are not included in this thesis. We have considered the substitute model, where all the censored observations are substituted with half the detection limit. As expected, this led to considerable bias in estimates en predictions, and was not suitable for analysis in any of the applications. More sophisticated methods for replacing the censored observations with alternative values do, however, exist. In multiple imputation (MI), the missing values are replaced with random variables from a suitable distribution. Several imputed datasets are generated in order to minimize the variance related with imputation. This method has been shown to provide valid statistical inference when less than half of the observations are censored (Lee et al., 2012).

Bivariate models were applied in order to estimate the correlation between the cytokines TNF-α and IL8. A significantly positive correlation was estimated in three of the four candidate models. However, none of the models were suitable for the data in question. The underlying assumptions in the one- and two-part models were not satisfied, and the four-part models resulted in overfitting. There are, however, many other approaches for multivariate analysis of data on this form. Exploring more methods in the multivariate realm could make it possible to analyze the diagnostic effects on the complete cytokine profiles, and not only on individual cytokines. Among other possibilities, Lee and Scott (2012) formulated an EM algorithm for multivariate analysis of data with interval censoring below a detection limit and zero inflation.

In this thesis, we used a frequentist approach for estimation of the model parameters. In further analysis, it would be interesting to try a Bayesian approach. This would open for incorporation of prior information, which has the potential to give better estimates and predictions. In particular, we would like to attempt to use integrated nested Laplace approximation (INLA), which has shown great success in a wide range of applications (Rue et al., 2017). In order to use this, the model must be expressed as a latent Gaussian model and the observations may only depend on a linear combination of latent nodes. The latter is not the case for the two-part model with interval censoring, but we see no reasons for why INLA should not be successful with the other univariate models.

References

Abonazel, M. R. (2018). A practical guide for creating Monte Carlo simulation studies using R. *International Journal of Mathematics and Computational Science*, 4(1):18–33.

Ajeganova, S. and Huizinga, T. W. (2015). Rheumatoid arthritis: Seronegative and seropositive RA: Alike but different? *Nature Reviews Rheumatology*, 11(1):8. http://doi.org/10.1038/nrrheum.2014.194.

Andersen, A., Benn, C. S., Jørgensen, M. J., and Ravn, H. (2013). Censored correlated cytokine concentrations: Multivariate Tobit regression using clustered variance estimation. *Statistics in medicine*, 32(16):2859–2874. https://doi.org/10.1002/sim.5696.

Azzalini, A. (2017). R package 'sn'. Version 1.5-0.

Baey, C., Cournède, P.-H., and Kuhn, E. (2017). Likelihood ratio test for variance components in nonlinear mixed effects models. *ArXiv e-prints*. http://arxiv.org/abs/1712.08567v1.

Berk, K. N. and Lachenbruch, P. A. (2002). Repeated measures with zeros. *Statistical Methods in Medical Research*, 11(4):303–316. https://doi.org/10.1191/0962280202sm293ra.

Bernhardt, P. W. (2018). Maximum likelihood estimation in a semicontinuous survival model with covariates subject to detection limits. *The international journal of biostatistics*, 14(2). https://doi.org/10.1515/ijb-2017-0058.

Bio-Rad Laboratories, Inc. (n.d.). Bio-Plex® Multiplex System Brochure, Ver E. Brochure.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer New York. https://doi.org/10.1007/978-1-4614-4818-1.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292. https://doi.org/10.1002/sim.2673.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Brooks/Cole, Cengage Learning, 2 edition.

Chai, H. S. and Bailey, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Stat Med*, 27(18):3643–55. https://doi.org/10.1002/sim.3210.

Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, 54(3/4):573–578. https://doi.org/10.2307/2335048.

Chu, H., Moulton, L. H., Mack, W. J., Passaro, D. J., Barroso, P. F., and Muñoz, A. (2005). Correlating two continuous variables subject to detection limits in the context of mixture distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(5):831–845. https://doi.org/10.1111/j.1467-9876.2005.00512.x.

Chu, H., Nie, L., and Zhu, M. (2008). On estimation of bivariate biomarkers with known detection limits. *Environmetrics: The official journal of the International Environmetrics Society*, 19(3):301–317. https://doi.org/10.1002/env.868.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to demand for durable goods. *Econometrica*, 39(5):829–844. https://doi.org/10.2307/1909582.

Crowther, J. R. (1995). *ELISA: Theory and practice*, volume 42. Springer Science & Business Media. https://doi.org/10.1385/0896032795.

Dagne, G. A. (2016). A growth mixture Tobit model: Application to AIDS studies. *Journal of Applied Statistics*, 43(7):1174–1185. https://doi.org/10.1080/02664763.2015.1092114.

Dagne, G. A. (2017). Joint two-part Tobit models for longitudinal and time-to-event data. *Statistics in Medicine*, 36(26). https://doi.org/10.1002/sim.7429.

Dagne, G. A. and Huang, Y. (2015). Bayesian two-part Tobit models with left-censoring, skewness and nonignorable missingness. *Journal of Biopharmaceutical Statistics*, 25(4):714–730. https://doi.org/10.1080/10543406.2014.920860.

Deshmane, S. L., Kremlev, S., Amini, S., and Sawaya, B. E. (2009). Monocyte chemoattractant protein-1 (MCP-1): An overview. *Journal of interferon & cytokine research*, 29(6):313–326. https://doi.org/10.1089/jir.2008.0027.

Fortunato, A. (2016). A new sensitive automated assay for procalcitonin detection: LIAISON® BRAHMS PCT® II GEN. *Practical laboratory medicine*, 6:1–7. https://doi.org/10.1016/j.plabm.2016.06.002.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, 2 edition.

Gasparini, A. (2018). rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, 3:739. https://doi.org/10.21105/joss.00739.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. https://doi.org/10.1198/016214506000001437.

Goldman, N. and Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026378.

Gordon, C. (2004). Pregnancy and autoimmune diseases. *Best Practice & Research Clinical Rheumatology*, 18(3):359 – 379. https://doi.org/10.1016/j.berh.2004.02.012.

Gould, W., Pitblado, J., and Sribney, W. (2006). *Maximum Likelihood Estimation with Stata*. Stata Press, 3 edition.

Gupta, V., Zhou, H., Reyes, C., and Wang, Q.-S. (2010). Multiplexing Across Cytokine Panels: Bio-Plex Pro. Human and Mouse Group I and Group II. Technical report, Bio-Rad Laboratories, Inc.

Gupta, V., Zimmerman, R., Zhan, T., Hamilton, T., Na, L., and Peng, J. (2014). Development and Validation of Bio-Plex Pro(TM) Human Chemokine Assays. Technical report, Bio-Rad Laboratories, Inc.

Hahn, E. D. and Soyer, R. (2005). Probit and logit models: Differences in the multivariate realm. *The Journal of the Royal Statistical Society, Series B*, pages 1–12.

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in quantitative methods for psychology*, 9(2):43. https://doi.org/10.20982/tqmp.09.2.p043.

Holt, M. M. and Eriksen, R. (2018). Er det forskjell i seroprevalens av borreliaantistoff i befolkningen mellom ulike geografiske områder av sør-trøndelag fylke? *Medical undergraduate research thesis at NTNU*.

Hornung, R. W. and Reed, L. D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene*, 5(1):46–51. https://doi.org/10.1080/1047322X.1990.10389587.

Jaffa, M. A., Gebregziabher, M., Garrett, S. M., Luttrell, D. K., Lipson, K. E., Luttrell, L. M., and Jaffa, A. A. (2018). Analysis of longitudinal semicontinuous data using marginalized two-part model. *Journal of translational medicine*, 16(1):301. https://doi.org/10.1186/s12967-018-1674-5.

Jore, S., Viljugrein, H., Hofshagen, M., Brun-Hansen, H., Kristoffersen, A. B., Nygård, K., Brun, E., Ottesen, P., Sævik, B. K., and Ytrehus, B. (2011). Multi-source analysis reveals latitudinal and altitudinal shifts in range of ixodes ricinus at its northern distribution limit. *Parasites & Vectors*, 4(1):84. https://doi.org/10.1186/1756-3305-4-84.

Kaune, A. and Kettrup, A. (1994). Treatment of values below the detection limits in correlation analysis of chlorinated dioxins and related compounds. *Chemosphere*, 29(9-11):1811–1818. https://doi.org/10.1016/0045-6535(94)90347-6.

Lee, G. and Scott, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829. https://doi.org/10.1016/j.csda.2012.03.003.

Lee, M., Kong, L., and Weissfeld, L. (2012). Multiple imputation for left-censored biomarker data based on gibbs sampling method. *Statistics in medicine*, 31(17):1838–1848. https://doi.org/10.1002/sim.4503.

Liu, L., Strawderman, R. L., Johnson, B. A., and O'Quigley, J. M. (2016). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Statistical methods in medical research*, 25(1):133–152. https://doi.org/10.1177/0962280212443324.

Lyles, R. H., Williams, J. K., and Chuachoowong, R. (2001). Correlating two viral load assays with known detection limits. *Biometrics*, 57(4):1238–1244. https://doi.org/10.1111/j.0006-341X.2001.01238.x.

Mahmud, S., Lou, W. W., and Johnston, N. W. (2010). A probit- log- skew-normal mixture model for repeated measures data with excess zeros, with application to a cohort study of paediatric respiratory symptoms. *BMC Med Res Methodol*, 10(55). https://doi.org/10.1186/1471-2288-10-55.

McInnes, I. B. and Schett, G. (2011). The pathogenesis of rheumatoid arthritis. *New England Journal of Medicine*, 365(23):2205–2219. http://doi.org/10.1056/NEJMra1004965.

Mok, C. and Lau, C. (2003). Pathogenesis of systemic lupus erythematosus. *Journal of clinical pathology*, 56(7):481–490. http://doi.org/10.1136/jcp.56.7.481.

Molenberghs, G. and Verbeke, G. (2004). *An Introduction to (Generalized (Non)Linear Mixed Models*. Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_4.

Morris, T. P., White, I. R., and Crowther, M. J. (2017). Using simulation studies to evaluate statistical methods. *arXiv preprint*. https://arxiv.org/abs/1712.03198.

Moulton, L. H. and Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, 51(4):1570–8. https://doi.org/10.2307/2533289.

Moulton, L. H. and Halsey, N. A. (1996). A mixed gamma model for regression analyses of quantitative assay data. *Vaccine*, 14(12):1154–1158. https://doi.org/10.1016/0264-410X(96)00017-5.

Owen, D. B. (1956). Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090. https://doi.org/10.1214/aoms/1177728074.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1):12–35. https://doi.org/10.1080/10618600.1995.10474663.

Powell, L. A. (2007). Approximating variance of demographic parameters using the delta method: A reference for avian biologists. *The Condor*, 109(4):949–954. https://doi.org/10.1650/0010-5422(2007)109[949:AVODPU]2.0.CO;2.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301 – 323. https://doi.org/10.1016/j.jeconom.2004.08.017.

Ridgley, L. A., Anderson, A. E., and Pratt, A. G. (2018). What are the dominant cytokines in early rheumatoid arthritis? *Current opinion in rheumatology*, 30(2):207. https://doi.org/10.1097/BOR.0000000000000470.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, 4(1):395–421. https://doi.org/10.1146/annurev-statistics-060116-054045.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114. http://doi.org/10.2307/3002019.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610. https://doi.org/10.1080/01621459.1987.10478472.

Smith, V. A., Preisser, J. S., Neelon, B., and Maciejewski, M. L. (2014). A marginalized two-part model for semicontinuous data. *Statistics in medicine*, 33(28):4891–4903. https://doi.org/10.1002/sim.6263.

Stanek, G., Wormser, G. P., Gray, J., and Strle, F. (2012). Lyme borreliosis. *The Lancet*, 379(9814):461–473. https://doi.org/10.1016/S0140-6736(11)60103-7.

Su, X. and Luo, S. (2017). Analysis of censored longitudinal data with skewness and a terminal event. *Communications in Statistics - Simulation and Computation*, 26(7):5378–5391. https://doi.org/10.1080/03610918.2016.1157181.

Swain, S. and Jena, P. (2016). Current understanding of rheumatoid arthritis therapy in pregnancy. *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, 5(10):3275–3279. http://dx.doi.org/10.18203/2320-1770.ijrcog20163402.

Thomopoulos, N. T. (2012). *Essentials of Monte Carlo simulation: Statistical methods for building simulation models*. Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4614-6022-0.

Tobin, J. (1958). Estimation of relationships for limited dependent-variables. *Econometrica*, 26(1):24–36. https://doi.org/10.2307/1907382.

Vestrheim, D. F., White, R. A., Aaberge, I. S., and Aase, A. (2016). Geographical differences in seroprevalence of borrelia burgdorferi antibodies in Norway, 2011–2013. *Ticks and tick-borne diseases*, 7(5):698–702. https://doi.org/10.1016/j.ttbdis.2016.02.020.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228. https://doi.org/10.1037/a0027127.

White, I. R. (2010). simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal*, 10(3):369–385. http://www.stata-journal.com/article.html?article=st0200.

Wit, E., Heuvel, E. v. d., and Romeijn, J.-W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236. https://doi.org/10.1111/j.1467-9574.2012.00530.x.

Xing, D., Huang, Y., Chen, H., Zhu, Y., Dagne, G. A., and Baldwin, J. (2017). Bayesian inference for two-part mixed-effects model using skew distributions, with application to longitudinal semicontinuous alcohol data. *Statistical methods in medical research*, 26(4):1838–1853. https://doi.org/10.1177/0962280215590284.

Zhang, J.-M. and An, J. (2007). Cytokines, inflammation and pain. *International Anesthesiology Clinics*, 45(2):27–37. http://dx.doi.org/10.1097/AIA.0b013e318034194e.

Østensen, M., Förger, F., and Villiger, P. M. (2006). Cytokines and pregnancy in rheumatic disease. *Annals of the New York Academy of Sciences*, 1069(1):353–363. http://dx.doi.org/10.1196/annals.1351.033.

Appendix

A Simulation Study - R Code

In Section A.1 we present the code for maximizing the likelihoods of the models used in the simulation study, followed by the code for calculating the mean CRPS in Section A.2. The framework of the simulation study is provided in Section A.3.

A.1 Likelihood Maximization

```
fit_model <-function(x, y, LOD, model, init, cont = "ln"){
# x: Data, y: responses, LOD: Limit of Detection,
# model: Which model to fit, init: Initial values
# cont: Which continuous distribution to use.
# (ln = lognormal, lsn = log-skew-normal)

  if(model == "ni"){ # Two-part
    return(optim(par = init, fn = optim_likelihood_ni, x = x, y = y,
    LOD = LOD, cont = cont,  method = "BFGS", hessian = TRUE,
    control = list("fnscale"=-1)))
  }else if(model == "i"){ # Two-part w/ interval censoring
    return(optim(par = init, fn = optim_likelihood_i, x = x, y = y,
    LOD = LOD, cont = cont, method = "BFGS", hessian = TRUE,
    control = list("fnscale"=-1)))
  }else if(model == "pi"){ # Substitute model
    return(optim(par = init[1:ceiling((length(init)+1)/2)],
    fn = optim_likelihood_pi, x = x, y = y, LOD = LOD, cont = cont,
    method = "BFGS", hessian = TRUE, control = list("fnscale"=-1)))
  }else if(model == "t"){ # Tobit model
    return(optim(par = init[1:ceiling((length(init)+1)/2)],
    fn = optim_likelihood_t, x = x, y = y, LOD = LOD, cont = cont,
    method = "BFGS", hessian = TRUE, control = list("fnscale"=-1)))
  }else if(model == "mt"){ # Marginalized two-part model
    return(optim(par = init, fn = optim_likelihood_mt, x = x, y = y,
    LOD = LOD, cont = cont, method = "BFGS", hessian = TRUE,
    control = list("fnscale"=-1)))
  }
}


### TWO-PART W/O INTERVAL CENSORING:
log_likelihood_ni= function(x, y, beta, gamma, sigma, LOD, cont, delta = NA){
```

```
# Log-likelihood of TP model

  censored = as.integer(is.na(y))
  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y)){
    y_i = y[i]
    x_i = x[i,]
    p = pnorm(sum(x_i * beta)) # probit link function
    if(censored[i]){ # censored
      update = log(p)
    }else{ # not censored
      if(cont == "ln"){ # Lognormal continuous part
        update = log(1 - p) + dnorm(x = y_i, mean = sum(x_i * gamma),
                                    sd = abs(sigma), log = TRUE)
      }else if(cont == "lsn"){ # Log-skew-normal continuous part
        update = log(1 - p) + dsn(x = y_i, xi = sum(x_i * gamma),
                                  alpha = delta/abs(sigma),
                                  omega = sqrt(sigma^2 + delta^2), log = TRUE)
      }
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_likelihood_ni = function(par, x, y, LOD, cont){
# Function to be maximized by optim for TP model

  if(cont == "lsn"){ # Log-skew-normal continuous part
    delta = par[length(par)]
    par = par[-length(par)]
  }else{ # Lognormal continuous part
    delta = NA
  }
  gamma = par[1:floor(length(par)/2)]
  sigma = par[ceiling(length(par)/2)]
  beta = par[(ceiling(length(par)/2)+1):length(par)]
  return(log_likelihood_ni(x, y, beta, gamma, sigma, LOD, cont, delta))
}


### TWO-PART MODELT WITH INTERVAL CENSORING:
log_likelihood_i= function(x, y, beta, gamma, sigma, LOD, cont, delta = NA){
```

```
# Log-likelihood of TPIC model

  censored = as.integer(is.na(y))
  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y)){
    y_i = y[i]
    x_i = x[i,]
    p = pnorm(sum(x_i * beta)) # probit link function
    if(censored[i]){ # censored
      if(cont == "ln"){ # Lognormal continuous part
        update = log(p + (1 - p) * pnorm(q = LOD, mean = sum(x_i * gamma),
                                         sd = abs(sigma)))
      }else if(cont == "lsn"){ # Log-skew-normal continuous part
        update = log(p + (1 - p) * psn(x = LOD, xi = sum(x_i * gamma),
                                       alpha = delta/abs(sigma),
                                       omega = sqrt(sigma^2 + delta^2)))
      }

    }else{ # not censored
      if(cont == "ln"){ # Lognormal continuous part
        update = log(1 - p) + dnorm(x = y_i, mean = sum(x_i * gamma),
                                    sd = abs(sigma), log = TRUE)
      }else if(cont == "lsn"){  # Log-skew-normal continuous part
        update = log(1 - p) + dsn(x = y_i, xi = sum(x_i * gamma),
                                  alpha = delta/abs(sigma),
                                  omega = sqrt(sigma^2 + delta^2), log = TRUE)
      }
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_likelihood_i = function(par, x, y, LOD, cont){
# Function to be maximized by optim for TPIC model

  if(cont == "lsn"){ # Log-skew-normal continuous part
    delta = par[length(par)]
    par = par[-length(par)]
  }else{ # Lognormal continuous part
    delta = NA
  }
  gamma = par[1:floor(length(par)/2)]
```

```
  sigma = par[ceiling(length(par)/2)]
  beta = par[(ceiling(length(par)/2)+1):length(par)]
  return(log_likelihood_i(x, y, beta, gamma, sigma, LOD, cont, delta))
}


### SUBSTITUE MODEL
log_likelihood_pi= function(x, y, gamma, sigma, LOD, cont, delta){
# Log-likelihood function of substitute model

  y[is.na(y)] = LOD - log(2) # plugin LOD/2
  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y)){
    y_i = y[i]
    x_i = x[i,]
    if(cont == "ln"){ # Lognormal distribution
      update = dnorm(x = y_i, mean = sum(x_i * gamma),
                     sd = abs(sigma), log = TRUE)
    }else if(cont == "lsn"){ # Log-skew-normal distribution
      update = dsn(x = y_i, xi = sum(x_i * gamma), alpha = delta/abs(sigma),
                   omega = sqrt(sigma^2 + delta^2), log = TRUE)
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_likelihood_pi = function(par, x, y, LOD, cont){
# Function to be maximized by optim for substitute model

  if(cont == "lsn"){ # Log-skew-normal distribution
    delta = par[length(par)]
    par = par[-length(par)]
  }else{ # Log-normal distribution
    delta = NA
  }
  gamma = par[1:(length(par)-1)]
  sigma = par[length(par)]
  return(log_likelihood_pi(x, y, gamma, sigma, LOD, cont, delta))
}


### TOBIT MODEL:
```

```
log_likelihood_t= function(x, y, gamma, sigma, LOD, cont, delta){
# Log-likelihood function of substitute model

  censored = as.integer(is.na(y))
  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y)){
    y_i = y[i]
    x_i = x[i,]
    if(censored[i]){ # censored
      if(cont == "ln"){ # Lognormal distribution
        update = pnorm(q = LOD, mean = sum(x_i * gamma),
                       sd = abs(sigma), log = TRUE)
      }else if(cont == "lsn"){ # Log-skew-normal distribution
        update = log(psn(x = LOD, xi = sum(x_i * gamma),
                         alpha = delta/abs(sigma),
                         omega = sqrt(sigma^2 + delta^2)))
      }
    }else{ # not censored
      if(cont == "ln"){ # Lognormal distribution
        update = dnorm(x = y_i, mean = sum(x_i * gamma),
                       sd = abs(sigma), log = TRUE)
      }else if(cont == "lsn"){ # Log-skew-normal distribution
        update = dsn(x = y_i, xi = sum(x_i * gamma),
                     alpha = delta/abs(sigma),
                     omega = sqrt(sigma^2 + delta^2), log = TRUE)
      }
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_likelihood_t = function(par, x, y, LOD, cont){
# Function to be maximized by optim for Tobit model

  if(cont == "lsn"){ # Log-skew-normal distribution
    delta = par[length(par)]
    par = par[-length(par)]
  }else{ # Lognormal distribution
    delta = NA
  }
  gamma = par[1:(length(par)-1)]
  sigma = par[length(par)]
```

```
  return(log_likelihood_t(x, y, gamma, sigma, LOD, cont, delta))
}
```

A.2 Calculating Mean CRPS

```
mean.CRPS <- function(y0, x0, m = c(1,2,3,4,5), gamma_est,
sigma_est, beta_est = rep(NA,length(gamma_est)), realLOD = NA){
  # x0: covariate
  # y0: observed response
  # Returns corresponding CRPS for all pairs (y0,x0)

  p_est = pnorm(x0 %*% beta_est)
  mu_est = x0 %*% gamma_est

  if(m == 1){ # TPIC
    s = mean(mapply(intfunc, y0 = y0, p_est = p_est, mu_est = mu_est,
              MoreArgs = list(C=C1, realLOD=realLOD, sigma_est=sigma_est)))
  }else if(m == 2){ # TP
    s = mean(mapply(intfunc, y0 = y0, p_est = p_est, mu_est = mu_est,
              MoreArgs = list(C=C2, realLOD=realLOD, sigma_est=sigma_est)))
  }else if(m == 3){ # Substitute
    s = mean(mapply(intfunc, y0 = y0, p_est = p_est, mu_est = mu_est,
              MoreArgs = list(C=C3, realLOD=realLOD, sigma_est=sigma_est)))
  }else if(m == 4){ # Tobit
    s = mean(mapply(intfunc, y0 = y0, p_est = p_est, mu_est = mu_est,
              MoreArgs = list(C=C4, realLOD=realLOD, sigma_est=sigma_est)))
  }
  return(s)
}


intfunc <- function(y0, p_est, mu_est, sigma_est, C, realLOD){
  return(integrate(C, lower = 0, upper = Inf, y0, p = p_est, mu = mu_est,
                   sigma_est = sigma_est, realLOD = realLOD)$value)
}

# TPIC
C1 <- function(z, y0, p, mu, sigma_est, realLOD){
  c  = numeric(length(z))

  c[which(z < 0)] = (0 - as.integer(z[which(z < 0)] >= y0))^2
  c[which(z >= 0 & z < realLOD)] = (p + (1-p)*plnorm(realLOD, mu, sigma_est) -
                            as.integer(z[which(z >= 0 & z < realLOD)] >= y0))^2
  c[which(z >= realLOD)] = (p + (1-p)*plnorm(z[which(z >= realLOD)],
                                          mu, sigma_est) -
                            as.integer(z[which(z >= realLOD)] >= y0))^2
```

```
  return(c)
}

# TP
C2 <- function(z, y0, p, mu, sigma_est, realLOD){
  c  = numeric(length(z))

  c[which(z < 0)] = (0 - as.integer(z[which(z < 0)] >= y0))^2
  c[which(z >= 0)] = (p + (1-p)*plnorm(z[which(z >= 0)], mu, sigma_est) -
                                as.integer(z[which(z >= 0)] >= y0))^2

  return(c)
}

#SUB
C3 <- function(z, y0, p, mu, sigma_est, realLOD){
  c  = (plnorm(z, mu, sigma_est) - as.integer(z >= y0))^2
  return(c)
}

# TOBIT
C4 <- function(z, y0, p, mu, sigma_est, realLOD){
  c  = numeric(length(z))

  c[which(z < 0)] = (0 - as.integer(z[which(z < 0)] >= y0))^2
  c[which(z >= 0 & z < realLOD)] = (plnorm(realLOD, mu, sigma_est) -
                    as.integer(z[which(z >= 0 & z < realLOD)] >= y0))^2
  c[which(z >= realLOD)] = (plnorm(z[which(z >= realLOD)], mu, sigma_est) -
                           as.integer(z[which(z >= realLOD)] >= y0))^2
  return(c)
}
```

A.3 Simulation Study

```
library(rsimsum)

set.seed(1729) # Random seed. Do not change this.

n_sim = 2000 # Number of repeated simulations
n_obs = 300 # Sample size of datasets
n_sen = 16 # Number of different scenarios
n_met = 4 # Number of methods tested
n_par = 5+4 # Number of estimated parameters(regression parm. + calculated parm.)
n_test = 300 # Size of test set for calculating CRPS
```

```
alpha = 0.05 # For tests and confidence intervals

methods = c("TP", "TPIC", "Substitute", "Tobit") # List of methods

# List of parameter names
par_name = c("gamma0", "gamma1", "sigma", "beta0", "beta1",
             "mcrps", "marg_mean0", "marg_mean1", "marg_eff_obs")

## Data frames for storing results:
# Summary of results
summ.all = data.frame(stat = character(), est = numeric(), mcse = numeric(),
                      method = character(), LOD = character(),
                      disc_prob = character(), beta1 = character(),
                      par = character(), lower = numeric(), upper = numeric())

# Raw data
data.all = data.frame(dataset = integer(), method = character(),
                      theta = numeric(), se = numeric(),par = character(),
                      LOD = character(), disc_prob = character(),
                      .dropbig = logical())

res.list = list()

# Varying factors:
LOD_real = rep(c(c(0.5,1,2,3)), each = 4)
LOD_vec = log(LOD_real) # LOD on log-scale
LOD_n = gsub(".", ",", as.character(LOD_real), fixed = TRUE)

beta0_vec = rep(c(-1.2,-0.7,-0.5,-0.2), 4)
disc_prob = rep(c(10, 20, 30, 40), 4)
dp_n = gsub(".", ",", as.character(disc_prob), fixed = TRUE)

# Constant factors:
proportion_diagnosed = 0.5
beta1 = -0.2
gamma0 = 1
gamma1 = 0.2 # continuous part
sigma = 0.7 # continuous part

true.crps = numeric(n_sen) # Compute CRPS of true distribution

# Store failed data for analysis
xfail = c()
yfail = c()
dfail = c()
```

```
LODfail = c()
beta0fail = c()

### ITERATE OVER SCENARIOS:
for(sen in 1:n_sen){

  pos_res_sen = 1
  res.simsum.sen =
  data.frame(dataset = rep(NA, n_sim*n_met*n_par), # Scenario number
             method = rep(NA, n_sim*n_met*n_par), # Name of the method
             theta = rep(NA, n_sim*n_met*n_par), # Parameter estimate
             se = rep(NA, n_sim*n_met*n_par), # Estimated standard error
             par = rep(NA, n_sim*n_met*n_par), # Name of parameter
             LOD = rep(NA, n_sim*n_met*n_par), # Limit of detection for dataset
             disc_prob = rep(NA, n_sim*n_met*n_par) ) # Discrete probability

  # Varying beta0
  beta0 = beta0_vec[sen]

  # Varying the LOD
  LOD = LOD_vec[sen]
  LOD_name = LOD_n[sen]

  # Calculate true marginal means and effects:
  Z_0 = (LOD - gamma0)/sigma
  C_0 = (1-pnorm(Z_0 - sigma))
  Z_1 = (LOD - gamma0 - gamma1)/sigma
  C_1 = (1-pnorm(Z_1 - sigma))
  marginal_effect = (1-pnorm(beta0 + beta1))/(1-pnorm(beta0)) *
                     C_1 / C_0 * exp(gamma1)
  marginal_mean0 = (1-pnorm(beta0)) * (1-pnorm(Z_0 - sigma)) *
                    exp(gamma0 + sigma^2/2)
  marginal_mean1 = (1-pnorm(beta0 + beta1)) * (1-pnorm(Z_1 - sigma)) *
                    exp(gamma0 + gamma1 + sigma^2/2)

  ### SIMULATIONS:
  for(i in 1:n_sim){

    beta = c(beta0, beta1)
    gamma = c(gamma0, gamma1)

    # TRAINING DATA:
    x1 = rbinom(n = n_obs, size = 1, p = proportion_diagnosed) #"diagnosis"
    x = cbind(1, x1)
```

```
p = pnorm(x %*% beta)
discrete_part = as.integer(runif(n_obs) < p)

# On log-scale:
errors = rnorm(n = n_obs, mean = 0, sd = sigma)
y = x %*% gamma + errors
mean(y <= LOD)
y[y <= LOD] = NA # censored
y[discrete_part == 1] = NA # true zeroes

# TEST DATA FOR CRPS:
x10 = rbinom(n = n_test, size = 1, p = proportion_diagnosed) #"diagnosis"
x0 = cbind(1, x10)

p0 = pnorm(x0 %*% beta)
discrete_part0 = as.integer(runif(n_test) < p0)

# On real-scale:
errors = rnorm(n = n_test, mean = 0, sd = sigma)
y0_log = x0 %*% gamma + errors
y0_log[y0_log <= LOD] = NA # censored
y0_log[discrete_part0 == 1] = NA # true zeroes
y0 = exp(y0_log)
y0[is.na(y0)] = 0

# Likelihood maximization:
par = c(gamma, sigma, beta)
solution_ni = fit_model(x = x, y = y, LOD = LOD, model = "ni", par)
solution_i = fit_model(x = x, y = y, LOD = LOD, model = "i", par)
solution_pi = fit_model(x = x, y = y, LOD = LOD, model = "pi", par)
par = c(solution_pi$par[1:(length(gamma)+1)], beta)
solution_t = fit_model(x = x, y = y, LOD = LOD, model = "t", par)

solutions = list(solution_ni, solution_i, solution_pi, solution_t)

# Theoretically optimal CRPS from true distribution:
tc = mean.CRPS(y0, x0, m = 1, gamma_est = gamma, sigma_est = sigma,
               beta_est = beta, realLOD = exp(LOD))
true.crps[sen] = true.crps[sen] + 1/n_sim * tc

# Storing results:
for(met in 1:n_met){
  #Checking for errors:
  if(solutions[[met]]$convergence != 0 |
     class(try(solve(-solutions[[met]]$hessian), silent = TRUE)) !=
```

```
        "matrix"){
  # Non-convergence
  res.simsum.sen[pos_res_sen:(pos_res_sen+n_par-1),] =
      cbind( rep(sen,n_par),
             rep(methods[met],n_par),
             rep(NA, n_par), rep(NA, n_par), par_name,
             rep(LOD_name,n_par), rep(dp_n[sen],n_par), FALSE)
  pos_res_sen = pos_res_sen + n_par

  # Store failed data for analysis
  xfail = rbind(xfail, x1)
  yfail = rbind(yfail, t(y))
  dfail = rbind(dfail, discrete_part)
  LODfail = c(LODfail, LOD)
  beta0fail = c(beta0fail, beta0)

}else if(sum(diag(solve(-solutions[[met]]$hessian)) < 0) > 0){
  # Negative standard errors
  print("Failed SE")
  print(methods[met])
  print(sen)
  res.simsum.sen[pos_res_sen:(pos_res_sen+n_par-1),] =
      cbind( rep(sen,n_par),
             rep(methods[met],n_par),
             rep(NA, n_par), rep(NA, n_par), par_name,
             rep(LOD_name,n_par), rep(dp_n[sen],n_par), FALSE)
  pos_res_sen = pos_res_sen + n_par

  # Store failed data for analysis
  xfail = rbind(xfail, x1)
  yfail = rbind(yfail, t(y))
  dfail = rbind(dfail, discrete_part)
  LODfail = c(LODfail, LOD)
  beta0fail = c(beta0fail, beta0)

}else{
  # Successfull solution

  C = solve(-solutions[[met]]$hessian) # Inverse negative Hessian
  SE = sqrt(diag(C)) # Standard errors
  solutions[[met]]$par[length(gamma)+1] =
      abs(solutions[[met]]$par[length(gamma)+1]) # Ensure right sign

  gamma.res = solutions[[met]]$par[1:length(gamma)]
  sigma.res = solutions[[met]]$par[length(gamma)+1]
```

```
if(length(SE) == length(gamma) + length(beta) + 1){ # Two-part models
  beta.res = solutions[[met]]$par[(length(gamma)+2):(2*length(gamma)+1)]

  # Calculate marginal effects and CPRS
  if(methods[met] == "TP"){
    marg_eff = (1 - pnorm(beta.res[1]+beta.res[2])) /
               (1 - pnorm(beta.res[1])) * exp(gamma.res[2])
    marg_mean0 = (1 - pnorm(beta.res[1])) *
                 exp(gamma.res[1] + sigma.res^2/2)
    marg_mean1 = (1 - pnorm(beta.res[1] + beta.res[2])) *
                 exp(gamma.res[1] + gamma.res[2] + sigma.res^2/2)

    mcrps = mean.CRPS(y0 = y0, x0 = x0, m = 2,
                      gamma_est = gamma.res, sigma_est=sigma.res,
                      beta_est = beta.res, realLOD = exp(LOD))

  }else if(methods[met] == "TPIC"){
    Z_0 = (LOD - gamma.res[1])/sigma.res
    C_0 = (1-pnorm(Z_0 - sigma.res))
    Z_1 = (LOD - gamma.res[1] - gamma.res[2])/sigma.res
    C_1 = (1-pnorm(Z_1 - sigma.res))
    marg_eff = (1-pnorm(beta.res[1] + beta.res[2])) /
               (1-pnorm(beta.res[1])) * C_1 / C_0 * exp(gamma.res[2])
    marg_mean0 = (1-pnorm(beta.res[1])) * (1-pnorm(Z_0 - sigma.res)) *
                 exp(gamma.res[1] + sigma.res^2/2)
    marg_mean1 = (1-pnorm(beta.res[1] + beta.res[2])) *
                 (1-pnorm(Z_1 - sigma.res)) *
                 exp(gamma.res[1] + gamma.res[2] + sigma.res^2/2)

    mcrps = mean.CRPS(y0 = y0, x0 = x0, m = 1,
                      gamma_est = gamma.res, sigma_est=sigma.res,
                      beta_est = beta.res, realLOD = exp(LOD))
  }

  res.simsum.sen[pos_res_sen:(pos_res_sen+n_par-1),] =
    cbind( rep(sen,n_par),
           rep(methods[met],n_par),
           c(solutions[[met]]$par,
             mcrps, marg_mean0, marg_mean1, marg_eff),
           c(SE, 1, 1, 1, 1),
           par_name, rep(LOD_name,n_par), rep(dp_n[sen],n_par), FALSE)
  pos_res_sen = pos_res_sen + n_par

}else{ # One-part models
```

```
        if(methods[met] == "Substitute"){
          marg_eff = exp(gamma.res[2])
          marg_mean0 = exp(gamma.res[1] + sigma.res^2/2)
          marg_mean1 = exp(gamma.res[1] + gamma.res[2] + sigma.res^2/2)

          mcrps = mean.CRPS(y0, x0, m = 3, gamma_est = gamma.res,
                            sigma_est=sigma.res, realLOD = exp(LOD))

        }else if(methods[met] == "Tobit"){
          Z_0 = (LOD - gamma.res[1])/sigma.res
          Z_1 = (LOD - gamma.res[1] - gamma.res[2])/sigma.res

          marg_eff = (1 - pnorm(Z_1 - sigma.res)) /
                     (1 - pnorm(Z_0 - sigma.res)) * exp(gamma.res[2])
          marg_mean0 = (1 - pnorm(Z_0 - sigma.res)) *
                       exp(gamma.res[1] + sigma.res^2/2)
          marg_mean1 = (1 - pnorm(Z_1 - sigma.res)) *
                       exp(gamma.res[1] + gamma.res[2] + sigma.res^2/2)

          mcrps = mean.CRPS(y0, x0, m = 4, gamma_est = gamma.res,
                            sigma_est=sigma.res, realLOD = exp(LOD))
        }


        res.simsum.sen[pos_res_sen:(pos_res_sen+n_par-1),] =
          cbind( rep(sen,n_par),
                 rep(methods[met],n_par),
                 c(solutions[[met]]$par, rep(NA,length(beta)),
                   mcrps, marg_mean0, marg_mean1, marg_eff),
                 c(SE, rep(NA,length(beta)), 1, 1, 1, 1),
                 par_name, rep(LOD_name,n_par), rep(dp_n[sen],n_par), FALSE)
        pos_res_sen = pos_res_sen + n_par
      }
    }
  }
}
# Calculating performance measures
res.simsum.sen$theta = as.numeric(res.simsum.sen$theta)
res.simsum.sen$se = as.numeric(res.simsum.sen$se)

# Identify solutions with SE larger than 10 SD from average SE
res.simsum.sen = cbind(res.simsum.sen,
                       .dropbig = rep(FALSE, dim(res.simsum.sen)[1]))
```

```
  # .dropbig = TRUE for extreme outliers
  for(param in c("beta0", "beta1", "gamma0", "gamma1", "sigma")){
    res.simsum.sen[res.simsum.sen$par == param, ] =
      dropbig(subset(res.simsum.sen, par == param),
              estvarname = "theta", se = "se", methodvar = "method",
              by = c("LOD", "disc_prob"), max = Inf, semax = 10, robust = FALSE)
  }
  data.all = rbind(data.all, res.simsum.sen) # Store all data

  # Remove extreme outliers
  if(sum(res.simsum.sen$.dropbig, na.rm = TRUE) > 0){
    res.simsum.sen[which(res.simsum.sen$.dropbig),]$theta = NA
    res.simsum.sen[which(res.simsum.sen$.dropbig),]$se = NA
  }

  # Calculate performance measures (without extreme outliers)
  s.sen = multisimsum(data = res.simsum.sen, estvarname = "theta", par = "par",
                      true = c(gamma0 = gamma[1], gamma1 = gamma[2],
                               beta0 = beta[1], beta1 = beta[2], sigma = sigma,
                               mcrps = 0, marg_mean0 = marginal_mean0,
                               marg_mean1 = marginal_mean1,
                               marg_eff_obs = marginal_effect),
                      se = "se", by = c("LOD", "disc_prob"),
                      methodvar = "method", ref = 'TPIC', x = TRUE)

  summ.all = rbind(summ.all, summary(s.sen)$summ)
  res.list[[sen]] = s.sen
}
```

B Longitudinal Analysis - SAS code

```
/* TPIC WITH SKEW */
proc nlmixed
data = multiplex
cov
df = 1000
alpha = 0.05
tech = QUANEW
update = BFGS;
parms /* Initial values */
g0 = 2.41 ga = 0 gr = 0.33 gs = 0.53 gt1 = 0.91 gt2 = 0.92
gt1r = -0.97 gt2r = -1.25 gt1s = -1.44 gt2s = -1.41
sigma = 0.51 delta = -1.61
b0 = -0.67 br = -0.95 bs = -0.53 bt1 = -1.03 bt2 = -0.34  ba = 0.02
bt1r = 1.17 bt2r = -0.37 bt1s = 0.94 bt2s = -0.05
s11 = 0.5 s12 = 0 s22 = 0.5;

bounds sigma > 0; bounds s11 > 0; bounds s22 > 0;

T = log(0.29); *LOD;
censored = (y=-100); *Indicator variable;

mu = g0 + gr*(RA+SN_RA) + gs*SLE + ga*age +
gt1*(time3+time4) + gt2*(time5+time6) +
gt1r*(time3+time4)*(RA+SN_RA) + gt2r*(time5+time6)*(RA+SN_RA) +
gt1s*(time3+time4)*SLE + gt2s*(time5+time6)*SLE + t1;

mu_pi = b0 + br*(RA+SN_RA) + bs*SLE + ba*age +
bt1*(time3+time4) + bt2*(time5+time6) +
bt1r*(time3+time4)*(RA+SN_RA) + bt2r*(time5+time6)*(RA+SN_RA) +
bt1s*(time3+time4)*SLE + bt2s*(time5+time6)*SLE + t2;

pi = CDF('normal', mu_pi);

* Estimate Owenst T-function;
h = (T - mu)/sqrt(sigma**2 + delta**2);
a = delta/sigma;
jmax = 50;
cut.point = 8;

aa = abs(a);
ah = abs(h);
if (aa = .I) then OT = sign(a) * 0.5*CDF('normal',-ah);
else if (aa = 0) then OT = 0;
```

```
else if (ah = .I) then OT = 0;
else if (aa <= 1) then do;
    th = ah;
    ta = aa;

    * Two ways to approximate integral based on cutpoint;
    low = (th <= cut.point);
    if low then do;
        matr = 0;
        cumb = 0;
            do j = 0 to jmax;
                cumb = cumb + (th**(2*j)) / ((2**j) * GAMMA(j+1));
                b1 = EXP(-0.5 * th**2) * cumb;
                matr = matr + (-1)**j * ta**(2*j+1) / (2*j+1) * (1-b1);
            end; *for;
        T.int = (ATAN(ta) - matr) / (2*CONSTANT("pi"));
        OT = SIGN(a) * T.int;
        end; *if;
    else OT = SIGN(a) * (ATAN(ta) * EXP(-0.5*(th**2)*ta/ATAN(ta)) *
              (1+0.00868*(th*ta)**4) / (2*CONSTANT("pi")));
end; *if;
else do; *(aa > 1);
    th = aa * ah;
    ta = 1 / aa;

    * Two ways to approximate integral based on cutpoint;
    low = (th <= cut.point);
    if low then do;
        matr = 0;
        cumb = 0;
            do j = 0 to jmax;
                cumb = cumb + (th**(2*j)) / ((2**j) * GAMMA(j+1));
                b1 = EXP(-0.5 * th**2) * cumb;
                matr = matr + (-1)**j * ta**(2*j+1) / (2*j+1) * (1-b1);
            end; *for;
        T.int = (ATAN(ta) - matr) / (2*CONSTANT("pi"));
        end; *if;
    else T.int = ATAN(ta) * EXP(-0.5*(th**2)*ta/ATAN(ta)) *
                 (1+0.00868*(th*ta)**4) / (2*CONSTANT("pi"));

OT = sign(a) * (0.5*CDF('normal',ah) +
     CDF('normal',aa*ah)*(0.5-CDF('normal',ah)) - T.int);
end;

w = sqrt(sigma**2 + delta**2);
```

```
if censored then p = pi + (1-pi)*(CDF('normal',(T-mu)/w) - 2*OT);
else p = (1-pi) * 2/w * PDF('normal',(y-mu)/w) *
         CDF('normal',delta/sigma*(y-mu)/w);
loglike = log(p);

model y ~ general (loglike);
random t1 t2 ~ normal([0,0],[s11, s12, s22]) subject = individual;
run;


/* TP WITH SKEW */
proc nlmixed
data = multiplex
cov
df = 1000
alpha = 0.05
gconv = 0
tech = QUANEW
update = BFGS;
parms /* Initial values */
g0 = 2.41 ga = 0 gr = 0.33 gs = 0.53 gt1 = 0.91 gt2 = 0.92
gt1r = -0.97 gt2r = -1.25 gt1s = -1.44 gt2s = -1.41
sigma = 0.51 delta = -1.61
b0 = -0.67 br = -0.95 bs = -0.53 bt1 = -1.03 bt2 = -0.34  ba = 0.02
bt1r = 1.17 bt2r = -0.37 bt1s = 0.94 bt2s = -0.05
s11 = 0.5 s12 = 0 s22 = 0.5;

bounds sigma > 0; bounds s11 > 0; bounds s22 > 0;

censored = (y=-100); *Indicator variable;

mu = g0 + gr*(RA+SN_RA) + gs*SLE + ga*age +
gt1*(time3+time4) + gt2*(time5+time6) +
gt1r*(time3+time4)*(RA+SN_RA) + gt2r*(time5+time6)*(RA+SN_RA) +
gt1s*(time3+time4)*SLE + gt2s*(time5+time6)*SLE + t1;

mu_pi = b0 + br*(RA+SN_RA) + bs*SLE + ba*age +
bt1*(time3+time4) + bt2*(time5+time6) +
bt1r*(time3+time4)*(RA+SN_RA) + bt2r*(time5+time6)*(RA+SN_RA) +
bt1s*(time3+time4)*SLE + bt2s*(time5+time6)*SLE + t2;

pi = CDF('normal', mu_pi);

w = sqrt(sigma**2 + delta**2);
if censored then p = pi;
```

```
else p = (1-pi) * 2/w * PDF('normal',(y-mu)/w) *
         CDF('normal',delta/sigma*(y-mu)/w);
loglike = log(p);

model y ~ general (loglike);
random t1 t2 ~ normal([0,0],[s11, s12, s22]) subject = individual;
run;


/* TOBIT WITH SKEW */
proc nlmixed
data = multiplex
cov
df = 1000
alpha = 0.05
gconv = 0
tech = QUANEW
update = BFGS;
parms /* Initial values */
g0 = 2.41 ga = 0 gr = 0.33 gs = 0.53 gt1 = 0.91 gt2 = 0.92
gt1r = -0.97 gt2r = -1.25 gt1s = -1.44 gt2s = -1.41
sigma = 0.51 delta = -1.61
s11 = 0.5;

bounds sigma >= 0; bounds s11 > 0;

T = log(0.29); *LOD;
censored = (y=-100); *Indicator variable;

mu = g0 + gr*(RA+SN_RA) + gs*SLE + ga*age +
gt1*(time3+time4) + gt2*(time5+time6) +
gt1r*(time3+time4)*(RA+SN_RA) + gt2r*(time5+time6)*(RA+SN_RA) +
gt1s*(time3+time4)*SLE + gt2s*(time5+time6)*SLE + t1;

* Estimate Owenst T-function;
h = (T - mu)/sqrt(sigma**2 + delta**2);
a = delta/sigma;
jmax = 50;
cut.point = 8;

aa = abs(a);
ah = abs(h);
if (aa = .I) then OT = sign(a) * 0.5*CDF('normal',-ah);
else if (aa = 0) then OT = 0;
else if (ah = .I) then OT = 0;
```

```
else if (aa <= 1) then do;
    th = ah;
    ta = aa;

    * Two ways to approximate integral based on cutpoint;
    low = (th <= cut.point);
    if low then do; *th < 8, ta <= 1;
        matr = 0;
        cumb = 0;
            do j = 0 to jmax;
                cumb = cumb + (th**(2*j)) / ((2**j) * GAMMA(j+1));
                b1 = EXP(-0.5 * th**2) * cumb;
                matr = matr + (-1)**j * ta**(2*j+1) / (2*j+1) * (1-b1);
            end; *for;
        T.int = (ATAN(ta) - matr) / (2*CONSTANT("pi"));
        OT = SIGN(a) * T.int;
        end; *if;
    else do; *th > 8, ta <= 1;
        if(-0.5*(th**2)*ta/ATAN(ta) < -600) then factor = 0;
        else if(-0.5*(th**2)*ta/ATAN(ta) > 600) then factor = .I;
        else factor = EXP(-0.5*(th**2)*ta/ATAN(ta));

        T.int = ATAN(ta) * factor * (1+0.00868*(th*ta)**4) / (2*CONSTANT("pi"));
    end;
end; *if;
else do; *(aa > 1);
    th = aa * ah;
    ta = 1 / aa;

    * Two ways to approximate integral based on cutpoint;
    low = (th <= cut.point);
    if low then do;
        matr = 0;
        cumb = 0;
            do j = 0 to jmax;
                cumb = cumb + (th**(2*j)) / ((2**j) * GAMMA(j+1));
                b1 = EXP(-0.5 * th**2) * cumb;
                matr = matr + (-1)**j * ta**(2*j+1) / (2*j+1) * (1-b1);
            end; *for;
        T.int = (ATAN(ta) - matr) / (2*CONSTANT("pi"));
        end; *if;
    else do;
        if(-0.5*(th**2)*ta/ATAN(ta) < -600) then factor = 0;
        else if(-0.5*(th**2)*ta/ATAN(ta) > 600) then factor = .I;
        else factor = EXP(-0.5*(th**2)*ta/ATAN(ta));
```

```
        T.int = ATAN(ta) * factor * (1+0.00868*(th*ta)**4) / (2*CONSTANT("pi"));
    end;

    OT = sign(a) * (0.5*CDF('normal',ah) +
         CDF('normal',aa*ah)*(0.5-CDF('normal',ah)) - T.int);
end;

w = sqrt(sigma**2 + delta**2);
if censored then p =  (CDF('normal',(T-mu)/w) - 2*OT);
else p = 2/w * PDF('normal',(y-mu)/w) * CDF('normal',delta/sigma*(y-mu)/w);
loglike = log(p);

model y ~ general (loglike);
random t1 ~ normal(0,s11) subject = individual;
run;

/* SUBSTITUTE WITH SKEW */
proc nlmixed
data = multiplex_sort
cov
df = 1000
alpha = 0.05
gconv = 0
tech= QUANEW
update = BFGS;
parms
g0 = 2.41 ga = 0 gr = 0.33 gs = 0.53 gt1 = 0.91 gt2 = 0.92
gt1r = -0.97 gt2r = -1.25 gt1s = -1.44 gt2s = -1.41
sigma = 0.51 delta = -1.61
s11 = 0.5;

bounds sigma >= 0; bounds s11 > 0;

T = log(0.29); *LOD;
censored = (y=-100); *Indicator variable;

mu = g0 + gr*(RA+SN_RA) + gs*SLE + ga*age +
gt1*(time3+time4) + gt2*(time5+time6) +
gt1r*(time3+time4)*(RA+SN_RA) + gt2r*(time5+time6)*(RA+SN_RA) +
gt1s*(time3+time4)*SLE + gt2s*(time5+time6)*SLE + t1;

w = sqrt(sigma**2 + delta**2);
if censored then p = 2/w * PDF('normal',(T-log(2)-mu)/w) *
                     CDF('normal',delta/sigma*(T-log(2)-mu)/w);
else p = 2/w * PDF('normal',(y-mu)/w) * CDF('normal',delta/sigma*(y-mu)/w);
```

```
loglike = log(p);

model y ~ general (loglike);
random t1 ~ normal(0,s11) subject = individual;
run;
```

C Bivariate Analysis - R code

```
fit_bivariate = function(y, par, T1, T2, parts){
# Returns MLE of parameters with specified number of model parts
# y: data points
# par:  initial values
# T1, T2: Detection limits
# parts: Number of model parts. 0 = only fully observed pairs.

  y1 = y[,1]
  y2 = y[,2]
  if(parts == 1){ # Tobit
    return(optim(par = par, fn = optim_bitobit, y1 = y1, y2 = y2,
                 T1 = T1, T2 = T2, method = "BFGS", hessian = TRUE,
                 control = list("fnscale"=-1)))
  }else if(parts == 2){ # Two-part
    return(optim(par = par, fn = optim_bitp, y1 = y1, y2 = y2,
                 T1 = T1, T2 = T2, method = "BFGS", hessian = TRUE,
                 control = list("fnscale"=-1)))
  }else if(parts == 4){ # Four-part
    return(optim(par = par, fn = optim_fp, y1 = y1, y2 = y2,
                 T1 = T1, T2 = T2, method = "BFGS", hessian = TRUE,
                 control = list("fnscale"=-1)))
  }else if(parts == 0){ # Only fully observed pairs
    return(optim(par = par, fn = optim_obs, y1 = y1, y2 = y2,
                 method = "BFGS", hessian = TRUE, control = list("fnscale"=-1)))
  }
}


####### TOBIT
log_likelihood_bitobit = function(y1,y2,mu1,mu2,sigma1,sigma2,mu_rho,T1,T2){
# Loglikelihood of bivariate tobit model
  sigma1 = abs(sigma1)
  sigma2 = abs(sigma2)
  rho = 2*pnorm(mu_rho)-1
```

```
  sigma12 = sqrt(sigma1^2*(1-rho^2))
  sigma21 = sqrt(sigma2^2*(1-rho^2))

  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y1)){
    y1i = y1[i]
    y2i = y2[i]

    mu12 = mu1 + rho*sigma1/sigma2 * (y2i- mu2)
    mu21 = mu2 + rho*sigma2/sigma1 * (y1i- mu1)

    if(!is.na(y1i) & !is.na(y2i)){
      update = dnorm(y1i, mean = mu12, sd = sigma12, log = TRUE) +
               dnorm(y2i, mean = mu2, sd = sigma2, log = TRUE)
    }else if(!is.na(y2i)){
      update = dnorm(y2i, mean = mu2, sd = sigma2, log = TRUE) +
               pnorm(T1, mean = mu12, sd = sigma12, log = TRUE)
    }else if(!is.na(y1i)){
      update = dnorm(y1i, mean = mu1, sd = sigma1, log = TRUE) +
               pnorm(T2, mean = mu21, sd = sigma21, log = TRUE)
    }else{
      Sigma = matrix(c(sigma1^2,rho*sigma1*sigma2,rho*sigma1*sigma2,sigma2^2),
                     ncol = 2)
      update = log(pmvnorm(lower = c(-Inf,-Inf), upper = c(T1,T2),
                           mean = c(mu1,mu2), sigma = Sigma))
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_bitobit = function(par, T1, T2, y1, y2){
# Function to be maximized by optim for Tobit model
  mu1 = par[1]
  mu2 = par[2]
  sigma1 = par[3]
  sigma2 = par[4]
  mu_rho = par[5]
  return(log_likelihood_bitobit(y1, y2,mu1,mu2,sigma1,sigma2,mu_rho,T1,T2))
}


#### TWO- PART
```

```
log_likelihood_bitp = function(y1,y2,mu1,mu2,sigma1,sigma2,mu_rho,mu_pi,T1,T2){
# Log-likelihood function bivariate two-part model
  sigma1 = abs(sigma1)
  sigma2 = abs(sigma2)
  pi = pnorm(mu_pi)
  rho = 2*pnorm(mu_rho)-1


  sigma12 = sqrt(sigma1^2*(1-rho^2))
  sigma21 = sqrt(sigma2^2*(1-rho^2))

  likelihood = 0
  for(i in 1:length(y1)){
    y1i = y1[i]
    y2i = y2[i]

    mu12 = mu1 + rho*sigma1/sigma2 * (y2i- mu2)
    mu21 = mu2 + rho*sigma2/sigma1 * (y1i- mu1)

    if(!is.na(y1i) & !is.na(y2i)){
      update = log(1-pi) + dnorm(y1i, mean = mu12, sd = sigma12, log = TRUE) +
               dnorm(y2i, mean = mu2, sd = sigma2, log = TRUE)
    }else if(!is.na(y2i)){
      update = log(1-pi) + dnorm(y2i, mean = mu2, sd = sigma2, log = TRUE) +
               pnorm(T1, mean = mu12, sd = sigma12, log = TRUE)
    }else if(!is.na(y1i)){
      update = log(1-pi) + dnorm(y1i, mean = mu1, sd = sigma1, log = TRUE) +
               pnorm(T2, mean = mu21, sd = sigma21, log = TRUE)
    }else{
      Sigma = matrix(c(sigma1^2,rho*sigma1*sigma2,rho*sigma1*sigma2,sigma2^2),
                     ncol = 2)
      update = log(pi + (1-pi)*pmvnorm(lower = c(-Inf,-Inf), upper = c(T1,T2),
                                       mean = c(mu1,mu2), sigma = Sigma))
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_bitp = function(par, T1, T2, y1, y2){
# Function to be maximized by optim for bivariate two-part model
  mu1 = par[1]
  mu2 = par[2]
  sigma1 = par[3]
  sigma2 = par[4]
```

```
  mu_rho = par[5]
  mu_pi = par[6]
  return(log_likelihood_bitp(y1, y2,mu1,mu2,sigma1,sigma2,mu_rho,mu_pi,T1,T2))
}


#### FOUR-PART

log_likelihood_fp = function(y1,y2,mu1,mu2,sigma1,sigma2,mu_rho,mu_pi1,mu_pi2,
                             mu_pi3,mu1L,sigma1L,mu2L,sigma2L,T1,T2){
# Log-likelihood function bivariate four-part model
  sigma1 = abs(sigma1)
  sigma2 = abs(sigma2)
  sigma1L = abs(sigma1L)
  sigma2L = abs(sigma2L)
  pi1 = exp(mu_pi1)/(exp(mu_pi1) + exp(mu_pi2) + exp(mu_pi3) + 1)
  pi2 = exp(mu_pi2)/(exp(mu_pi1) + exp(mu_pi2) + exp(mu_pi3) + 1)
  pi3 = exp(mu_pi3)/(exp(mu_pi1) + exp(mu_pi2) + exp(mu_pi3) + 1)
  #print(pi1 + pi2 + pi3)
  rho = 2*pnorm(mu_rho)-1

  sigma12 = sqrt(sigma1^2*(1-rho^2))
  sigma21 = sqrt(sigma2^2*(1-rho^2))

  likelihood = 0

  # Iterate over all observations
  for(i in 1:length(y1)){
    y1i = y1[i]
    y2i = y2[i]

    mu12 = mu1 + rho*sigma1/sigma2 * (y2i- mu2)
    mu21 = mu2 + rho*sigma2/sigma1 * (y1i- mu1)

    if(!is.na(y1i) & !is.na(y2i)){ # Both observed
      update = log(1-pi1-pi2-pi3) + dnorm(y1i, mean = mu12,
                                          sd = sigma12, log = TRUE) +
                dnorm(y2i, mean = mu2, sd = sigma2, log = TRUE)
    }else if(!is.na(y1i)){ # y1 Observed
      update = log((1-pi1-pi2-pi3)*dnorm(y1i, mean = mu1, sd = sigma1) *
                     pnorm(T2, mean = mu21, sd = sigma21) +
                     pi1*dnorm(y1i,mean=mu1L,sd=sigma1L))
    }else if(!is.na(y2i)){ # y2 Observed
      update = log((1-pi1-pi2-pi3)*dnorm(y2i, mean = mu2, sd = sigma2) *
                   pnorm(T1, mean = mu12, sd = sigma12) +
```

```
                    pi2*dnorm(y2i,mean=mu2L,sd=sigma2L))
    }else{ # Both censored
      Sigma = matrix(c(sigma1^2,rho*sigma1*sigma2,rho*sigma1*sigma2,sigma2^2),
                     ncol = 2)
      update = log(pi3 + pi1*pnorm(T1, mean = mu1L, sd = sigma1L) +
                    pi2*pnorm(T2, mean = mu2L, sd=sigma2L) +
                    (1-pi1-pi2-pi3)*pmvnorm(lower = c(-Inf,-Inf),
                                            upper = c(T1,T2), mean = c(mu1,mu2),
                                            sigma = Sigma))
    }
    likelihood = likelihood + update
  }
  return(likelihood)
}

optim_fp = function(par, T1, T2, y1, y2){
# Function to be maximized by optim for bivariate four-part model
  mu_pi1 = par[1]
  mu_pi2 = par[2]
  mu_pi3 = par[3]
  mu1 = par[4]
  mu2 = par[5]
  sigma1 = par[6]
  sigma2 = par[7]
  mu_rho = par[8]
  mu1L = par[9]
  sigma1L = par[10]
  mu2L = par[11]
  sigma2L = par[12]

  return(log_likelihood_fp(y1,y2,mu1,mu2,sigma1,sigma2,mu_rho,mu_pi1,mu_pi2,
                           mu_pi3,mu1L,sigma1L,mu2L,sigma2L,T1,T2))
}
```

D Application to Cytokine Data - Results

Table 8.1: Resulting maximum likelihood parameter estimates from fitting the models on the data on the cytokine MCP1. The left column is the full TP model without skew, and the right column is the simplified TP model without age in the continuous part and diagnosis in the discrete part. Wald-type 95% confidence intervals are included in parenthesis for the fixed effects.

Parameter	TP	TP
γ_0	3.66 *(2.26, 5.05)*	3.39 *(2.97, 3.79)*
γ_{age}	-0.01 *(-0.05, 0.03)*	—
γ_{RA}	0.71 *(0.15, 1.28)*	0.63 *(0.10, 1.17)*
γ_{SLE}	0.94 *0.37, 1.50)*	0.90 *(0.37, 1.43)*
γ_{t_2}	1.56 *(1.20, 1.91)*	1.56 *(1.20, 1.91)*
γ_{t_3}	1.66 *(1.20, 2.12)*	1.66 *(1.19, 2.12)*
$\gamma_{t_2 \times \text{RA}}$	-1.70 *(-2.27, -1.13)*	-1.70 *(-2.27, -1.13)*
$\gamma_{t_3 \times \text{RA}}$	-1.27 *(-1.89, -0.64)*	-1.27 *(-1.90, -0.64)*
$\gamma_{t_2 \times \text{SLE}}$	-1.66 *(-2.17, -1.15)*	-1.67 *(-2.18, -1.16)*
$\gamma_{t_3 \times \text{SLE}}$	-0.97 *(-1.59, -0.36)*	-0.97 *(-1.59, -0.36)*
β_0	-3.77 *(-6.46, -1.09)*	-4.11 *(-6.61, -1.61)*
β_{RA}	-1.88 *(-6.01, 2.25)*	—
β_{SLE}	-0.57 *(-2.81, 1.67)*	—
σ	0.78	0.78
s_{11}	0.60	0.56
s_{12}	-1.49	-1.06
s_{22}	4.97	3.85
$\ell(\hat{\boldsymbol{\theta}})$	-433.4	-434.2
AIC	900.8	896.4

Table 8.2: Resulting maximum likelihood parameter estimates from fitting the models on the data on the cytokine TNF-α. The left column is the full TPIC model without skew, and the right column is the simplified TPIC model without age and correlation between the random effects. Wald-type 95% confidence intervals are included in parenthesis for the fixed effects.

Parameter	TPIC	TPIC
γ_0	1.25 *(-0.38, 2.89)*	2.03 *(1.52, 2.54)*
β_0	-0.62 *(-2.76, 1.51)*	0.24 *(-0.39, 0.88)*
γ_{age}	-0.01 *(-0.06, 0.04)*	—
β_{age}	0.03 *(-0.04, 0.10)*	—
γ_{RA}	0.18 *(-0.47, 0.84)*	0.21 *(-0.39, 0.80)*
β_{RA}	-1.22 *(-2.08, -0.36)*	-1.18 *(-2.06, -0.31)*
γ_{SLE}	0.26 *(-0.42, 0.95)*	0.20 *(-0.43, 0.83)*
β_{SLE}	-0.72 *(-1.57, 0.13)*	-0.73 *(-1.61, 0.14)*
γ_{t_2}	0.92 *(0.41, 1.44)*	0.91 *(0.42, 1.40)*
β_{t_2}	-1.44 *(-2.18, -0.71)*	-1.45 *(-2.20, -0.70)*
γ_{t_3}	1.16 *(0.46, 1.87)*	1.09 *(0.48, 1.70)*
β_{t_3}	-0.53 *(-1.37, 0.31)*	-0.53 *(-1.37, 0.32)*
$\gamma_{t_2 \times \text{RA}}$	-0.74 *(-1.46, -0.02)*	-0.79 *(-1.45, -0.13)*
$\beta_{t_2 \times \text{RA}}$	1.83 *(0.68, 2.98)*	1.83 *(0.65, 3.00)*
$\gamma_{t_3 \times \text{RA}}$	-1.11 *(-1.96, -0.25)*	-1.13 *(-1.90, -0.36)*
$\beta_{t_3 \times \text{RA}}$	-0.21 *(-1.56, 1.14)*	-0.28 *(-1.73, 1.16)*
$\gamma_{t_2 \times \text{SLE}}$	-1.20 *(-1.91, -0.50)*	-1.30 *(-1.95, -0.65)*
$\beta_{t_2 \times \text{SLE}}$	1.42 *(0.40, 2.44)*	1.40 *(0.35, 2.46)*
$\gamma_{t_3 \times \text{SLE}}$	-0.99 *(-1.86, -0.12)*	-0.94 *(-1.72, -0.15)*
$\beta_{t_3 \times \text{SLE}}$	0.01 *(-1.17, 1.19)*	0.03 *(-1.17, 1.24)*
σ	0.77	0.22
δ	—	-1.34
s_{11}	0.49	0.42
s_{12}	-0.02	—
s_{22}	0.89	0.95
$\ell(\hat{\boldsymbol{\theta}})$	-442.7	-437.1
AIC	933.5	918.3