SEMIPARAMETRIC CUMULATIVE PROBABILITY MODELS FOR SKEWED, CENSORED, AND CLUSTERED CONTINUOUS RESPONSE DATA

By

Yuqi Tian

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May 13, 2022

Nashville, Tennessee

Approved:

Frank E. Harrell, Ph.D.

Bryan E. Shepherd, Ph.D.

Jonathan J. Schildcrout, Ph.D.

Peter F. Rebeiro, Ph.D.

To my parents and my husband Haoyu.

ACKNOWLEDGMENTS

First and most directly, I would like to sincerely thank my advisor, Dr. Bryan Shepherd, for his tremendous guidance, consistent support, and excellent expertise along the way. You are the best advisor I could ever ask for, who has taught me about research, collaboration, and also being a good person. I would not have made it here if it were not for your help and support.

I am extremely grateful to my dissertation committee members, Dr. Frank Harrell, Dr. Bryan Shepherd, Dr. Jonathan Schildcrout, and Dr. Peter Rebeiro for their valuable suggestions and insightful feedback on my research projects and dissertation. I also want to thank Dr. Jonathan Schildcrout, who co-advised the Chapter 3 project, for his constant inspiration and help. Special thanks belong to Dr. Chun Li, who is a professor at the University of Southern California and not officially on my dissertation committee, for his inspiration and guidance during our weekly meetings. Dr. Bryan Shepherd and Dr. Chun Li's NIH grant funding (R01 AI093234) has financially supported most of my graduate study and offered me freedom to pursue my research projects unencumbered.

I have been fortunate to be part of the department and learn from all faculty, students, and staff in the department. Thank you for all your support that has carried me through. Big thanks to each member of my cohort, Megan Murray, Alese Halvorson, Michale Williams, Cassie Hennessy, Jess Brown, Coleman Harris, Ying Ji, Shenglai He, and Ben Lane. It has been an exciting and pleasant journey with you.

Lastly, I would like to express my deepest love and gratitude to my beloved parents Shimei Tang and Mao Tian. Thank you for your unconditional love, unwavering supporting, and always encouraging me to follow my dreams. Ultimate thanks go to my husband, Haoyu, for his support, patience, kindness, wisdom and joy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Continuous response variables are very common and often skewed. Regression analysis is one of the most widely used types of continuous data analysis. One often needs to transform continuous response variables for regression modeling assumptions. However, finding the optimal transformation is challenging and results may vary with the choice of transformation. Log-transformation and square-root transformation are often recommended for right-skewed data to improve normality and homoscedasticity. Box and Cox (1964) proposed a more general way for monotonic transformation that works for positive data. However, this two-stage parametric method does not consider the model uncertainty based on the transformation and is unable to always lead to ideal transformation. It is desirable to use more robust models that are invariant to transformation of response variables. Nonparametric and semiparametric approaches are usually more robust in this sense, but they may lack interpretability and computational efficiency. Liu et al. (2017) suggested applying the cumulative probability model (CPM), a semiparametric ordinal regression model, to continuous response variables to avoid pre-transformation of response variables. They showed that CPMs perform well in this setting. CPMs incorporate rank information only and are thus invariant to any monotonic transformation of response variable. In addition, CPMs model the cumulative distribution function (CDF) from which other quantities (e.g. expectations and quantiles) can be derived. Harrell (2020) implemented a computationally efficient algorithm for CPMs in the `orm()` function in the **rms** R package that is able to handle thousands of ordinal levels.

The CPM can be regarded as a semiparametric transformation model that models the CDF and requires link function specification (Zeng and Lin, 2006). CPMs assume that after some unspecified monotonic transformation, the response variable is a linear function of covariates and follows a distribution corresponding to the link function specified. The monotonic transformation is estimated nonparametrically as a step function by treating each response value as a distinct category. Hothorn et al. (2018) recently proposed a parametric linear transformation model, the most likely transformation (MLT) model, that estimates the transformation parametrically with basis functions. It is of interest to understand the strengths and limitations of the semiparametric and parametric linear transformation models. In Chapter 2, we compare the two novel transformation models, CPMs and MLTs, for fitting continuous response variables. We ran extensive simulations under different scenarios and compared both methods by analyzing data from an HIV biomarker study.

To complicate things, in addition to being skewed, continuous response variables can also be censored. A continuous variable subject to detection limits (DLs) can only be measured within a certain rage. To

investigate the association between a continuous response variable subject to DLs and covariates by regression analysis, many approaches explicitly or implicitly make parametric assumptions on the distribution outside the DLs. Dichotomizing a continuous response variable at a DL and then fitting logistic regression is a commonly used approach, but it may lead to information loss (Jiamsakul et al., 2017). With multiple DLs, more information is lost because a response variable is dichotomized at its smallest lower DL or largest upper DL. Replacing values outside DLs with a single constant is another common approach, but the results are sensitive to the constant used (Baccarelli et al., 2005). More sophisticated methods often make parametric assumptions on the distribution of values outside DLs, and biased results are possible when distributional assumptions are inappropriate. Nonparametric approaches for DLs, although usually more robust, do not allow for adjusting covariates. In Chapter 3, we introduce a new method to address DLs in the response variables based on CPMs. We also propose a new estimator for the conditional quantile derived from a CPM that is more interpretable when response variables are subject to DLs. Two examples are presented to demonstrate the proposed method. One investigates the association between covariates and a biomarker subject to a lower DL. The other example uses data from a large multi-cohort study of viral load after starting antiretroviral therapy (ART) among people with HIV. The response variable, viral load, is subject to multiple DLs that vary across sites and over time. We implement our method in an R package **multipleDL**.

When a continuous response variable is repeatedly measured for a subject or the continuous responses come in clusters, it is more challenging to model the clustered continuous response data due to correlation within clusters. Methods for cross-sectional continuous response variables assume observations are independent of each other. Therefore, cross-sectional methods cannot be directly applied on clustered data. Liang and Zeger (1986) and Zeger and Liang (1986) proposed the generalized estimating equation (GEE) method that combines generalized linear models and quasi-likelihood methods for longitudinal data analysis, where the quasi-likelihood method only requires specification of the first two multivariate moments of the response variable. GEE methods require correctly specifying the marginal regression model and a working correlation structure for association within clusters that does not necessarily need to be correct. To fit continuous response variables, GEE methods still have similar parametric assumption as generalized linear models and may require transformation of response variables. Again, inappropriate transformation might lead to biased results and results are sensitive to the choice of transformation in this case. Therefore, it is important to have robust approaches for clustered continuous response data. We study the extension of CPMs on analyzing clustered continuous response variables based on GEE methods for ordinal response variables in Chapter 4. Two feasible and computationally efficient approaches are proposed and demonstrated by simulations. We apply our approaches on two data examples. One uses data from The Lung Health Study to investigate the contribution of a single nucleotide polymorphism to lung function decline. The other studies predictors of

CD4:CD8 ratios in an HIV study. An R package **cpmgee** has been developed.

CHAPTER 2

An Empirical Comparison of Two Novel Transformation Models

This chapter is from An Empirical Comparison of Two Novel Transformation Models published in *Statistics in Medicine* and has been reproduced with the permission of the publisher and my co-authors Torsten Hothorn, Chun Li, Frank Harrell, and Bryan Shepherd.

2.1 Introduction

We often transform continuous response variables to meet modeling assumptions, but it is not easy to find the optimal transformation. Box and Cox modified a method proposed by Tukey (Box and Cox, 1964; Tukey et al., 1957) that provides a family of power transformations to create a monotonic function of the responses. The Box-Cox transformation is widely used to improve normality and homoscedasticity. However, the Box-Cox transformation only works for positive response variables. It is generally implemented in a two-stage manner (1. select transformation, 2. fit model to transformed response) that ignores the model uncertainty regarding the choice of transformation, and it is still a parametric procedure that may result in sub-optimal transformations.

Two transformation models have recently been proposed: semiparametric cumulative probability models (CPMs) (Liu et al., 2017) and parametric most likely transformation models (MLTs) (Hothorn et al., 2018). Both approaches model the cumulative distribution function and require specifying a link function, which implicitly assumes the response variable follows a known distribution after some monotonic transformation. However, the two approaches estimate the transformation differently. With CPMs, an ordinal regression model is fit, which essentially treats each realization of the response as a unique ordered category and encodes the empirical CDF into the intercepts, and therefore nonparametrically estimates the transformation; CPMs belong to the class of semiparametric linear transformation models (Zeng and Lin, 2007; De Neve et al., 2019). In contrast, with MLTs, the transformation is parameterized using flexible basis functions. Conditional expectations and quantiles are readily derived from both methods on the outcome's original scale. Both methods have been shown to be robust and flexible, and have good performance in estimation (Liu et al., 2017; Hothorn et al., 2018).

The goal of this paper is to compare the CPM and MLT methods to each other to better understand the advantages and disadvantages of each. In Section 2, we give a brief introduction to linear transformation models, cumulative probability models and most likely transformation models. In Section 3, we describe a wide range of simulation scenarios to compare the methods and in Section 4 we present simulation results.

In Section 5 we illustrate and contrast both methods using data from a study of biomarkers among persons living with HIV. Finally, we provide discussions and conclusions in Section 6.

2.2 Review of Methods

2.2.1 Linear Transformation Models

Let Y designate a continuous response variable. The goal is to model some aspect of the distribution of Y as a function of a vector of covariates, X. It may be difficult to directly model Y, so the analyst may instead want to model a transformation of the outcome, $Y^* = h(Y)$, where $h(\cdot)$ is a monotonic transformation. A linear transformation model assumes $h(Y) = Y^* = \beta^T X + \varepsilon$, where $\varepsilon \sim F_\varepsilon$ is a known distribution. Let $H(\cdot) \equiv h^{-1}(\cdot)$. Then

$$Y = H(Y^*) = H(\beta^T X + \varepsilon), \text{ where } \varepsilon \sim F_\varepsilon. \tag{2.1}$$

The linear transformation model (2.1) can be rewritten as a cumulative probability model (CPM). The conditional cumulative distribution function of Y can be expressed as

$$\begin{aligned} F(y \mid X) &= P(Y \leq y \mid X) \\ &= P[H(\beta^T X + \varepsilon) \leq y \mid X] \\ &= P[\varepsilon \leq H^{-1}(y) - \beta^T X \mid X] \\ &= F_\varepsilon[h(y) - \beta^T X]. \end{aligned}$$

Let $G = F_\varepsilon^{-1}$ be a link function. Then

$$G[F(y \mid X)] = h(y) - \beta^T X. \tag{2.2}$$

2.2.2 Semiparametric Cumulative Probability Models

A semiparametric linear transformation model leaves the transformation, $h(y)$, unspecified, estimating it nonparametrically with a step function (Zeng and Lin, 2006). The partial likelihood approach to the Cox model can also be interpreted as a member of this class. We first consider the situation of no ties in the outcome. Without loss of generality, assume $y_1 < y_2 < \cdots < y_n$. Then for the observed values $\{y_i; i = 1, 2, \ldots, n\}$, the semiparametric CPM is

$$G[F(y_i \mid X)] = \alpha_i - \beta^T X, \tag{2.3}$$

where $\alpha_i = h(y_i)$.

Since $\alpha(\cdot)$ is an increasing function, $\alpha_1 < \alpha_2 < \cdots < \alpha_n$. The semiparametric likelihood can then be approximated as

$$L^*(\beta, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \left[F_\varepsilon(\alpha_i - \beta^T x_i) - F_\varepsilon(\alpha_{i-1} - \beta^T x_i) \right], \tag{2.4}$$

where an auxiliary parameter $\alpha_0 (< \alpha_1)$ is added in the model. L^* is maximized when $\hat{\alpha}_0 = -\infty$ and $\hat{\alpha}_n = +\infty$ because F_ε is increasing, so in practice $\hat{\alpha}_0$ and $\hat{\alpha}_n$ are fixed to these values and maximization of L^* is with respect to the other parameters (Liu et al., 2017).

The semiparametric cumulative probability model (2.3) is equivalent to the "cumulative link model" commonly used for the analysis of ordered categorical data (Walker and Duncan, 1967; McCullagh, 1980), and the likelihood (2.4) is equivalent to the multinomial likelihood used for these ordinal models. In fact, maximizing (2.4) to obtain nonparametric maximum likelihood estimators (NPMLEs) for $(\beta, \boldsymbol{\alpha})$ can be done by treating continuous Y as if it were a discrete ordinal variable with n categories. The approach also works seamlessly if Y is a mixture of continuous and discrete data or if there are ties (Liu et al., 2017).

Although in theory, semiparametric CPMs can be fit using algorithms for cumulative link models, in practice, most commonly used software programs employ algorithms that can handle only a relatively small number of discrete ordinal categories. However, this need not be the case, as large portions of the score equation and Hessian matrix are zero permitting computational simplifications. The `orm()` function in the **rms** package in R statistical software allows efficient maximization of (2.4) for continuous Y with thousands of distinct levels (Harrell Jr, 2015; Harrell Jr et al., 2016).

With the NPMLEs $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$, one can estimate the conditional CDF, $\hat{F}(y_i \mid X) = F_\varepsilon(\hat{\alpha}_i - \hat{\beta} X)$. From the estimated conditional CDF, one can estimate conditional expectations and conditional quantiles. The delta method can be used to derive the standard error for the conditional CDF and the conditional expectation. Confidence intervals for conditional quantiles can be obtained using linear interpolation of the inverse of confidence intervals for the conditional CDF. Details are in Liu et al. (2017). The probability index (PI), defined as $P(Y_1 < Y_2 | X_1, X_2)$ for independent and identically distributed copies (Y_1, X_1) and (Y_2, X_2), and its confidence interval can also be readily obtained from CPMs (Acion et al., 2006; De Neve et al., 2019).

2.2.3 Most Likely Transformation Models

The motivation behind most likely transformation models is similar to that of semiparametric CPMs. After some transformation, $h(y)$, the outcome is assumed to be linearly associated with covariates with errors following a known distribution, F_ε, leading to the linear transformation model (2.1). This can then be re-

written as the cumulative probability model (3.1). MLTs differ from semiparametric CPMs in the manner that the unknown transformation function, $h(y)$, is modeled. Rather than nonparametrically estimating $h(y)$, it is flexibly modeled using basis functions. Specifically, $h(y) = a(y)^T \vartheta$, where a is a vector of appropriate basis functions and ϑ is a vector of coefficients. The conditional cumulative probability model then becomes

$$G[F(y \mid X)] = a(y)^T \vartheta - \beta^T X, \tag{2.5}$$

where as before, $G = F_\varepsilon^{-1}$.

The choice of basis function is problem-specific and depends on the scale of Y. For continuous outcomes, the basis functions can be any polynomial or splines basis. Bernstein polynomials of order M can be applied on the support of y, $[l, u]$, as

$$h(y) = a_{\text{Bs},M}(y)^T \vartheta = \sum_{m=0}^{M} \vartheta_m f_{\text{Be}(m+1, M-m+1)}(\tilde{y})/(M+1), \tag{2.6}$$

where $\tilde{y} = \frac{y-l}{u-l} \in [0,1]$ and $f_{\text{Be}(m,M)}$ is the probability density function of a Beta distribution with parameters m and M. In theory, the Bernstein polynomials can approximate any function on an interval as long as M is big enough. Polynomial basis functions and log basis functions can also be used in suitable cases. The monotonicity of h can be ensured by constrained optimization.

A more general class of transformation models are conditional transformation models of the form

$$G[F(y \mid X)] = c(y,x)^T \vartheta, \tag{2.7}$$

where the unknown transformation function now depends both on y and x, and c is a vector of basis functions conditioning on x (Hothorn and Zeileis, 2017). Although the MLT framework handles such transformations, unless noted otherwise, we will consider models of the form (2.5) rather than of the form (2.7).

Estimation proceeds using maximum likelihood. The likelihood of a datum $C = (\underline{y}, \bar{y}]$, where $(\underline{y}, \bar{y}]$ is a short interval around y, for a given transformation function h is (Lindsey, 1996):

$$L(h|Y \in (\underline{y}, \bar{y}]) = F_\varepsilon(h(\bar{y})) - F_\varepsilon(h(\underline{y})). \tag{2.8}$$

For absolute continuous responses, the log-density is used as log-likelihood and the maximum likelihood estimator of h is called most likely transformation (Hothorn et al., 2018).

The **mlt** R package is an implementation of most likely transformation models in R (Hothorn, 2018). A variety of increasingly complex transformation models can be built and evaluated in a computationally

efficient way by this package. In the rest of the paper, MLT refers to the theoretical method rather than the package. As with semiparametric CPMs, conditional expectations, quantiles, and probability indices and their confidence intervals can be computed after fitting MLT models.

2.3 Simulation Plan

2.3.1 Simulation Set-up

We compared semiparametric CPMs and MLTs using a wide variety of simulation scenarios. The basic structure for our simulations was the following:

$$\begin{aligned} Y^* &= X\beta + Z\gamma + \varepsilon, \\ \varepsilon &\sim F_\varepsilon(.), \\ Y &= H(Y^*). \end{aligned}$$

For the primary simulation setting, we set $\beta = 1, \gamma = 0$ with no Z included in the model, $X \sim$ *Binomial*$(p = 0.5), \varepsilon \sim N(0,1)$, and $H(y) =$ Inv-$\chi^2(\Phi(y), 5)$, where Φ is the probability density function of the standard normal distribution and Inv-$\chi^2(\cdot, 5)$ is the inverse of the CDF for a chi-square distribution with 5 degrees of freedom. $H(\cdot)$ was chosen in this manner so that there would be no obvious closed form transformation function h. All other simulations were some variation from this primary simulation setting.

For each setting, we varied the sample size from 50, 100, 500 to 1000 and specified the number of simulation replications at 10,000. CPMs and MLTs were fit with the same specified link function. MLTs were generally fit using Bernstein polynomials with $M = 10$ unless stated otherwise.

Modifications of the primary simulation setting included the following:

- $\beta = 0$ and 0.5.
- $X \sim$ *Binomial*$(p = 0.3)$, *Uniform*(-1, 1), and $N(0,1)$.
- $\varepsilon \sim N(0,1)$, *Logistic*$(0, 3/\pi^2)$, and *Gompertz*$(0,1)$.
- $Z \sim N(0,1)$ and $N(X,1)$; $\gamma = 1$.
- Multiple covariates $Z_1, \ldots, Z_6$, with $Z_1, Z_2, Z_3 \sim N(0,1), Z_4 \sim N(X,1), Z_5 \sim N(Z_1 + X, 1)$, and $Z_6 \sim N(Z_3 - Z_4, 1)$; $\gamma = \{1,1,1,1,1,1\}$.
- $H(y) = y$, $\exp(y)$, and Inv-*logistic*$(\Phi(y))$.

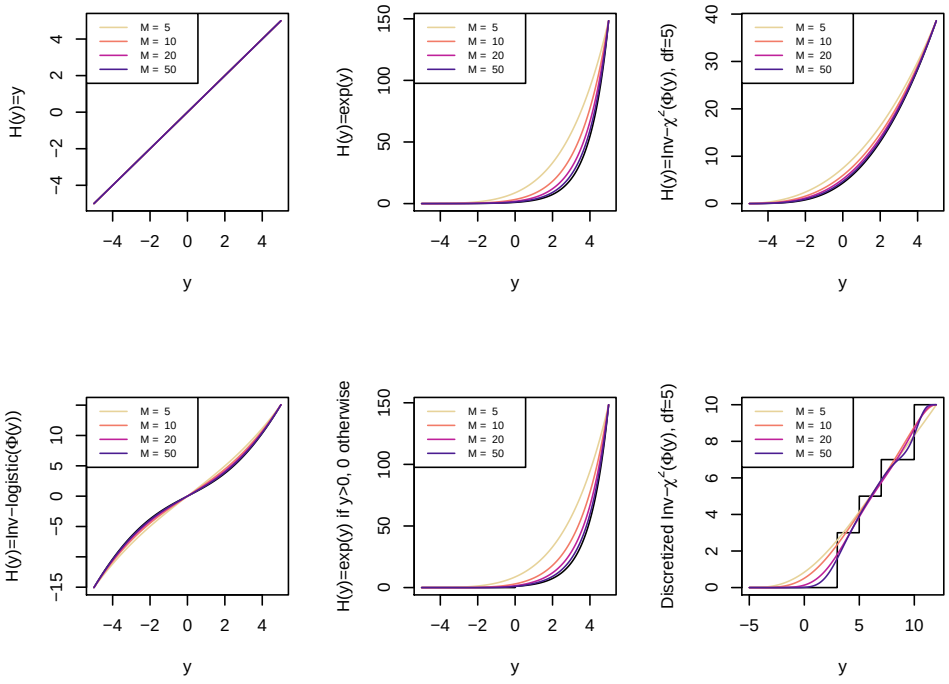

Figure 2.1: Transformation functions used in simulation and corresponding Bernstein polynomials approximation with order M

We also evaluated the two methods with data simulated from a mixed distribution, corresponding to a setting with a detection limit or left censoring:

$$H(y) = \begin{cases} \exp(y) \text{ if } y > 0 \\ 0 \text{ if } y \leq 0 \end{cases}$$

We also considered settings where Y was a discretized version of Y^* using 5, 10, 20, and 50 categories based on quantiles of the distribution (see details in Supplementary Materials).

Figure 2.1 illustrates the different transformation functions considered in these simulations. The Figure also includes curves illustrating how well the Bernstein polynomials approximate the transformation functions.

Note that the CDF in `orm()` is in the form $G_1[1-F(y|X)] = \alpha_{orm} + \beta_{orm}X$, which can be transformed to (3.1) if $G(t) = -G_1(1-t), \alpha = -\alpha_{orm}$ and $\beta = \beta_{orm}$. For symmetric error distributions, we use the same link function in `orm()` as in the CPM and $\alpha = -\alpha_{orm}$. For nonsymmetric error distributions, its complementary version can be used (Liu et al., 2017).

2.3.2 Evaluations

To evaluate the two methods, we estimated bias, mean squared error (MSE) and coverage of 95% confidence intervals for β, as well as conditional expectations, conditional quantiles, and conditional cumulative distribution functions. We also computed the out-of-sample log-likelihood based on the fitted model parameters for a separate data set of the same size sampled from the simulation distribution. For the purpose of comparing the out-of-sample log-likelihoods, the responses in MLT were categorized into short intervals $(\underline{y_i}, \bar{y_i}]$ based on CPM categorization, which were the distinct observed responses of the original data. The likelihood was then calculated as $L(H) = \prod_i \left[F_\varepsilon(H(\bar{y_i})) - F_\varepsilon(H(\underline{y_i}))\right]$.

Under correct link function specification, for $\varepsilon \sim N(0,1)$, the probit link function was used and when ε followed a logistic distribution, the logit link function was used. We used the cloglog link function if ε followed a Gompertz distribution.

Ordinary linear regression was also evaluated and compared with the two methods for simple transformations $H(y) = y$ and $H(y) = \exp(y)$. All simulations and analyses were performed in R version 3.4.4 (**?**); complete code is available at http://biostat.mc.vanderbilt.edu/ArchivedAnalyses and an abbreviated version is available at https://github.com/harrelfe/rscripts/blob/master/sim-continuous-ordinal.r.

2.4 Simulation Results

In general, CPMs and MLTs were quite comparable when models are correctly specified (i.e., correct link function and linear terms). Bias was close to 0, MSE was low, and the coverage probability of 0.95 confidence intervals tended to be 0.95 with increasing sample sizes. CPMs tended to have a slightly smaller bias for β than MLTs as the sample size increased. In terms of the conditional mean, CPMs generally had a smaller bias than MLTs, especially in large sample sizes, but MSEs were very close. Neither one had obvious advantages in estimating conditional quantiles. Both methods performed better estimating conditional medians than more extreme quantiles. CPMs generally had better performance in estimating conditional CDFs with a smaller bias, particularly in large sample sizes. MLTs tended to have slightly narrower confidence intervals. With continuous Y, the out-of-sample log-likelihood was larger in MLTs probably because it directly maximizes the likelihood whereas CPMs maximize an approximated multinomial likelihood. Details for specific simulation scenarios are provided below and in Supplementary Material.

2.4.1 The Primary Setting and its Modifications

Simulation results under the primary setting, with the order of Bernstein basis varying from $M = 5$ to $M = 10$, are shown in Figure 2.2 and reported in Table 2.3 (in Supplementary Material). For β estimation, CPMs and MLTs performed similarly, resulting in minimal bias and 95% coverage that improved with increasing sample

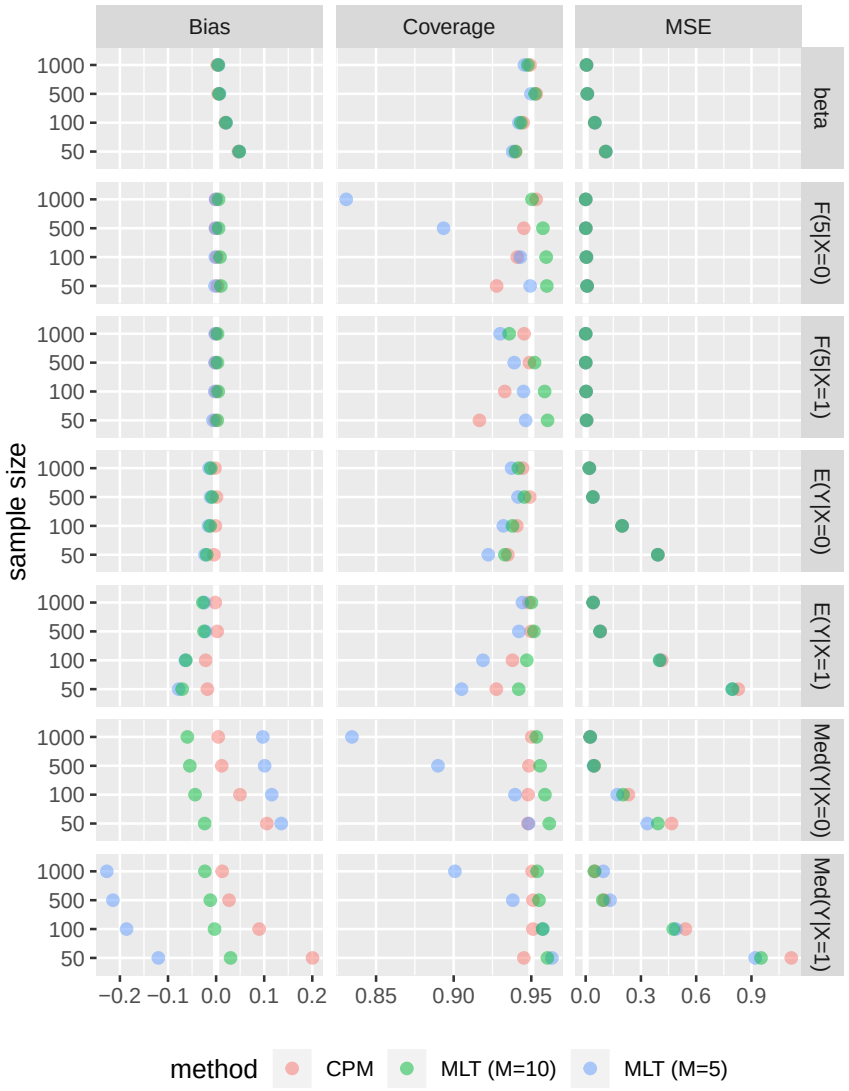

Figure 2.2: Simulation results under the primary setting

sizes. CPMs had slightly less bias and similar to lower coverage than MLT with $M = 10$ when estimating conditional expectations. For estimating conditional CDFs and medians, MLTs with $M = 5$ generally underperformed MLTs of $M = 10$ and CPMs. Coverage of MLT with $M = 10$ was slightly better than that of CPMs for conditional CDFs and slightly worse for conditional medians. At large samples, estimates of conditional medians were less biased for CPMs than MLT with $M = 10$, but more biased at small sample sizes. For most of the remaining simulations, CPMs were only compared with MLTs with $M = 10$.

For simple transformations $H(y) = y$ and $H(y) = \exp(y)$, ordinary linear regression after the correct transformation (i.e., no transformation and log-transformation, respectively) had, not surprisingly, the best performance in estimating β with much smaller bias, particularly at small sample sizes. The other two

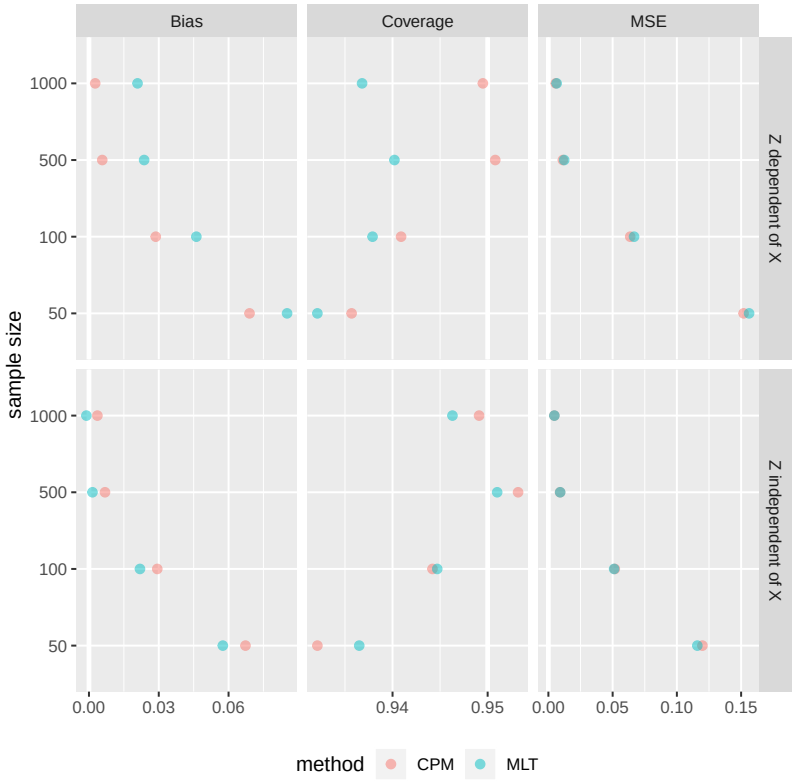

Figure 2.3: Simulation results when including covariate Z, which is dependent and independent of X

methods had similar respectable performance with coverage near 95% at all sample sizes and MSE 10% to 30% larger than the correctly specified linear regression, with MSE getting closer with larger sample sizes. The results of β estimation for transformation $H(y) = y$ are in Table 2.1. With moderate and large sample sizes, the results of estimated conditional expectation were very similar. More detailed results are shown in Supplementary Material Figure 2.10, Figure 2.11, Figure 2.12, Table 2.4, Table 2.6, and Table 2.6.

When including a covariate Z, the results are shown in Figure 2.3. If Z was independent of X, MLTs had a slightly smaller bias in β and the differences between the two methods decreased as the sample size got larger. MSEs and confidence interval coverage rates were similar. However, if Z was dependent on X, the CPM had slightly better performance in estimating β than the MLT. In both scenarios, the MLT had a larger out-of-sample log-likelihood. When including multiple covariates, some of them being independent of X while others being dependent on X, CPMs generally outperformed MLTs although only by a small amount. (See detailed results in Table 2.7, Table 2.8, Table 2.9, and Figure 2.13 in Supplementary Material.)

CPMs performed slightly better than MLTs when using the correct link function for $\varepsilon \sim \mathit{Logistic}(0, \frac{3}{\pi^2})$ (See Table 2.10 and Figure 2.14 in Supplementary Material). Results were similar using correct link function for $\varepsilon \sim \mathit{Gompertz}$ (See Table 2.11 and Figure 2.15 in Supplementary Material). Results were similar when us-

ing different distributions for X (see Supplementary Material Table 2.12, Table 2.13, Table 2.14, Figure 2.16, Figure 2.17, and Figure 2.18). When changing the value of β, the results were similar (see Supplementary Material Table 2.15, Table 2.20, Figure 2.19 and Figure 2.20).

2.4.2 Link Function Misspecification

Under minor or moderate link function misspecification, the bias of the estimated β was slightly smaller in CPMs. Results were similar in other evaluation criteria (See Table 2.17 and Figure 2.21 in Supplementary Material). With severe link function misspecification, MLTs tended to have slightly better performance in estimating β (See Table 2.18, Table 2.19, Figure 2.22, and Figure 2.23 in Supplementary Material). MLTs always had larger out-of-sample log-likelihood under model misspecification.

2.4.3 Mixture of Discrete and Continuous Responses

For the mixture of discrete and continuous responses corresponding to the setting where values below zero were set to zero, we compared CPMs and two MLT models, one treating the responses as ordinary continuous responses and the second properly treating the zero values as left censored responses. For β estimation, the results are shown in Figure 2.4. For small sample sizes, the uncensored MLT had the smallest bias while the censored MLT and CPM had better confidence interval coverage rates. However, the uncensored MLT performed the worst when the sample size became large. CPM had the smallest bias in large sample sizes and it also had the largest out-of-sample log-likelihood in all sample sizes. See Table 2.20 in Supplementary Material for more detailed results.

2.4.4 Discretization of Continuous Response

If continuous responses are discretized into categories, the MLT can handle them as ordered factors (i.e., resulting in identical estimation to CPMs) or as continuous responses. Simulation results are in Figure 2.5. CPMs, in general, performed better than MLTs (Bernstein polynomials with $M = 5$) treating the discrete data as continuous. Such advantages were more obvious as the sample size increased. MLTs outperformed CPMs for estimated β in sample sizes when the number of categories was small; while CPMs always had better confidence interval coverage rates for β. CPMs also had larger out-of-sample log-likelihood for all cases.

2.4.5 Computation Time

The average computing time for the primary setting based on 100 replications is shown in Table 2.2. In general, both methods are quite fast for moderate sample sizes. On average, CPMs ran much faster in small sample sizes while MLTs were faster in large sample sizes. MLTs with $M = 10$ took longer to run than MLTs

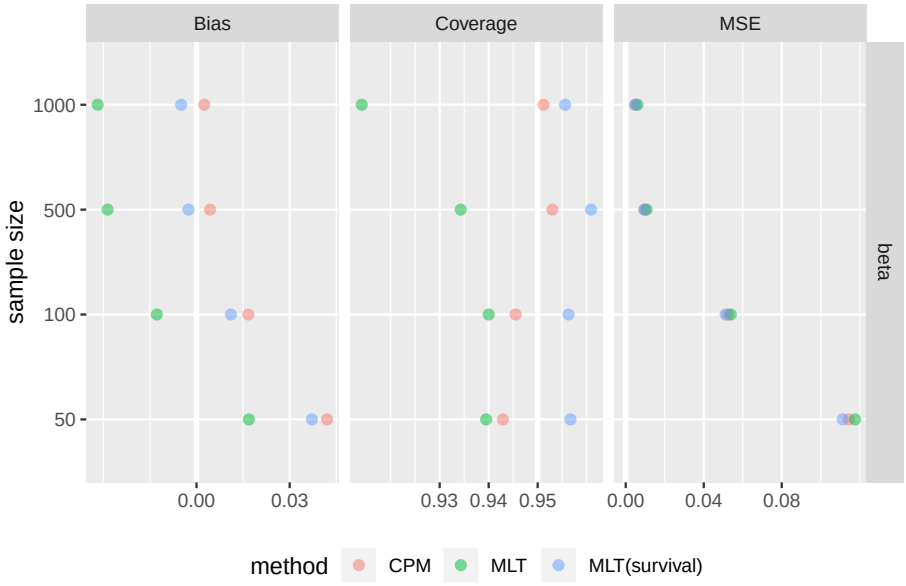

Figure 2.4: Simulation results for mixture of discrete and continuous responses comparing CPM and MLT treating response as ordinary continuous responses and censoring responses.

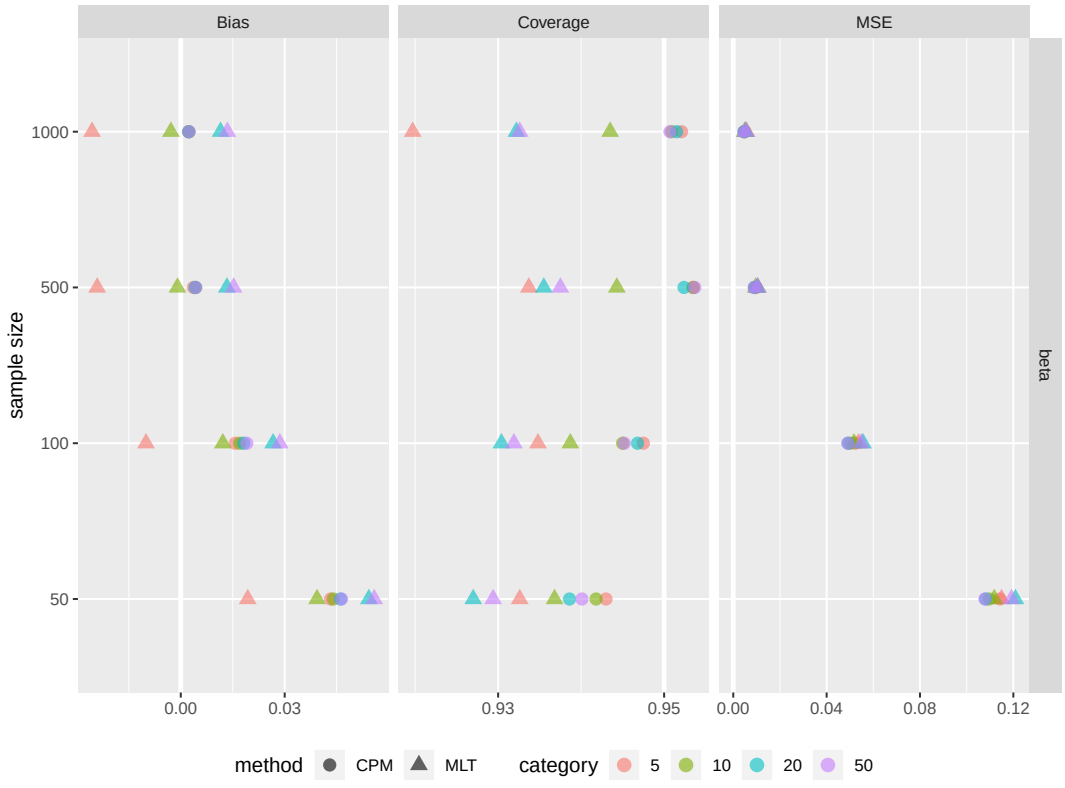

Figure 2.5: Simulation results for discretized continuous responses into 5, 10, 20 and 50 categories.

with $M = 5$. This simulation and all other simulations were performed on a 64 bit Linux server equipped with 2 Intel Xeon X5647 processors running at 2.93GHz, 96Gb of memory.

2.5 Application Examples

To further compare the two models, we applied them to a biomarker study among people living with HIV (PLWH). The risks of diabetes and cardiovascular disease are higher for PLWH than the general population. There is interest in assessing the association between body mass index (BMI) and biomarkers of inflammation and metabolism among PLWH. We used data from 216 HIV-positive adults on antiretroviral therapy (ART) with no history of diabetes or myocardial infarction and with a viral load less than or equal to 400 copies/mL from the Vanderbilt Lipoatrophy and Neuropathy Cohort (LiNC; n=147) (Koethe et al., 2012) and the Adiposity and Immune Activation Cohort (AIAC; n=69) (Koethe et al., 2016). We estimated the association between BMI and five inflammation biomarkers: Interleuken 6 (IL-6), high sensitivity C-reactive protein (hsCRP), Interleuken 1 β (IL-1-β), soluble CD14 (sCD14) and leptin. The study over-sampled overweight patients; the median BMI was 29.3 kg/m^2; the range was 17.8 to 57.4. The analysis adjusted for age, sex, race, study location, CD4 cell count, and smoking status. Probit link functions were used for all biomarkers.

Figure 2.6 shows the distribution of IL-6, which is right skewed and has a lower detection limit; those below the detection limit (3%) were recorded as having a value of 0. The estimated transformation functions are shown in Figure 2.6 and are similar for the CPM and MLT analyses. Because it is parametrically estimated by basis functions, the transformation function is a smooth curve for MLT. The transformation function for CPM is a step function. The estimated conditional mean and median as a function of BMI are also very similar for the two models. The estimated PI for IL-6 for a 10 kg/m^2 difference in BMI is 0.64 (95% CI 0.58-0.69) for both CPM and MLT analyses, further demonstrating the similarity between models. This suggests that for a 10 kg/m^2 difference in BMI, the subject with the higher BMI will have a 0.64 probability of having a higher IL-6.

As shown in Figure 2.7, the distribution of hsCRP is extremely right-skewed. The estimated transformation is similar between the CPM and MLT analyses, but it is not as close as it was in the analyses with IL-6 as the outcome. Hence, the conditional expectation and the conditional median as a function of BMI are comparable, but slightly different, under the two transformation models. The probability indices for a 10 kg/m^2 increase in BMI are 0.59 (95% CI 0.54-0.65) and 0.60 (95% CI 0.55-0.65) for CPM and MLT, respectively. Interestingly, if one initially log-transforms hsCRP, fits MLT, and then transforms back to the original scale, then the MLT estimates are much more similar to those of the CPM; Figure 2.8 shows the conditional expectation. Notice that CPM is invariant to any pre-transformation transformation; i.e., estimates of β, expectations, quantiles, and probability indices are identical whether or not one applies an initial

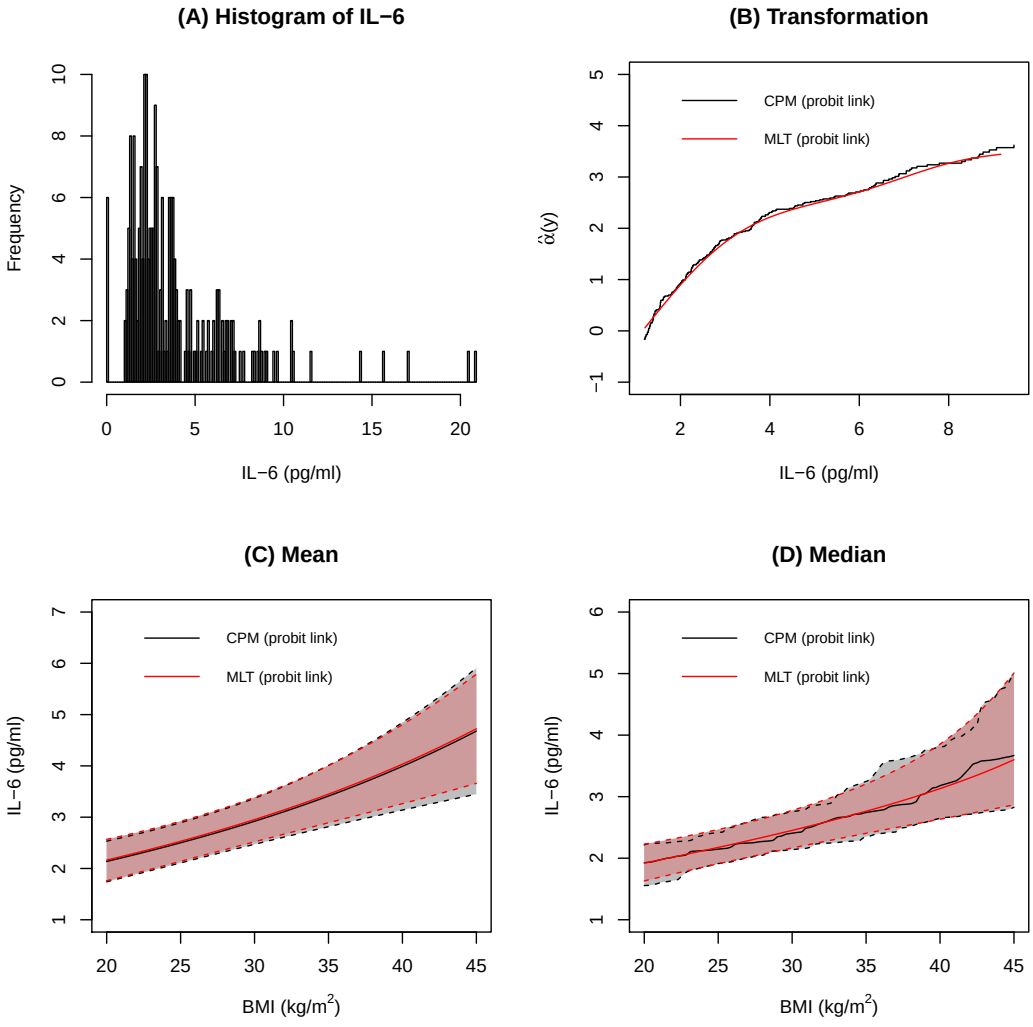

Figure 2.6: Results for IL-6. A: The distribution of IL-6. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

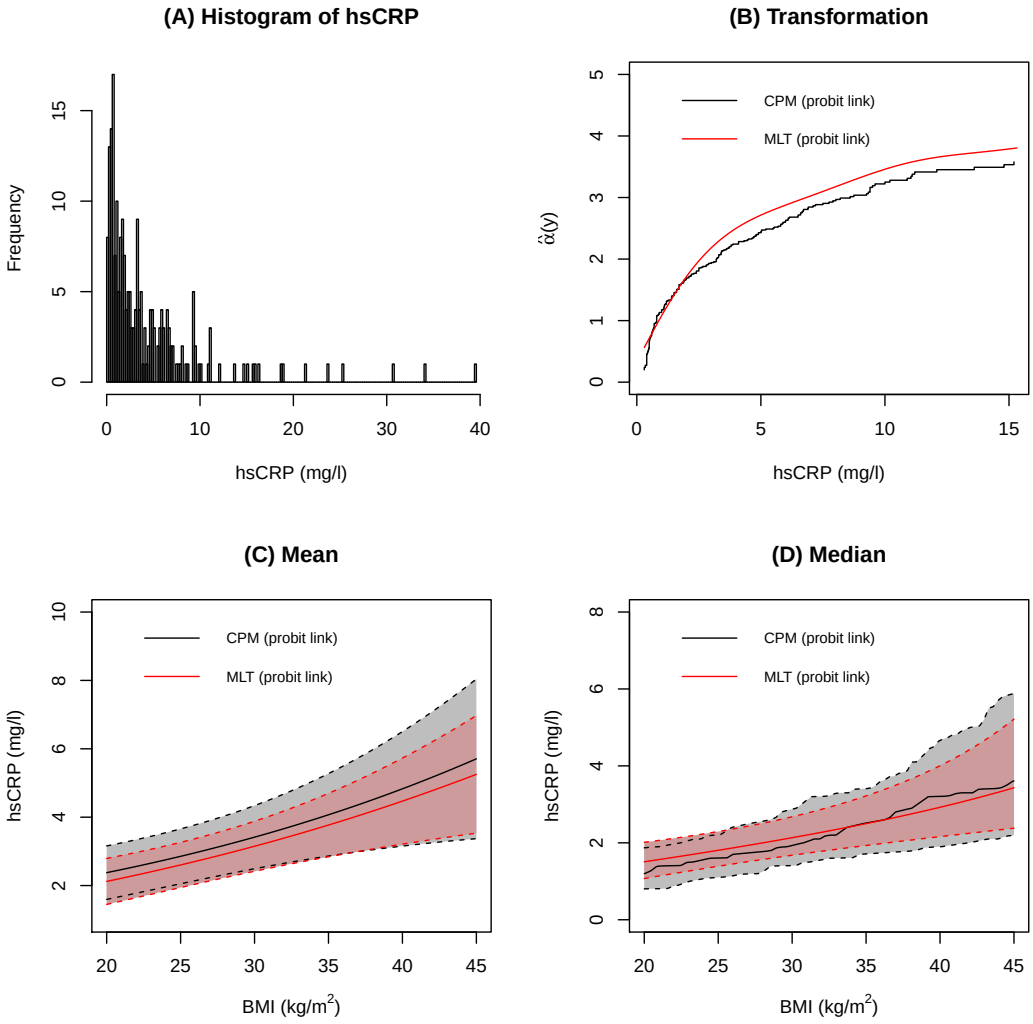

Figure 2.7: Results for hsCRP. A: The distribution of hsCRP. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

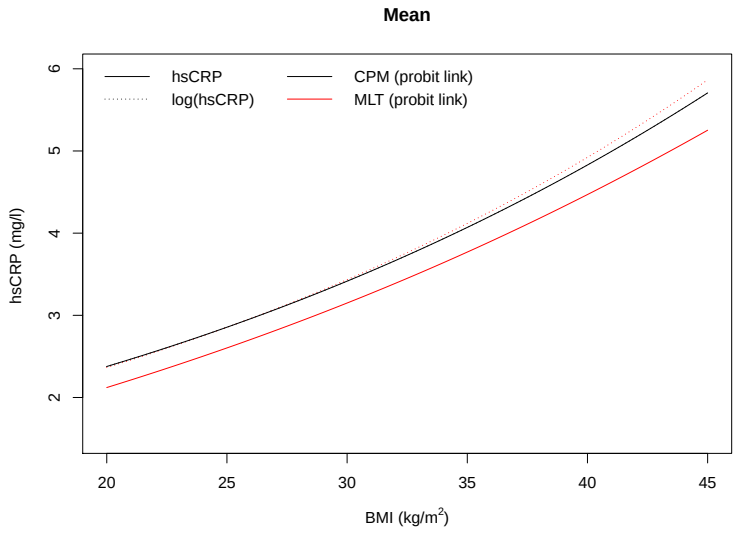

Figure 2.8: The comparison of the estimated conditional mean on the original scale and the transformed log scale

transformation. This is an advantage of CPM over MLT.

The distribution of IL-1-β is shown in Figure 2.9. It is right skewed and a large portion (39%) are below the assay detection limit and assigned the value 0. We applied CPM and MLT with left censoring to the data. The estimated transformation functions of the two models are somewhat similar. There is a flat line in the transformation function for CPM, which corresponds to the gap around 1 pg/ml in the histogram; CPM is flexible enough to capture this. The estimated conditional expectations are similar between the two models, with MLT generating a narrower confidence interval. The results for the conditional median are also similar for the two models. The PIs for 10 kg/m^2 difference in BMI is 0.50 (95% CI 0.44-0.56 for CPM and 0.44-0.55 for MLT) for both CPM and MLT, suggesting there is little association between BMI and IL-1-β.

We also fit CPM and MLT models to assess the association between BMI and the biomarkers leptin and sCD14. Leptin was positively associated with BMI and sCD14 was negatively associated. In both cases, results from the CPM and MLT models were almost identical (similar to the IL-6 results); details are in Figure 2.25 and Figure 2.26 in Supplementary Material.

2.6 Discussion

In this paper, we have reviewed two novel transformation models, CPMs and MLTs, and we have compared them under a variety of simulation settings. The paper also serves as a validation of the two software implementations in `orm()` and `mlt()`. Both methods directly model the conditional CDF from which other characteristics of the distribution can be derived easily. Both models are linear transformation models, in that

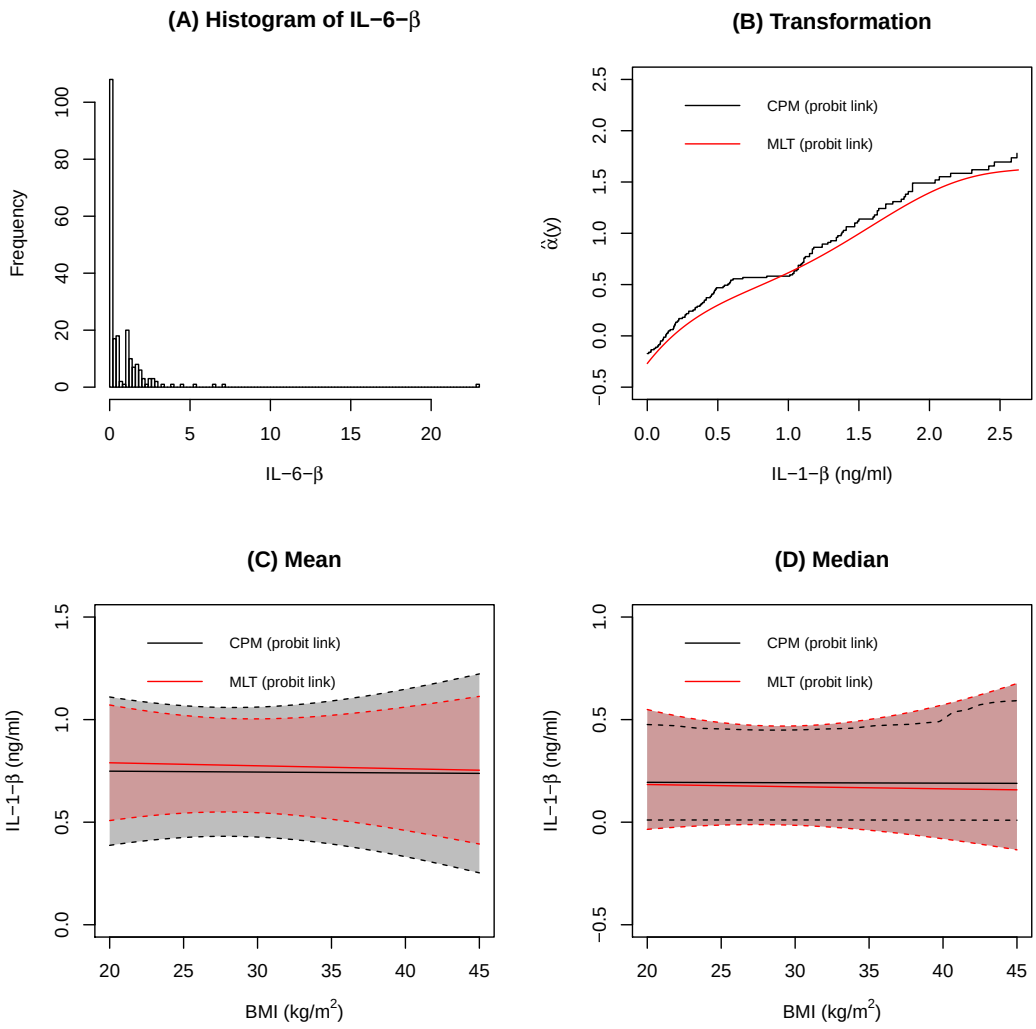

Figure 2.9: Results for IL-1-β. A: The distribution of IL-1-β. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

they assume that after some transformation, the association between response and predictors can be characterized linearly with errors following a known distribution. The main difference between the two methods lies in the estimation of the transformation. CPMs are semiparametric transformation models; each distinct observed response is treated as a category and an ordinal regression model is fit which essentially models the transformation (or equivalently the intercept when written as a cumulative probability model) with a step function. With MLTs, the transformation is parametrically modeled using flexible basis functions. MLT also allows for easy set-up of more complex models featuring covariate-dependent effects using the low-dimensional parameterization of ϑ (Hothorn and Zeileis, 2017; Hothorn, 2019).

We ran extensive simulations to compare the two methods under different settings. The two methods had similar results in most cases and both methods handled complex transformations quite well. We had expected to see more gains in efficiency using MLT and more benefits in terms of robustness using CPMs; if this was the case, only minor differences were seen. MLTs were slightly more efficient. With larger sample sizes, the bias for MLTs occasionally slightly increased; this is presumably because MLTs are slightly misspecified with small orders (e.g. $M = 10$) and we kept the order constant irrespective of the sample size in our simulations. We ran another simulation using $M = 15$ with sample size of 1000 under the primary setting. The bias of conditional medians are -0.018 for $X = 0$ and -0.017 for $X = 1$, which are much smaller than the bias using $M = 10$. As illustrated with the biomarker data, CPMs are invariant to any monotonic transformation of the outcome, which can be considered an advantage. The CPM and MLT approaches handle censoring differently, with CPMs assigning values below a detection limit the lowest rank value, whereas MLTs assume that they follow a distribution informed by data above the detection limit. Resulting conditional expectations, therefore, are slightly different with MLTs treating censored values as something less than the detection limit whereas CPMs compute the expectation as the value after transforming the data back to the original scale (i.e., expectations will use the numeric value assigned to values below the detection limit). For computation time, MLT is significantly faster for large sample sizes with large numbers of distinct response values.

It should be emphasized that CPMs are semiparametric linear transformation models (SLTMs). SLTMs have been advocated for use with time-to-event outcomes and its parametric counterpart `mlt()` was employed to estimate Cox models with time-varying effects in Hothorn (2020). Some attempts have been made to use these models with continuous data (De Neve et al., 2019; Zeng and Lin, 2006), but computation has been a limiting factor. By recognizing that ordinal "cumulative link models" are a special case of SLTMs and that algorithms applying ordinal models can be sped up using a few simple tricks implemented in the function `orm` of the R package **rms**, SLTMs can now be efficiently estimated as CPMs. It should be noted that most measurements in biomedical research are discrete to within the resolution of the measurement method. Results from semiparametric models treating the responses as discrete can in a sense be considered more

Table 2.1: Simulation results for β estimation of transformation $H(y) = y$

Sample Size	Method	Bias	MSE	Coverage (%)
n=50	CPM	0.0465	0.1077	0.9400
	MLT	0.0457	0.1068	0.9412
	Linear Regression	0.0003	0.0824	0.9399
n=100	CPM	0.0192	0.0491	0.9448
	MLT	0.0183	0.0487	0.9456
	Linear Regression	-0.0039	0.0403	0.9463
n=500	CPM	0.0045	0.0090	0.9532
	MLT	0.0043	0.0090	0.9537
	Linear Regression	0.0002	0.0080	0.9526
n=1000	CPM	0.0024	0.0046	0.9492
	MLT	0.0022	0.0046	0.9498
	Linear Regression	0.0001	0.0040	0.9518

Table 2.2: Average computation time (in seconds) for CPM, MLT($M = 5$), and MLT($M = 10$) for the primary simulation setting using different sample sizes and based on 100 replications

Sample Size	CPM	MLT($M = 5$)	MLT($M = 10$)
50	0.0349	0.1326	0.1729
100	0.0261	0.1360	0.1844
500	0.2909	0.2318	0.3121
1000	0.8703	0.3995	0.4416
10000	63.7773	2.8190	4.0533

accurate than continuous methods that approximate discrete responses using a smooth probability density function.

Although not the focus of this manuscript, diagnostics and goodness-of-fit can be assessed for both methods using probability-scale residuals and/or the probability integral transformation (Cox and Snell, 1968; Li and Shepherd, 2012; Shepherd et al., 2016). Since CPMs and MLTs both model the CDF, under proper specification with continuous responses, probability-scale residuals will be approximately uniformly distributed. Probability-scale residuals can also be used in residual-by-predictor plots and partial regression plots to investigate whether covariates are correctly included in the CPM or MLT models. Link functions can be selected based on an approach by Genter and Farewell (Genter and Farewell, 1985). Details and examples are in Liu et al. (2017).

Future studies might consider even more flexible versions of transformation models. For example, it may be worthwhile to develop CPMs that permit different relationships and different distributions for different covariate levels. Extensions of both approaches to handle correlated or longitudinal data, using a similar approach to Manuguerra and Heller,would also be beneficial (Manuguerra and Heller, 2010).

2.7 Supplementary Material

Table 2.3: Simulation results for the primary setting

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04642	0.10774	0.9399
		MLT($M=5$)	0.04800	0.11051	0.9380
		MLT($M=10$)	0.04814	0.10882	0.9398
	n=100	CPM	0.01915	0.04906	0.9448
		MLT($M=5$)	0.02083	0.05062	0.9420
		MLT($M=10$)	0.02032	0.04961	0.9433
	n=500	CPM	0.00452	0.00899	0.9532
		MLT($M=5$)	0.00712	0.00937	0.9498
		MLT($M=10$)	0.00629	0.00915	0.9524
	n=1000	CPM	0.00244	0.00459	0.9492
		MLT($M=5$)	0.00502	0.00478	0.9456
		MLT($M=10$)	0.00427	0.00467	0.9478
$F(5\|X=0)$	n=50	CPM	0.00274	0.00851	0.9277
		MLT($M=5$)	-0.00221	0.00787	0.9493
		MLT($M=10$)	0.00992	0.00785	0.9600
	n=100	CPM	0.00129	0.00422	0.9410
		MLT($M=5$)	-0.00154	0.00400	0.9430
		MLT($M=10$)	0.00823	0.00393	0.9596
	n=500	CPM	-0.00093	0.00084	0.9452
		MLT($M=5$)	-0.00134	0.00082	0.8935
		MLT($M=10$)	0.00521	0.00079	0.9575
	n=1000	CPM	-0.00050	0.00041	0.9531
		MLT($M=5$)	-0.00100	0.00041	0.8308
		MLT($M=10$)	0.00510	0.00040	0.9505
$F(5\|X=1)$	n=50	CPM	-0.00159	0.00533	0.9166
		MLT($M=5$)	-0.00635	0.00496	0.9464
		MLT($M=10$)	0.00252	0.00502	0.9605
	n=100	CPM	0.00002	0.00266	0.9329
		MLT($M=5$)	-0.00280	0.00251	0.9450
		MLT($M=10$)	0.00453	0.00251	0.9586
	n=500	CPM	-0.00115	0.00051	0.9488
		MLT($M=5$)	-0.00223	0.00050	0.9390
		MLT($M=10$)	0.00288	0.00048	0.9521
	n=1000	CPM	-0.00065	0.00026	0.9454
		MLT($M=5$)	-0.00178	0.00026	0.9300
		MLT($M=10$)	0.00300	0.00025	0.9358
$E(Y\|X=0)$	n=50	CPM	-0.00438	0.39292	0.9349
		MLT($M=5$)	-0.02293	0.39143	0.9223
		MLT($M=10$)	-0.01963	0.39142	0.9330
	n=100	CPM	-0.00138	0.19771	0.9407
		MLT($M=5$)	-0.01571	0.19880	0.9320
		MLT($M=10$)	-0.01226	0.19779	0.9379
	n=500	CPM	0.00083	0.03881	0.9490
		MLT($M=5$)	-0.01148	0.03949	0.9415
		MLT($M=10$)	-0.00808	0.03912	0.9456
	n=1000	CPM	-0.00230	0.01965	0.9444
		MLT($M=5$)	-0.01428	0.02008	0.9373

Table 2.3: Simulation results for the primary setting *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT($M = 10$)	-0.01097	0.01988	0.9417
$E(Y\|X = 1)$	n=50	CPM	-0.01814	0.82857	0.9275
		MLT($M = 5$)	-0.07790	0.79589	0.9051
		MLT($M = 10$)	-0.07050	0.79704	0.9419
	n=100	CPM	-0.02173	0.41321	0.9379
		MLT($M = 5$)	-0.06392	0.40291	0.9189
		MLT($M = 10$)	-0.06286	0.40137	0.9471
	n=500	CPM	0.00237	0.07998	0.9499
		MLT($M = 5$)	-0.02227	0.07857	0.9420
		MLT($M = 10$)	-0.02523	0.07775	0.9517
	n=1000	CPM	-0.00177	0.04069	0.9484
		MLT($M = 5$)	-0.02446	0.04048	0.9443
		MLT($M = 10$)	-0.02749	0.04008	0.9501
$F^{-1}(0.1\|X = 0)$	n=50	CPM	0.20035	0.24566	0.8743
		MLT($M = 5$)	0.04432	0.18880	0.9752
		MLT($M = 10$)	0.18626	0.19364	0.9755
	n=100	CPM	0.10556	0.11364	0.9560
		MLT($M = 5$)	0.02878	0.09116	0.9691
		MLT($M = 10$)	0.15908	0.10185	0.9730
	n=500	CPM	0.02232	0.02161	0.9482
		MLT($M = 5$)	0.01338	0.01782	0.8539
		MLT($M = 10$)	0.13383	0.03273	0.9210
	n=1000	CPM	0.01065	0.01097	0.9467
		MLT($M = 5$)	0.00990	0.00882	0.6541
		MLT($M = 10$)	0.13064	0.02451	0.8324
$F^{-1}(0.1\|X = 1)$	n=50	CPM	0.22521	0.56862	0.9450
		MLT($M = 5$)	0.31949	0.48065	0.9541
		MLT($M = 10$)	0.15119	0.42908	0.9591
	n=100	CPM	0.10398	0.25717	0.9483
		MLT($M = 5$)	0.24913	0.24292	0.9589
		MLT($M = 10$)	0.06296	0.19640	0.9502
	n=500	CPM	0.02582	0.04755	0.9497
		MLT($M = 5$)	0.21077	0.07880	0.9669
		MLT($M = 10$)	0.00438	0.03683	0.8879
	n=1000	CPM	0.01335	0.02417	0.9486
		MLT($M = 5$)	0.20332	0.05912	0.9612
		MLT($M = 10$)	-0.00774	0.01913	0.7867
$F^{-1}(0.5\|X = 0)$	n=50	CPM	0.10563	0.46642	0.9479
		MLT($M = 5$)	0.13547	0.33453	0.9483
		MLT($M = 10$)	-0.02376	0.39242	0.9616
	n=100	CPM	0.04961	0.23222	0.9480
		MLT($M = 5$)	0.11572	0.17180	0.9395
		MLT($M = 10$)	-0.04373	0.20316	0.9588
	n=500	CPM	0.01172	0.04694	0.9485
		MLT($M = 5$)	0.10096	0.04239	0.8899
		MLT($M = 10$)	-0.05468	0.04496	0.9557
	n=1000	CPM	0.00457	0.02290	0.9503
		MLT($M = 5$)	0.09721	0.02563	0.8345

Table 2.3: Simulation results for the primary setting *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT($M = 10$)	-0.05971	0.02459	0.9533
$F^{-1}(0.5\|X = 1)$	n=50	CPM	0.20048	1.11656	0.9452
		MLT($M = 5$)	-0.12001	0.92052	0.9633
		MLT($M = 10$)	0.03007	0.95229	0.9604
	n=100	CPM	0.08966	0.54157	0.9511
		MLT($M = 5$)	-0.18618	0.48877	0.9574
		MLT($M = 10$)	-0.00326	0.47608	0.9573
	n=500	CPM	0.02691	0.10205	0.9511
		MLT($M = 5$)	-0.21455	0.13279	0.9381
		MLT($M = 10$)	-0.01198	0.09274	0.95509
	n=1000	CPM	0.01260	0.05013	0.9506
		MLT($M = 5$)	-0.22718	0.09557	0.9008
		MLT($M = 10$)	-0.02327	0.04707	0.9538
$F^{-1}(0.8\|X = 0)$	n=50	CPM	0.06688	1.05641	0.9514
		MLT($M = 5$)	-0.22179	0.93013	0.9534
		MLT($M = 10$)	-0.08754	0.97143	0.9579
	n=100	CPM	0.03294	0.52462	0.9488
		MLT($M = 5$)	-0.23006	0.50334	0.9519
		MLT($M = 10$)	-0.04745	0.48855	0.9557
	n=500	CPM	0.00787	0.10244	0.9507
		MLT($M = 5$)	-0.24076	0.14714	0.9542
		MLT($M = 10$)	-0.02601	0.09556	0.9419
	n=1000	CPM	0.00213	0.05195	0.9468
		MLT($M = 5$)	-0.24694	0.10570	0.9515
		MLT($M = 10$)	-0.03046	0.04867	0.9186
$F^{-1}(0.8\|X = 1)$	n=50	CPM	0.31165	2.47150	0.9609
		MLT($M = 5$)	0.08370	1.70604	0.9587
		MLT($M = 10$)	0.03272	1.89087	0.9628
	n=100	CPM	0.14366	1.16903	0.9554
		MLT($M = 5$)	0.18623	0.90953	0.9593
		MLT($M = 10$)	0.03091	0.99242	0.9651
	n=500	CPM	0.04875	0.22490	0.9517
		MLT($M = 5$)	0.30724	0.27373	0.8984
		MLT($M = 10$)	0.02032	0.20574	0.9606
	n=1000	CPM	0.01889	0.11247	0.9512
		MLT($M = 5$)	0.30983	0.18657	0.8150
		MLT($M = 10$)	-0.00706	0.10641	0.9509
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-186.896	
		MLT($M = 5$)		-170.377	
		MLT($M = 10$)		-171.092	
	n=100	CPM		-455.055	
		MLT($M = 5$)		-420.825	
		MLT($M = 10$)		-420.056	
	n=500	CPM		-3055.871	
		MLT($M = 5$)		-2852.102	
		MLT($M = 10$)		-2852.031	
	n=1000	CPM		-6785.049	

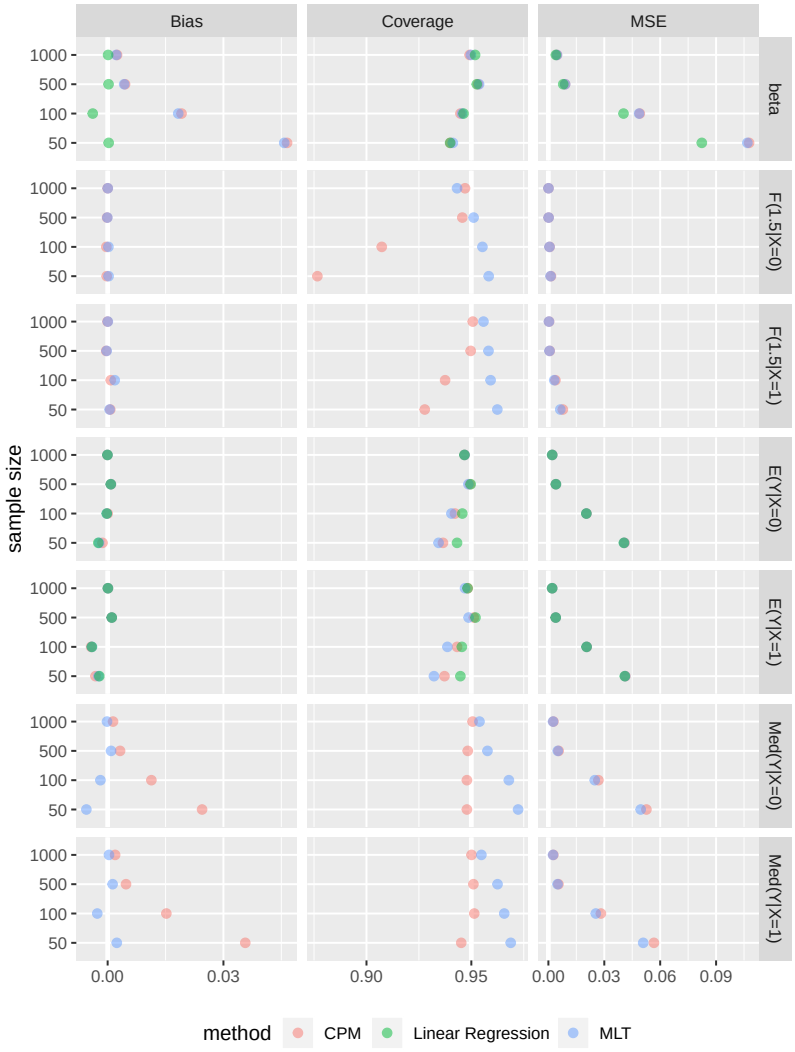

Figure 2.10: Simulation results for $H(y) = y$

Table 2.3: Simulation results for the primary setting *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT($M = 5$)		-6392.950	
		MLT($M = 10$)		-6376.261	

Table 2.4: Simulation results for $H(y) = y$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		CPM	0.0465	0.1077	0.9400
	n=50	MLT	0.0457	0.1068	0.9412

β

Table 2.4: Simulation results for $H(y) = y$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		Linear Regression	0.0003	0.0824	0.9399
	n=100	CPM	0.0192	0.0491	0.9448
		MLT	0.0183	0.0487	0.9456
		Linear Regression	-0.0039	0.0403	0.9463
	n=500	CPM	0.0045	0.0090	0.9532
		MLT	0.0043	0.0090	0.9537
		Linear Regression	0.0002	0.0080	0.9526
	n=1000	CPM	0.0024	0.0046	0.9492
		MLT	0.0022	0.0046	0.9498
		Linear Regression	0.0001	0.0040	0.9518
$F(5\|X=0)$	n=50	CPM	-0.00024	0.00138	0.8766
		MLT	0.00025	0.00118	0.9583
	n=100	CPM	-0.00034	0.00068	0.9073
		MLT	0.00020	0.00059	0.9553
	n=500	CPM	-0.00014	0.00013	0.9457
		MLT	-0.00008	0.00012	0.9511
	n=1000	CPM	0.00001	0.00007	0.9470
		MLT	0.0004	0.00006	0.9432
$F(5\|X=1)$	n=50	CPM	0.00068	0.00776	0.9278
		MLT	0.00051	0.00638	0.9624
	n=100	CPM	0.00082	0.00379	0.9375
		MLT	0.00184	0.00314	0.9592
	n=500	CPM	-0.00038	0.00072	0.9497
		MLT	-0.00025	0.00061	0.9582
	n=1000	CPM	-0.00003	0.00037	0.9507
		MLT	0.00005	0.00031	0.9558
$E(Y\|X=0)$	n=50	CPM	-0.00133	0.04083	0.9365
		MLT	-0.00221	0.04065	0.9344
		Linear Regression	-0.00242	0.04052	0.9432
	n=100	CPM	0.00001	0.02045	0.9422
		MLT	-0.00021	0.02038	0.9405
		Linear Regression	-0.00022	0.02037	0.9457
	n=500	CPM	0.00084	0.00403	0.9493
		MLT	0.00083	0.00403	0.9486
		Linear Regression	0.00082	0.00403	0.9497
	n=1000	CPM	-0.00006	0.00205	0.9468
		MLT	-0.00005	0.00204	0.9468
		Linear Regression	-0.00005	0.00204	0.9467
$E(Y\|X=1)$	n=50	CPM	-0.00314	0.04129	0.9372
		MLT	-0.00237	0.04117	0.9322
		Linear Regression	-0.00215	0.04102	0.9448
	n=100	CPM	-0.00428	0.02059	0.9432
		MLT	-0.00409	0.02053	0.9386
		Linear Regression	-0.00408	0.02050	0.9455
	n=500	CPM	0.00105	0.00394	0.9512
		MLT	0.00106	0.00394	0.9487
		Linear Regression	0.00107	0.00393	0.9520
	n=1000	CPM	0.00008	0.00202	0.9483

Table 2.4: Simulation results for $H(y) = y$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	0.00007	0.00202	0.9470
		Linear Regression	0.00007	0.00202	0.9482
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.10268	0.09696	0.8819
		MLT	0.04438	0.07366	0.9420
	n=100	CPM	0.05572	0.04877	0.9566
		MLT	0.02392	0.03649	0.9299
	n=500	CPM	0.01184	0.01010	0.9481
		MLT	0.00476	0.00721	0.8975
	n=1000	CPM	0.00541	0.00519	0.9468
		MLT	0.00176	0.00367	0.8756
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.06788	0.07780	0.94519
		MLT	0.03050	0.07139	0.94849
	n=100	CPM	0.03064	0.03743	0.9482
		MLT	0.01176	0.03518	0.9317
	n=500	CPM	0.00753	0.00720	0.9494
		MLT	0.00436	0.00678	0.8200
	n=1000	CPM	0.00351	0.00369	0.9488
		MLT	0.00233	0.00348	0.6570
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.02447	0.05262	0.94789
		MLT	-0.00551	0.04954	0.97230
	n=100	CPM	0.01133	0.02685	0.9479
		MLT	-0.00187	0.02490	0.9679
	n=500	CPM	0.00322	0.00551	0.9483
		MLT	0.00089	0.00506	0.9577
	n=1000	CPM	0.00140	0.00270	0.9506
		MLT	-0.00016	0.00250	0.9539
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.03566	0.05664	0.9452
		MLT	0.00237	0.05088	0.9688
	n=100	CPM	0.01521	0.02819	0.9514
		MLT	-0.00269	0.02545	0.9657
	n=500	CPM	0.00476	0.00540	0.9510
		MLT	0.00129	0.00494	0.9625
	n=1000	CPM	0.00194	0.00266	0.9501
		MLT	0.00035	0.00249	0.9548
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.00518	0.06243	0.9514
		MLT	-0.02582	0.05878	0.9412
	n=100	CPM	0.00274	0.03120	0.9486
		MLT	-0.01281	0.02936	0.9152
	n=500	CPM	0.00121	0.00610	0.9508
		MLT	-0.00118	0.00575	0.7699
	n=1000	CPM	0.00036	0.00310	0.9463
		MLT	-0.00071	0.00292	0.6111
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.04268	0.07296	0.9616
		MLT	-0.02176	0.05473	0.9347
	n=100	CPM	0.01890	0.03555	0.9555
		MLT	-0.01163	0.02827	0.8977
	n=500	CPM	0.00639	0.00696	0.9522
		MLT	0.00109	0.00577	0.5443

Table 2.4: Simulation results for $H(y) = y$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=1000	CPM	0.00161	0.00350	0.9512
		MLT	0.00019	0.00299	0.2415
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-186.896	
		MLT		-171.731	
	n=100	CPM		-455.055	
		MLT		-417.689	
	n=100	CPM		-3055.871	
		MLT		-2848.799	
	n=100	CPM		-6785.049	
		MLT		-6367.617	

Table 2.5: Simulation results for $H(y) = \exp(y)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04642	0.10774	0.9399
		MLT	0.04863	0.11454	0.9353
		Linear Regression	0.00021	0.08238	0.9399
	n=100	CPM	0.01915	0.04906	0.9448
		MLT	0.02065	0.05218	0.9386
		Linear Regression	-0.00386	0.04032	0.9463
	n=500	CPM	0.00452	0.00899	0.9532
		MLT	0.00500	0.00963	0.9454
		Linear Regression	0.00024	0.00796	0.9526
	n=1000	CPM	0.00244	0.00459	0.9492
		MLT	-0.00221	0.00491	0.9421
		Linear Regression	0.00012	0.00402	0.9518
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-196.409	
		MLT		-179.517	
	n=100	CPM		-450.390	
		MLT		-430.551	
	n=500	CPM		-3062.587	
		MLT		-2877.188	
	n=1000	CPM		-6789.501	
		MLT		-6434.118	

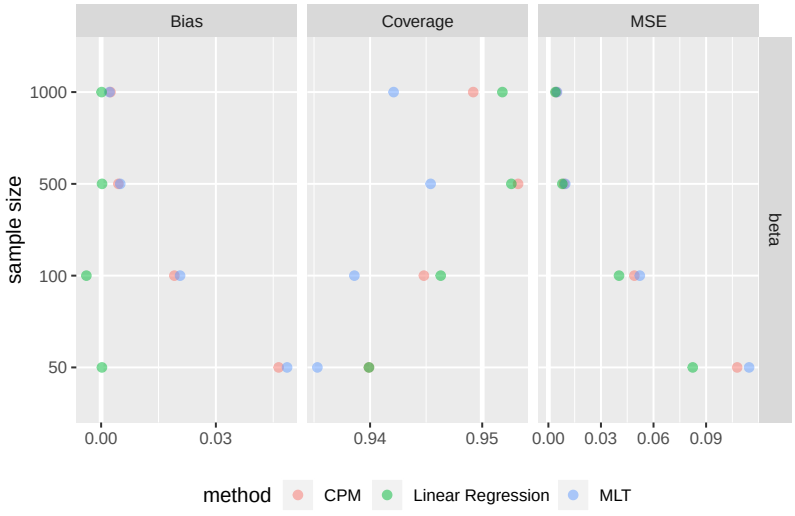

Figure 2.11: Simulation results for $H(y) = \exp(y)$

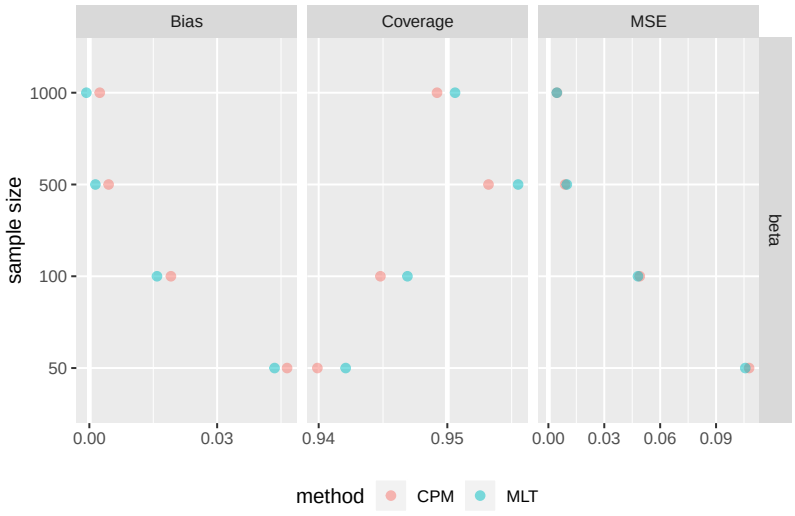

Figure 2.12: Simulation results for $H(y) = \text{Inv-}logistic(\Phi(y))$

Table 2.6: Simulation results for $H(y) = \text{Inv-}logistic(\Phi(y))$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04642	0.10774	0.9399
		MLT	0.04348	0.10569	0.9421
	n=100	CPM	0.01915	0.04906	0.9448
		MLT	0.01588	0.04811	0.9469
	n=500	CPM	0.00452	0.00899	0.9532
		MLT	0.00142	0.00990	0.9555
	n=1000	CPM	0.00244	0.00459	0.9492
		MLT	-0.00070	0.00454	0.9506
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-186.8960	
		MLT		-172.0136	
	n=100	CPM		-455.0545	
		MLT		-417.9264	
	n=500	CPM		-3055.871	
		MLT		-2847.904	
	n=1000	CPM		-6785.049	
		MLT		-6372.934	

Table 2.7: Simulation results for including covariate $Z \sim N(0,1)$, which is independent of X

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.06728	0.11989	0.9321
		MLT	0.05754	0.11591	0.9365
	n=100	CPM	0.02936	0.05169	0.9442
		MLT	0.02193	0.05116	0.9447
	n=500	CPM	0.00689	0.00911	0.9532
		MLT	0.00151	0.00917	0.9510
	n=1000	CPM	0.00363	0.00462	0.9491
		MLT	-0.00111	0.00464	0.9463
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-178.060	
		MLT		-159.152	
	n=100	CPM		-409.099	
		MLT		-371.022	
	n=500	CPM		-2927.662	
		MLT		-2735.722	
	n=1000	CPM		-6490.808	
		MLT		-6104.455	

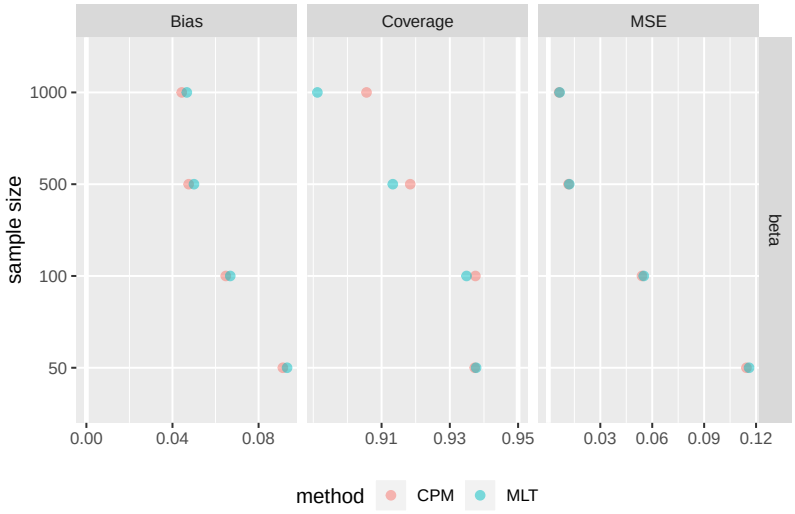

Figure 2.13: Simulation results for including multiple covariates $Z_1, Z_2, Z_3 \sim N(\mathbf{0}, \mathbf{I}), Z_4 \sim N(X, 1), Z_5 \sim N(Z_1 + X, 1), Z_6 \sim N(Z_3 - Z_4, 1)$

Table 2.8: Simulation results for including covariate $Z \sim N(X, 1)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.06902	0.15194	0.9357
		MLT	0.08516	0.15621	0.9321
	n=100	CPM	0.02871	0.06368	0.9409
		MLT	0.04616	0.06668	0.9379
	n=500	CPM	0.00571	0.01128	0.9508
		MLT	0.02372	0.01229	0.9402
	n=1000	CPM	0.00268	0.00574	0.9495
		MLT	0.02088	0.00643	0.9368
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-171.942	
		MLT		-153.132	
	n=100	CPM		-406.006	
		MLT		-371.460	
	n=500	CPM		-2844.186	
		MLT		-2651.443	
	n=1000	CPM		-6375.575	
		MLT		-5983.449	

Table 2.9: Simulation results for including multiple covariates $Z_1, Z_2, Z_3 \sim N(\mathbf{0}, \mathbf{I}), Z_4 \sim N(X, 1), Z_5 \sim N(Z_1 + X, 1), Z_6 \sim N(Z_3 - Z_4, 1)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=50	CPM	0.09129	0.11445	0.9373
		MLT	0.09324	0.11598	0.9377
β					

Table 2.9: Simulation results for including multiple covariates $Z_1, Z_2, Z_3 \sim N(\mathbf{0}, \mathbf{I}), Z_4 \sim N(X, 1), Z_5 \sim N(Z_1 + X, 1), Z_6 \sim N(Z_3 - Z_4, 1)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=100	CPM	0.06477	0.05410	0.9375
		MLT	0.06684	0.05513	0.9349
	n=500	CPM	0.04753	0.01166	0.9184
		MLT	0.04999	0.01206	0.9133
	n=1000	CPM	0.04430	0.00619	0.9056
		MLT	0.04666	0.00649	0.8912
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-186.919	
		MLT		-170.907	
	n=100	CPM		-449.502	
		MLT		-413.414	
	n=500	CPM		-3055.861	
		MLT		-2856.913	
	n=1000	CPM		-6783.180	
		MLT		-6376.370	

Table 2.10: Simulation results for using the correct logit link function for $\varepsilon \sim Logistic(0, \frac{3}{\pi^2})$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04694	0.10712	1.0000
		MLT	0.05690	0.10856	0.9999
	n=100	CPM	0.02094	0.04841	0.9998
		MLT	0.03347	0.04987	1.0000
	n=500	CPM	0.00423	0.00922	0.9997
		MLT	0.01474	0.00954	0.9997
	n=1000	CPM	0.00209	0.00458	0.9996
		MLT	0.01253	0.00481	0.9996
$F(5\|X=0)$	n=50	CPM	-0.00001	0.00896	0.9312
		MLT	0.00944	0.00827	0.9659
	n=100	CPM	0.00069	0.00441	0.9403
		MLT	0.01211	0.00425	0.9602
	n=500	CPM	-0.00005	0.00087	0.9507
		MLT	0.01146	0.00094	0.9504
	n=1000	CPM	-0.00006	0.00044	0.9467
		MLT	0.01126	0.00053	0.9306
$F(5\|X=1)$	n=50	CPM	-0.00046	0.00475	0.9126
		MLT	0.00170	0.00437	0.9636
	n=100	CPM	0.00056	0.00240	0.9306
		MLT	0.00405	0.00227	0.9531
	n=500	CPM	-0.00008	0.00047	0.9452
		MLT	0.00439	0.00046	0.9284
	n=1000	CPM	-0.00006	0.00023	0.9482
		MLT	0.00435	0.00024	0.8942

Table 2.10: Simulation results for using the correct logit link function for $\varepsilon \sim Logistic(0, \frac{3}{\pi^2})$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
$E(Y\|X=0)$	n=50	CPM	0.00901	0.37500	0.9373
		MLT	-0.05876	0.36693	0.9402
	n=100	CPM	0.00359	0.18297	0.9414
		MLT	-0.06609	0.18365	0.9328
	n=500	CPM	0.00626	0.03743	0.9475
		MLT	-0.05312	0.03960	0.9345
	n=1000	CPM	0.00602	0.01843	0.9473
		MLT	-0.05231	0.02077	0.9266
$E(Y\|X=1)$	n=50	CPM	0.00399	0.86735	0.9228
		MLT	-0.14146	0.75629	0.9050
	n=100	CPM	-0.01119	0.42757	0.9339
		MLT	-0.14248	0.37896	0.9292
	n=500	CPM	0.00414	0.08364	0.9467
		MLT	-0.11026	0.08245	0.9385
	n=1000	CPM	0.00383	0.04241	0.9466
		MLT	-0.10741	0.04733	0.9266
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.19870	0.28072	0.9009
		MLT	0.15886	0.21725	0.9647
	n=100	CPM	0.10327	0.13827	0.9692
		MLT	0.10953	0.10873	0.9662
	n=500	CPM	0.02211	0.02724	0.9492
		MLT	0.09981	0.02952	0.9711
	n=1000	CPM	0.01136	0.01370	0.9483
		MLT	0.09743	0.01918	0.9512
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.21597	0.59737	0.9539
		MLT	0.14212	0.46472	0.9634
	n=100	CPM	0.10216	0.27626	0.9484
		MLT	0.09233	0.21145	0.9544
	n=500	CPM	0.01758	0.05236	0.9500
		MLT	0.02975	0.04010	0.9311
	n=1000	CPM	0.00869	0.02579	0.9490
		MLT	0.02391	0.02018	0.8978
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.11908	0.40509	0.9534
		MLT	-0.02145	0.34071	0.9648
	n=100	CPM	0.05550	0.19342	0.9523
		MLT	-0.03446	0.16103	0.9636
	n=500	CPM	0.01147	0.03881	0.9487
		MLT	-0.05361	0.03521	0.9628
	n=1000	CPM	0.00662	0.01901	0.9490
		MLT	-0.05364	0.01896	0.9578
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.17551	0.93895	0.9518
		MLT	0.02756	0.79539	0.9648
	n=100	CPM	0.06786	0.43809	0.9518
		MLT	-0.00326	0.38323	0.9576
	n=500	CPM	0.00953	0.08394	0.9504
		MLT	0.00346	0.07501	0.9551
	n=1000	CPM	0.00176	0.04187	0.9511
		MLT	0.00166	0.03785	0.9530

Table 2.10: Simulation results for using the correct logit link function for $\varepsilon \sim Logistic(0, \frac{3}{\pi^2})$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.09922	0.94830	0.9514
		MLT	-0.06426	0.86592	0.9627
	n=100	CPM	0.04129	0.45649	0.9543
		MLT	-0.07018	0.43292	0.9605
	n=500	CPM	0.01556	0.09357	0.9492
		MLT	-0.02872	0.08995	0.9458
	n=1000	CPM	0.00889	0.04634	0.9466
		MLT	-0.03072	0.04510	0.9265
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.36291	2.53392	0.9574
		MLT	0.08217	1.90985	0.9625
	n=100	CPM	0.14527	1.11688	0.9568
		MLT	0.03524	0.91880	0.9649
	n=500	CPM	0.02104	0.20628	0.9539
		MLT	0.03492	0.18837	0.9568
	n=1000	CPM	0.00504	0.10591	0.9469
		MLT	0.02743	0.09906	0.9380
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-189.186	
		MLT		-173.154	
	n=100	CPM		-455.262	
		MLT		-417.729	
	n=500	CPM		-3053.475	
		MLT		-2852.847	
	n=1000	CPM		-6802.809	
		MLT		-6392.715	

Table 2.11: Simulation results for using the correct cloglog link function for $\varepsilon \sim Gompertz$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.06104	0.12733	0.9437
		MLT	0.06460	0.12534	0.9473
	n=100	CPM	0.03525	0.05469	0.9459
		MLT	0.03718	0.05403	0.9479
	n=500	CPM	0.00680	0.00994	0.9504
		MLT	0.00795	0.00988	0.9499
	n=1000	CPM	0.00368	0.00496	0.9476
		MLT	0.00458	0.00492	0.9476
$F(5\|X=0)$	n=50	CPM	0.00356	0.00698	0.9229
		MLT	0.00714	0.00631	0.9548
	n=100	CPM	0.00185	0.00345	0.9382
		MLT	0.00411	0.00310	0.9625
	n=500	CPM	0.00060	0.00070	0.9461
		MLT	0.00182	0.00063	0.9577
	n=1000	CPM	0.00077	0.00034	0.9495

Table 2.11: Simulation results for using the correct cloglog link function for $\varepsilon \sim$ *Gompertz (continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	0.00181	0.00031	0.9523
$F(5\|X=1)$	n=50	CPM	-0.00506	0.00729	0.9214
		MLT	-0.00404	0.00671	0.9601
	n=100	CPM	-0.00432	0.00352	0.9352
		MLT	-0.00343	0.00324	0.9587
	n=500	CPM	-0.00078	0.00068	0.9442
		MLT	-0.00019	0.00063	0.9552
	n=1000	CPM	-0.00017	0.00034	0.9491
		MLT	0.00037	0.00031	0.9528
$E(Y\|X=0)$	n=50	CPM	-0.01111	0.25880	0.9277
		MLT	-0.03566	0.26265	0.9220
	n=100	CPM	-0.00865	0.12678	0.9428
		MLT	-0.03064	0.12896	0.9344
	n=500	CPM	-0.00671	0.02554	0.9480
		MLT	-0.02723	0.02658	0.9387
	n=1000	CPM	-0.00618	0.01268	0.9440
		MLT	-0.02643	0.01351	0.9318
$E(Y\|X=1)$	n=50	CPM	0.00648	0.62496	0.9269
		MLT	0.00869	0.62310	0.9367
	n=100	CPM	0.01666	0.30593	0.9408
		MLT	0.01659	0.30524	0.9423
	n=500	CPM	-0.00251	0.05932	0.9521
		MLT	-0.00502	0.05911	0.9504
	n=1000	CPM	-0.00370	0.02983	0.9490
		MLT	-0.00653	0.02973	0.9471
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.18037	0.18072	0.9252
		MLT	0.22361	0.18316	0.9705
	n=100	CPM	0.09172	0.08122	0.9712
		MLT	0.20968	0.10825	0.9684
	n=500	CPM	0.01696	0.01469	0.9486
		MLT	0.19575	0.05098	0.9365
	n=1000	CPM	0.00799	0.00708	0.9508
		MLT	0.19605	0.04478	0.8661
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.24974	0.46851	0.9495
		MLT	0.33606	0.38757	0.9685
	n=100	CPM	0.13925	0.20991	0.9493
		MLT	0.26547	0.19314	0.9631
	n=500	CPM	0.03264	0.03712	0.9495
		MLT	0.19311	0.05936	0.9612
	n=1000	CPM	0.02289	0.01818	0.9488
		MLT	0.18532	0.04499	0.9511
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.08669	0.40232	0.9506
		MLT	-0.04023	0.33103	0.9640
	n=100	CPM	0.04574	0.20179	0.9510
		MLT	-0.05537	0.17342	0.9671
	n=500	CPM	0.00738	0.04079	0.9474
		MLT	-0.06849	0.04094	0.9659
	n=1000	CPM	0.00190	0.01989	0.9490

Table 2.11: Simulation results for using the correct cloglog link function for $\varepsilon \sim$ *Gompertz (continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.07029	0.02270	0.9594
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.22305	1.09728	0.9485
		MLT	0.08381	0.93684	0.9624
	n=100	CPM	0.13725	0.53565	0.9489
		MLT	0.05743	0.46698	0.9581
	n=500	CPM	0.02949	0.10164	0.9506
		MLT	-0.00474	0.09177	0.9544
	n=1000	CPM	0.01880	0.05075	0.9500
		MLT	-0.01397	0.04618	0.9537
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.05483	0.65794	0.9482
		MLT	-0.08004	0.60558	0.9561
	n=100	CPM	0.02389	0.32129	0.9531
		MLT	-0.04480	0.29221	0.9612
	n=500	CPM	0.00143	0.06535	0.9504
		MLT	-0.02236	0.05848	0.9398
	n=1000	CPM	-0.00034	0.03224	0.9488
		MLT	-0.02328	0.02913	0.9189
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.20617	1.59207	0.9623
		MLT	0.00263	1.20935	0.9559
	n=100	CPM	0.12090	0.75820	0.9543
		MLT	0.05491	0.62127	0.9620
	n=500	CPM	0.03115	0.14890	0.9510
		MLT	0.03568	0.13167	0.9574
	n=1000	CPM	0.01950	0.07565	0.9512
		MLT	0.02443	0.06840	0.9413
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-191.801	
		MLT		-178.390	
	n=100	CPM		-446.725	
		MLT		-416.666	
	n=500	CPM		-3054.302	
		MLT		-2876.370	
	n=1000	CPM		-6805.474	
		MLT		-6473.184	

Table 2.12: Simulation results for $X \sim$ *Uniform*$(0,1)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04475	0.30304	0.9378
		MLT	0.04285	0.29959	0.9390
	n=100	CPM	0.02040	0.13539	0.9473
		MLT	0.01791	0.13496	0.9486
	n=500	CPM	0.00541	0.02444	0.9551
		MLT	0.00430	0.02446	0.9554

Table 2.12: Simulation results for $X \sim Uniform(0,1)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=1000	CPM	0.00194	0.01274	0.9473
		MLT	0.00111	0.01278	0.9463
$F(5\|X=0)$	n=50	CPM	0.00092	0.01456	0.9111
		MLT	0.00650	0.01385	0.9492
	n=100	CPM	0.00051	0.00720	0.9347
		MLT	0.00564	0.00685	0.9560
	n=500	CPM	-0.00061	0.00139	0.9487
		MLT	0.00343	0.00134	0.9547
	n=1000	CPM	-0.00059	0.00071	0.9476
		MLT	0.00312	0.00069	0.9411
$F(5\|X=1)$	n=50	CPM	0.00374	0.00900	0.8983
		MLT	0.00771	0.00875	0.9534
	n=100	CPM	0.00190	0.00428	0.9278
		MLT	0.00620	0.00424	0.9535
	n=500	CPM	-0.00064	0.00081	0.9495
		MLT	0.00266	0.00079	0.9571
	n=1000	CPM	-0.00028	0.00043	0.9443
		MLT	0.00272	0.00042	0.9428
$E(Y\|X=0)$	n=50	CPM	0.02610	0.80669	0.9279
		MLT	0.00520	0.80453	0.9186
	n=100	CPM	0.01026	0.38698	0.9409
		MLT	-0.00222	0.38872	0.9345
	n=500	CPM	0.00196	0.07398	0.9499
		MLT	-0.00678	0.07468	0.9470
	n=1000	CPM	-0.00015	0.03831	0.9477
		MLT	-0.00855	0.03875	0.9446
$E(Y\|X=1)$	n=50	CPM	-0.01906	1.64840	0.9223
		MLT	-0.07521	1.54617	0.9206
	n=100	CPM	-0.01865	0.80486	0.9379
		MLT	-0.06616	0.76725	0.9361
	n=500	CPM	0.00324	0.15570	0.9517
		MLT	-0.03166	0.14910	0.9519
	n=1000	CPM	-0.00342	0.08022	0.9485
		MLT	-0.03664	0.07791	0.9507
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.12730	0.86396	0.9421
		MLT	0.02457	0.75358	0.9487
	n=100	CPM	0.05858	0.40703	0.9517
		MLT	-0.01387	0.36659	0.9572
	n=500	CPM	0.01294	0.07815	0.9507
		MLT	-0.03928	0.07561	0.9575
	n=1000	CPM	0.00551	0.03984	0.9511
		MLT	-0.04265	0.04005	0.9571
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.23511	2.04371	0.9427
		MLT	0.06687	1.80262	0.9514
	n=100	CPM	0.10632	0.93012	0.9474
		MLT	0.00751	0.84960	0.9509
	n=500	CPM	0.03332	0.16934	0.9528
		MLT	-0.00820	0.16173	0.9555

Table 2.12: Simulation results for $X \sim Uniform(0,1)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=1000	CPM	0.01483	0.08708	0.9499
		MLT	-0.02251	0.08409	0.9503
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-191.684	
		MLT		-175.702	
	n=100	CPM		-458.573	
		MLT		-422.812	
	n=500	CPM		-3089.845	
		MLT		-2886.741	
	n=1000	CPM		-6858.044	
		MLT		-6445.935	

Table 2.13: Simulation results for $X \sim N(0,1)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.05438	0.04242	0.9346
		MLT	0.03896	0.03795	0.9464
	n=100	CPM	0.02823	0.01821	0.9420
		MLT	0.01465	0.01682	0.9498
	n=500	CPM	0.00623	0.00318	0.9446
		MLT	-0.00398	0.00305	0.9471
	n=1000	CPM	0.00284	0.00152	0.9509
		MLT	-0.00694	0.00153	0.9475
$F(5\|X=0)$	n=50	CPM	0.00170	0.00643	0.9372
		MLT	0.01273	0.00528	0.9714
	n=100	CPM	0.00073	0.00320	0.9418
		MLT	0.00993	0.00260	0.9704
	n=500	CPM	-0.00053	0.00062	0.9497
		MLT	0.00713	0.00053	0.9637
	n=1000	CPM	-0.00045	0.00031	0.9499
		MLT	0.00712	0.00028	0.9648
$F(5\|X=1)$	n=50	CPM	-0.00422	0.00593	0.9069
		MLT	0.00659	0.00526	0.9654
	n=100	CPM	-0.00281	0.00293	0.9256
		MLT	0.00719	0.00261	0.9601
	n=500	CPM	-0.00129	0.00059	0.9389
		MLT	0.00736	0.00056	0.9408
	n=1000	CPM	-0.00072	0.00028	0.9464
		MLT	0.00783	0.00030	0.9159
$E(Y\|X=0)$	n=50	CPM	-0.01140	0.22060	0.9356
		MLT	-0.03473	0.21979	0.9400
	n=100	CPM	-0.00528	0.10869	0.9443
		MLT	-0.02495	0.10871	0.9436
	n=500	CPM	-0.00077	0.02154	0.9465

Table 2.13: Simulation results for $X \sim N(0,1)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.01904	0.02185	0.9426
	n=1000	CPM	-0.00307	0.01051	0.9529
		MLT	-0.02127	0.01092	0.9427
$E(Y\|X=1)$	n=50	CPM	0.03037	0.92104	0.9277
		MLT	-0.10792	0.84272	0.9475
	n=100	CPM	0.02781	0.43177	0.9427
		MLT	-0.08295	0.40661	0.9543
	n=500	CPM	0.00896	0.08647	0.9467
		MLT	-0.06748	0.08490	0.9455
	n=1000	CPM	0.00312	0.04216	0.9504
		MLT	-0.06911	0.04400	0.9426
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.12767	0.36809	0.9505
		MLT	-0.09699	0.29458	0.9725
	n=100	CPM	0.06259	0.17759	0.9521
		MLT	-0.10707	0.15958	0.9701
	n=500	CPM	0.01362	0.03439	0.9462
		MLT	-0.10659	0.04192	0.9739
	n=1000	CPM	0.00369	0.01671	0.9533
		MLT	-0.10892	0.02666	0.9771
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.31356	1.42888	0.9473
		MLT	-0.00825	1.07147	0.9685
	n=100	CPM	0.17049	0.64786	0.9513
		MLT	-0.05418	0.51316	0.9657
	n=500	CPM	0.03959	0.12480	0.9443
		MLT	-0.08739	0.11372	0.9596
	n=1000	CPM	0.02007	0.06016	0.9487
		MLT	-0.09519	0.06124	0.9556
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-184.301	
		MLT		-169.515	
	n=100	CPM		-420.930	
		MLT		-387.442	
	n=500	CPM		-2935.062	
		MLT		-2737.588	
	n=1000	CPM		-6532.848	
		MLT		-6151.616	

Table 2.14: Simulation results for $X \sim Binomial(1, p = 0.3)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04705	0.13165	0.9427
		MLT	0.04915	0.13183	0.9454
	n=100	CPM	0.02280	0.05775	0.9458
		MLT	0.02476	0.05775	0.9471

Table 2.14: Simulation results for $X \sim Binomial(1, p = 0.3)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=500	CPM	0.00486	0.01053	0.9496
		MLT	0.00678	0.01061	0.9500
	n=1000	CPM	0.00313	0.00541	0.9491
		MLT	0.00506	0.00545	0.9483
$F(5\|X=0)$	n=50	CPM	0.00066	0.00651	0.9326
		MLT	0.00627	0.00579	0.9621
	n=100	CPM	0.00075	0.00322	0.9412
		MLT	0.00557	0.00288	0.9586
	n=500	CPM	-0.00100	0.00063	0.9507
		MLT	0.00238	0.00056	0.9531
	n=1000	CPM	-0.00051	0.00032	0.9492
		MLT	0.00227	0.00029	0.9362
$F(5\|X=1)$	n=50	CPM	0.00017	0.00797	0.8993
		MLT	0.00289	0.00766	0.9554
	n=100	CPM	0.00010	0.00390	0.9260
		MLT	0.00271	0.00372	0.9537
	n=500	CPM	-0.00102	0.00075	0.9460
		MLT	0.00088	0.00071	0.9556
	n=1000	CPM	-0.00070	0.00039	0.9452
		MLT	0.00078	0.00037	0.9441
$E(Y\|X=0)$	n=50	CPM	-0.00508	0.28263	0.9357
		MLT	-0.02122	0.28146	0.9350
	n=100	CPM	-0.00730	0.14215	0.9439
		MLT	-0.01922	0.14183	0.9402
	n=500	CPM	0.00085	0.02771	0.9525
		MLT	-0.00717	0.02778	0.9495
	n=1000	CPM	-0.00282	0.01404	0.9495
		MLT	-0.01035	0.01415	0.9475
$E(Y\|X=1)$	n=50	CPM	-0.02920	1.38866	0.9197
		MLT	-0.08721	1.31227	0.9305
	n=100	CPM	-0.01793	0.68027	0.9350
		MLT	-0.06541	0.64494	0.9396
	n=500	CPM	0.00235	0.13261	0.9491
		MLT	-0.03070	0.12714	0.9538
	n=1000	CPM	0.00004	0.06822	0.9465
		MLT	-0.03066	0.06604	0.9512
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.10374	0.36472	0.9524
		MLT	-0.02964	0.31474	0.9675
	n=100	CPM	0.04564	0.17627	0.9491
		MLT	-0.04900	0.16112	0.9656
	n=500	CPM	0.01087	0.03514	0.9499
		MLT	-0.04602	0.03464	0.9673
	n=1000	CPM	0.00432	0.01727	0.9513
		MLT	-0.04767	0.01855	0.9600
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.25922	1.82671	0.9453
		MLT	0.06664	1.57711	0.9559
	n=100	CPM	0.11981	0.83894	0.9467
		MLT	0.00599	0.74058	0.9564

Table 2.14: Simulation results for $X \sim Binomial(1, p = 0.3)$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=500	CPM	0.03466	0.15817	0.9505
		MLT	-0.00916	0.14763	0.9574
	n=1000	CPM	0.01771	0.07873	0.9486
		MLT	-0.01840	0.07557	0.9561
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-188.865	
		MLT		-172.420	
	n=100	CPM		-452.227	
		MLT		-417.286	
	n=500	CPM		-3068.055	
		MLT		-2864.786	
	n=1000	CPM		-6810.171	
		MLT		-6400.164	

Table 2.15: Simulation results for $\beta = 0.5$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.02284	0.09650	0.9384
		MLT	0.02276	0.09564	0.9409
	n=100	CPM	0.00787	0.04431	0.9463
		MLT	0.00690	0.04427	0.9462
	n=500	CPM	0.00251	0.00827	0.9530
		MLT	0.00210	0.00829	0.9534
	n=1000	CPM	0.00133	0.00419	0.9509
		MLT	0.00101	0.00421	0.9508
$F(5\|X=0)$	n=50	CPM	0.00115	0.00827	0.9307
		MLT	0.00534	0.00756	0.9587
	n=100	CPM	0.00033	0.00410	0.9390
		MLT	0.00346	0.00377	0.9554
	n=500	CPM	-0.00083	0.00081	0.9465
		MLT	0.00093	0.00075	0.9476
	n=1000	CPM	-0.00057	0.00040	0.9485
		MLT	0.00085	0.00038	0.9432
$F(5\|X=1)$	n=50	CPM	-0.00226	0.00809	0.9273
		MLT	0.00140	0.00748	0.9584
	n=100	CPM	-0.00042	0.00404	0.9380
		MLT	0.00281	0.00370	0.9603
	n=500	CPM	-0.00171	0.00077	0.9496
		MLT	0.00014	0.00071	0.9590
	n=1000	CPM	-0.00126	0.00039	0.9485
		MLT	0.00024	0.00036	0.9571
$E(Y\|X=0)$	n=50	CPM	-0.00492	0.39055	0.9352
		MLT	-0.01941	0.38509	0.9336
	n=100	CPM	-0.00120	0.19607	0.9424

Table 2.15: Simulation results for $\beta = 0.5$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.00979	0.19508	0.9395
	n=500	CPM	0.00081	0.03862	0.9480
		MLT	-0.00459	0.03864	0.9458
	n=1000	CPM	-0.00222	0.01957	0.9455
		MLT	-0.00724	0.01962	0.9434
$E(Y\|X=1)$	n=50	CPM	-0.01557	0.58278	0.9293
		MLT	-0.03880	0.56750	0.9358
	n=100	CPM	-0.01874	0.29075	0.9385
		MLT	-0.03684	0.28586	0.9420
	n=500	CPM	0.00190	0.05634	0.9495
		MLT	-0.00933	0.05541	0.9497
	n=1000	CPM	-0.00170	0.02855	0.9486
		MLT	-0.01202	0.02831	0.9490
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.17264	0.21298	0.9119
		MLT	0.18992	0.17960	0.9526
	n=100	CPM	0.08944	0.10012	0.9513
		MLT	0.16004	0.09478	0.9231
	n=500	CPM	0.01819	0.01915	0.9465
		MLT	0.13211	0.03092	0.8368
	n=1000	CPM	0.00965	0.00960	0.9483
		MLT	0.12785	0.02306	0.7668
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.17586	0.32541	0.9466
		MLT	0.17742	0.25404	0.9584
	n=100	CPM	0.08356	0.15205	0.9484
		MLT	0.11744	0.11929	0.9427
	n=500	CPM	0.02076	0.02842	0.9522
		MLT	0.07558	0.02551	0.7755
	n=1000	CPM	0.00956	0.01445	0.9449
		MLT	0.06649	0.01456	0.5432
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.10025	0.46589	0.9466
		MLT	-0.02077	0.41633	0.9604
	n=100	CPM	0.04586	0.22615	0.9495
		MLT	-0.03269	0.21341	0.9589
	n=500	CPM	0.01015	0.04523	0.9477
		MLT	-0.03065	0.04486	0.9545
	n=1000	CPM	0.00367	0.02228	0.9499
		MLT	-0.03132	0.02294	0.9465
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.14988	0.74505	0.9459
		MLT	0.02800	0.66157	0.9595
	n=100	CPM	0.06338	0.36033	0.9477
		MLT	-0.00047	0.32757	0.9600
	n=500	CPM	0.02242	0.06762	0.9525
		MLT	-0.00266	0.06159	0.9587
	n=1000	CPM	0.01071	0.03343	0.9524
		MLT	-0.01044	0.03105	0.9562
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.09239	1.06059	0.9515
		MLT	-0.07025	0.94719	0.9623
	n=100	CPM	0.04724	0.52680	0.9501

Table 2.15: Simulation results for $\beta = 0.5$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.04584	0.47602	0.9567
	n=500	CPM	0.01153	0.09914	0.9508
		MLT	-0.02804	0.09490	0.9511
	n=1000	CPM	0.00438	0.05124	0.9470
		MLT	-0.03026	0.04971	0.9446
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.21087	1.68003	0.9562
		MLT	-0.00870	1.40610	0.9656
	n=100	CPM	0.09878	0.80246	0.9522
		MLT	-0.02060	0.72827	0.9652
	n=500	CPM	0.03519	0.15369	0.9500
		MLT	-0.02163	0.14466	0.9665
	n=1000	CPM	0.01440	0.07515	0.9529
		MLT	-0.03716	0.07263	0.9673
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-192.250	
		MLT		-176.341	
	n=100	CPM		-460.401	
		MLT		-424.847	
	n=500	CPM		-3094.909	
		MLT		-2891.795	
	n=1000	CPM		-6871.982	
		MLT		-6460.752	

Table 2.16: Simulation results for $\beta = 0$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.00051	0.09294	0.9378
		MLT	0.00039	0.09163	0.9398
	n=100	CPM	-0.00418	0.04321	0.9463
		MLT	-0.00396	0.04267	0.9477
	n=500	CPM	0.00019	0.00807	0.9521
		MLT	0.00028	0.00804	0.9524
	n=1000	CPM	0.00019	0.00405	0.9508
		MLT	0.00019	0.00405	0.9508
$F(5\|X=0)$	n=50	CPM	-0.00021	0.00815	0.9311
		MLT	0.00195	0.00761	0.9612
	n=100	CPM	-0.00037	0.00405	0.9390
		MLT	0.00097	0.00380	0.9600
	n=500	CPM	-0.00112	0.00080	0.9481
		MLT	0.00034	0.00076	0.9520
	n=1000	CPM	-0.00066	0.00040	0.9483
		MLT	0.00075	0.00038	0.9424
$F(5\|X=1)$	n=50	CPM	-0.00070	0.00839	0.9269
		MLT	0.00150	0.00777	0.9586

Table 2.16: Simulation results for $\beta = 0$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=100	CPM	0.00095	0.00411	0.9386
		MLT	0.00220	0.00385	0.9631
	n=500	CPM	-0.00149	0.00079	0.9481
		MLT	-0.00006	0.00074	0.9525
	n=1000	CPM	-0.00103	0.00039	0.9503
		MLT	0.00038	0.00037	0.9412
$E(Y\|X=0)$	n=50	CPM	-0.01021	0.38727	0.9321
		MLT	-0.02377	0.38216	0.9325
	n=100	CPM	-0.00252	0.19634	0.9388
		MLT	-0.01225	0.19412	0.9402
	n=500	CPM	0.00039	0.03855	0.9477
		MLT	-0.00541	0.03841	0.9451
	n=1000	CPM	-0.00253	0.01956	0.9450
		MLT	-0.00774	0.01953	0.9443
$E(Y\|X=1)$	n=50	CPM	-0.00818	0.39449	0.9320
		MLT	-0.02181	0.38806	0.9329
	n=100	CPM	-0.01445	0.19697	0.9390
		MLT	-0.02345	0.19509	0.9395
	n=500	CPM	0.00188	0.03800	0.9504
		MLT	-0.00368	0.03777	0.9474
	n=1000	CPM	-0.00120	0.01922	0.9491
		MLT	-0.00640	0.01918	0.9473
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.14353	0.18357	0.9429
		MLT	0.17897	0.16066	0.9550
	n=100	CPM	0.07530	0.08754	0.9483
		MLT	0.14613	0.08275	0.9099
	n=500	CPM	0.01548	0.01700	0.9452
		MLT	0.11519	0.02523	0.4498
	n=1000	CPM	0.00715	0.00845	0.9459
		MLT	0.10978	0.01797	0.1263
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.14853	0.18764	0.9425
		MLT	0.18166	0.16313	0.9503
	n=100	CPM	0.07010	0.08777	0.9484
		MLT	0.14173	0.08235	0.9082
	n=500	CPM	0.01723	0.01617	0.9514
		MLT	0.11642	0.02499	0.4541
	n=1000	CPM	0.00843	0.00850	0.9435
		MLT	0.11084	0.01825	0.1300
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.10486	0.46978	0.9476
		MLT	0.00160	0.43062	0.9610
	n=100	CPM	0.05100	0.22548	0.9465
		MLT	-0.00835	0.21472	0.9575
	n=500	CPM	0.01174	0.04522	0.9478
		MLT	-0.00869	0.04292	0.9503
	n=1000	CPM	0.00486	0.02222	0.9477
		MLT	-0.00960	0.02147	0.9510
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.10718	0.47960	0.9412
		MLT	0.00554	0.44035	0.9587

Table 2.16: Simulation results for $\beta = 0$ *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=100	CPM	0.04277	0.23224	0.9501
		MLT	-0.01656	0.21920	0.9579
	n=500	CPM	0.01472	0.04384	0.9485
		MLT	-0.00540	0.04191	0.9541
	n=1000	CPM	0.00798	0.02196	0.9502
		MLT	-0.00641	0.02110	0.9530
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.12923	1.10047	0.9543
		MLT	-0.05079	0.96618	0.9666
	n=100	CPM	0.06987	0.54428	0.9484
		MLT	-0.04266	0.50030	0.9613
	n=500	CPM	0.01783	0.10171	0.9512
		MLT	-0.02934	0.09465	0.9534
	n=1000	CPM	0.00376	0.05132	0.9515
		MLT	-0.02821	0.04861	0.9396
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.13690	1.12511	0.9530
		MLT	-0.04482	0.98421	0.9682
	n=100	CPM	0.05761	0.53678	0.9489
		MLT	-0.05491	0.49127	0.9612
	n=500	CPM	0.02539	0.10174	0.9528
		MLT	-0.02072	0.09390	0.9534
	n=1000	CPM	0.01155	0.05045	0.9480
		MLT	-0.02073	0.04707	0.9403
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-196.430	
		MLT		-180.425	
	n=100	CPM		-461.252	
		MLT		-424.277	
	n=500	CPM		-3107.826	
		MLT		-2904.348	
	n=1000	CPM		-6908.245	
		MLT		-6496.046	

Table 2.17: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = logit

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.77590	0.92688	0.7388
		MLT	0.80261	0.97609	0.7280
	n=100	CPM	0.72863	0.67946	0.5302
		MLT	0.75260	0.71829	0.5098
	n=500	CPM	0.70479	0.52417	0.0080
		MLT	0.72678	0.55637	0.0066
	n=1000	CPM	0.70144	0.50599	0
		MLT	0.72287	0.5368	0
$F(5\|X=0)$	n=50	CPM	0.00898	0.00909	0.9233

Table 2.17: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = logit *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	0.02035	0.00882	0.9612
	n=100	CPM	0.00848	0.00457	0.9338
		MLT	0.01942	0.00456	0.958
	n=500	CPM	0.00713	0.00095	0.9369
		MLT	0.01718	0.0012	0.9598
	n=1000	CPM	0.00766	0.00050	0.9298
		MLT	0.01714	0.00070	0.9626
$F(5\|X=1)$	n=50	CPM	-0.00542	0.00515	0.9101
		MLT	-0.00269	0.00486	0.9642
	n=100	CPM	-0.00449	0.00256	0.9268
		MLT	-0.00123	0.00240	0.9620
	n=500	CPM	-0.00619	0.00053	0.9341
		MLT	-0.00298	0.00046	0.9618
	n=1000	CPM	-0.00582	0.00029	0.9252
		MLT	-0.00286	0.00024	0.9574
$E(Y\|X=0)$	n=50	CPM	0.05750	0.40320	0.9381
		MLT	-0.01167	0.40194	0.9374
	n=100	CPM	0.06342	0.20642	0.9475
		MLT	0.0257	0.20336	0.9434
	n=500	CPM	0.06761	0.04439	0.9420
		MLT	0.01392	0.04040	0.9500
	n=1000	CPM	0.06498	0.02428	0.9289
		MLT	0.01260	0.02041	0.9486
$E(Y\|X=1)$	n=50	CPM	-0.06659	0.83691	0.9229
		MLT	-0.14046	0.81146	0.8982
	n=100	CPM	-0.07391	0.42048	0.9287
		MLT	-0.13820	0.41514	0.9280
	n=500	CPM	-0.05200	0.08317	0.9370
		MLT	-0.10688	0.08793	0.9316
	n=1000	CPM	-0.05601	0.04402	0.9334
		MLT	-0.11009	0.05105	0.9166
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.24638	0.27051	0.8866
		MLT	0.20722	0.22113	0.9734
	n=100	CPM	0.15492	0.12948	0.9687
		MLT	0.18939	0.12302	0.9768
	n=500	CPM	0.07665	0.02760	0.9308
		MLT	0.1723	0.04683	0.9560
	n=1000	CPM	0.06554	0.01541	0.9137
		MLT	0.17092	0.03778	0.9314
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.16833	0.57776	0.9553
		MLT	0.12658	0.43810	0.9599
	n=100	CPM	0.03875	0.26215	0.9540
		MLT	0.03123	0.19651	0.9491
	n=500	CPM	-0.04642	0.05197	0.9491
		MLT	-0.03290	0.03766	0.8597
	n=1000	CPM	-0.05900	0.02867	0.9337
		MLT	0.01350	0.01795	0.7775
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.08316	0.46017	0.9521

Table 2.17: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = logit *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.06783	0.39497	0.9621
	n=100	CPM	0.02036	0.22882	0.9483
		MLT	-0.09296	0.20723	0.9555
	n=500	CPM	-0.02391	0.04698	0.9467
		MLT	-0.09524	0.04603	0.9437
	n=1000	CPM	-0.03179	0.02365	0.9417
		MLT	-0.11717	0.03399	0.9169
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.23651	1.14031	0.9472
		MLT	0.08698	0.97437	0.9633
	n=100	CPM	0.13560	0.55622	0.9501
		MLT	0.06205	0.48468	0.9570
	n=500	CPM	0.08089	0.10870	0.9433
		MLT	0.06139	0.09642	0.9438
	n=1000	CPM	0.06738	0.05495	0.9373
		MLT	0.05190	0.04926	0.9378
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.01151	1.11193	0.9556
		MLT	-0.18056	1.06432	0.9621
	n=100	CPM	-0.02666	0.55764	0.9526
		MLT	-0.13868	0.54058	0.9580
	n=500	CPM	-0.05658	0.11304	0.9509
		MLT	-0.11033	0.11396	0.9570
	n=1000	CPM	-0.06268	0.05955	0.9451
		MLT	-0.11203	0.06343	0.9535
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.16403	2.32095	0.9651
		MLT	-0.16465	1.82775	0.9684
	n=100	CPM	0.00087	1.11914	0.9606
		MLT	-0.14446	0.96128	0.9631
	n=500	CPM	-0.09039	0.22339	0.9483
		MLT	-0.11637	0.20761	0.9556
	n=1000	CPM	-0.11838	0.12235	0.9377
		MLT	-0.13346	0.11813	0.9507
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-187.049	
		MLT		-170.972	
	n=100	CPM		-457.143	
		MLT		-423.316	
	n=500	CPM		-3060.906	
		MLT		-2860.515	
	n=1000	CPM		-6788.465	
		MLT		-6383.818	

Figure 2.14: Simulation results for using the correct logit link function for $\varepsilon \sim Logistic(0, \frac{3}{\pi^2})$

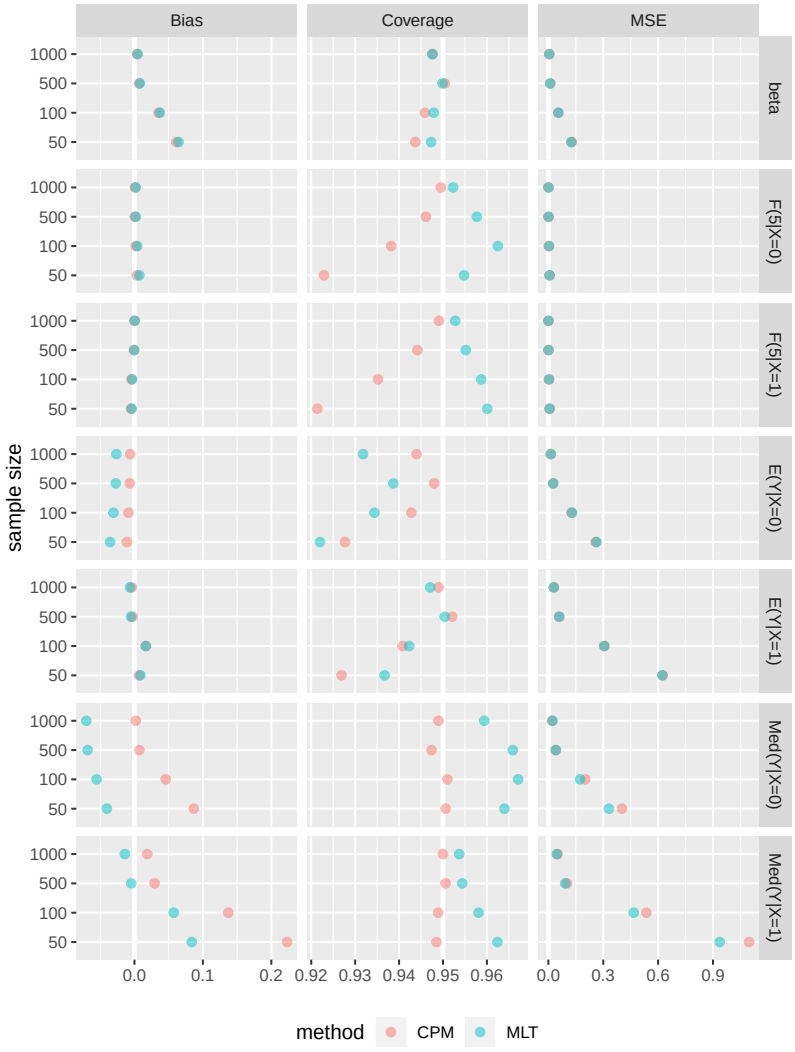

Figure 2.15: Simulation results for using the correct cloglog link function for $\varepsilon \sim \textit{Gompertz}$

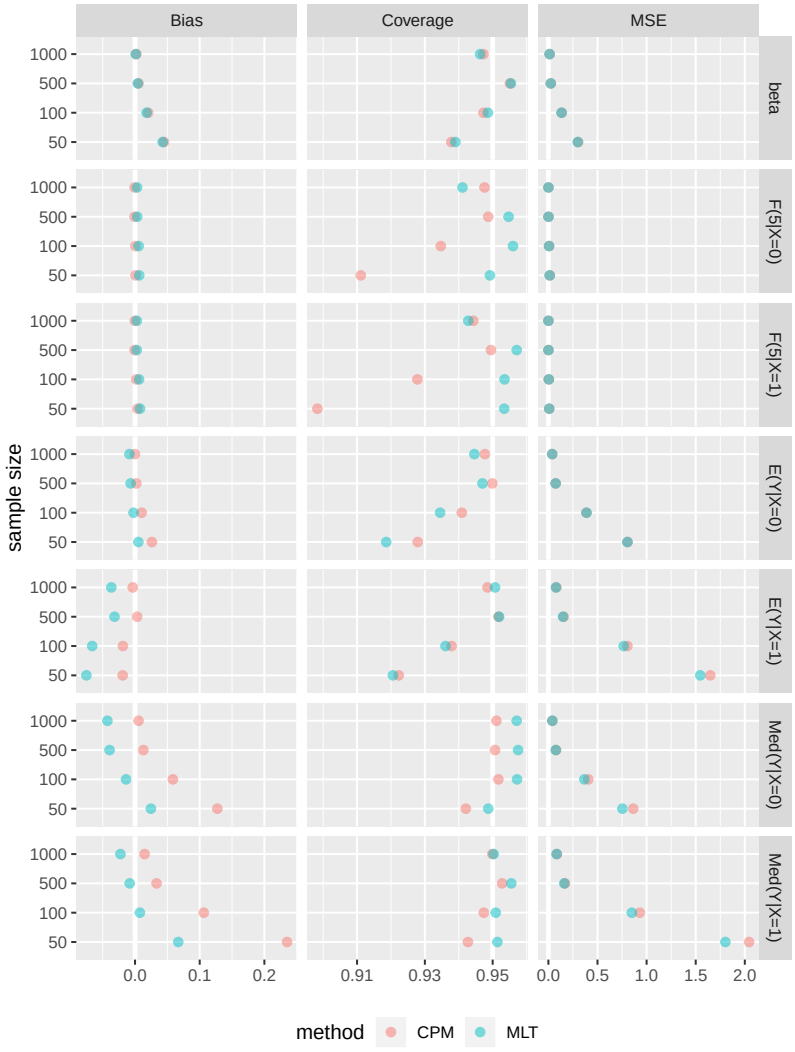

Figure 2.16: Simulation results for $X \sim Uniform(0, 1)$

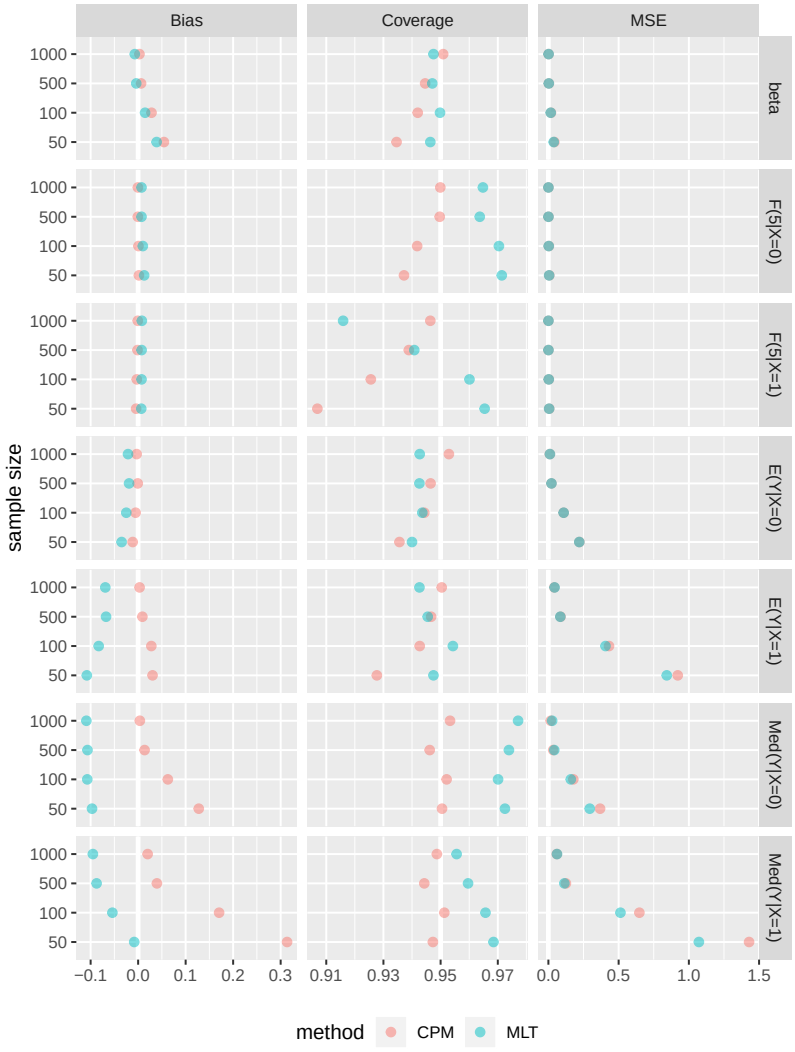

Figure 2.17: Simulation results for $X \sim N(0, 1)$

Figure 2.18: Simulation results for $X \sim Binomial(1, p = 0.3)$

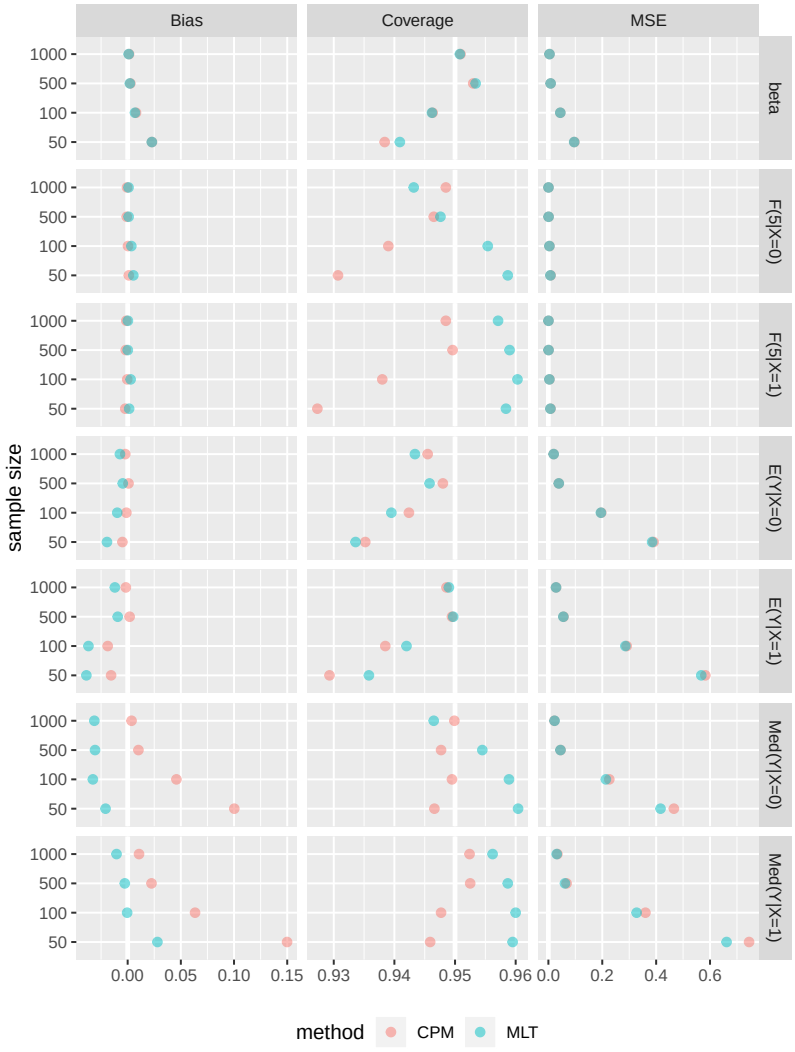

Figure 2.19: Simulation results for $\beta = 0.5$

Figure 2.20: Simulation results for $\beta = 0$

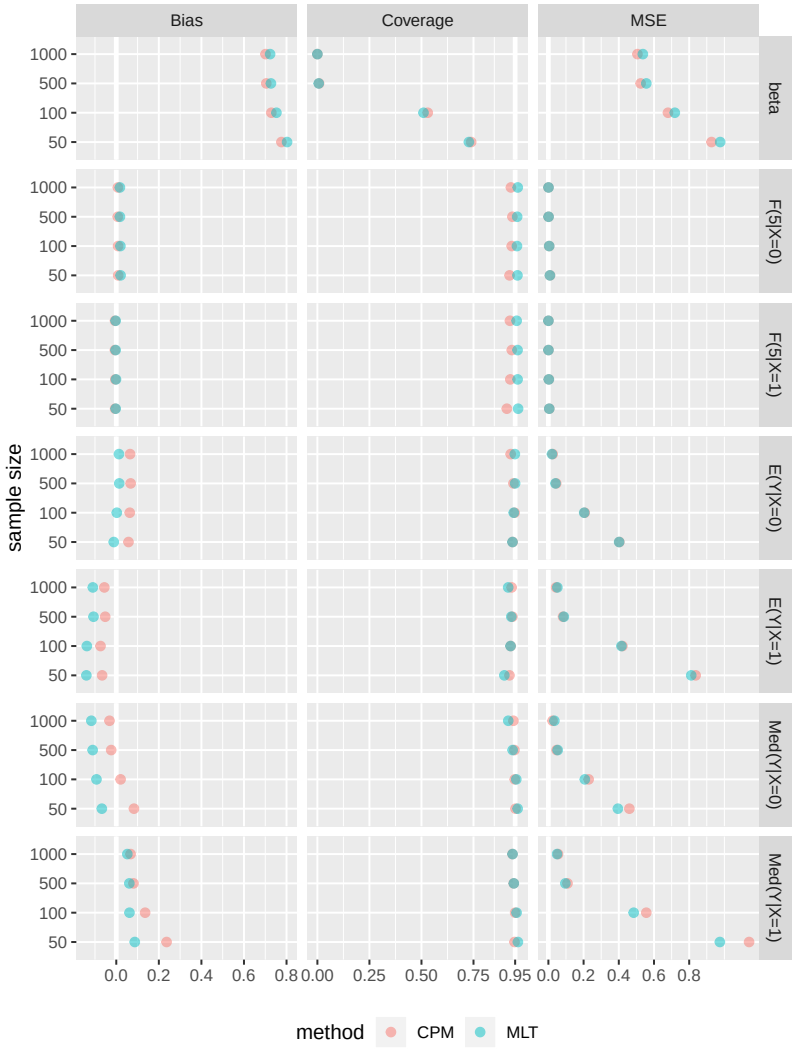

Figure 2.21: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = logit

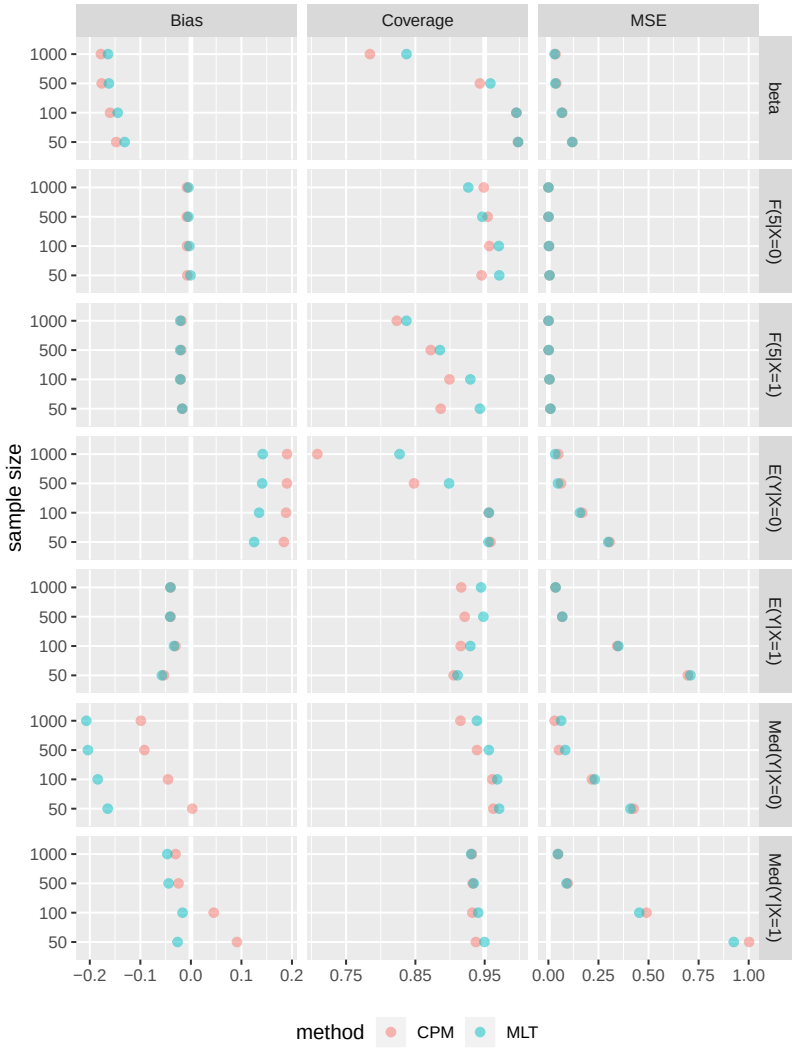

Figure 2.22: Simulation results for misspecification: $\varepsilon \sim$ *Gompertz*, link function = logit

Table 2.18: Simulation results for misspecification: $\varepsilon \sim$ *Gompertz*, link function = logit

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	-0.14776	0.12024	0.9981
		MLT	-0.13106	0.11882	0.9982
	n=100	CPM	-0.15993	0.06967	0.9955
		MLT	-0.14478	0.06661	0.9960
	n=500	CPM	-0.17641	0.03944	0.9431
		MLT	-0.16223	0.03495	0.9582
	n=1000	CPM	-0.17795	0.0359	0.7844
		MLT	-0.16401	0.0313	0.8372
$F(5\|X=0)$	n=50	CPM	-0.00712	0.00600	0.9456
		MLT	-0.00075	0.00544	0.9710
	n=100	CPM	-0.00770	0.00299	0.9566
		MLT	-0.00299	0.00271	0.9704
	n=500	CPM	-0.00812	0.00066	0.9543
		MLT	-0.00475	0.00057	0.9466
	n=1000	CPM	-0.00782	0.00035	0.9487
		MLT	-0.00474	0.00029	0.9263
$F(5\|X=1)$	n=50	CPM	-0.01680	0.01034	0.8866
		MLT	-0.01738	0.00985	0.9431
	n=100	CPM	-0.01989	0.00532	0.8992
		MLT	-0.02125	0.00513	0.9293
	n=500	CPM	-0.01898	0.00134	0.8722
		MLT	-0.02119	0.00138	0.8854
	n=1000	CPM	-0.01868	0.00084	0.8232
		MLT	-0.02107	0.00091	0.8371
$E(Y\|X=0)$	n=50	CPM	0.18398	0.30572	0.9585
		MLT	0.12510	0.29897	0.9555
	n=100	CPM	0.18823	0.16896	0.9557
		MLT	0.13496	0.15764	0.9562
	n=500	CPM	0.19011	0.06298	0.8481
		MLT	0.14102	0.04795	0.8986
	n=1000	CPM	0.19031	0.04959	0.7086
		MLT	0.14215	0.03422	0.8272
$E(Y\|X=1)$	n=50	CPM	-0.05325	0.69629	0.9053
		MLT	-0.05738	0.70941	0.9108
	n=100	CPM	-0.03053	0.34354	0.9155
		MLT	-0.03348	0.34924	0.9293
	n=500	CPM	-0.04014	0.06853	0.9212
		MLT	-0.04129	0.06964	0.9481
	n=1000	CPM	-0.03995	0.03540	0.9162
		MLT	-0.04083	0.03596	0.9448
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.12732	0.15789	0.9225
		MLT	0.14124	0.14994	0.9581
	n=100	CPM	0.03308	0.06982	0.9614
		MLT	0.11766	0.07672	0.9487
	n=500	CPM	-0.04725	0.01558	0.9157
		MLT	0.09770	0.02186	0.9203
	n=1000	CPM	-0.05758	0.00978	0.8788
		MLT	0.09731	0.01557	0.9107

Table 2.18: Simulation results for misspecification: $\varepsilon \sim$ *Gompertz*, link function = logit *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
$F^{-1}(0.1\|X=1)$	n=50	CPM	0.66103	1.09235	0.8809
		MLT	0.66107	0.92133	0.9197
	n=100	CPM	0.55765	0.62829	0.8317
		MLT	0.58788	0.57164	0.8837
	n=500	CPM	0.44938	0.26269	0.5118
		MLT	0.50103	0.29183	0.6847
	n=1000	CPM	0.43999	0.22455	0.2352
		MLT	0.49108	0.26145	0.45980
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.00281	0.42524	0.9624
		MLT	-0.16451	0.40993	0.9709
	n=100	CPM	-0.04526	0.21744	0.9609
		MLT	-0.18431	0.23166	0.9681
	n=500	CPM	-0.09203	0.05276	0.9389
		MLT	-0.20370	0.08403	0.9559
	n=1000	CPM	-0.09889	0.03120	0.9151
		MLT	-0.20677	0.06359	0.9388
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.09116	1.00208	0.9373
		MLT	-0.02631	0.92564	0.9497
	n=100	CPM	0.04523	0.49025	0.9321
		MLT	-0.01648	0.45395	0.9406
	n=500	CPM	-0.02448	0.09638	0.9327
		MLT	-0.04396	0.09079	0.9340
	n=1000	CPM	-0.03022	0.04889	0.9314
		MLT	-0.04658	0.04694	0.9305
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.34849	0.88943	0.9429
		MLT	0.16858	0.75153	0.9578
	n=100	CPM	0.32264	0.48415	0.9363
		MLT	0.20811	0.39554	0.9393
	n=500	CPM	0.30268	0.16950	0.8337
		MLT	0.23338	0.12507	0.7428
	n=1000	CPM	0.30186	0.13007	0.7098
		MLT	0.23475	0.09079	0.5032
$F^{-1}(0.8\|X=1)$	n=50	CPM	-0.22993	1.55672	0.9686
		MLT	-0.53588	1.45754	0.9479
	n=100	CPM	-0.31622	0.82751	0.9469
		MLT	-0.46692	0.81245	0.9224
	n=500	CPM	-0.40681	0.30843	0.8121
		MLT	-0.43961	0.31670	0.7970
	n=1000	CPM	-0.41821	0.24745	0.6539
		MLT	-0.43679	0.25499	0.6746
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-192.958	
		MLT		-179.062	
	n=100	CPM		-451.006	
		MLT		-420.752	
	n=500	CPM		-3064.242	
		MLT		-2883.105	
	n=1000	CPM		-6824.644	

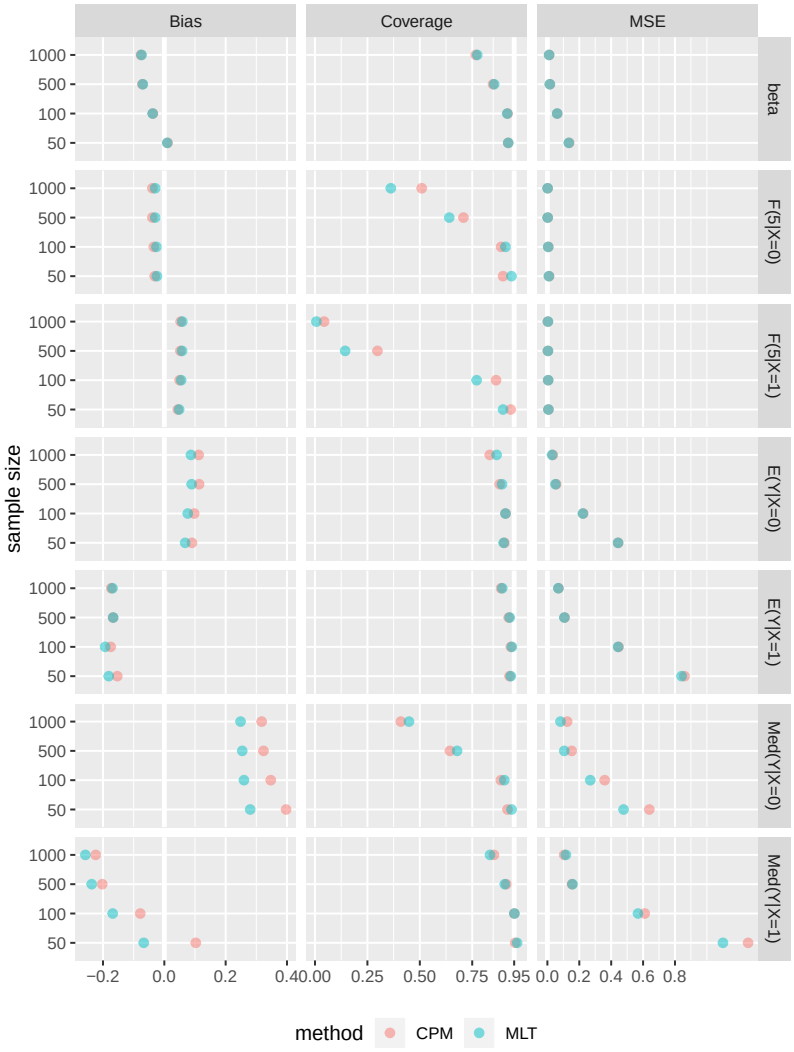

Figure 2.23: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = cloglog

Table 2.18: Simulation results for misspecification: $\varepsilon \sim$ *Gompertz*, link function = logit *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT		-6490.792	

Table 2.19: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = cloglog

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=50	CPM	0.01065	0.13672	0.9213
		MLT	0.00937	0.13420	0.9219
	n=100	CPM	-0.03729	0.06188	0.9191
β					

Table 2.19: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = cloglog *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.03800	0.06089	0.9166
	n=500	CPM	-0.07157	0.01580	0.8493
		MLT	-0.06990	0.01546	0.8549
	n=1000	CPM	-0.07619	0.01114	0.7670
		MLT	-0.07417	0.01083	0.7740
$F(5\|X=0)$	n=50	CPM	-0.03124	0.01061	0.8970
		MLT	-0.02419	0.00960	0.9373
	n=100	CPM	-0.03422	0.00594	0.8877
		MLT	-0.02594	0.00513	0.9087
	n=500	CPM	-0.03902	0.00246	0.7082
		MLT	-0.02971	0.00175	0.6405
	n=1000	CPM	-0.03886	0.00197	0.5092
		MLT	-0.02974	0.00131	0.3618
$F(5\|X=1)$	n=50	CPM	0.04417	0.00644	0.9338
		MLT	0.04839	0.00657	0.8970
	n=100	CPM	0.04985	0.00471	0.8637
		MLT	0.05506	0.00514	0.7711
	n=500	CPM	0.05241	0.00317	0.2981
		MLT	0.05795	0.00376	0.1438
	n=1000	CPM	0.05334	0.00306	0.0435
		MLT	0.05871	0.00365	0.0067
$E(Y\|X=0)$	n=50	CPM	0.09018	0.44355	0.9035
		MLT	0.06796	0.44324	0.9001
	n=100	CPM	0.09747	0.22445	0.9089
		MLT	0.07646	0.22254	0.9086
	n=500	CPM	0.11348	0.05512	0.8811
		MLT	0.08916	0.05072	0.8922
	n=1000	CPM	0.11206	0.03387	0.8337
		MLT	0.08679	0.02914	0.8665
$E(Y\|X=1)$	n=50	CPM	-0.15323	0.85954	0.9283
		MLT	-0.18113	0.84165	0.9342
	n=100	CPM	-0.17479	0.44633	0.9344
		MLT	-0.19302	0.44309	0.9392
	n=500	CPM	-0.16722	0.10775	0.9243
		MLT	-0.16666	0.10637	0.9290
	n=1000	CPM	-0.17296	0.07006	0.8876
		MLT	-0.16891	0.06832	0.8934
$F^{-1}(0.1\|X=0)$	n=50	CPM	0.36060	0.34293	0.8858
		MLT	0.36352	0.31390	0.9906
	n=100	CPM	0.28030	0.18741	0.9584
		MLT	0.35684	0.21606	0.9918
	n=500	CPM	0.21304	0.06779	0.7473
		MLT	0.34896	0.13938	0.9623
	n=1000	CPM	0.20398	0.05289	0.5481
		MLT	0.34843	0.13023	0.9195
$F^{-1}(0.1\|X=1)$	n=50	CPM	-0.48997	0.59171	0.8319
		MLT	-0.43024	0.43626	0.7981
	n=100	CPM	-0.62493	0.55468	0.6362

Table 2.19: Simulation results for misspecification: $\varepsilon \sim N(0,1)$, link function = cloglog *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.52008	0.38478	0.5335
	n=500	CPM	-0.72610	0.55887	0.0260
		MLT	-0.58516	0.36313	0.0033
	n=1000	CPM	-0.73961	0.56236	0.0002
		MLT	-0.59653	0.36624	0.0000
$F^{-1}(0.5\|X=0)$	n=50	CPM	0.39705	0.63930	0.9183
		MLT	0.28004	0.47767	0.9378
	n=100	CPM	0.34717	0.35933	0.8866
		MLT	0.25974	0.26879	0.9032
	n=500	CPM	0.32349	0.15247	0.6437
		MLT	0.25407	0.10572	0.6779
	n=1000	CPM	0.31751	0.12413	0.4094
		MLT	0.24862	0.08228	0.4485
$F^{-1}(0.5\|X=1)$	n=50	CPM	0.10261	1.25842	0.9548
		MLT	-0.06703	1.10094	0.9640
	n=100	CPM	-0.07837	0.61030	0.9504
		MLT	-0.16823	0.56816	0.9514
	n=500	CPM	-0.20233	0.15407	0.9125
		MLT	-0.23678	0.15753	0.9053
	n=1000	CPM	-0.22376	0.10526	0.8536
		MLT	-0.25741	0.11619	0.8349
$F^{-1}(0.8\|X=0)$	n=50	CPM	0.15682	1.13436	0.9183
		MLT	-0.00113	1.02389	0.9306
	n=100	CPM	0.11847	0.55292	0.9228
		MLT	0.03498	0.50474	0.9287
	n=500	CPM	0.10465	0.11636	0.9125
		MLT	0.07027	0.10189	0.8749
	n=1000	CPM	0.10081	0.06350	0.8990
		MLT	0.06676	0.05266	0.7988
$F^{-1}(0.8\|X=1)$	n=50	CPM	0.62461	3.08822	0.9621
		MLT	0.43530	2.43644	0.9686
	n=100	CPM	0.45054	1.52173	0.9532
		MLT	0.40836	1.36050	0.9691
	n=500	CPM	0.33662	0.37300	0.9220
		MLT	0.33763	0.36804	0.9218
	n=1000	CPM	0.30132	0.22027	0.8873
		MLT	0.28898	0.21728	0.8539
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-189.412	
		MLT		-173.534	
	n=100	CPM		-453.361	
		MLT		-420.154	
	n=500	CPM		-3061.024	
		MLT		-2860.836	
	n=1000	CPM		-6799.685	
		MLT		-6405.667	

Table 2.20: Simulation results for mixture of discrete and continuous distribution: $H(y) = 0$ if $y \leq 0$ else $H(y) = \exp(y)$

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04205	0.11432	0.9429
		MLT	0.01687	0.11762	0.9395
		MLT(survival)	0.03716	0.11123	0.9567
	n=100	CPM	0.01670	0.05241	0.9455
		MLT	-0.01277	0.05390	0.9400
		MLT(survival)	0.01190	0.05124	0.9563
	n=500	CPM	0.00437	0.00969	0.9530
		MLT	-0.02862	0.01072	0.9343
		MLT(survival)	-0.00258	0.00952	0.9609
	n=1000	CPM	0.00251	0.00488	0.9512
		MLT	-0.003180	0.00603	0.9141
		MLT(survival)	-0.00493	0.00483	0.9556
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-128.763	
		MLT		-142.921	
	n=100	CPM		-337.646	
		MLT		-357.800	
	n=500	CPM		-2337.830	
		MLT		-2373.479	
	n=1000	CPM		-4791.826	
		MLT		-5104.161	

Table 2.21: Simulation results for discretizing continuous responses into 5 categories: 0, 3, 5, 7, 10

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04324	0.11425	0.9430
		MLT	0.01932	0.11494	0.9326
	n=100	CPM	0.01575	0.05235	0.9475
		MLT	-0.01004	0.05366	0.9348
	n=500	CPM	0.00364	0.00975	0.9535
		MLT	-0.02414	0.01057	0.9337
	n=1000	CPM	0.00230	0.00493	0.9521
		MLT	-0.02562	0.00572	0.9197
$E(Y\|X=0)$	n=50	CPM	-0.00607	0.36455	0.9354
		MLT	-0.00865	0.37464	0.8757
	n=100	CPM	0.00074	0.18621	0.9401
		MLT	0.00096	0.19135	0.8796
	n=500	CPM	0.00291	0.03688	0.9463
		MLT	0.00556	0.03784	0.8848
	n=1000	CPM	-0.00048	0.01852	0.9479
		MLT	0.00220	0.01901	0.8839
$E(Y\|X=1)$	n=50	CPM	-0.00184	0.37449	0.9341
		MLT	0.00423	0.38376	0.9025
	n=100	CPM	-0.00886	0.18647	0.9430

Table 2.21: Simulation results for discretizing continuous responses into 5 categories: 0, 3, 5, 7, 10 *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
		MLT	-0.00569	0.19101	0.9007
	n=500	CPM	0.00482	0.03550	0.9519
		MLT	0.00594	0.03633	0.9044
	n=1000	CPM	0.00274	0.01832	0.9506
		MLT	0.00362	0.01875	0.9059
				Value	
Out-of-sample Log-likelihood	n=50	CPM		-75.732	
		MLT		-88.033	
	n=100	CPM		-159.409	
		MLT		-180.299	
	n=500	CPM		-762.421	
		MLT		-824.768	
	n=1000	CPM		-1494.791	
		MLT		-1648.610	

Table 2.22: Simulation results for discretizing continuous responses into 10 categories: 0, 2, 3, 4, 5, 6, 7, 8, 10, 12

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04405	0.10969	0.9418
		MLT	0.03934	0.11182	0.9368
	n=100	CPM	0.01706	0.05014	0.9450
		MLT	0.01214	0.05170	0.9387
	n=500	CPM	0.00429	0.00926	0.9535
		MLT	-0.00096	0.00959	0.9443
	n=1000	CPM	0.00234	0.00470	0.9509
		MLT	-0.00285	0.00488	0.9435
$E(Y\|X=0)$	n=50	CPM	-0.00705	0.36468	0.9374
		MLT	-0.01228	0.36708	0.9005
	n=100	CPM	-0.00137	0.18439	0.9435
		MLT	-0.00375	0.18634	0.9068
	n=500	CPM	0.00167	0.03647	0.9485
		MLT	0.00190	0.03700	0.9154
	n=1000	CPM	-0.00078	0.01830	0.9484
		MLT	-0.00028	0.01855	0.9122
$E(Y\|X=1)$	n=50	CPM	-0.00380	0.44249	0.9357
		MLT	-0.00508	0.44741	0.9033
	n=100	CPM	-0.01078	0.21997	0.9435
		MLT	-0.01143	0.22248	0.9027
	n=500	CPM	0.00554	0.04194	0.9516
		MLT	0.00607	0.04242	0.9034
	n=1000	CPM	0.00183	0.02158	0.9481
		MLT	0.00262	0.02180	0.9034
				Value	

Out-of-sample Log-likelihood

Table 2.22: Simulation results for discretizing continuous responses into 10 categories: 0, 2, 3, 4, 5, 6, 7, 8, 10, 12 *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=50	CPM		-110.317	
		MLT		-116.119	
	n=100	CPM		-227.991	
		MLT		-236.425	
	n=500	CPM		-1095.694	
		MLT		-1152.974	
	n=1000	CPM		-2171.780	
		MLT		-2283.256	

Table 2.23: Simulation results for discretizing continuous responses into 20 categories: 0, 1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 9, 10, 11, 12, 15

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04613	0.10813	0.9386
		MLT	0.05430	0.12089	0.9270
	n=100	CPM	0.01818	0.04923	0.9468
		MLT	0.02666	0.05560	0.9304
	n=500	CPM	0.00416	0.00908	0.9524
		MLT	0.01331	0.01039	0.9355
	n=1000	CPM	0.00226	0.00462	0.9515
		MLT	0.01148	0.00532	0.9322
$E(Y\|X=0)$	n=50	CPM	-0.02196	1.02983	0.9345
		MLT	-0.00975	1.05967	0.9080
	n=100	CPM	-0.00840	0.51849	0.9416
		MLT	0.00244	0.53628	0.9203
	n=500	CPM	0.00255	0.10262	0.9470
		MLT	0.01201	0.10619	0.9266
	n=1000	CPM	-0.00153	0.05143	0.9465
		MLT	0.00760	0.05314	0.9270
$E(Y\|X=1)$	n=50	CPM	0.00560	1.08025	0.9327
		MLT	-0.01249	1.11214	0.9158
	n=100	CPM	-0.01182	0.53736	0.9423
		MLT	-0.02783	0.55479	0.9183
	n=500	CPM	0.00970	0.10215	0.9501
		MLT	-0.00438	0.10553	0.9233
	n=1000	CPM	0.00412	0.05249	0.9508
		MLT	-0.00989	0.05449	0.9236
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-142.016	
		MLT		-143.775	
	n=100	CPM		-300.009	
		MLT		-309.736	
	n=500	CPM		-1439.589	
		MLT		-1493.155	

Table 2.23: Simulation results for discretizing continuous responses into 20 categories: 0, 1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 9, 10, 11, 12, 15 *(continued)*

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
	n=1000	CPM		-2841.223	
		MLT		-2964.074	

Table 2.24: Simulation results for discretizing continuous responses into 50 categories: 0, 1, 1.4, 1.6, 1.9, 2.1, 2.3, 2.5, 2.7, 2.9, 3.1, 3.3, 3.5, 3.7, 3.8, 4.0, 4.2, 4.4, 4.6, 4.8, 5.0, 5.1, 5.3, 5.5 5.7, 5.9, 6.1, 6.4, 6.6, 6.8, 7.0, 7.3, 7.6, 7.8, 8.2, 8.5, 8.8, 9.0, 9.4, 9.8, 10.2, 10.6, 11.0, 11.5, 12.0, 13.0, 13.5, 14.5, 16.0, 18.0

Measure	Sample Size	Method	Bias	MSE	Coverage Rate
β	n=50	CPM	0.04641	0.10784	0.9401
		MLT	0.05585	0.11921	0.9294
	n=100	CPM	0.01906	0.04905	0.9452
		MLT	0.02859	0.05482	0.9319
	n=500	CPM	0.00432	0.00901	0.9537
		MLT	0.01526	0.01027	0.9375
	n=1000	CPM	0.00237	0.00460	0.9507
		MLT	0.01344	0.00530	0.9326
$E(Y\|X=0)$	n=50	CPM	-0.05463	6.53464	0.9349
		MLT	0.02347	6.68861	0.9159
	n=100	CPM	-0.02541	3.28806	0.9409
		MLT	0.05314	3.39546	0.9254
	n=500	CPM	0.00495	0.64939	0.9477
		MLT	0.07668	0.67764	0.9298
	n=1000	CPM	-0.00568	0.32671	0.9468
		MLT	0.06408	0.34119	0.9307
$E(Y\|X=1)$	n=50	CPM	0.00706	6.62922	0.9341
		MLT	-0.39988	7.39843	0.9064
	n=100	CPM	-0.02841	3.29753	0.9423
		MLT	-0.39039	3.74305	0.9104
	n=500	CPM	0.02244	0.62732	0.9513
		MLT	-0.30736	0.76786	0.9096
	n=1000	CPM	0.00791	0.32268	0.9487
		MLT	-0.32699	0.44968	0.8822
Out-of-sample Log-likelihood				Value	
	n=50	CPM		-168.705	
		MLT		-164.335	
	n=100	CPM		-383.900	
		MLT		-382.161	
	n=500	CPM		-1920.775	
		MLT		-1947.334	
	n=1000	CPM		-3775.912	
		MLT		-3873.933	

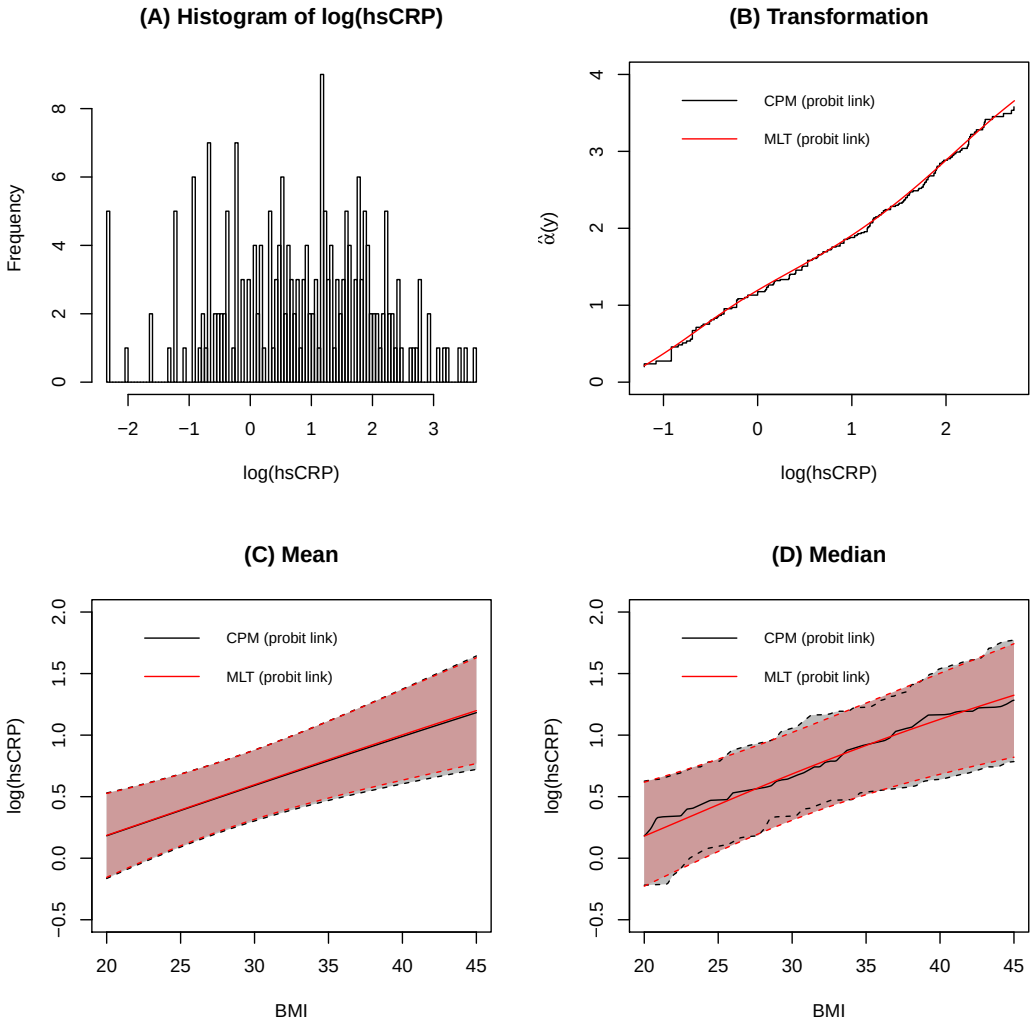

Figure 2.24: Results for log-transformed hsCRP. A: The distribution of log-transformed hsCRP. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

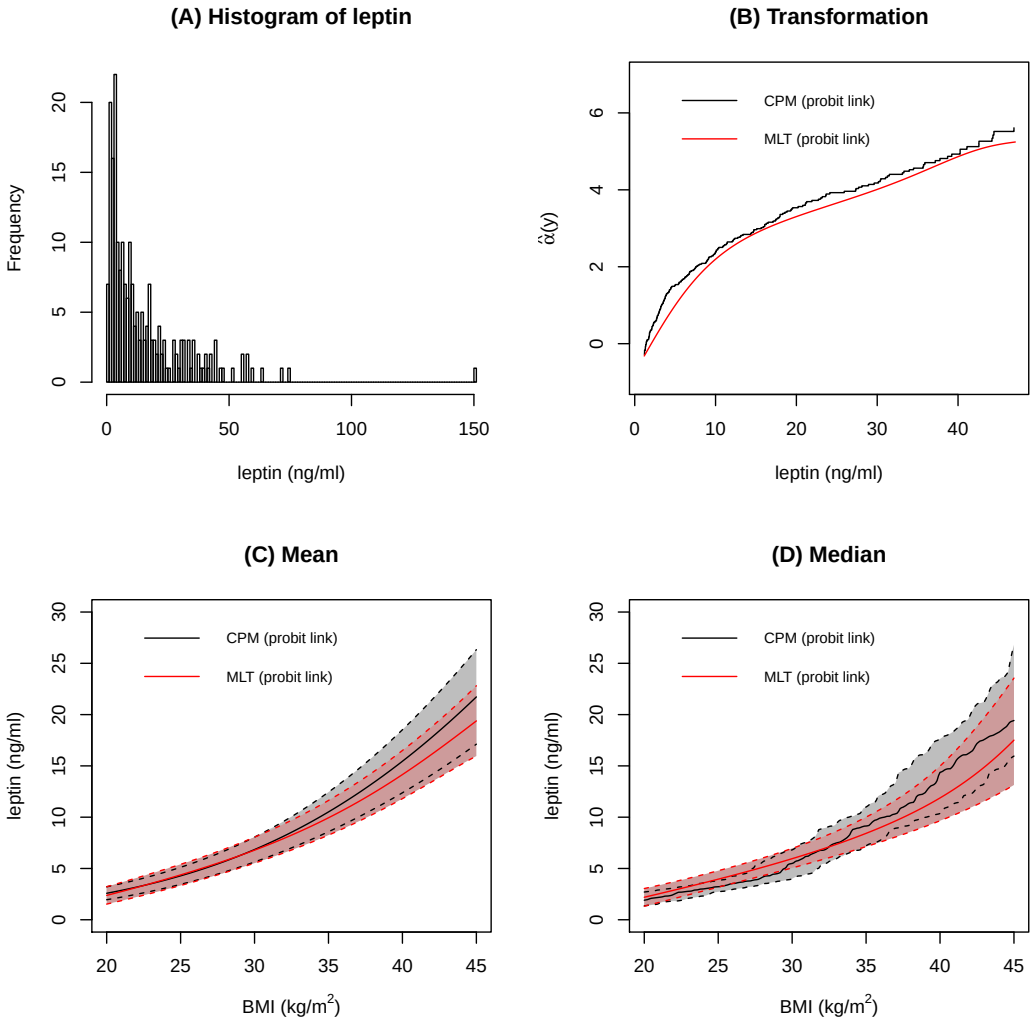

Figure 2.25: Results for leptin. A: The distribution of leptin. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

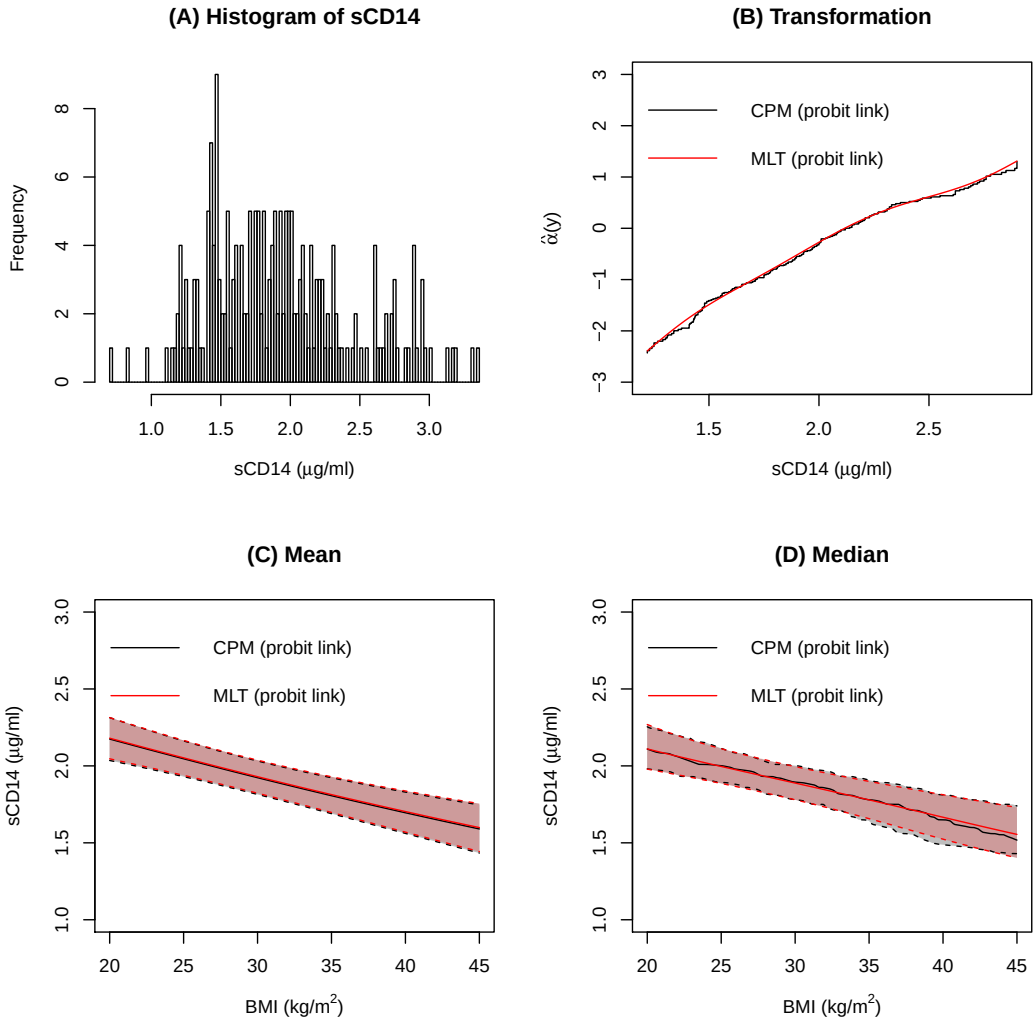

Figure 2.26: Results for sCD14. A: The distribution of sCD14. B: The estimated transformation functions. C: The estimated conditional means and their confidence intervals. Other covariates are at their most frequent level or median level. D: The estimated conditional medians and their confidence intervals. Other covariates are at their most frequent level or median level.

CHAPTER 3

Addressing Detection Limits with Semiparametric Cumulative Probability Models

3.1 Introduction

Detection limits (DLs) are not uncommon in biomedical research and other fields. For example, radiation doses may only be detected above a certain threshold (Wing et al., 1991), antibody concentrations may not be measured below certain levels (Wu et al., 2001), and X-rays may have lower limits of detection (Pan et al., 2017). In HIV research, viral load can only be detected above certain levels. To complicate matters, DLs often vary by assay and may change over time. For example, HIV viral load assays have had lower DLs at 400, 300, 200, 50, and 20 copies/mL depending on the commercial assay and year of application (Steegen et al., 2007).

Different types of analysis methods to handle DLs have been proposed for different purposes. In this manuscript, we will focus on studying the association between an outcome variable and covariates, where the outcome variable is subject to DLs. This is typically achieved with some sort of regression model. One common and simple method is to dichotomize the outcome as detectable or undetectable, and then to perform logistic regression (Jiamsakul et al., 2017). While it can be useful for some purposes, the dichotomization leads to information loss since the observed values inside the DLs are treated as if they are the same. Another common approach for handling DLs is substitution, where all nondetects are imputed with a single constant and a linear regression model is fit. The imputed constant may be, for example, the DL itself, DL/2, DL/$\sqrt{2}$ (Hornung and Reed, 1990; Lubin et al., 2004; Helsel, 2011), or the expectation of the measurement conditional on being outside the DL under some assumed parametric model (Garland et al., 1993). For example, DL/2 corresponds to the expectation of a uniform distribution between 0 and the DL. Although simple, these substitution approaches typically result in biased estimation, underestimated variances, and thus sometimes wrong conclusions (Baccarelli et al., 2005; Fiévet and Della Vedova, 2010). In a third approach, one explicitly makes parametric assumptions on the distribution of the data, both within and outside the DLs. Parameters of interest can then be estimated by maximizing the censored data likelihood. Such a maximum likelihood approach is efficient and consistent when the distribution is correctly specified, but may perform poorly when distributional assumptions are incorrect. To compound the problem, there is typically no way to examine model fit outside the DLs; goodness-of-fit of a parametric model inside DLs does not ensure goodness-of-fit outside the DLs (Baccarelli et al., 2005; Harel et al., 2014). A related, fourth approach for addressing DLs is to multiply impute values outside the DLs (Little and Rubin, 2019; Harel and Zhou, 2007). This approach

may be computationally expensive, still requires parametric assumptions that can only be verified inside DLs, and may be particularly problematic with high rates of censoring or small sample sizes (Lubin et al., 2004; Zhang et al., 2009).

To avoid strong parametric assumptions, nonparametric methods such as Kaplan–Meier, score and rank-based methods have been proposed in two-sample comparisons (Helsel, 2011). Zhang et al. (2009) explored the use of the Wilcoxon rank sum test, other weighted rank tests, Gehan and Peto-Peto tests, and a novel nonparametric method for location-shift inference with DLs. Although attractive for two-sample tests, these nonparametric methods do not permit inclusion of covariates.

In this manuscript, we propose a new approach for analyzing data subject to detection limits. Data with DLs effectively follow a mixture distribution, where those below a lower DL can be thought of as belonging to a discrete category, those above an upper DL belonging to another discrete category, while those inside the DLs are continuous. Whether discrete or continuous, the values are orderable. In earlier work, Liu et al. (2017) showed that continuous response variables can be modeled using a popular model for ordinal outcomes, namely the cumulative probability model (CPM), also known as the 'cumulative link model' (Agresti, 2003). CPMs are a type of semiparametric linear transformation model, in which the continuous response variable after some unspecified monotonic transformation is assumed to follow a linear model, and the transformation is nonparametrically estimated (Zeng and Lin, 2007). These models are very flexible and can handle a wide variety of outcomes, including variables with DLs. Importantly, when fitting CPMs to data with DLs, minimal assumptions are made on the distribution of the response variable outside the DLs as these models are based on ranks, and values below/above DLs are simply the lowest/highest rank values. Because of their relationship to the Wilcoxon rank sum test (McCullagh, 1980), the CPM can be thought of as a semiparametric extension to permit covariates to the approaches that Zhang et al. (2009) found effective for handling DLs in two-sample comparisons. Finally, as will be shown, because CPMs model the conditional cumulative distribution function (CDF), it is easy to extract many different measures of conditional association from a single fitted model, including conditional quantiles, conditional probabilities, odds ratios, and probabilistic indexes, which permits flexible and compatible interpretation.

In Section 2, we review the CPM, illustrate its use for simple settings where there is only a single set of DLs, and then show how CPMs can be extended to address multiple DLs. We also propose a new method for estimating the conditional quantile from a CPM. In Section 3, we illustrate and demonstrate the advantages of the proposed approach using real data from two studies. The first study aims to measure the association between covariates and a biomarker whose values are below a DL in approximately 15% of observations. The second example is a large multi-cohort study of viral load (VL) after starting antiretroviral therapy among persons with HIV, where most observations are below DLs, but the DLs vary across sites and change over

time. In Section 4, we demonstrate the performance of our method with simulations. The final section contains a discussion of the strengths and limitations of our method and future work.

3.2 Methods

3.2.1 Cumulative Probability Models

Transformation is often needed for regression of a continuous outcome variable Y to satisfy model assumptions, but specifying the correct transformation can be difficult. In a linear transformation model, the outcome is modeled as $Y = H(\beta^T X + \varepsilon)$, where $H(\cdot)$ is an unknown monotonically increasing transformation, X is a vector of covariates, and ε follows a known distribution with CDF F_ε. This linear transformation model can be equivalently expressed in terms of the conditional CDF,

$$F(y|X) \equiv \Pr(Y \leq y|X) = \Pr[\varepsilon \leq H^{-1}(y) - \beta^T X|X] = F_\varepsilon[H^{-1}(y) - \beta^T X].$$

Let $G = F_\varepsilon^{-1}$ and $\alpha = H^{-1}$; $\alpha(\cdot)$ is monotonically increasing but otherwise unknown. Then

$$G[F(y|X)] = \alpha(y) - \beta^T X, \tag{3.1}$$

where G serves as a link function and the model becomes a cumulative probability model (CPM). The intercept function $\alpha(y)$ is the transformation of the response variable such that $\alpha(Y) = \beta^T X + \varepsilon$. The β coefficients indicate the association between the response variable and covariates: fixing other covariates, a positive/negative β_j means that an increase in X_j is associated with a stochastic increase/decrease in the distribution of the response variable.

In the CPM (3.1), the intercept function $\alpha(y)$ can be nonparametrically estimated with a step function (Zeng and Lin, 2007; Liu et al., 2017). This allows great model flexibility. Consider an iid dataset $\{(y_i, x_i) : i = 1, \ldots, n\}$. The nonparametric likelihood is

$$\prod_{i=1}^{n} \left[F(y_i|x_i) - F(y_i^-|x_i)\right], \tag{3.2}$$

where $F(y_i^-|x_i) = \lim_{t \uparrow y_i} F(t|x_i)$. In nonparametric maximum likelihood estimation, the probability mass given any x will be distributed over the discrete set of observed outcome values. Thus we only need to consider functions for $\alpha(\cdot)$ such that $F(y|x_i)$ is a discrete distribution over the observed values. Let J be the number of distinct outcome values, denoted as $a_1 < \cdots < a_J$. Let $S = \{a_1, \ldots, a_J\}$. These serve as the anchor points for the nonparametric likelihood. Let $\alpha_j = \alpha(a_j)$; then $\alpha_1 < \cdots < \alpha_J$. The nonparametric likelihood

(3.2) can be written as

$$L(\beta, \boldsymbol{\alpha}) = \prod_{i:y_i=a_1} F_\varepsilon(\alpha_1 - \beta^T x_i) \times \prod_{j=2}^{J-1} \prod_{i:y_i=a_j} \left[F_\varepsilon(\alpha_j - \beta^T x_i) - F_\varepsilon(\alpha_{j-1} - \beta^T x_i)\right] \times \prod_{i:y_i=a_J} \left[1 - F_\varepsilon(\alpha_{J-1} - \beta^T x_i)\right]. \tag{3.3}$$

Maximizing (4.4), we obtain the nonparametric maximum likelihood estimates (NPMLEs), $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{J-1})$. Note the multinomial form of the likelihood (4.4); because the probabilities in a multinomial likelihood add to one, α_J is not estimated. Note also that the likelihood in (4.4) is identical to that of cumulative link models for ordinal data if the outcome Y is treated as ordinal with categories $\{a_1, \ldots, a_J\}$. Liu et al. (2017) and Tian et al. (2020) have shown that CPMs can be fit to and work well for continuous and mixed types of responses. CPMs have also been shown to be consistent and asymptotically normal, with variance consistently estimated with the inverse of the information matrix under mild conditions including boundedness of the outcome variable (Li et al., 2022b). The NPMLEs and their estimated variances can be efficiently computed with the `orm()` function in the **rms** package in R (Harrell, 2020), which takes advantage of the tridiagonal nature of the Hessian matrix using Cholesky decomposition (Liu et al., 2017).

CPMs have several nice features. Some widely used regression methods model only one aspect of the conditional distributions (e.g., conditional mean for linear regression and conditional quantile for quantile regression). With the NPMLEs $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$, we can estimate the conditional CDFs as $\hat{F}(y|x) = F_\varepsilon(\hat{\alpha}_j - \hat{\beta}^T x)$ where j is the index such that $a_j = \max\{a \in S : a \leq y\}$; standard errors can be obtained by the delta method. Since conditional CDFs are directly modeled, other characteristics of the distribution, such as the conditional quantiles and conditional expectations, can be easily derived (Liu et al., 2017). Depending on the choice of link function, β may be interpretable; for example, with the logit link function, $\exp(\beta)$ is an odds ratio. Probabilistic indexes (De Neve et al., 2019), which are defined as $\Pr(Y_1 < Y_2 | X_1, X_2)$, can also be easily derived; for example, with the logit link, $P(Y_1 < Y_2 | X_1, X_2) = \left[1 + \exp\left(-(X_2 - X_1)^T \beta\right)\right]^{-1}$. With the transformation $\alpha(\cdot)$ nonparametrically estimated, CPMs are invariant to any monotonic transformation of the outcome; therefore, no pre-transformation is needed. With a single binary covariate and the logit link function, the score test for the CPM is nearly identical to the Wilcoxon rank sum test (McCullagh, 1980); see Supplemental Material. Because only the order of the outcome values but not the specific values matter when estimating β in the CPM, it can handle any ordinal, continuous, or mixture of ordinal and continuous distributions, which can be useful for analyzing data with DLs.

3.2.2 Single Detection Limits

In this subsection, we first present our method for the simple scenario that there is a single lower DL and/or a single upper DL. We will describe the general approach for multiple DLs in the next subsection.

Consider a dataset with a lower DL, l, and an upper DL, u. The outcome Y is observed if it is inside the DLs (i.e., $l \leq Y \leq u$) or censored if it is outside the DLs. The J distinct values of the observed outcomes are denoted as $l \leq a_1 < \cdots < a_J \leq u$. When there are no observations outside the DLs, these values are treated as ordered categories in CPMs and they are the anchor points in the nonparametric likelihood (4.4), and correspondingly there are $J-1$ alpha parameters, $\alpha_1 < \cdots < \alpha_{J-1}$. With observations outside the DLs, the likelihood (4.4) needs to be modified accordingly.

When there are observations below the lower DL, we do not know their values except that they are $< l$. As there is no way to distinguish them, we treat them as a single category, denoted as a_0. Note that a_0 is not a value but a symbol for the additional category below a_1. The nonparametric likelihood for a subject outcome censored at the lower DL l is

$$\Pr(Y_i < l | X_i = x_i) = F_\varepsilon(\alpha_0 - \beta^T x_i),$$

where α_0 is the extra alpha parameter corresponding to category a_0 such that $\alpha_0 < \alpha_1$. Because a_1, the previously lowest category, now has a category below it, the nonparametric likelihood for a subject with $y_i = a_1$ becomes

$$F_\varepsilon(\alpha_1 - \beta^T x_i) - F_\varepsilon(\alpha_0 - \beta^T x_i).$$

Similarly, when there are observations above the upper DL, they are also treated as a single category, denoted as a_{J+1}, which is a symbol for the additional category above a_J. The nonparametric likelihood for a subject censored at the upper DL u is

$$\Pr(Y_i > u | X_i = x_i) = 1 - F_\varepsilon(\alpha_J - \beta^T x_i),$$

Because a_J is no longer the highest category, α_J will need to be estimated, and the likelihood for a subject with $y_i = a_J$ is now

$$F_\varepsilon(\alpha_J - \beta^T x_i) - F_\varepsilon(\alpha_{J-1} - \beta^T x_i).$$

Put together, with observed data subject to a single lower DL and a single upper DL, the CPM likelihood

is

$$L(\beta, \boldsymbol{\alpha}) = \prod_{i:y_i=a_0} F_\varepsilon(\alpha_0 - \beta^T x_i) \times \prod_{j=1}^{J} \prod_{i:y_i=a_j} \left[F_\varepsilon(\alpha_j - \beta^T x_i) - F_\varepsilon(\alpha_{j-1} - \beta^T x_i)\right] \times \prod_{i:y_i=a_{J+1}} \left[1 - F_\varepsilon(\alpha_J - \beta^T x_i)\right], \tag{3.4}$$

which is equivalent to (4.4) except with two new anchor points, a_0 and a_{J+1}. Therefore, (3.4) is maximized in an identical manner to (4.4), with outcomes below the lower DL and outcomes above the upper DL simply assigned to categories a_0 and a_{J+1}, respectively.

In summary, when there are data censored below the lower DL, we add a new anchor point $a_0 < a_1$ and a new parameter α_0; when there are data censored above the upper DL, we add a new anchor point $a_{J+1} > a_J$ and a new parameter α_J. The alpha parameters to be estimated are $(\alpha_1, \ldots, \alpha_{J-1})$ when there are no DLs or no data censored at DLs, $(\alpha_0, \alpha_1, \ldots, \alpha_{J-1}, \alpha_J)$ when both categories a_0 and a_{J+1} are added, $(\alpha_0, \alpha_1, \ldots, \alpha_{J-1})$ when only a_0 is added, and $(\alpha_1, \ldots, \alpha_{J-1}, \alpha_J)$ when only a_{J+1} is added.

In practice, one can fit the NPMLE in these settings using the `orm()` function by setting outcomes below the lower DL to some arbitrary number $< l$ and outcomes above the upper DL to some arbitrary number $> u$. Note that unlike single imputation approaches for dealing with DLs, the CPM estimation procedure is invariant to the choice of these numbers assigned to values outside the DLs. The CPM (3.1) assumes that after some unspecified transformation, the outcome follows a linear model both within and outside the DLs. In contrast, parametric approaches to deal with DLs assume the full distribution of the outcome conditional on covariates is known, both within and outside DLs. Hence, CPMs make much weaker assumptions than fully parametric approaches.

3.2.3 Multiple Detection Limits

We now consider the general situation where data may be collected from multiple study sites. A site may have no DL, only one DL, or both lower and upper DLs. Each site may have different lower DLs and different upper DLs, which may change over time.

Every subject has a vector X of covariates and three underlying random variables (Y, C_L, C_U), where Y is the true outcome and $C_L < C_U$ are the lower and upper DLs. When there is no upper DL, $C_U = \infty$, and when there is no lower DL, $C_L = -\infty$. C_L and C_U are assumed to be independent of Y conditional on X; the vector X may contain variables for study sites or calendar time. This non-informative censoring assumption is typically plausible as DLs are determined by available equipment / assays. We assume the CPM (3.1) holds for all subjects. Due to DLs, we may not always observe Y. Instead we only observe (Z, Δ), where $Z = \max(\min(Y, C_U), C_L)$ and Δ is a variable indicating whether Y is observed or censored at a DL: $\Delta = 1$

and $Z = Y$ if Y is observed, $\Delta = L$ and $Z = C_L$ if $Y < C_L$, and $\Delta = U$ and $Z = C_U$ if $Y > C_U$.

Given a dataset $\{(z_i, \delta_i; x_i)\}$ $(i = 1, \ldots, n)$, we first determine how many anchor points are needed to support the nonparametric likelihood of the CPM. Let J be the number of distinct values of z_i among those with $\delta_i = 1$; they are denoted as $a_1 < \cdots < a_J$. For data without any DLs, these points are the anchor points for the nonparametric likelihood, and they are effectively treated as ordered categories in a CPM. Let $S = \{a_1, \cdots, a_J\}$ be the set of these values. When there are data with $\delta_i = L$, let l be the smallest z_i with $\delta_i = L$. Similarly, when there are data with $\delta_i = U$, let u be the largest z_i with $\delta_i = U$. If $l \leq a_1$, we add a category into S below a_1, denoted as a_0; note that it is not a value but a symbol for the additional category in S below a_1. Similarly, if $u \geq a_J$, we add a_{J+1} into S, which is a symbol for the additional category above a_J. Depending on the data, the number of ordered categories can be J, $J+1$, or $J+2$.

Consider the situation where both a_0 and a_{J+1} have been added to S (i.e., $S = \{a_0, a_1, \ldots, a_J, a_{J+1}\}$). When $\delta_i = 1$, the nonparametric likelihood for $(z_i, 1)$ is

$$F_\varepsilon(\alpha_j - \beta^T x_i) - F_\varepsilon(\alpha_{j-1} - \beta^T x_i), \tag{3.5}$$

where j is the index such that $a_j = z_i$. When $\delta_i = L$, the nonparametric likelihood for (z_i, L) is

$$\Pr(Y < z_i | x_i) = \begin{cases} F_\varepsilon(\alpha_0 - \beta^T x_i), & (z_i = l) \\ F_\varepsilon(\alpha_j - \beta^T x_i), & (z_i \neq l) \end{cases} \tag{3.6}$$

where j is the index such that $a_j = \max\{a \in S : a < z_i\}$ when $z_i \neq l$. When $\delta_i = U$, the nonparametric likelihood for (z_i, U) is

$$\Pr(Y > z_i | x_i) = \begin{cases} 1 - F_\varepsilon(\alpha_J - \beta^T x_i), & (z_i = u) \\ 1 - F_\varepsilon(\alpha_{j-1} - \beta^T x_i), & (z_i \neq u) \end{cases} \tag{3.7}$$

where j is the index such that $a_j = \min\{a \in S : a > z_i\}$ when $z_i \neq u$. The overall nonparametric likelihood is the product of these individual likelihoods over all subjects. Note that if there are no uncensored observations between two lower (or upper) DLs, the two DLs are effectively treated as the same DL. We show a toy example to illustrate the approach to handle multiple DLs described in Table 3.1.

Slight modifications will be applied when no or only one additional category is added to S. When there is no need to add a_0 to S (i.e., when $l > a_1$ or there are no lower DLs), only the second row in the likelihood (3.6) for (z_i, L) will be employed, and the likelihood for $(z_i, 1)$ with $z_i = a_1$ is $F_\varepsilon(\alpha_1 - \beta^T x_i)$. When there is no need to add a_{J+1} to S (i.e., when $u < a_J$ or there are no upper DLs), only the second row in the likelihood

Table 3.1: The likelihood contribution for each observation in a toy example with multiple detection limits. Suppose there are eight observations from two sites. The lower and upper DLs of one site are 3 and 9, respectively, and those of the other site are 5 and 12. Suppose the dataset is $\{(3,L;x_1),(4,1;x_2),(6,1;x_3),(9,U;x_4)\}$ from site 1 and $\{(5,L;x_5),(7,1;x_6),(10,1;x_7),(12,U;x_8)\}$ from site 2. There are thus $J=4$ uncensored values, and we include two additional categories corresponding to those below the smallest DL and those above the largest DL, such that $S=\{a_0,a_1,\ldots,a_5\}=\{\text{`<3'},4,6,7,10,\text{`>12'}\}$.

z_i	δ_i	j	Likelihood
3	L	0	$F_\varepsilon(\alpha_0-\beta^T x_1)$
4	1	1	$F_\varepsilon(\alpha_1-\beta^T x_2)-F_\varepsilon(\alpha_0-\beta^T x_2)$
5	L	1	$F_\varepsilon(\alpha_1-\beta^T x_5)$
6	1	2	$F_\varepsilon(\alpha_2-\beta^T x_3)-F_\varepsilon(\alpha_1-\beta^T x_3)$
7	1	3	$F_\varepsilon(\alpha_3-\beta^T x_6)-F_\varepsilon(\alpha_2-\beta^T x_6)$
9	U	4	$1-F_\varepsilon(\alpha_3-\beta^T x_4)$
10	1	4	$F_\varepsilon(\alpha_4-\beta^T x_7)-F_\varepsilon(\alpha_3-\beta^T x_7)$
12	U	5	$1-F_\varepsilon(\alpha_4-\beta^T x_8)$

(3.7) for (z_i,U) will be employed, and the likelihood for $(z_i,1)$ with $z_i=a_J$ is $1-F_\varepsilon(\alpha_{J-1}-\beta^T x_i)$.

We have developed an R package, **multipleDL** available at https://github.com/YuqiTian35/multipleDL, which uses the `optimizing()` function in the **rstan** package to maximize the likelihood (Stan Development Team, 2020).

3.2.4 Interpretable Quantities and Conditional Quantiles

Interpretation of results after fitting CPMs to outcomes with DLs is similar to settings without DLs. Depending on the link function, β may be directly interpretable. The conditional CDF, probabilistic indexes, and conditional quantiles are also easily derived. Note, however, that without additional assumptions on the distribution of the outcome outside DLs, conditional expectations cannot be estimated.

We now describe how to infer conditional quantiles from a CPM fitted on data with DLs. The conditional CDF from a CPM for a given x can be computed as $\hat{F}(y|x)=F_\varepsilon(\hat{\alpha}_j-\hat{\beta}^T x)$ where j is the index such that $a_j=\max\{a\in S:a\le y\}$; if there is no $a\in S$ such that $a\le y$, then $\hat{F}(y|x)=0$. For ease of presentation, we fix x and let $P_j=\hat{F}(a_j|x)$ $(j=0,1,\ldots,J,J+1)$; for convenience, let $P_{-1}=0$. Our goal is to define a quantile function $\hat{Q}(p)$, where $0<p<1$, for the conditional distribution given x.

The quantile function for a CDF $F(\cdot)$ is typically defined as $Q(p)=\inf\{z:F(z)\ge p\}$. A plug-in estimator for an estimated CDF, $\hat{F}$, is $\hat{Q}_0(p)=\inf\{z:\hat{F}(z)\ge p\}$. When applied to our setting, $\hat{Q}_0(p)=a_j$ when $P_{j-1}<p\le P_j$. This estimator may not be suitable for CPMs because $\hat{F}(\cdot)$ is a step function and therefore $\hat{Q}_0(p)$ only takes values at the anchor points, which can be undesirable for continuous outcomes, especially when there is a large gap between adjacent anchor points.

Liu et al. (2017) proposed to estimate quantiles for CPMs with linear interpolation. Specifically, given

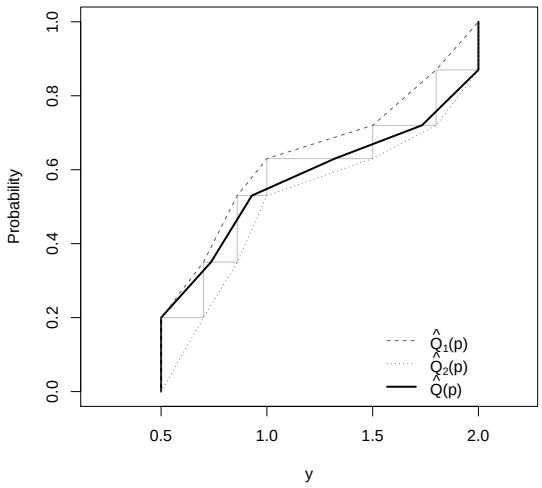

Figure 3.1: Illustration of three approaches for conditional quantiles. The data set has a lower DL 0.5, an upper DL 2, and five observed values of y: 0.7, 0.86, 1, 1.5, 1.8. Thus $S = \{\text{`<0.5'}, 0.7, 0.86, 1, 1.5, 1.8, \text{`>2'}\}$. The dashed lines are for $\hat{Q}_1(p)$, the dotted lines are for $\hat{Q}_2(p)$, the solid black lines are for $\hat{Q}(p)$, and the solid gray lines are for the empirical CDF. Here, $\hat{Q}(p) = \hat{Q}_1(p) =$ '<0.5' when $p < \hat{F}(0.5|x)$, and $\hat{Q}(p) = \hat{Q}_2(p) =$ '>2' when $p > \hat{F}(2|x)$.

a fixed p, let $j = j(p)$ be the index such that $P_{j-1} < p \leq P_j$. When $p > P_0$, $j \geq 1$ and define $\hat{Q}_1(p) = a_{j-1} + \frac{p-P_{j-1}}{P_j - P_{j-1}}(a_j - a_{j-1})$, which is a linear interpolation between a_{j-1} and a_j. When $p \leq P_0$, $\hat{Q}_1(p)$ is set to be a_0. Recall that a_0 is not a value but a symbol for being below the lower DL, l; we thus relabel it as '$<l$', so when $p \leq P_0$, $\hat{Q}_1(p) =$ '$<l$'. For the linear interpolation between a_0 and a_1, we set a_0 to be l. Similarly, a_{J+1} is labeled '$>u$' and assigned the value u for the linear interpolation between a_J and a_{J+1}. $\hat{Q}_1(p)$ is illustrated as the dashed lines in Figure 3.1. An alternative definition is to interpolate between a_j and a_{j+1}: $\hat{Q}_2(p) = a_j + \frac{p-P_{j-1}}{P_j - P_{j-1}}(a_{j+1} - a_j)$ when $p < P_J$ and $\hat{Q}_2(p) = a_{J+1} =$ '$>u$' when $p \geq P_J$. $\hat{Q}_2(p)$ is illustrated as the dotted lines in Figure 3.1. For continuous data without DLs, $\hat{Q}_1(p)$ and $\hat{Q}_2(p)$ converge as the sample size increases. However, they are problematic for continuous data with DLs because $\hat{Q}_1(p) < a_{J+1}$ for all $p < 1$ and $\hat{Q}_2(p) > a_0$ for all $p > 0$ even though there are non-zero estimated probabilities at the lower DL a_0 and upper DL a_{J+1}.

We propose a new quantile estimator as a weighted average between $\hat{Q}_1(p)$ and $\hat{Q}_2(p)$,

$$\hat{Q}(p) = (1-w)\hat{Q}_1(p) + w\hat{Q}_2(p), \tag{3.8}$$

where $w = w(p) = \frac{p-P_0}{P_J - P_0}$ when $P_0 < p < P_J$, 0 when $p \leq P_0$, and 1 when $p \geq P_J$. This definition is shown as the black curve in Figure 3.1. Note that $\hat{Q}(p)$ equals $\hat{Q}_1(p) =$ '$<l$' when $p \leq P_0$, and equals $\hat{Q}_2(p) =$ '$>u$'

when $p \geq P_J$. It can be shown that similar to $\hat{Q}_1(p)$ and $\hat{Q}_2(p)$, $\hat{Q}(p)$ is also piecewise linear with transition points at P_j $(j = 0, 1, \ldots, J)$.

In situations where there is only a lower DL or an upper DL, our definition of $\hat{Q}(p)$ is similar. Confidence intervals for the conditional quantiles can be estimated by applying a weighted linear interpolation to the confidence intervals of the conditional CDF similar to the above procedure (Liu et al., 2017).

3.3 Applications

In this section, we illustrate our method with two datasets, one from a biomarker study with a single lower DL and the other from a multi-center study with multiple DLs varying within and across centers.

3.3.1 Single Detection Limit

Our first example uses data from a study investigating the relationship between HIV, diabetes, obesity, and various biomarkers. Data were collected on 161 adults, some of whom were highly overweight (body mass index (BMI) ranged from 22 to 58 kg/m^2). Several biomarkers were measured. Here, we focus on interleuken 4 (IL-4), a cytokine that is related to T-cell production and metabolism and has been seen to limit lipid accumulation in mice (Tsao et al., 2014). We examine the association between IL-4 and BMI, controlling for age, sex, HIV-status, and diabetes-status. Our measures of IL-4 had a single lower DL of 0.019 pg/ml, and 24 subjects (15%) had IL-4 values below the DL. Figure 3.2 shows the distribution of IL-4 is right skewed and has a lower detection limit as 0.019. After log-transformation, the response variable is approximately normally distributed except for value below the DL.

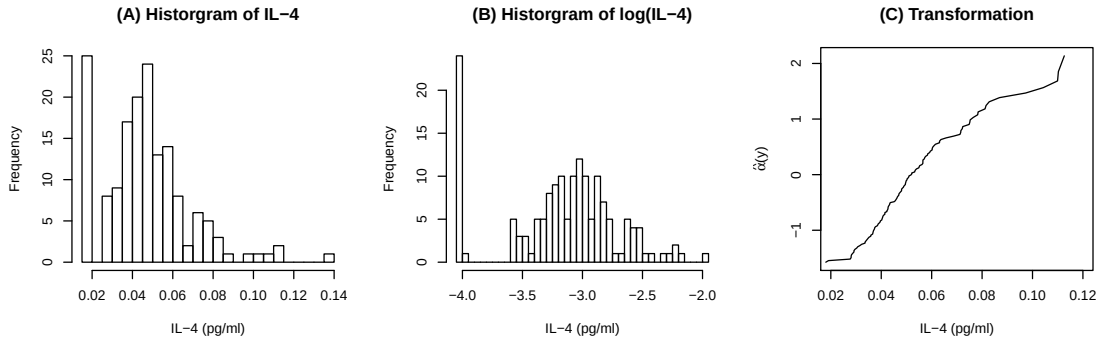

Figure 3.2: (A) The distribution of IL-4. (B) The distribution of log-transformed IL-4. (C) The estimated transformation function.

We fit a CPM as described in Section 2.2, using the logit link; results are in Table 3.2. No transformation of IL-4 was needed. With the logit link function, the β parameters can be interpreted as log odds ratios. BMI was found to be negatively associated with IL-4 (p-value 0.023). Holding other covariates constant, a 5

Table 3.2: Application results for IL-4: the estimated odds ratios with 95% confidence intervals, and p-values

Predictor	Odds Ratio	95% (CI)	P-value	PI
Age (per 10 years)	1.16	(0.89, 1.51)	0.271	0.525
Sex			0.689	
Female (reference)	1			
Male	1.14	(0.59, 2.22)	0.689	0.522
BMI (per 5 kg/m^2)	0.78	(0.62, 0.97)	0.023	0.264
Status			$<$0.001	
HIV positive, insulin sensitive (reference)	1			
HIV positive, pre-diabetic	2.15	(1.03, 4.50)	0.041	0.625
HIV positive, diabetic	1.03	(0.44, 2.44)	0.945	0.505
HIV negative, diabetic	1.82	(0.70, 4.74)	0.218	0.599

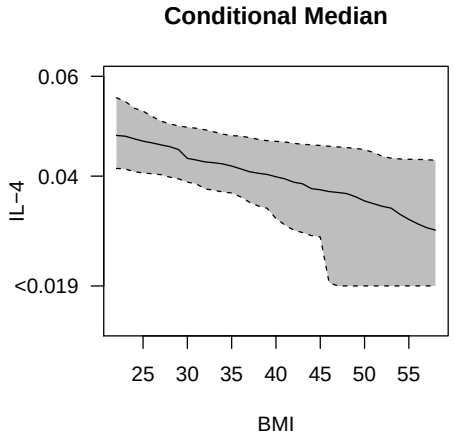

Figure 3.3: The conditional median obtained by CPMs varying BMI while fixing other covariates at median/mode levels

kg/m^2 increase in BMI corresponded to a 22% decrease in the odds of having a higher IL-4 value (adjusted odds ratio 0.78, 95% confidence interval (CI) of $(0.62, 0.97)$). The corresponding probabilistic index was 0.264, meaning that holding other variables constant, a 5 kg/m^2 increase in BMI was associated with a 0.736 $(= 1 - 0.264)$ probability of having a lower IL-4. The median IL-4 conditional on BMI and controlling for all other covariates at their median/mode levels was estimated from the CPM and is shown in Figure 3.3. The conditional median decreased as BMI increased, with the 95% CI including the category '$<$0.019' for those with a very large BMI. Note that '$<$0.019' is the smallest ordered category indicating for values below the DL. Other quantiles and quantities can also be easily derived from the CPM; for example, Figure 3.4 shows the 90th percentile of IL-4 as a function of BMI, and the probabilities of IL-4 being greater than 0.019 (the DL) and greater than 0.05 as functions of BMI.

It is worth comparing results from the CPM to other potential analysis approaches. (i) The most common approach in practice would be to singly impute those values below the DL; given the skewed nature of the

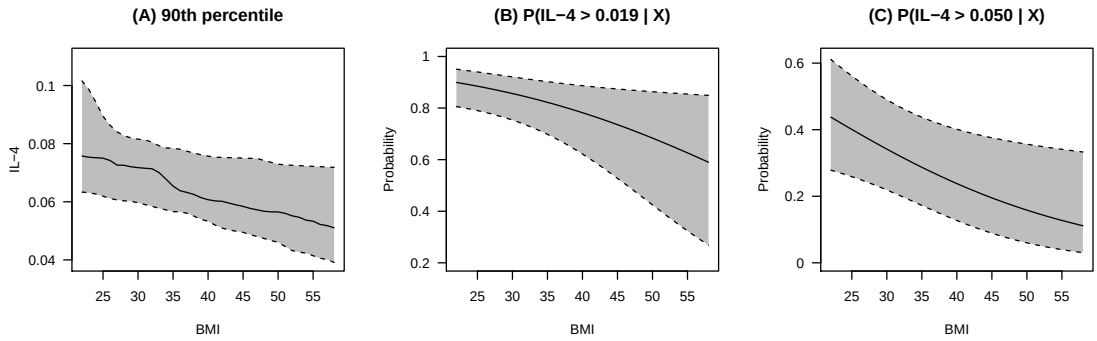

Figure 3.4: We can obtain conditional quantities by CPMs. Other covariates set to corresponding median or mode. (A) The conditional 90th percentile of IL-4 as a function of BMI. (B) The probabilities of IL-4 being greater than 0.019 (the DL). (C) The probabilities of IL-4 being greater than 0.05.

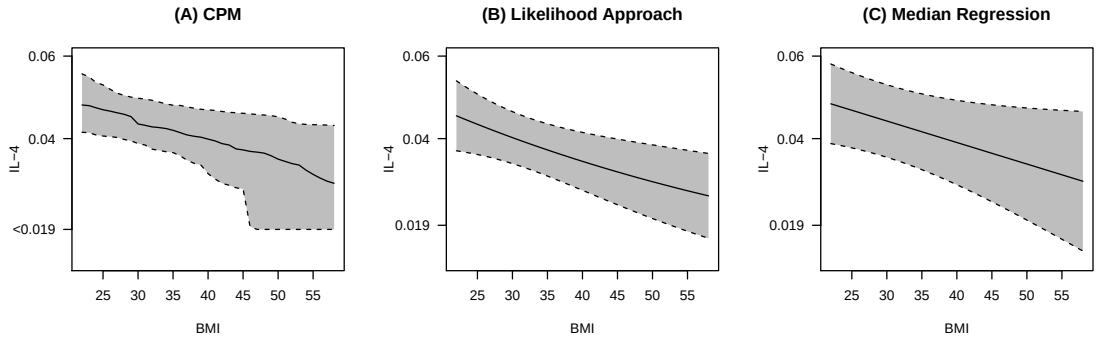

Figure 3.5: (A) The conditional median obtained by CPMs fixing other covariates. (B) The conditional median obtained by the likelihood approach (after log-transformation). (C) The conditional median obtained by median regression.

data, one would then likely log-transform the data and fit a linear regression model. The result can vary depending on the choice of the imputed number: if one imputes with the DL itself (0.019) vs. 0.001, the log-transformed IL-4 is estimated to decrease 0.013 pg/ml vs. 0.032 pg/ml, respectively, per 5 kg/m^2 increase in BMI, with different statistical significance (p-value 0.020 vs. 0.073). (ii) A more sophisticated approach might be to assume the data are log-normally distributed and perform a likelihood-based analysis, which results in an estimated change on the log-scale of -0.015 pg/ml per 5 kg/m^2 BMI increase (p-value 0.018). The conditional median as a function of BMI could also be extracted from this analysis, and is in Figure 3.5(B). The curve of conditional median as a function of BMI is similar to what was estimated with the CPM (Figure 3.5(A)), but it is slightly lower and its confidence bands are tighter than those of the CPM. The tighter bands reflect the parametric assumption that the data are truly log-normally distributed. In contrast, the CPM does not require transformation of the data, and it non-parametrically estimates the best transformation. (iii) One could also directly estimate the conditional median as a function of BMI using quantile regression (Koenker and Hallock, 2001). This estimated curve is in Figure 3.5(C), which closely matches that estimated from the CPM. However, the confidence bands for median regression are wider than those of the CPM, and the 95% CI for the slope contains 0. One could argue that the CPM is assuming more than median regression (which only assumes a linear relationship between the median on the original outcome scale and the covariates); hence the narrower confidence bands. However, the CPM is able to yield several additional quantities (e.g., other quantiles, odds ratios, excedence probabilities) from a single model that cannot be obtained from median regression. Also, the confidence bands obtained by the CPM do not go below the DL. (iv) Finally, one could dichotomize IL-4 into "undetectable" and "detectable" and fit a logistic regression model. However, logistic regression was not able to provide stable estimation for this dichotomization. One could consider other dichotomizations, but the choice is arbitrary. In fact, a beta coefficient in the CPM can be thought of as a weighted average of the log-odds ratios for logistic regression models that consider all possible orderable dichotomizations of the outcome.

3.3.2 Multiple Detection Limits

We illustrate our approach to handle multiple DLs with data from a multi-center HIV study. The data include 5301 adults living with HIV starting antiretroviral therapy (ART) at one of 5 study centers in Latin America between 2000 and 2018. Viral load (VL) measures the amount of virus circulating in a person with HIV. A high VL after ART initiation may indicate non-adherence or an ineffective ART regimen that should be switched. We study the association between VL at approximately 6 months after ART initiation and variables measured at ART initiation (baseline). The DLs for the outcome VL differed by site and calendar time. Figure 3.6 shows the most frequent lower DL values for each year and at each site. There are five distinct

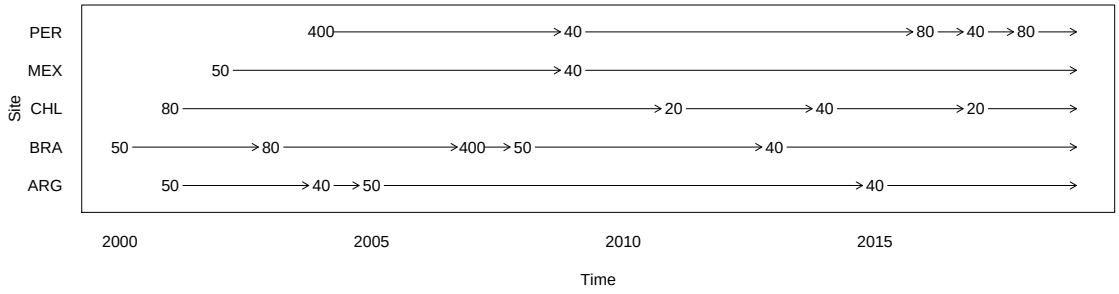

Figure 3.6: The changes of most frequent DL values every year at each study site over time.

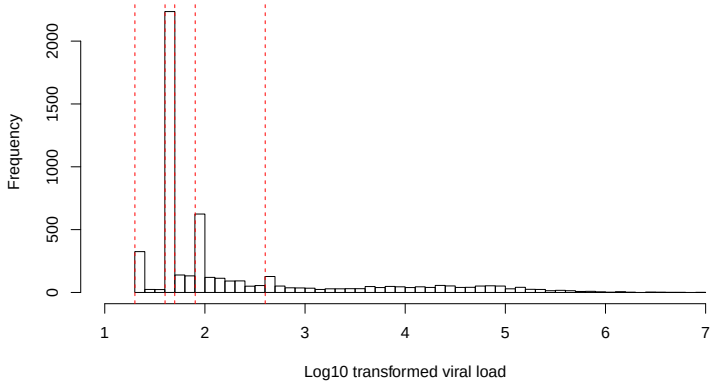

Figure 3.7: Distribution of the log10 transformed 6-month VL. The transformed DLs are shown in red dashed lines; 2992 (56%) of the patients were censored at one of these DLs.

lower DLs in this database: 20, 40, 50, 80, and 400 copies/mL. A total of 2992 (56%) patients had 6-month VL censored at a DL: 45%, 54%, 52%, 65%, and 57% at study sites in Argentina, Brazil, Chile, Mexico, and Peru, respectively. Figure 3.7 shows the distribution of log-transformed 6-month VL and lower DLs.

A traditional analysis in the HIV literature would dichotomize VL as detectable and undetectable and perform logistic regression (Jiamsakul et al., 2017). There are a few issues that make this analysis less than ideal. First, all VLs above the DL (nearly half of all observations) would be collapsed into a "detectable" category resulting in well-known loss of information due to dichotomizing continuous variables (Fedorov et al., 2009). Second, because the DL varies with time and by site, the analyst is forced to dichotomize at the largest DL (in this case 400 copies/mL) or else perform an analysis where values above the DL at one site are treated differently than they would be treated at another site. For example, a VL of 300 copies/mL measured in Mexico in 2005 would be measured as '<400' that same year in Peru; assigning this value as '<400'

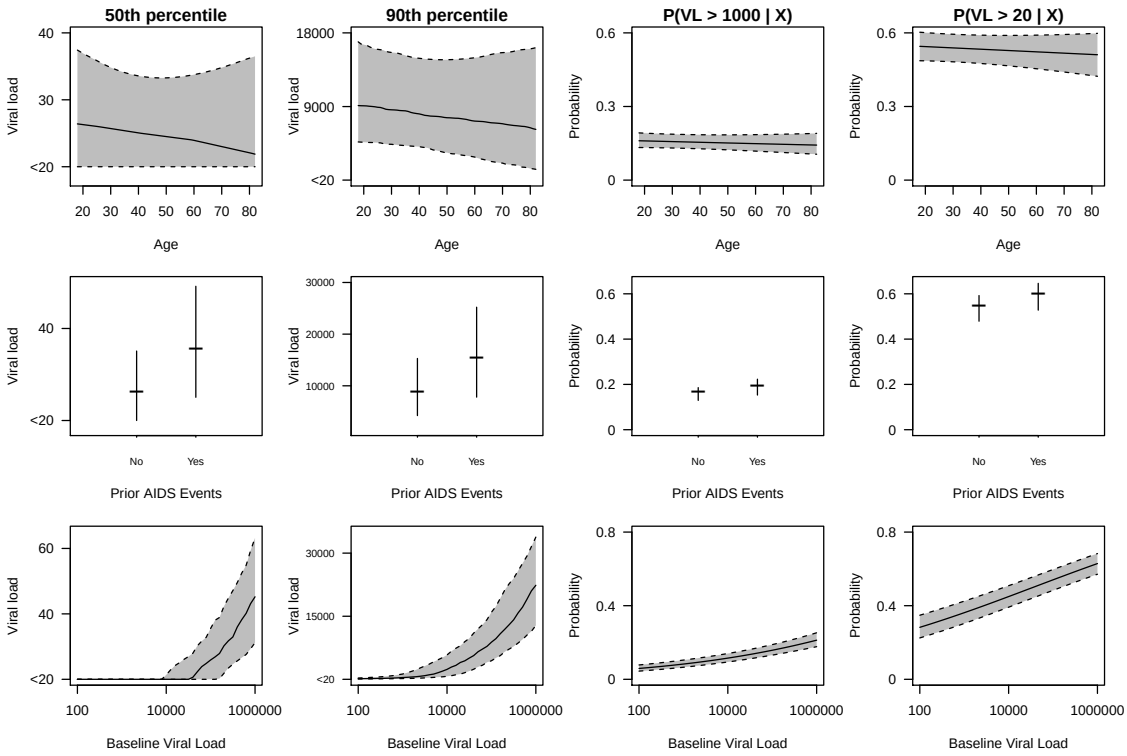

Figure 3.8: The estimated conditional 50th and 90th percentiles of 6-month VL and the conditional probability of 6-month VL being greater than 1000 and 20 as functions of age (top row), prior AIDS events (middle row), and baseline VL (bottom row) while keeping other covariates at their medians (for continuous variables) or modes (for categorical variables) based on our method.

results in lost information but leaving it as "detectable" would make the outcome variable different across time and sites. A more parametric analysis might assume that the VL follow a specified distribution (e.g., log-normal distribution) and fit the censored data likelihood or multiply impute values below the DL from the assumed distribution to obtain estimated regression coefficients. However, distributional assumptions for values below the DL are strong and untestable, and given that over half of the response variables are below the DL, these assumptions would have a large impact on results.

In contrast, the CPM uses all available information (i.e., does not dichotomize the response variable) and makes much weaker assumptions than the fully parametric approaches. Similar to the parametric approaches, the CPM assumes non-informative censoring conditional on covariates (which is reasonable, given that DLs are determined by equipment / assays independent of true values) and that all observations follow a common distribution conditional on covariates, which permits borrowing information across sites and time. Unlike the fully parametric approach, however, the CPM does not fully specify this distribution. Rather, the CPM assumes that response variables follow a linear model with known error distribution after some unspecified transformation.

Table 3.3: The β coefficients in CPMs can be interpreted as log odds ratios. We show the odds ratio (95% confidence interval) and p-value for the predictors included in the model.

Predictor	Odds Ratio (95% CI)	P-value
Age (per 10 years)	0.98 (0.93, 1.03)	0.418
Sex		0.201
Male (reference)	1	
Female	0.90 (0.76, 1.06)	
Study center		<0.001
Peru (reference)	1	
Argentina	1.26 (0.98, 1.61)	
Brazil	1.07 (0.91, 1.26)	
Chile	1.07 (0.90, 1.26)	
Mexico	0.59 (0.49, 0.70)	
Route of infection		<0.001
Homosexual/Bisexual (reference)	1	
Heterosexual	0.96 (0.83, 1.10)	
Other/Unknown	0.79 (0.62, 1.01)	
Prior AIDS event		0.001
No (reference)	1	
Yes	1.24 (1.09, 1.41)	
Baseline CD4 (per 1 square root cells/μL)	1.09 (1.08, 1.10)	<0.001
Baseline VL (per 1 $\log_{10}$ copies/mL)	1.44 (1.34, 1.54)	<0.001
ART regimen		0.034
NNRTI-based (reference)	1	
INSTI-based	0.55 (0.40, 0.75)	
PI-based	1.10 (0.95, 1.29)	
Other	2.57 (1.28, 5.16)	
Months to VL measure	0.95 (0.92, 0.98)	0.002
Calendar year	0.89 (0.88, 0.91)	<0.001

We applied our method in Section 2.3 to fit a CPM of the 6-month VL on baseline variables with the logit link. Results are shown in Table 3.3. With the logit link, the β parameters can be interpreted as log odds ratios and are presented as odds ratios in the table along with 95% CIs. P-values are likelihood ratio test p-values. The results suggest that study center, route of infection, prior AIDS event, baseline CD4 count, baseline VL, ART regimen, the time from ART initiation until the VL measurement, and calendar year are all associated with VL at 6 months. Holding other variables fixed, a 10-fold increase in VL at baseline is associated with a 44% increase in the odds of having a higher VL at 6 months (95% CI 34% to 54%).

Quantiles and cumulative probabilities are also easily extracted from the CPM. The first row of Figure 3.8 are the estimated conditional 50th and 90th percentiles of 6-month VL and the conditional probabilities for 6-month VL being greater than 1000 and 20 copies/mL as functions of age. The plots show that VL at 6 months is fairly similar across age after fixing the other covariates. The smallest DL is 20 copies/mL, and all VL less than this DL belong to the smallest ordered category, which we label as '<20'. The second row of Figure 3.8 contains the estimated conditional quantiles and probabilities as functions of whether a patient had an AIDS event prior to starting ART. People with a prior AIDS event (36%) tended to have a higher VL at 6 months. The third row of Figure 3.8 are the estimated conditional quantiles and probabilities as functions of baseline VL. People with a higher baseline VL tended to have a higher VL at 6 months.

Figure 3.9 and Table 3.4 show the results from a similar CPM, except with continuous covariates expanded using restricted cubic splines to relax linearity assumptions and increase model flexibility. The results are fairly similar. Figure 3.9 plots the estimated conditional quantiles and probabilities of the outcome being above 1000 and 20. The conditional median and its confidence intervals all fall into the smallest level "<20". For age, the general trend is the same as in the model without splines but for patients younger than 40 years old, age increase is correlated with slightly increase in 6-month viral load. The estimated conditional quantiles and probabilities as a function of prior AIDS events shown in the third row of Figure 3.9 are very similar to the results in model without splines. In the second row of Figure 3.9, patients with large baseline viral load values tend to have much higher 6-month viral load measures. Table 3.4 shows the odds ratio and 95% confidence intervals using splines for all predictors. The p-values in the table are based on the likelihood ratio tests.

Table 3.4: Application results for viral load: the estimated odds ratios with 95% confidence intervals, and p-values from the model with splines on continuous covariates.

Predictor	Odds Ratio	95% CI	P-value
Age			0.959
20	1.00	0.82-1.23	
30 (reference)	1		

Continued on next page

Table 3.4 – continued from previous page

Predictor	Odds Ratio	95% CI	P-value
40	1.00	0.87-1.14	
50	0.98	0.84-1.13	
60	0.94	0.76-1.16	
Sex			0.213
Male (reference)	1		
Female	0.90	0.76-1.06	
Study center			<0.001
Peru (reference)	1		
Argentina	1.17	0.91-1.50	
Brazil	0.99	0.83-1.17	
Chile	0.96	0.81-1.15	
Mexico	0.57	0.47-0.68	
Route of infection			0.099
Homosexual/Bisexual (reference)	1		
Heterosexual	0.93	0.81-1.07	
Other/Unknown	0.77	0.60-0.98	
Prior AIDS event			0.075
No (reference)	1		
Yes	1.13	0.99-1.29	
Baseline CD4			<0.001
50	0.66	0.60-0.73	
100	0.65	0.60-0.69	
200 (reference)	1		
300	1.71	1.59-1.85	
Baseline VL			<0.001
100	0.74	0.52-1.04	
1000	0.85	0.73-1.01	
10000 (reference)	1		
100000	1.33	1.18-1.50	
ART regimen			<0.001
NNRTI-based (reference)	1		
INSTI-based	0.60	0.43-0.83	
PI-based	111	0.95-1.29	
Other	2.51	1.26-5.00	
Time			<0.001
3 months	1.57	1.20-2.04	
6 months (reference)	1		
9 months	1.05	0.91-1.22	
Calendar year			<0.001
2005	1.50	1.33-1.68	
2010 (reference)	1		
2015	0.45	0.38-0.52	

For comparisons, we also analyzed the data using competing approaches described earlier. First, we fit logistic regression to 6-month VL values dichotomized as <400 vs. ≥ 400 copies/mL, corresponding to the highest DL. Results are in Table 3.5. The CPM and the logistic regression model gave similar estimates of the beta coefficients (which are log odds ratios), although there were some differences in the estimates and

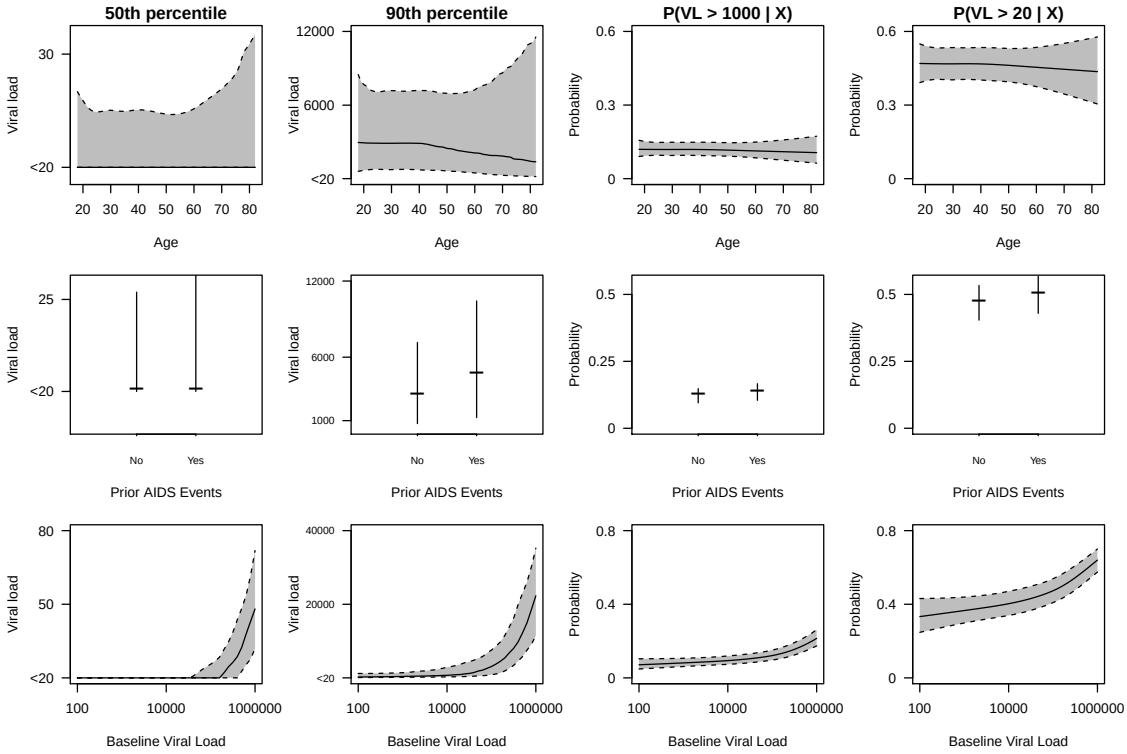

Figure 3.9: The estimated conditional 50th and 90th percentile of age, prior AIDS events, and 6-month viral load and the conditional probability of the 6-month viral load being greater than 1000 and 20 as functions of age, prior AIDS event, and baseline viral load while keeping other covariates at their medians (for continuous variables) or modes (for categorical variables) level based on the model with splines.

Table 3.5: The odds ratio analyzing the data by logistic regression and point estimation obtained by the fully likelihood approach. For logistic regression, the responses are dichotomized into two levels: $<$400 and $\geq$400. We show odds ratios with 95% confidence intervals, and p-values of parameter estimation for predictors included in the model. The point estimation with 95% confidence intervals, and p-values of parameter estimation for predictors included in the model by the fully likelihood approach.

	Logistic Regression		Fully Likelihood Approach	
	Odds Ratio (CI)	P-value	Estimation (CI)	P-value
Age (per 10 years)	0.91 (0.85, 0.98)	0.013	-0.03 (-0.09, 0.03)	0.276
Sex		0.739		0.271
Male (reference)	1		1	
Female	0.96 (0.77, 1.21)	-0.10	(-0.28, 0.08)	
Study center		$<$0.001		$<$0.001
Peru (reference)	1		1	
Argentina	1.43 (1.04, 1.97)		0.26 (-0.01, 0.53)	
Brazil	1.21 (0.97, 1.51)		0.10 (-0.08, 0.28)	
Chile	1.08 (0.85, 1.37)		0.08 (-0.11, 0.27)	
Mexico	0.58 (0.45, 0.75)		-0.54 (-0.73, -0.35)	
Route of infection		0.563		0.197
Homosexual/Bisexual (ref)	1		1	
Heterosexual	0.99 (0.81, 1.21)		-0.04 (-0.20, 0.11)	
Other/Unknown	0.84 (0.61, 1.17)		-0.24 (-0.51, 0.02)	
Prior AIDS event		0.009		$<$0.001
No (reference)	1		1	
Yes	1.27 (1.06, 1.52)		0.25 (0.1, 0.39)	
Baseline CD4 (per 1 square root cells/μL)	1.13 (1.12, 1.15)	$<$0.001	0.10 (0.09, 0.11)	$<$ 0.001
Baseline VL (per 1 $\log_{10}$ copies/mL)	1.04 (0.95, 1.14)	0.357	0.35 (0.27, 0.42)	$<$ 0.001
ART regimen		$<$0.001		$<$0.001
NNRTI-based (reference)	1		1	
INSTI-based	0.44 (0.27, 0.72)		-0.60 (-0.93, -0.28)	
PI-based	0.95 (0.77, 1.16)		0.07 (-0.10, 0.23)	
Other	3.58 (1.49, 8.60)		1.06 (0.30, 1.82)	
Months to VL measure	1.01 (1.00, 1.03)	0.068	-0.03 (-0.06, 0.01)	0.140
Calendar year	0.86 (0.84, 0.87)	$<$0.001	-0.12 (-0.13, -0.10)	$<$0.001

the CIs from CPMs tend to be narrower, as expected. In logistic regression, the log odds ratios are based on the single undetectable vs. detectable dichotomization, while those in CPMs are based on dichotomizations at each response value. Second, we fit a full likelihood based model assuming the outcome variable was normally distributed after $\log_{10}(\cdot)$ transformation. Note that even the $\log_{10}$-transformed 6-month VL were still quite skewed (shown in Figure 3.7), and hence the assumptions of this fully parametric approach were questionable. The parameters in this approach and those from the CPM are not directly comparable because they are on different scales, however, the directions of associations were similar.

3.4 Simulations

Extensive simulations of CPMs with continuous data have been reported elsewhere (Liu et al., 2017; Tian et al., 2020). Here we present a limited set of simulations investigating the performance of CPMs with data subject to single and multiple DLs.

3.4.1 Single Detection Limits

Data were generated for sample sizes of $n = 100$ and $n = 500$ such that the outcome Y followed a normal linear model after log-transformation in the following manner:

$$Y = \exp(Y^*), \text{ where } Y^* = X\beta + \varepsilon, \beta = 1, X \sim N(0,1), \text{ and } \varepsilon \sim N(0,1).$$

Various scenarios of DLs of Y were considered: 1. No DL. 2. One lower DL at 0.25 (censoring rate 16.3%). 3. One upper DL at 4 (censoring rate 16.3%). 4. One lower DL at 0.25 and one upper DL at 4 (censoring rate 32.7%). 5. One lower DL at 4 (censoring rate 83.7%). In addition, we considered a scenario with a more complicated transformation: 6. One lower DL at 0.0625 and

$$Y = \begin{cases} \exp(2Y^*) \text{ if } Y^* < \log(0.25) \\ \sqrt{\exp(Y^*)} \text{ if } \log(0.25) \leq Y^* < \log(2) \\ \exp(Y^*) \text{ if } Y^* \geq \log(2). \end{cases}$$

Note that the Y in scenario 6 is a monotonic transformation of that in scenario 2 with exactly the same censoring rate.

CPMs were fit to the observed data $\{X, Y\}$ without any knowledge of the correct transformation or Y^*. We simulated 1,000 replications under each scenario. Percent bias, root mean squared error (RMSE), and coverage of 95% CIs were estimated with respect to β, conditional medians for $X = \{0,1\}$, and conditional CDFs at $y = 1.5$ for $X = \{0,1\}$.

Table 3.6 shows results under correctly specified models (i.e., probit link function and X correctly included). CPMs resulted in nearly unbiased estimation and good CI coverage. As the sample size increased, both the bias and RMSE decreased. Note that estimation of the condition medians was "perfect" in scenario 5 because the true conditional medians were below the lower DL due to the high censoring rate and the estimated conditional medians were always '<4', the lowest outcome category corresponding to below the DL. The estimate of β was more variable in scenario 5 because of the high censoring rate. The estimation of β in scenario 6, where data were generated from the complicated transformation, was exactly the same as that in scenario 2 because the same seed was used in all simulation scenarios and the order information above

the DL was identical between scenarios 2 and 6. However, the conditional medians and CDFs depend on the scale of the outcome, and their estimates differed between scenarios 2 and 6.

Table 3.7 shows results under scenario 2 with $n = 1000$ comparing CPMs with some widely used methods for handling DLs, specifically single imputation with $l/2$, single imputation with $l/\sqrt{2}$, multiple imputation, and fully parametric maximum likelihood estimation (MLE). For all non-CPM approaches, we first correctly assumed that the outcome variable followed a log-normal distribution. With the imputation approaches, unobserved values were imputed, then a linear regression model was fit on the log-transformed outcome to obtain the β estimate, and median regression was used to estimate conditional medians. In multiple imputation, the correct tail distribution was used for imputing data and 10 iterations were performed for each data set. As expected, the MLE performed the best with the lowest bias and RMSE, and highest efficiency because the distributional assumptions matched the true distribution. The performance of multiple imputation was similar to that of the MLE, but with higher RMSE. As a semiparametric method, the CPM also resulted in minimal bias and correct coverage, but had slightly larger variance and RMSE. In contrast, the single imputation estimators were biased and tended to have poor coverage, especially for estimating β. We also evaluated the comparator methods under misspecification of the transformation. We simulated datasets with $X \sim N(5,1)$, $Y^* = X\beta + \varepsilon$, $\beta = 1$, $\varepsilon \sim N(0,1)$, $Y = Y^{*2}$, $n = 1000$, and $l = 13.12$ (approximately 17% censored). The non-CPM approaches assumed a normal linear model after an incorrectly specified log-transformation. As shown in the bottom half of Table 3.7, only the CPM was able to properly estimate β and the conditional medians, because pre-transformation and strict distributional assumptions are not needed for fitting CPMs.

Finally, the Table 3.8 shows the performance of CPMs for the data generated in scenario 2 under moderate and severe link function misspecification (i.e., fitting CPMs with logit and loglog link functions, respectively). Link function misspecification is equivalent to misspecification of the distribution of ε because $F_\varepsilon = G^{-1}$. The performance of CPMs was reasonable with moderate link function misspecification with bias under 6% and coverage of 95% CI close to 0.95 with $n = 100$, although as low as 0.91 with $n = 500$. With severe link function misspecification, performance of CPMs was noticeably worse, with bias as high as 12% and coverage as low as 0.60 for the conditional median at $X = 1$.

3.4.2 Multiple Detection Limits

To illustrate the use of CPMs with multiple detection limits, we simulated data from 3 study sites. The data were generated in a similar way as in Section 4.1, but different DLs were applied at different sites and the distribution of the covariate X was allowed to vary across sites in some scenarios. Specifically, we considered the following 5 scenarios:

Table 3.6: Simulation results for single DLs

		n=100			n=500		
Parameter	Truth	Bias(%)	RMSE	Coverage	Bias(%)	RMSE	Coverage
Scenario 1							
β	1	2.803	0.133	0.944	0.638	0.057	0.945
$Q(0.5\|X=0)$	1	-0.388	0.140	0.951	-0.124	0.063	0.951
$Q(0.5\|X=1)$	2.718	1.552	0.494	0.949	0.321	0.218	0.951
$F(1.5\|X=0)$	0.658	0.117	0.054	0.949	0.059	0.024	0.951
$F(1.5\|X=1)$	0.276	-1.429	0.060	0.949	-0.383	0.026	0.945
Scenario 2							
β	1	2.665	0.138	0.945	0.585	0.057	0.948
$Q(0.5\|X=0)$	1	-0.240	0.142	0.953	0.028	0.063	0.948
$Q(0.5\|X=1)$	2.718	1.445	0.498	0.953	0.406	0.222	0.946
$F(1.5\|X=0)$	0.658	0.005	0.054	0.946	-0.085	0.024	0.950
$F(1.5\|X=1)$	0.276	-0.479	0.061	0.948	0.368	0.028	0.943
Scenario 3							
β	1	2.710	0.139	0.943	0.581	0.058	0.948
$Q(0.5\|X=0)$	1	-0.460	0.141	0.951	-0.020	0.063	0.949
$Q(0.5\|X=1)$	2.718	0.803	0.477	0.954	0.310	0.223	0.945
$F(1.5\|X=0)$	0.658	0.0147	0.054	0.946	-0.083	0.024	0.951
$F(1.5\|X=1)$	0.276	-0.487	0.062	0.948	0.381	0.028	0.941
Scenario 4							
β	1	2.544	0.139	0.945	0.538	0.058	0.951
$Q(0.5\|X=0)$	1	-0.243	0.141	0.954	0.028	0.063	0.949
$Q(0.5\|X=1)$	2.718	1.017	0.477	0.953	0.358	0.223	0.947
$F(1.5\|X=0)$	0.658	0.004	0.054	0.947	-0.086	0.024	0.950
$F(1.5\|X=1)$	0.276	-0.285	0.062	0.948	0.432	0.028	0.943
Scenario 5							
β	1	7.315	0.276	0.946	1.330	0.101	0.948
$Q(0.5\|X=0)$	1	0*	0	1	0	0	1
$Q(0.5\|X=1)$	2.718	0	0	1	0	0	1
$F(1.5\|X=0)$	0.658	0.183	0.026	0.954	-0.029	0.010	0.949
$F(1.5\|X=0)$	0.276	-0.189	0.069	0.952	-0.169	0.030	0.949
Scenario 6							
β	1	2.665	0.138	0.945	0.585	0.057	0.948
$Q(0.5\|X=0)$	1	-0.841	0.071	0.951	-0.503	0.032	0.945
$Q(0.5\|X=1)$	0.368	-0.312	0.542	0.953	-0.529	0.222	0.946
$F(1.5\|X=0)$	0.654	0.254	0.048	0.947	0.056	0.022	0.953
$F(1.5\|X=1)$	0.500	0.061	0.069	0.949	0.536	0.032	0.946

The results of zero bias and RMSE when there is a high censoring rate are because the true conditional medians are below the lower DL and the estimated conditional medians were always '<4', the lowest outcome category corresponding to below the DL.

Table 3.7: Comparison of methods under correct and incorrect model specifications

Method	Truth	Bias(%)	Empirical SE	RMSE	Coverage
Correct model specification					
CPM					
β	1	0.258	0.040	0.040	0.951
$Q(0.5\|X=0)$	1	0.226	0.045	0.045	0.953
$Q(0.5\|X=1)$	2.718	0.490	0.155	0.156	0.949
Single imputation with $dl/\sqrt{2}$					
β	1	-10.294	0.028	0.107	0.057
$Q(0.5\|X=0)$	1	3.614	0.039	0.053	0.859
$Q(0.5\|X=1)$	2.718	-4.580	0.144	0.190	0.883
Single imputation with $dl/2$					
β	1	-4.247	0.030	0.052	0.732
$Q(0.5\|X=0)$	1	0.620	0.039	0.040	0.955
$Q(0.5\|X=1)$	2.718	-1.700	0.147	0.155	0.949
Multiple imputation					
β	1	0.269	0.035	0.036	0.964
$Q(0.5\|X=0)$	1	1.341	0.039	0.041	0.947
$Q(0.5\|X=1)$	2.718	-1.576	0.148	0.154	0.945
MLE					
β	1	0.031	0.035	0.035	0.952
$Q(0.5\|X=0)$	1	-0.006	0.032	0.032	0.952
$Q(0.5\|X=1)$	2.718	0.075	0.123	0.123	0.952
Model misspecification					
CPM					
β	1	0.257	0.040	0.040	0.951
$Q(0.5\|X=0)$	24.997	0.070	0.444	0.444	0.953
$Q(0.5\|X=1)$	35.996	0.127	0.680	0.681	0.949
Single imputation with $dl/\sqrt{2}$					
β	1	-61.821	0.011	0.618	0.000
$Q(0.5\|X=0)$	24.997	-3.568	0.389	0.973	0.405
$Q(0.5\|X=1)$	35.996	-0.818	0.665	0.727	0.935
Single imputation with $dl/2$					
β	1	-55.739	0.014	0.558	0.000
$Q(0.5\|X=0)$	24.997	-4.833	0.445	1.287	0.262
$Q(0.5\|X=1)$	35.996	0.240	0.698	0.703	0.887
Multiple imputation					
β	1	-59.690	0.016	0.597	0.000
$Q(0.5\|X=0)$	25.000	-3.040	0.382	0.851	1.000
$Q(0.5\|X=1)$	35.996	-1.014	0.660	0.754	1.000
MLE					
β	1	-62.108	0.012	0.621	0.000
$Q(0.5\|X=0)$	24.997	-4.838	0.303	1.247	0.021
$Q(0.5\|X=1)$	35.996	-3.468	0.535	1.358	0.451

Table 3.8: Simulation results for model misspecification with one lower DL at 0.25

		n=100			n=500		
Parameter	Truth	Bias(%)	RMSE	Coverage	Bias(%)	RMSE	Coverage
Logit Link							
$Q(0.5\|X=0)$	1	0.006	0.142	0.949	0.287	0.064	0.942
$Q(0.5\|X=1)$	2.718	2.827	0.514	0.947	1.965	0.233	0.936
$F(1.5\|X=0)$	0.658	0.923	0.057	0.944	0.869	0.024	0.940
$F(1.5\|X=1)$	0.276	-5.221	0.062	0.939	-4.862	0.030	0.912
Loglog Link							
$Q(0.5\|X=0)$	1	-4.116	0.232	0.953	-3.241	0.042	0.932
$Q(0.5\|X=1)$	2.718	-10.673	0.491	0.887	-11.861	0.192	0.603
$F(1.5\|X=0)$	0.658	-0.4602	0.045	0.961	-1.081	0.020	0.947
$F(1.5\|X=1)$	0.276	8.092	0.076	0.900	10.164	0.010	0.809

1. Lower DLs 0.16, 0.30, and 0.50 for the 3 sites (about 10%, 20%, and 30% censored), and X is independent of DLs/sites.

2. Upper DLs 0.16, 0.30, and 0.50 for the 3 sites (about 90%, 80%, and 70% censored), and X is independent of DLs/sites.

3. Lower DLs 0.16, 0.30, and 0.50 for the 3 sites (about 17%, 20%, and 20% censored), and $X \sim N(\mu_x, 1)$ where $\mu_x = -0.5, 0$, and 0.5 for site 1, 2, and 3, respectively.

4. Upper DLs 0.16, 0.30, and 0.50 for the 3 sites (about 83%, 80%, and 80% censored), and $X \sim N(\mu_x, 1)$ where $\mu_x = -0.5, 0$, and 0.5 for site 1, 2, and 3, respectively.

5. Lower DLs 0.2, 0.3, and -∞ (13%, 20%, and 0% censored) and upper DLs at ∞, 4, and 3.5 (0%, 19%, and 16% censored) for the 3 sites, and X is independent of DLs/sites.

We considered two sample sizes, $n = 150$ and $n = 900$, with the samples sizes distributed equally across sites. In scenarios 2 and 4, because of the high censoring rates, we estimated the quantiles at $p = 0.03$ (i.e., 3rd percentile) and CDFs at $y = 0.05$. Results from fitting the CPM based on 10,000 replications are shown Table 3.9. In summary, estimates had very low bias and confidence intervals had proper coverage in all simulation scenarios.

3.5 Discussion

In this paper, we have described an approach to address detection limits in response variables using CPMs. CPMs have several advantages over existing methods for addressing DLs. They make minimal distributional assumptions, they yield interpretable parameters, and they are invariant to the value assigned to measures outside DLs. Any values outside the lowest/highest DLs are simply assigned to the lowest/highest ordinal

Table 3.9: Simulation results for multiple DLs

	n=50 for each site				n=300 for each site			
Parameter	Bias(%)	Bias	RMSE	Coverage	Bias(%)	Bias	RMSE	Coverage
Scenario 1								
β	1.871	0.019	0.111	0.948	0.249	0.003	0.044	0.948
$Q(0.5\|X=0)$	-0.071	-0.001	0.12	0.948	0.001	-0.001	0.047	0.953
$Q(0.5\|X=1)$	1.137	0.031	0.404	0.952	0.042	0.001	0.164	0.947
$F(1.5\|X=0)$	0.258	0.002	0.044	0.950	0.134	0.001	0.017	0.954
$F(1.5\|X=1)$	-1.115	-0.003	0.050	0.950	-0.435	-0.001	0.020	0.952
Scenario 2								
β	3.706	0.037	0.186	0.948	0.455	0.005	0.068	0.954
$Q(0.03\|X=0)$	3.440	0.005	0.037	0.953	0.152	0.000	0.014	0.952
$Q(0.03\|X=1)$	-1.952	-0.008	0.071	0.953	1.795	0.007	0.046	0.9545
$F(0.05\|X=0)$	11.276	0.000	0.000	0.974	-2.917	0.000	0.000	0.953
$F(0.05\|X=1)$	79.059	0.000	0.000	0.961	-3.762	0.000	0.000	0.944
Scenario 3								
β	1.891	0.019	0.106	0.958	0.248	0.003	0.041	0.962
$Q(0.5\|X=0)$	-0.202	-0.002	0.117	0.964	-0.092	-0.001	0.047	0.969
$Q(0.5\|X=1)$	0.939	0.026	0.395	0.950	0.041	0.001	0.160	0.949
$F(1.5\|X=0)$	0.223	0.002	0.045	0.963	0.159	0.001	0.017	0.962
$F(1.5\|X=1)$	-1.221	-0.003	0.050	0.948	-0.383	-0.001	0.020	0.955
Scenario 4								
β	1.926	0.019	0.175	0.945	0.206	0.002	0.065	0.954
$Q(0.03\|X=0)$	4.032	0.006	0.039	0.955	0.332	0.001	0.014	0.951
$Q(0.03\|X=1)$	7.982	0.033	0.125	0.950	2.096	0.009	0.049	0.950
$F(0.05\|X=0)$	21.028	0.000	0.000	0.965	-1.349	0.000	0.000	0.952
$F(0.05\|X=1)$	110.371	0.000	0.000	0.956	-0.940	0.000	0.028	0.943
Scenario 5								
β	1.838	0.018	0.111	0.957	0.250	0.003	0.044	0.960
$Q(0.5\|X=0)$	-0.046	-0.001	0.115	0.948	0.440	0.004	0.047	0.951
$Q(0.5\|X=1)$	2.019	0.054	0.412	0.968	0.171	0.005	0.165	0.963
$F(1.5\|X=0)$	0.406	0.003	0.04	0.945	-0.059	-0.000	0.229	0.960
$F(1.5\|X=1)$	-0.059	-0.000	0.050	0.960	-0.319	-0.001	0.020	0.961

The percent bias in scenario 4 is relatively high due to the small true values.

categories, and estimation proceeds naturally. CPMs are also easily extended to handle multiple DLs. From simulation studies we saw that CPMs performed well, even with high censoring rates and relatively small sample sizes. We also illustrated the use of CPMs with two quite different HIV datasets with censored response data. Similar datasets with limits of detection are quite common in biomedical research; the CPM is an effective analysis tool in these settings.

CPMs have some limitations. Although CPMs do not make distributional assumptions on the response variable, the link function must still be specified, which corresponds to making an assumption on the distribution of the response variable after an unspecified transformation. Performance can be poor with severe link function misspecification; however, CPMs appear to be fairly robust to moderate misspecification. In addition, because we do not make distributional assumptions outside DLs, we are not able to estimate conditional expectations after fitting a CPM; however, with DLs, conditional quantiles are probably more reasonable statistics to report anyway.

Further research could consider extensions of CPMs to handle clustered or longitudinal data with DLs. It may be of interest to study the use of these models with right-censored failure time data (i.e., survival data), where each observation is potentially subject to a different censoring time; the current manuscript only considered situations with a relatively small number of potential censoring times (i.e., upper DLs).

3.6 Supplementary Material

Let $a_1 < \cdots < a_J$ be the unique outcome values in a data set. In a CPM,

$$G[\Pr(Y \le a_j|X,\alpha,\beta)] = \alpha_j - \beta^T X \quad (j = 1,...,J-1).$$

For observation i with $X_i = x_i$, let $\gamma_{i,j} = \Pr(Y_i \le a_j|x_i,\alpha,\beta)$. For convenience, let $\gamma_{i,0} = 0$ and $\gamma_{i,J} = 1$. Let $p_{i,j} = \gamma_{i,j} - \gamma_{i,j-1}$ be the multinomial probability at a_j ($j = 1,\ldots,J$).

Consider the logit link function, $G(p) = \log\frac{p}{1-p}$. Then for $j = 1,\ldots,J-1$, $\gamma_{i,j} = \frac{e^{\phi_{i,j}}}{1+e^{\phi_{i,j}}}$, where $\phi_{i,j} = \alpha_j - \beta^T x_i$, and $\frac{\partial\gamma_{i,j}}{\partial\beta} = \frac{\partial\gamma_{i,j}}{\partial\phi_{i,j}}\frac{\partial\phi_{i,j}}{\partial\beta} = -x_i\gamma_{i,j}(1-\gamma_{i,j})$. Note that this equation also holds for $j = 0$ and $j = J$. Then

$$\begin{aligned}
\frac{\partial p_{i,j}}{\partial\beta} = \frac{\partial\gamma_{i,j}}{\partial\beta} - \frac{\partial\gamma_{i,j-1}}{\partial\beta} &= -x_i[\gamma_{i,j}(1-\gamma_{i,j}) - \gamma_{i,j-1}(1-\gamma_{i,j-1})] \\
&= -x_i[(\gamma_{i,j-1}+p_{i,j})(1-\gamma_{i,j}) - \gamma_{i,j-1}(1-\gamma_{i,j}+p_{i,j})] \\
&= -x_i[p_{i,j}(1-\gamma_{i,j}) - \gamma_{i,j-1}p_{i,j}] \\
&= x_i p_{i,j}(\gamma_{i,j}+\gamma_{i,j-1}-1).
\end{aligned}$$

Given data $\{(x_i,y_i)\}$, the log-likelihood is $l(\alpha,\beta) = \sum_i \log p_{i,j(i)}$, where $j(i)$ is the index such that $a_{j(i)} = y_i$. The score function with respect to β is

$$\frac{\partial l}{\partial\beta} = \sum_i \frac{1}{p_{i,j(i)}}\frac{\partial p_{i,j(i)}}{\partial\beta} = \sum_i x_i(\gamma_{i,j(i)} + \gamma_{i,j(i)-1} - 1).$$

Now consider the situation where there is a single binary X, coded as 0 and 1. Let n_k be the number of observations with $X = k$ ($k = 0,1$), and $n = n_0 + n_1$. Under the null of $\beta = 0$, the CPM estimate of α_j is $\hat\alpha_j = G(\hat P_j)$, where $\hat P_j$ is the CDF of the empirical distribution of $\{y_i\}$ at a_j. As a result, $\hat\gamma_{i,j(i)} = \hat P_{j(i)}$. In this situation, the numerator of the score test statistic is

$$S = \left.\frac{\partial l}{\partial\beta}\right|_{\hat\alpha|\beta=0} = \sum_{i:x_i=1}(\hat P_{j(i)} + \hat P_{j(i)-1} - 1) = \sum_{i:x_i=1}(\hat P_{j(i)} + \hat P_{j(i)-1}) - n_1.$$

Let k_j be the number of observations with $y_i = a_j$, R_j be the midrank for a_j, and $R_1 = \sum_{i:x_i=1} R_{j(i)}$ be the sum of the midranks for the observations with $X = 1$. Then $R_j = n\hat P_j - \frac{k_j-1}{2} = n\hat P_{j-1} + \frac{k_j+1}{2}$. Since $\frac{n}{2}(\hat P_{j(i)} + \hat P_{j(i)-1}) = \frac{1}{2}(R_{j(i)} + \frac{k_j-1}{2} + R_{j(i)} - \frac{k_j+1}{2}) = R_{j(i)} - \frac{1}{2}$, we have

$$\frac{n}{2}S = R_1 - \frac{n_1}{2} - \frac{nn_1}{2} = R_1 - \frac{n_1(n+1)}{2}.$$

Note that $(R_1) = \frac{n_1(n+1)}{2}$ under the null of $\beta = 0$.

In comparison, the Wilcoxon–Mann–Whitney statistic is $U = R_1 - \frac{n_1(n_1+1)}{2}$, which has mean $\mu_U = \frac{n_1 n_0}{2}$. The corresponding test statistic, $z = \frac{U-\mu_U}{\sigma_U}$, has numerator $U - \mu_U = R_1 - \frac{n_1(n+1)}{2}$. Thus the numerator of the Wilcoxon test statistic differs from that of the score test statistic by a constant multiplier $\frac{n}{2}$.

CHAPTER 4

Analyzing Clustered Continuous Response Variables with Ordinal Regression Models

4.1 Introduction

Analyses of quantitative response variables are often challenged by distributions that do not follow standard parametric assumptions. A common approach in such settings is to transform the response variables so that model assumptions are satisfied. However, response transformations are often ad hoc and parameters associated with the models can be difficult to interpret on their natural, untransformed scale. For example, numerous studies of participants living with HIV model associations with CD4:CD8 ratio, a biomarker that measures the strength of an individual's immune system. CD4:CD8 ratio tends to be right-skewed (Figure 4.1), and there is no standard accepted transformation. Researchers have analyzed CD4:CD8 ratio with no transformation (Castilho et al., 2016), log-transformation (Sauter et al., 2016), square-root transformation (da Silva et al., 2018), fifth-root transformation (Gras et al., 2019), and various categorizations (Petoumenos et al., 2017; Serrano-Villar et al., 2017). Finding the appropriate transformation can be challenging and results may be sensitive to the choice of transformation.

A compelling approach to tackling the challenges associated with non-standard response distribution modeling is to treat continuous response variables as if they were ordinal using cumulative probability models (CPMs), also known as cumulative link models (Liu et al., 2017). The CPM is a semi-parametric linear transformation model (Zeng and Lin, 2006) that assumes the response variable follows a linear model following an unspecified transformation is applied. Rather than making an assumption about the appropriate transformation to apply, CPM fitting uses the data to estimate the transformation non-parametrically by a step function. The CPM is invariant to any monotonic transformation of the response variable because only order information is incorporated in regression parameter estimation. Therefore, no pre-transformation of the response variable is needed. Regression parameters from CPMs are interpretable, and because the cumulative distribution function (CDF) is modeled, conditional (on covariates) means and quantiles can be extracted from the CPM fit. The use of CPMs for cross-sectional continuous response variables, even with thousands of unique outcomes, is computationally feasible with applications of sparse matrix calculations and it has been implemented in Harrell's `orm()` function in the **rms** R package (Harrell, 2020).

Clustered continuous data are common in practice and important for studying associations over time. The generalized estimating equation (GEE) procedure proposed by Liang and Zeger (1986) and Zeger and Liang (1986) estimates marginal regression parameters for clustered responses. GEE methods extend generalized

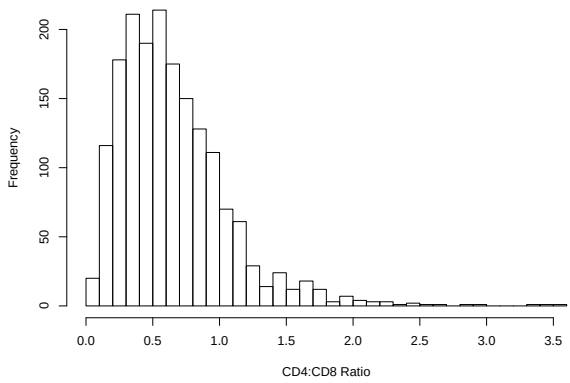

Figure 4.1: The histogram of CD4:CD8 ratio measured at first follow-up visit for people living with HIV and on antiretroviral therapy for a year with a suppressed viral load at the Vanderbilt Comprehensive Care Clinic (VCCC) between 1998 and 2012.

linear models (GLMs) and quasi-likelihood methods to correlated data. Even though valid inferences are possible with GEE when second and higher order moments are misspecified, GEE for correlated is challenged by non-standard distributions in the same way linear regression is for scalar response data. Inspired by Liu et al. (2017), in this paper, we discuss CPMs for clustered continuous response variables again to avoid specifying a transformation. Specifically, we demonstrate that 1) CPMs can be fit to correlated data using GEE for ordinal data, and 2) GEE for ordinal data can be applied to non-standard, quantitative response distributions. We will propose two practical approaches to fit ordinal GEE models to continuous data, depending on the specified GEE working correlation structure. With our approach, which requires no transformation of the response variable, we can obtain parameter and CDF estimates, from which estimates of the mean and quantiles as functions of covariates can be derived. We are unaware of existing methods and software that implement ordinal data GEE to settings with large numbers (i.e., hundreds or thousands) of unique levels.

In Section 2, we review CPMs for cross-sectional continuous response variables. In Section 3, we demonstrate how CPMs for clustered data can be fit using GEE for ordinal response variables, and we propose practical estimation techniques. We illustrate the performance of the methods by simulation in Section 4. In Section 5, we apply our methods to data from two studies. The first investigates predictors of CD4:CD8 ratio in a longitudinal cohort of people living with HIV. The second evaluates the genetic contribution of a single nucleotide polymorphism to lung function decline in a cohort of smokers with chronic obstructive pulmonary disease (COPD). Finally, we discuss strengths and limitations of the proposed methods and potential future directions in Section 6.

4.2 Review of Methods

Liu et al. (2017) proposed fitting cross-sectional continuous response variables using CPMs. Let Y be a continuous response variable, and $Y^* = h(Y)$ be some transformation of Y, where $h(\cdot)$ is an unknown monotonic function. Let $\boldsymbol{X}$ be a vector of covariates and ε be the error term. We assume the relationship between the transformed variable and covariates is linear $Y^* = \boldsymbol{\beta}^T\boldsymbol{X} + \varepsilon$, where ε follows a known distribution F_ε and $\boldsymbol{\beta}$ is a vector of regression parameters. Then,

$$Y = h^{-1}(Y^*) = h^{-1}(\boldsymbol{\beta}^T\boldsymbol{X} + \varepsilon). \tag{4.1}$$

Let $G = F_\varepsilon^{-1}$ be a link function corresponding to the distribution of ε. (4.1) can be expressed as a CPM:

$$\begin{aligned} F(y|\boldsymbol{X}) &= P(Y \le y|\boldsymbol{X}) \\ &= P\left(h^{-1}(\boldsymbol{\beta}^T\boldsymbol{X} + \varepsilon) \le y|\boldsymbol{X}\right) \\ &= P\left(\varepsilon \le h(y) - \boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{X}\right) \\ &= F_\varepsilon\left(h(y) - \boldsymbol{\beta}^T\boldsymbol{X}\right), \text{ which implies} \\ G[F(y|\boldsymbol{X})] &= h(y) - \boldsymbol{\beta}^T\boldsymbol{X}. \end{aligned}$$

The intercept function $h(y) = G[F(y|\boldsymbol{X} = \boldsymbol{0})]$ represents the reference distribution of the link function transformed CDF when $\boldsymbol{X} = \boldsymbol{0}$, and $\boldsymbol{\beta}^T\boldsymbol{X}$ represents shifts in this that depend on the values of $\boldsymbol{X}$.

Assume there are N i.i.d subjects. Denote $y_{(j)}$ as the jth smallest observed response value ($j = 1, \ldots, J$). Rather than specifying a function form for $h(\cdot)$, we can estimate it using a step function with $\gamma_j = h(y_{(j)})$. Such a model, where $h(\cdot)$ is estimated nonparametrically, is referred to as a semi-parametric linear transformation model (Zeng and Lin, 2006). For each $\{y_i, i = 1, \ldots, N\}$, we have the CPM

$$G[F(y_i|\boldsymbol{x}_i)] = G[F(y_{(j)}|\boldsymbol{x}_i)] = \gamma_j - \boldsymbol{\beta}^T\boldsymbol{x}_i. \tag{4.2}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})^T$. Then are able to identify the likelihood with

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i:y_i = y_{(j)}} \left[F(y_i|\boldsymbol{x}_i) - F(y_i^-|\boldsymbol{x}_i)\right], \tag{4.3}$$

where $F(y_i^-|\boldsymbol{x}_i) = \lim_{t \uparrow y_i} F(t|\boldsymbol{x}_i)$. The "nonparametric" likelihood can be obtained by substituting $F(y_i^-|\boldsymbol{x}_i) =$

$F(y_{(j)}^{-}|\boldsymbol{x}_i)$ with $F(y_{(j-1)}|\boldsymbol{x}_i)$ as follows

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i:y_i=y_{(j)}} \left[G^{-1}(\gamma_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - G^{-1}(\gamma_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) \right], \tag{4.4}$$

where $-\infty \equiv \gamma_0 < \gamma_1 < \cdots < \gamma_{J-1} < \gamma_J \equiv \infty$. Nonparametric maximum likelihood estimates (NPMLEs) of $\boldsymbol{\theta}$ can then be estimated.

The CPM in (4.2) is identical to the cumulative link model often used for ordered categorical data and the likelihood in (4.4) is identical to the multinomial likelihood used to estimate parameters of cumulative link models for ordinal data (Snell, 1964; McCullagh, 1980; Agresti, 2010). Therefore, a semi-parametric linear transformation model can be fit using an ordinal CPM where each distinct value of continuous Y is treated as its own ordinal category. With truly continuous Y, there will be N such categories. In summary, with CPMs, a continuous response variable is a linear function of covariates after an unspecified monotonic transformation is applied. The transformation is estimated nonparametrically with a step function.

CPMs have a number of attractive properties for fitting continuous response variables (Liu et al., 2017; Tian et al., 2020). First, since only ordinal information is used for estimating $\boldsymbol{\beta}$, CPMs are invariant to any monotonic transformation of response variables, which means no transformation of response variables is needed. They also work well with continuous response variables subject to detection limits even with high censoring rates and small sample sizes (**?**). It has been shown that under some mild conditions, CPMs result in estimates that are consistent and asymptotically normal (**?**), whose variance can be estimated as the inverse of the information matrix. The estimated CDF conditioning on covariates is $\hat{F}(y|\boldsymbol{X}) = G^{-1}(\hat{\gamma}_j - \hat{\boldsymbol{\beta}}^T \boldsymbol{X})$, where j is the index such that $y_{(j)} = \max\{i \in \{1,\ldots,J\} : y_{(i)} \leq y\}$. Other quantities, such as quantiles and expectation conditional on covariates can be easily derived. The expectation conditional on covariates can be estimated as $\hat{E}(Y|\boldsymbol{X}) = \sum_{j=1}^{J} \sum_{i:y_i=y_{(j)}} y_{(j)} \left[\hat{F}(y_{(j)}|\boldsymbol{X}) - \hat{F}(y_{(j-1)}|\boldsymbol{X}) \right]$. Standard errors for CDFs and expectations can be calculated using the delta method (Liu et al., 2017). Quantiles conditional on covariates and their confidence intervals are based on the linear interpolation of inverse of the CDFs (**?**).

Until recently, the use of CPMs for continuous responses was rare due to lack of unawareness and computational limitations. Harrell's `orm()` function in the **rms** package in R is a computationally efficient implementation of CPMs that can be fit with tens of thousands of distinct responses. The `orm()` function takes advantage of the sparse structure of the Hessian matrix which allows for efficient inversion by Cholesky decomposition in a Newton-Raphson algorithm (Harrell, 2020; Liu et al., 2017).

4.3 Methods

4.3.1 CPMs for Clustered Continuous Response Variables

We now consider the setting with clustered continuous response variables where some transformation of response variables may also be needed before analysis. We would like to extend CPMs to handle clustered continuous response variables, so we do not have to specify the transformation.

We first introduce some notation and setup. Suppose there are N subjects, and subject i has T_i observations for $i = 1,\ldots,N$. Denote the response of subject i at time t as Y_{it}, and $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{iT_i})^T$. Across all subjects, $\boldsymbol{Y} = (\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N)^T$ has a total of J distinct values; with truly continuous $\boldsymbol{Y}$, $J = \sum_{i=1}^{N} T_i$. Let $Z_{itj} = I(Y_{it} \leq y_{(j)})$ and $\mu_{itj} = E(Z_{itj}|\boldsymbol{x}_{it}) = P(Y_{it} \leq y_{(j)}|\boldsymbol{x}_{it})$, where $y_{(j)}$ corresponds to the jth smallest value among the J levels of the response variable. Let the vector of binary indicator variables for subject i at time t be $\boldsymbol{Z}_{it} = (Z_{it1},\ldots,Z_{it(J-1)})^T$, and $\boldsymbol{\mu}_{it} = (\mu_{it1},\ldots,\mu_{it(J-1)})^T$. For subject i, let $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T,\ldots,\boldsymbol{Z}_{iT_i}^T)^T$ and $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T,\ldots,\boldsymbol{\mu}_{iT_i}^T)^T$. Covariates for subject i at time t are represented as $\boldsymbol{x}_{it}$.

Suppose Y_{it} has a linear relationship with the covariates $\boldsymbol{x}_{it}$ after some unspecified monotonic transformation $h(\cdot)$ that leads to a linear transformation model

$$Y_{it} = h^{-1}(Y_{it}^*) = h^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_{it} + \varepsilon_{it}), \tag{4.5}$$

where ε_{it} follows a specified distribution corresponding to the link function $G = F_\varepsilon^{-1}$. We assume that ε_{it} is independent of $\varepsilon_{i't'}$ for $i \neq i'$, but not independent if $i = i'$. Based on the linear transformation model, we have

$$\begin{aligned}
\mu_{itj} &= P(Y_{it} \leq y_{(j)}|\boldsymbol{x}_{it}) \\
&= P(h^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_{it} + \varepsilon_{it}) \leq y_{(j)}|\boldsymbol{x}_{it}) \\
&= P(\varepsilon_{it} \leq h(y_{(j)}) - \boldsymbol{\beta}^T\boldsymbol{x}_{it}|\boldsymbol{x}_{it}) \\
&= F_\varepsilon(h(y_{(j)}) - \boldsymbol{\beta}^T\boldsymbol{x}_{it}), \text{ which implies} \\
G(\mu_{itj}) &= h(y_{(j)}) - \boldsymbol{\beta}^T\boldsymbol{x}_{it}.
\end{aligned}$$

Therefore, similar to (4.2), the CPM for a clustered continuous response variable is:

$$G(\mu_{itj}) = \gamma_j - \boldsymbol{\beta}^T\boldsymbol{x}_{it}, \tag{4.6}$$

where $G(\cdot)$ is the specified link function and $\gamma_j = h(y_{(j)})$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$ are marginal parameters that do not conditional on random effects. The interpretation of $\boldsymbol{\beta}$ depends on the link function specified. For example, $\boldsymbol{\beta}$ is interpreted as a log odds ratio with the logit link and a hazard ratio with the

complementary log-log link (cloglog link). The intercepts $\boldsymbol{\gamma}$ are the link function transformed CDFs with all covariates equal to 0, which also represents the transformation needed for the response variable to be modeled by a linear model.

With clustered data, we cannot directly apply nonparametric maximum likelihood estimation to fit CPMs due to within cluster correlation. Because the CPM has been parameterized as an expectation, μ_{itj}, to obtain estimates of (4.6), we can use GEE techniques. GEE was proposed to model longitudinal data with generalized linear models and quasi-likelihood methods. It only requires correct specification of the marginal model for the response mean while the within cluster correlation is modeled with a working correlation in terms of association parameters (Liang and Zeger, 1986; Zeger and Liang, 1986). The estimation of association parameters can be improved by introducing a second estimating function based on a response dependence model (Prentice, 1988; Prentice and Zhao, 1991; Carey et al., 1993). GEE methods for longitudinal ordinal responses have been proposed, where μ_{itj}, the mean of the binary indicator for the response Z_{itj}, is modeled (Heagerty and Zeger, 1996; Lipsitz et al., 1994; Huang et al., 2002; Parsons et al., 2006; Touloumis et al., 2013).

Specifically, we estimate $\boldsymbol{\theta}$ in (4.6) using GEE methods for ordinal response data and solve the estimating equation

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\alpha}) = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4.7}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{W}_i = \boldsymbol{S}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{S}_i^{\frac{1}{2}}$. $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix of $\boldsymbol{Z}_i$ in terms of the association parameters $\boldsymbol{\alpha}$ and $\boldsymbol{S}_i$ is a $T_i(J-1) \times T_i(J-1)$ block matrix with elements based on the variance of Z_{itj}. $\boldsymbol{W}_i^{-1}$ can be considered as a weight matrix for subject i. More efficiency is gained as the working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ gets closer to the true correlation structure of $\boldsymbol{Z}_i$. The structure of $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is assumed by the analyst and $\boldsymbol{\alpha}$ can then be estimated with a second estimating function that will be described in more detail in Section 4.3.3.

The variance of the estimate of $\boldsymbol{\theta}$ can be estimated by

$$V_{\boldsymbol{\theta}} = \left[\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \mathrm{Cov}(\boldsymbol{Z}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right] \left[\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{D}_i\right]^{-1} \tag{4.8}$$

with plug-in estimates, where $\mathrm{Cov}(\boldsymbol{Z}_i) = (\boldsymbol{Z}_i - \boldsymbol{\mu}_i)(\boldsymbol{Z}_i - \boldsymbol{\mu}_i)^T$.

Since $\mu_{itj} = F(y_{(j)}|\boldsymbol{x}_{it})$, the marginal CDF conditional on covariates, is modeled, other quantities conditional on covariates can be readily obtained from a fitted CPM. The CDF conditional on covariates $\boldsymbol{X}$ can be calculated as $\hat{F}(y|\boldsymbol{X}) = G^{-1}(\hat{\gamma}_j - \hat{\boldsymbol{\beta}}^T \boldsymbol{X})$, where j is the index such that $y_{(j)} = \max\{i \in \{1,\ldots,J\} : y_{(i)} \leq y\}$..

We can derive its standard error by the delta method. Quantiles conditional on covariates along with their confidence intervals can be obtained from the linear interpolation of the inverse of the CDF and its corresponding confidence intervals. The expectation conditional on covariates $\boldsymbol{X}$ can be calculated as $\hat{E}(Y|\boldsymbol{X}) = \sum_{i=1}^{J}\sum_{i,t:y_{it}=y_{(j)}} y_{(j)}\left[\hat{F}(y_{(j)}|\boldsymbol{X}) - \hat{F}(y_{(j-1)}|\boldsymbol{X})\right]$. Similar to $\hat{F}(y|\boldsymbol{X})$, the standard error of the expectation can also be obtained by the delta method.

Although the model described has attractive features, it is very challenging to fit ordinal GEE methods on clustered continuous response variables due to computation limitations. More specifically, for each observation Y_{it}, we need $J-1$ indicators $Z_{itj} = I(Y_{it} \leq y_{(j)})$, and J is usually a large number for continuous data, which implies that $\boldsymbol{W}_i$ and $\boldsymbol{D}_i$ in (4.7) and (4.8) can be enormous. In the following subsections, we will introduce two feasible and computationally efficient implementations to analyze clustered continuous response variables based on CPMs with specific working correlation structures.

4.3.2 CPMs with Independence Working Correlation

Independence working correlation structures do not require estimating $\boldsymbol{\alpha}$ because $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is set to $\boldsymbol{I}$, the identity matrix. In addition to potentially improving computation efficiency, there are settings where using an independence working correlation structure is recommended. For example, if $E(Y_{it}|\boldsymbol{X}_{it}) \neq E(Y_{it}|\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{iT_i})$, one must use an independence working correlation to get unbiased estimates of marginal parameters (Pepe and Anderson, 1994). There are many examples in practice where the cross-sectional conditional expectation is not equal to the full conditional expectation, particularly with time-varying covariates (e.g., Lauderdale et al. (2008); Schildcrout et al. (2020)).

With independence working correlation, solving (4.7) and plugging estimates into (4.8) to estimate the variance is equivalent to treating the data as unclustered, computing the NPMLEs as described in Section 2, and then correcting the variance using a sandwich-variance estimate. CPMs with robust covariance is ordinal GEE with independence working correlation structure. This equivalence is known but a detailed proof is in the Supplementary Material. CPMs can be efficiently fit to clustered continuous responses with thousands of distinct values, since this approach permits fitting CPMs to cross-sectional response data in a computationally efficient manner.

Specifically, we fit CPMs to clustered continuous response variables by maximizing the likelihood

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{J} \prod_{i,t:y_{it}=y_{(j)}} \left[F(y_{it}|\boldsymbol{x}_{it}) - F(y_{it}^{-}|\boldsymbol{x}_{it})\right] \\ &= \prod_{j=1}^{J} \prod_{i,t:y_{it}=y_{(j)}} \left[G^{-1}(\gamma_j - \boldsymbol{\beta}^T\boldsymbol{x}_{it}) - G^{-1}(\gamma_{j-1} - \boldsymbol{\beta}^T\boldsymbol{x}_{it})\right] \\ &= \prod_{j=1}^{J} \prod_{i,t:y_{it}=y_{(j)}} \left(\mu_{itj} - \mu_{it(j-1)}\right). \end{aligned} \tag{4.9}$$

To correct for correlated responses within each cluster, we use the Huber sandwich estimator to estimate the covariance (Freedman, 2006). Since the clusters are independent, we group terms within clusters and then treat clusters as independent units. Let

$$l(\boldsymbol{\theta}) = \log\left(L(\boldsymbol{\theta})\right) = \sum_{j=1}^{J} \sum_{i,t:y_{it}=y_{(j)}} \log\left(f_{itj}\right)$$

be the log-likelihood of (4.9), where $f_{itj} = \mu_{itj} - \mu_{it(j-1)}$. The first and second partial derivatives of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are given by

$$\begin{aligned} l'(\boldsymbol{\theta}) &= \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{J} \sum_{i,t:y_{it}=y_{(j)}} \frac{\partial \log\left(f_{itj}\right)}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{J} \sum_{i,t:y_{it}=y_{(j)}} g_{itj}, \\ l''(\boldsymbol{\theta}) &= \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \sum_{j=1}^{J} \sum_{i,t:y_{it}=y_{(j)}} \frac{\partial^2 \log\left(f_{itj}\right)}{\partial \boldsymbol{\theta}^2}. \end{aligned}$$

The Huber sandwich estimator for the covariance based on the estimated parameters $\hat{\boldsymbol{\theta}}$ is

$$\left[l''(\hat{\boldsymbol{\theta}})\right]^{-1} \left[\sum_{i=1}^{N} \left(\sum_{t=1}^{T_i} \hat{g}_{itj}\right) \left(\sum_{t=1}^{T_i} \hat{g}_{itj}\right)^T\right] \left[l''(\hat{\boldsymbol{\theta}})\right]^{-1}. \tag{4.10}$$

The consistency and asymptotic normality of estimates using this approach to clustered continuous response data can be shown under mild regularity conditions in a manner similar to that of Li et al. (2022b). Details are in the Supplementary Material.

The point estimations and corresponding robust covariance of CPMs can be obtained by `orm()` and `robcov()` functions in the **rms** package in R respectively (Harrell, 2020).

4.3.3 CPMs with Exchangeable/AR1 Working Correlation

Though computationally efficient, CPMs with independence working correlation structure can be statistically inefficient if the within cluster correlation is high. GEE methods for ordinal response variables allow for more

complicated working correlation structures to improve efficiency. There have been many ways proposed for specifying and estimating $\boldsymbol{\alpha}$ in such settings. Lipsitz et al. (1994) estimated the association parameter by Pearson residuals; Heagerty and Zeger (1996) extended alternating logistic regression for binary longitudinal outcomes to ordinal longitudinal outcomes using pairwise log-odds ratio parameters as the association parameters (Lipsitz et al., 1991; Carey et al., 1993); Touloumis et al. (2013) described the association as local odds ratios based on Goodman's row and column effects models.

For computation efficiency, we utilize a framework proposed by Parsons et al. (2006, 2009) that specifies the association parameter α as a correlation and estimates the parameter iteratively by minimizing the logarithm of the determinant of the covariance matrix of the regression parameters. This method, which Parsons et al. (2009) denoted as "repolr", uses the covariance matrix to estimate $\boldsymbol{\alpha}$, where the dimension is manageable. Other methods require all pairs of observations in the second estimating equation, which is extremely computationally intensive for continuous response data. In repolr, $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is constructed as $\boldsymbol{R}_i(\alpha) = \boldsymbol{K}_i(\alpha) \otimes \boldsymbol{C}$, where $\boldsymbol{K}_i(\alpha)$ is a $T_i \times T_i$ matrix of within cluster correlation and $\boldsymbol{S}$ is a $(J-1) \times (J-1)$ matrix of correlations between $\boldsymbol{Z}_{it}$. By assumption, $\boldsymbol{C}$ is the same for every pair of binary indicators within one cluster.

$$\boldsymbol{C} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1(J-1)} \\ \vdots & \ddots & \vdots \\ \rho_{(J-1)1} & \cdots & \rho_{(J-1)(J-1)} \end{bmatrix},$$

where ρ_{pq} is expected correlation between Z_{itp} and Z_{itq} for $i = 1, \ldots, N$. With the logit link, $\rho_{pq} = \rho_{qp} = \exp(\gamma_p - \gamma_q)^{\frac{1}{2}}$ where $p < q$ (Kenward et al., 1994). Two most widely-used structures for $\boldsymbol{K}(\alpha)$ are exchangeable (also called uniform or compound symmetric) and first-order autoregressive (AR1) structures (Diggle et al., 2002). For exchangeable structure, $\boldsymbol{K}_{(p,q)}(\alpha) = 1$ if $p = q$ and $\boldsymbol{K}_{(p,q)}(\alpha) = \alpha$ otherwise; for AR1 structure, $\boldsymbol{K}_{(p,q)}(\alpha) = 1$ for $p = q$ and $\boldsymbol{K}_{(p,q)}(\alpha) = \alpha^{|t_p - t_q|}$ otherwise. The additional estimating equation for the association parameter α in repolr is

$$\frac{\partial \log |V_{\boldsymbol{\theta}}(\alpha)|}{\partial \alpha} = 0, \tag{4.11}$$

which is equivalent to estimating α by minimizing $\log |V_{\boldsymbol{\theta}}(\alpha)|$. Hence, this equation estimates α to minimize the confidence region size of the $\boldsymbol{\theta}$ parameter estimates. The algorithm iterates between solving (4.7) for $\hat{\boldsymbol{\theta}}$ and solving (4.11) for $\hat{\alpha}$ until convergence. This approach can be applied with the `repolr()` function in the **repolr** package in R (Parsons, 2017) for complete data and the logit link.

With continuous response variables, it may still be expensive to run a fully-iterated repolr model; hence, we propose a more computationally efficient one-step repolr.

Others have proposed one-step GEE estimators to reduce computational burden (Lipsitz et al., 2017). In our setting, instead of iterating between the two estimating equations (4.7) and (4.11) until convergence, we start with an estimate of $\boldsymbol{\theta}$ under independence working correlation structure $\hat{\boldsymbol{\theta}}_I$, and then obtain the association parameter $\hat{\boldsymbol{\alpha}}$ by plugging $\hat{\boldsymbol{\theta}}_I$ into (4.8), and finally solve (4.7) to get $\hat{\boldsymbol{\theta}}$ which is asymptotically equivalent to the fully-iterated GEE estimator (Lipsitz et al., 2017). Specifically,

1. Estimate $\hat{\boldsymbol{\theta}}_I$ by minimizing (4.9)
2. Estimate $\hat{\alpha}$ by solving (4.11) with $V_{\hat{\boldsymbol{\theta}}_I}(\alpha)$
3. Estimate $\hat{\boldsymbol{\theta}}$ by solving (4.7) with $\hat{\alpha}$.

We have built an R package, **cpmgee** (available at https://github.com/YuqiTian35/cpmgee), that applies this one-step estimation procedure for exchangeable and AR1 working correlation structures. This pacakge also fits CPMs with independence working correlation.

Although this one-step repolr can substantially reduce the computational burden, computation with exchangeable and AR1 working correlation structures may still be an issue with large numbers of continuous outcomes. We may need to further reduce the number of distinct values in the response by binning. Specifically, the $N' = \sum_{i=1}^{N} T_i$ observations can be divided into M_b bins, where the value assigned to each observation in the bin is the median value for observations in that bin. Approximately equal-quantile binning can be achieved by expressing N' as

$$N' = M_b q + r = (M_b - r)q + r(q+1), \tag{4.12}$$

where q is the integer quotient of $\frac{N'}{M_b}$. In this way, $M_b - r$ bins have q observations, and r bins have $q+1$ observations. Rounding to a certain decimal place is another way to reduce the number of distinct values. More strategies for binning and rounding for cross-sectional CPMs with very large sample sizes are provided elsewhere (Li et al., 2022a).

4.4 Simulations

We studied the performance of our estimators applying CPMs with independence, exchangeable, and AR1 working correlation structures to continuous clustered data under various simulation settings. Responses were generated in the following manner for subject i at time t:

$$Y_{it} = \text{Inv-}\chi^2\left(\frac{\Phi(Y_{it}^*)}{2},\ \text{df} = 5\right),\ \text{and}\ Y_{it}^* = X_i\beta_X + T_{it}\beta_T + \varepsilon_{it},$$

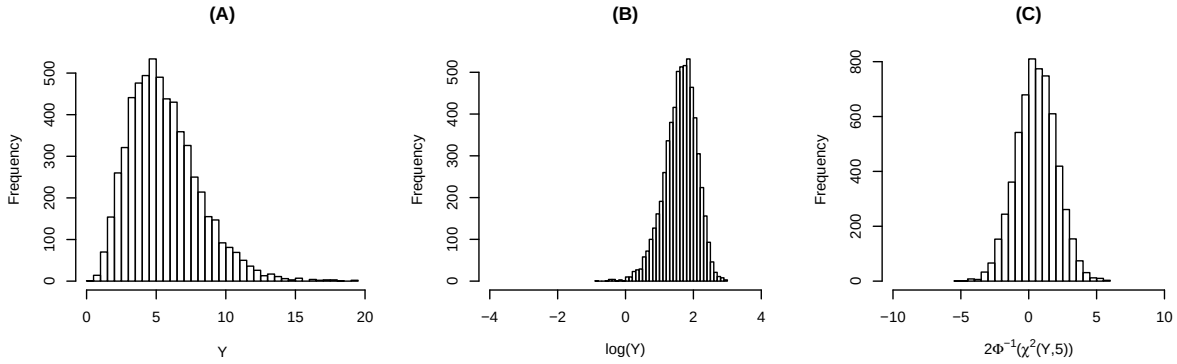

Figure 4.2: (A) Histogram of the response variable. (B) Histogram of the log-transformed response variable. (C) Histogram of the response variable with the correct transformation $2\Phi^{-1}(\chi^2(Y, \text{df} = 5))$ for a linear model.

where Inv-$\chi^2(\cdot,$ df=5$)$ is the inverse of the CDF for a chi-square distribution with 5 degrees of freedom and $\Phi(\cdot)$ is the probability density function of the standard normal distribution. The transformation has been used in earlier work (Tian et al., 2020) and was chosen because it corresponds to no closed-form transformation.

In the primary setting, we set the sample size N to be 1000. Each subject had at least 2 and at most 6 observations, where dropouts were missing completely at random. X_i is a time-invariant covariate following the standard normal distribution. T_{it} represents time, a time-variant covariate, and was set to be $0, 0.2, \ldots, 1$. A logistic residual distribution was used and the correlation structure was exchangeable with $\alpha = 0.7$. We set $\beta_X = 1$ and $\beta_T = 1$.In Figure 4.2, we show histograms of the response variable after different transformation based on one simulated data set. The response variable on its original scale is right-skewed. A natural choice for right-skewed data is log-transformation. However, the log-transformed response variable is slightly left-skewed. The correct transformation $2\Phi^{-1}(\chi^2(\cdot,$ df=5$))$ is not a function one would typically consider for transformation. For CPMs with exchangeable and AR1 working correlation structures, we fit models using equal-quantile binning with $M_b = 300$.

Besides the primary setting, we also looked into scenarios with smaller α, identity transformation (i.e., $Y = Y^*$), complete data, different M_b for equal-quantile binning, rounding with different decimal places, different sample sizes, different cluster sizes, different time effects, different correlation structures, link function misspecification, and fully-iterated repolr models. (Simulation results for some settings are in Supplementary Material.)

We simulated 1000 replications under each scenario and evaluated the results by percent bias, root mean squared error (RMSE), and coverage of 95% CIs with respect to covariate coefficients. We also compared our two methods with standard GEE methods for continuous data with the correctly transformed response

variable. We also investigated the performance of estimates of $E(Y|X=1,T=0.2)$, $Q(0.5|X=1,T=0.2)$, and $F(5|X=1,T=0.2)$ that were estimated from the fitted CPMs. To evaluate the relative efficiency (RE) of regression parameters, we divided the empirical variance obtained from a CPM method by the empirical variance of standard GEE for continuous response variables with the correct transformation. With the correct transformation and correlation structure, the standard GEE method is optimal.

4.4.1 The Primary Setting and its Modifications

Simulation results under the primary setting and two modifications are shown in Table 4.1. For the primary setting, CPMs performed quite well with low bias and generally good coverage for β_X, β_T, $E(Y|X=1,T=0.2)$, $Q(0.5|X=1,T=0.2)$, and $F(5|X=1,T=0.2)$. CPMs with an independence working correlation structure had minimal bias and coverage near 0.95. Estimates of β_T from CPMs with a properly specified exchangeable working correlation structure tended to be slightly more biased ($\sim$3%) and have lower than normal coverage (0.91) but were much more efficient than those using independence working correlation (RMSE of 0.075 vs. 0.091). The relative efficiency of CPMs with a exchangeable working correlation was fairly close to that of the gold standard GEE estimator where the correct transformation and correlation structure were correctly assumed.

When the within cluster correlation was low ($\alpha = 0.3$), the performance of CPMs remained good and had similar efficiency to standard GEE. CPMs with exchangeable working correlation was even slightly more efficient in estimating the time-varying covariate and this might be due to variance reduction from binning. With identity transformation, the results of CPMs for β_X and β_T were identical to the primary setting because both methods are invariant to monotonic transformation of the response variable. The conditional mean was estimated with low bias and correct coverage rates using both working correlation structures.

4.4.2 Equal-quantile Binning and Rounding

In the primary simulation setting, when applying CPMs with exchangeable working correlation, we used equal-quantile binning with $M_b = 300$. To investigate the sensitivity of results to this choice, we repeated simulations using different binning/rounding strategies. Table 4.2 shows results. As M_b increased, we observed fairly similar performance with slightly higher bias in coefficient estimation and slightly lower bias in conditional quantities. With larger M_b might lead to poorer estimation of coefficient parameter and better estimation of conditional quantities due to increasing number of intercepts estimated while fixing the sample and cluster size. Rounding to 0 decimal place resulted in 169 categories in the response variable on average. There was great information loss by rounding to 0 decimal place and therefore the performance particularly of $Q(0.5|X=1,T=0.2)$ and $F(5|X=1,T=0.2)$ was poor. Also, rounding to a decimal place tends to be

Table 4.1: Simulation results for the primary setting and its modifications

Scenario	Method	Metric	β_X	β_T	$E(Y\|X{=}1,T{=}0.2)$	$Q(0.5\|X{=}1,T{=}0.2)$	$F(5\|X{=}1,T{=}0.2)$
		Bias(%)	-0.010	0.087	-	-	-
	GEE	RMSE	0.050	0.060	-	-	-
	(correct)	Coverage	0.953	0.944	-	-	-
		RE	baseline	baseline	-	-	-
		Bias(%)	0.129	0.270	-0.009	-0.074	-0.169
$\alpha = 0.7$	CPM	RMSE	0.054	0.091	1.232	1.199	0.171
$Y{=}\text{Inv-}\chi^2\left(\frac{\Phi(Y^*)}{2},5\right)$	(ind)	Coverage	0.957	0.942	0.956	0.958	0.956
		RE	1.129	2.279	-	-	-
		Bias(%)	0.234	2.983	-0.181	-0.270	-0.077
	CPM	RMSE	0.052	0.075	1.224	1.191	0.170
	(ex)	Coverage	0.957	0.910	0.948	0.956	0.958
		RE	1.047	1.310	-	-	-
		Bias(%)	-0.061	0.127	-	-	-
	GEE	RMSE	0.040	0.089	-	-	-
	(correct)	Coverage	0.955	0.943	-	-	-
		RE	baseline	baseline	-	-	-
		Bias(%)	0.063	0.254	-0.015	-0.054	-0.069
$\alpha = 0.3$	CPM	RMSE	0.041	0.092	1.236	1.204	0.171
$Y{=}\text{Inv-}\chi^2\left(\frac{\Phi(Y^*)}{2},5\right)$	(ind)	Coverage	0.959	0.946	0.957	0.952	0.959
		RE	1.063	1.073	-	-	-
		Bias(%)	0.160	2.929	-0.196	-0.249	0.023
	CPM	RMSE	0.041	0.091	1.227	1.195	0.170
	(ex)	Coverage	0.961	0.936	0.953	0.943	0.959
		RE	1.041	0.943	-	-	-
		Bias(%)	-0.010	0.087	-	-	-
	GEE	RMSE	0.050	0.060	-	-	-
	(correct)	Coverage	0.953	0.944	-	-	-
		RE	baseline	baseline	-	-	-
		Bias(%)	0.129	0.270	0.005	-0.248	-0.006
$\alpha = 0.7$	CPM	RMSE	0.054	0.091	0.708	0.707	0.011
$Y = Y^*$	(ind)	Coverage	0.957	0.942	0.957	0.958	0.954
		RE	1.129	2.279	-	-	-
		Bias(%)	0.234	2.983	-0.266	-0.803	0.008
	CPM	RMSE	0.052	0.075	0.706	0.702	0.011
	(ex)	Coverage	0.957	0.910	0.955	0.956	0.932
		RE	1.047	1.310	-	-	-

Table 4.2: Simulation results for equal-quantile binning and rounding with exchangeable correlation structure

Scenario	Metric	β_X	β_T	$E(Y\|X=1,T=0.2)$	$Q(0.5\|X=1,T=0.2)$	$F(5\|X=1,T=0.2)$
Binning $M_b =$ 50	Bias(%)	0.174	0.757	-0.039	-0.053	-0.233
	RMSE	0.052	0.068	1.205	1.166	0.171
	Coverage	0.958	0.942	0.929	0.923	0.935
Binning $M_b =$ 100	Bias(%)	0.187	1.193	-0.316	-0.493	-0.174
	RMSE	0.052	0.069	1.217	1.181	0.171
	Coverage	0.957	0.936	0.945	0.948	0.953
Binning $M_b =$ 200	Bias(%)	0.208	2.069	-0.197	-0.311	-0.112
	RMSE	0.052	0.071	1.223	1.189	0.170
	Coverage	0.957	0.924	0.946	0.952	0.958
Rounding 0 decimal place	Bias(%)	0.196	0.799	-0.015	-7.316	-20.965
	RMSE	0.052	0.070	1.231	0.940	0.222
	Coverage	0.959	0.937	0.952	0.244	0.004
Rounding 1 decimal place	Bias(%)	0.210	3.180	-0.123	-0.693	-2.147
	RMSE	0.052	0.076	1.229	1.175	0.176
	Coverage	0.957	0.907	0.953	0.942	0.943

a sub-optimal choice for such right-skewed responses because many values were rounded at the lower end to a single value. There were 498 ordinal levels on average if the response variable was rounded to 1 decimal place. The performance for rounding to 1 decimal place was good.

4.4.3 Sample Size and Cluster Size

Table 4.3: Simulation results for different sample sizes and cluster sizes

Scenario	Method	Metric	β_X	β_T	$E(Y\|X=1,T=0.2)$	$Q(0.5\|X=1,T=0.2)$	$F(5\|X=1,T=0.2)$
$N=100$ $M=6$	GEE (correct)	Bias(%)	0.551	-0.313	-	-	-
		RMSE	0.166	0.184	-	-	-
		Coverage	0.917	0.952	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	2.114	1.504	0.289	0.469	-0.757
		RMSE	0.182	0.288	1.325	1.299	0.180
		Coverage	0.940	0.945	0.939	0.939	0.947
		RE	1.196	2.433	-	-	-
	CPM (ex)	Bias(%)	3.122	40.440	-1.114	-0.873	0.985
		RMSE	0.181	0.516	1.260	1.233	0.175
		Coverage	0.920	0.556	0.920	0.942	0.946
		RE	1.159	3.016	-	-	-
$N=200$ $M=6$	GEE (correct)	Bias(%)	0.497	0.099	-	-	-
		RMSE	0.112	0.131	-	-	-
		Coverage	0.938	0.948	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	1.090	0.873	0.085	0.164	-0.505
		RMSE	0.126	0.195	1.266	1.240	0.174
		Coverage	0.940	0.952	0.939	0.946	0.934
		RE	1.263	2.200	-	-	-
	CPM	Bias(%)	1.595	16.625	-0.505	-0.418	0.179
		RMSE	0.122	0.242	1.251	1.221	0.174

Table 4.3: Simulation results for different sample sizes and cluster sizes *(continued)*

Scenario	Method	Metric	β_X	β_T	$E(Y\|X{=}1,T{=}0.2)$	$Q(0.5\|X{=}1,T{=}0.2)$	$F(5\|X{=}1,T{=}0.2)$
	(ex)	Coverage	0.946	0.790	0.933	0.939	0.939
		RE	1.166	1.797	-	-	-
$N=500$ $M=6$	GEE (correct)	Bias(%)	-0.259	0.034	-	-	-
		RMSE	0.074	0.084	-	-	-
		Coverage	0.940	0.949	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.074	0.444	-0.049	-0.068	-0.035
		RMSE	0.082	0.122	1.236	1.210	0.171
		Coverage	0.941	0.955	0.942	0.941	0.944
		RE	1.211	2.137	-	-	-
	CPM (ex)	Bias(%)	0.192	5.862	-0.318	-0.360	0.262
		RMSE	0.079	0.114	1.223	1.196	0.170
		Coverage	0.940	0.911	0.927	0.934	0.931
		RE	1.134	1.370	-	-	-
$N=1000$ $M=3$	GEE (correct)	Bias(%)	0.011	-0.013	-	-	-
		RMSE	0.052	0.142	-	-	-
		Coverage	0.950	0.943	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.281	0.157	-0.006	-0.049	-0.267
		RMSE	0.054	0.167	1.233	1.201	0.171
		Coverage	0.956	0.950	0.944	0.957	0.957
		RE	1.067	1.387	-	-	-
	CPM (ex)	Bias(%)	0.283	2.863	-0.083	-0.144	-0.364
		RMSE	0.053	0.154	1.229	1.196	0.171
		Coverage	0.955	0.938	0.945	0.956	0.957
		RE	1.050	1.137	-	-	-
$N=1000$ $M=12$	GEE (correct)	Bias(%)	0.041	-0.083	-	-	-
		RMSE	0.050	0.023	-	-	-
		Coverage	0.950	0.954	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.221	0.029	0.026	-0.022	-0.212
		RMSE	0.056	0.046	1.235	1.203	0.171
		Coverage	0.949	0.954	0.955	0.945	0.950
		RE	1.235	4.075	-	-	-
	CPM (ex)	Bias(%)	0.461	2.455	-0.365	-0.401	-0.144
		RMSE	0.053	0.043	1.216	1.186	0.170
		Coverage	0.947	0.886	0.940	0.945	0.954
		RE	1.115	2.310	-	-	-

We conducted additional simulations varying the number of clusters, N, from 100 to 500. The cluster size is of interest as well. Let $M = \max\{T_i\}$ be the largest cluster size. Performances of the two methods were evaluated with smaller ($M = 3$) and larger ($M = 12$) cluster sizes while other settings were the same as the primary settings ($N = 1000, M = 6$). Results are shown in Table 4.3. When $N = 100$, CPMs with independence working correlation had good performance while CPMs with exchangeable working correlation had substantial bias. The bias decreased and efficiency gains increased as the sample size increased. With

Table 4.4: Simulation results for the AR1 correlation structure

Method	Metric	β_X	β_T	$E(Y\|X=1,T=0.2)$	$Q(0.5\|X=1,T=0.2)$	$F(5\|X=1,T=0.2)$
GEE (correct)	Bias(%)	-0.065	0.144	-	-	-
	RMSE	0.046	0.099	-	-	-
	Coverage	0.954	0.943	-	-	-
	RE	baseline	baseline	-	-	-
CPM (ind)	Bias(%)	0.087	0.277	0	-0.062	-0.111
	RMSE	0.048	0.109	1.231	1.198	0.170
	Coverage	0.960	0.946	0.958	0.958	0.950
	RE	1.115	1.169	-	-	-
CPM (AR1)	Bias(%)	0.089	0.530	-0.107	-0.190	-0.114
	RMSE	0.048	0.103	1.226	1.193	0.170
	Coverage	0.958	0.946	0.950	0.951	0.950
	RE	1.079	1.051	-	-	-
CPM (ex)	Bias(%)	0.203	3.007	-0.176	-0.256	-0.033
	RMSE	0.048	0.107	1.223	1.190	0.170
	Coverage	0.961	0.946	0.950	0.947	0.958
	RE	1.004	1.040	-	-	-

large N, performance of CPMs was good regardless of cluster size. However, the RE of standard GEE over CPMs seemed to be greater as the number of clusters increased.

4.4.4 First-order Autoregressive (AR1) Correlation Structure

We generated residuals with AR1 correlation structure with $\alpha = 0.7$, and fit both AR1 and exchangeable working correlation structures keeping other settings the same as the primary setting. The results are in Table 4.4. CPM methods were almost as efficient as continuous GEE methods, especially with the correct AR1 working correlation structure. If fitting exchangeable working correlation, CPMs method still had small bias and correct coverage rates.

4.4.5 Link Function Misspecification

We look into the performance of our approaches with link function misspecification. The residuals were generated with standard normal distributions and we still fit models with the logit link. Results are shown in Table 4.5. Regression parameters were transformed to the same scale. CPMs methods are generally robust to moderate link function misspecification (Liu et al., 2017; Tian et al., 2020). The bias of regression parameters is larger than that in correctly specified models. Mean and median estimation are still good. The results for CDF is less satisfying under link function misspecification.

4.5 Applications

To illustrate the use of CPM methods proposed, we applied them on two real data sets. The first studies CD4:CD8 ratios among people living with HIV CD4:CD8 ratios. The second considers lung function among

Table 4.5: Simulation results for the link function misspecification

Method	Metric	β_X	β_T	$E(Y\|X{=}1,T{=}0.2)$	$Q(0.5\|X{=}1,T{=}0.2)$	$F(5\|X{=}1,T{=}0.2)$
GEE (correct)	Bias(%)	-0.009	0.057	-	-	-
	RMSE	0.028	0.033	-	-	-
	Coverage	0.952	0.944	-	-	-
	RE	baseline	baseline	-	-	-
CPM (ind)	Bias(%)	-3.983	-3.879	0.272	0.503	-6.193
	RMSE	0.052	0.066	1.220	1.223	0.271
	Coverage	0.793	0.876	0.956	0.953	0.842
	RE	1.414	2.654	-	-	-
CPM (ex)	Bias(%)	-4.602	-2.726	0.237	0.342	-5.338
	RMSE	0.056	0.050	1.219	1.216	0.270
	Coverage	0.722	0.890	0.952	0.954	0.864
	RE	1.341	1.597	-	-	-

smokers with mild COPD.

4.5.1 CD4:CD8 Ratio

CD4:CD8 ratio is the ratio of CD4 lymphocyte count (cells/mm^3) to CD8 lymphocyte count (cells/mm^3). It has been associated with immune senescence, inflammation, and comorbidities for people living with HIV (Castilho et al., 2016). As highlighted in the Introduction, CD4:CD8 ratio tends to be right-skewed and there is no standard transformation. To study the relationship between CD4:CD8 ratio and age, an observational cohort study was conducted among people living with HIV who had been on antiretroviral therapy (ART) for one year, had a suppressed viral load, and received treatment at the Vanderbilt Comprehensive Care Clinic (VCCC) between 1998 and 2012 (Castilho et al., 2016). In the current analysis, we are interested in factors associated with CD4:CD8 ratio during one year of follow-up, i.e., during the second year after starting ART. CD4:CD8 ratio was collected longitudinally during routine clinical visits. Our study included 1763 subjects with a mean of 2.9 CD4:CD8 measurements (median = 3; range = 1-7), and 3862 unique values in the outcome.

CPMs with independence working correlation is able to handle 3862 ordinal levels, while CPMs with exchangeable or AR1 working correlation requires binning or rounding due to computation limitations. For the latter, We divided the outcome to 1000 bins and rounded to 2 decimal places. The equal-quantile binning resulted in 979 ordinal levels because of ties on the original scales. The 2 decimal place rounding led to 234 levels. The logit link was used in all models. The time-invariant covariates considered were calendar year at baseline (one year after ART initiation), race, baseline age, sex, probable route of infection, hepatitis C virus (HCV) infection status, and hepatitis B virus (HBV) infection status. Time (in years) after baseline was the only time-varying covariate.

Table 4.6: Odds ratio estimates of higher CD4:CD8 ratios with 95% confidence intervals from CPMs with independence working correlation and CPMs with exchangeable working correlation with binning and rounding (1000-bin equal-quantile binning that led to 979 levels; rounding to 2 decimal place) are shown. Variance ratios are calculated by the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances of the the log-odds ratios from CPMs with independence working correlation.

Predictor	Independence	Exchangeable (Binning)	VR	Exchangeable (Rounding)	VR
Time (years)	1.217 (1.082, 1.370)	1.226 (1.135, 1.325)	0.429	1.226 (1.134, 1.324)	0.429
Enrollment Year	1.011 (0.984, 1.038)	1.013 (0.989, 1.038)	0.813	1.014 (0.989, 1.038)	0.814
Race					
African American	(Reference)				
Caucasian	1.014 (0.829, 1.239)	1.069 (0.886, 1.291)	0.881	1.063 (0.881, 1.283)	0.881
Hispanic	0.679 (0.464, 0.992)	0.730 (0.501, 1.064)	0.983	0.721 (0.495, 1.051)	0.984
Other	0.724 (0.467, 1.123)	0.740 (0.489, 1.119)	0.890	0.732 (0.484, 1.108)	0.889
Baseline Age (10 years)	0.670 (0.608, 0.735)	0.677 (0.620, 0.740)	0.877	0.676 (0.619, 0.738)	0.877
Sex					
Male	(Reference)				
Female	1.724 (1.321, 2.249)	1.803 (1.400, 2.322)	0.902	1.800 (1.398, 2.318)	0.902
Route					
Heterosexual	(Reference)				
Injection Drug Use	0.991 (0.675, 1.455)	0.931 (0.644, 1.347)	0.927	0.931 (0.644, 1.348)	0.928
MSM	0.904 (0.700, 1.167)	0.902 (0.709, 1.148)	0.885	0.898 (0.706, 1.142)	0.885
Other/Unknown	0.793 (0.466, 1.351)	0.858 (0.535, 1.377)	0.789	0.851 (0.530, 1.366)	0.789
HCV	0.824 (0.596, 1.138)	0.808 (0.599, 1.088)	0.850	0.805 (0.597, 1.085)	0.851
HBV	0.993 (0.660, 1.492)	0.919 (0.642, 1.314)	0.771	0.920 (0.643, 1.317)	0.770

Odds ratio estimates and 95% confidence intervals from the fitted CPMs are shown in Table 4.6. The results suggested that time, race, baseline age and sex are significantly associated with CD4:CD8 ratio. Results are fairly similar across all three fitted CPMs. For example, fixing other variables, a 10-year increase in baseline age is associated with 33% decrease in the odds of having higher CD4:CD8 ratio.

There were some differences in efficiency of estimates across different CPM estimating procedures. The variance ratios in Table 4.6 are the ratios of the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances of the log-odds ratios from CPMs with independence working correlation. The variances for the estimated log-odds ratio for the time-varying covariate, time, for the two exchangeable working correlation models was 0.429 times that for the independence working correlation model. We saw variance ratios ranging from 0.78 to 0.93 for time-invariant covariates.

Other quantities can be estimated from the fitted CPMs. Conditional means, and medians of CD4:CD8 and the conditional probabilities of CD4:CD8 being greater than 1 are shown as a function of time since baseline in Figure 4.3 with other covariates fixed at their median (for continuous covariates) or mode (for categorical covariates) levels. CD4:CD8 ratio above 1 is considered normal for people without HIV (Petoumenos et al., 2017). Results from the three models are generally very close. We also included the conditional mean obtained by a standard GEE model without transforming the response data for purpose of comparison; results

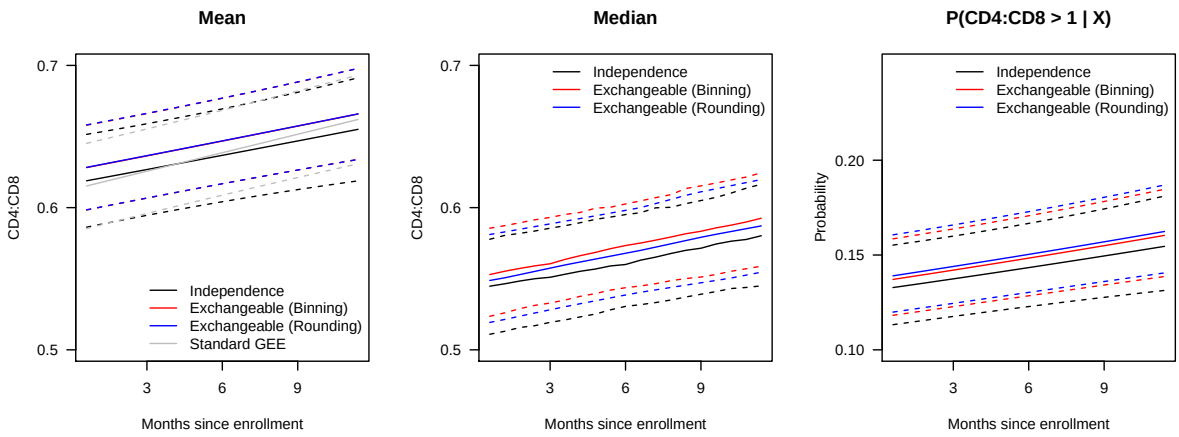

Figure 4.3: The estimated conditional mean, median of CD4:CD8 and the conditional probability of the ratio being greater than 1 (normal range) as functions of years since enrollment while fixing other covariates at their medians (for continuous covariates) or modes (for categorical covariates).

from this analysis are also fairly similar.

4.5.2 The Lung Health Study

For the second application, we used data from The Lung Health Study on smokers with mild COPD. The Lung Health Study is a randomized clinical trial collecting data from 10 centers in the United States and Canada from 1986 to 1994 (Schildcrout et al., 2020). We are interested in the genetic contributions of a single nucleotide polymorphism (SNP), rs12194741, on chromosome 6 to lung function decline over 5 years (Hansel et al., 2013). Lung function was quantified as the amount of air in liters one can force from the lung in the first second of exhalation (FEV1). rs12194741 was represented by a binary indicator for whether there was at least 1 copy of the T allele at rs12194741. The interaction of rs12194741 and visit was used to evaluate the genetic contribution to lung function decline. Data were collected from participants' annual visits over a 5-year follow-up period. In this analysis, we included participants who were smokers for all their visits and had at least 2 observations. There were 2562 subjects included and 1694 (66%) completed 5 visits. Baseline adjustment covariates included age, study site, body mass index (BMI, weight(kg)/height(m^2)), lifetime smoking status (in pack years), and average number of cigarettes smoked per day over the previous year. BMI change from baseline and study visit were included as time-varying covariates. The distribution of the responses, FEV1, was fairly symmetric (Figure 4.4 in Supplementary Material). 61% (1567) of participants had at least 1 copy of the T allele at rs12194741.

We applied both CPMs with independence and exchangeable working correlation with the logit link on the data and compared the results and efficiency. No binning or rounding was needed to fit the models as

Table 4.7: We show odds ratios estimates for higher FEV1 with 95% confidence intervals from CPMs with independence and exchangeable working correlation without adjusting for the baseline FEV1. The last column shows the variance ratios (VRs) calculated by the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances the log-odds ratios from CPMs with independence working correlation.

Predictor	Independence	Exchangeable	VR
Visit:rs12194741	0.965 (0.941, 0.989)	0.971 (0.955, 0.987)	0.431
Visit	0.859 (0.842, 0.877)	0.854 (0.843, 0.866)	0.445
rs12194741	1.120 (0.971, 1.291)	1.118 (0.973, 1.284)	0.950
BMI Change (per 5 kg/m^2)	0.651 (0.529, 0.801)	0.666 (0.587, 0.754)	0.365
Baseline Age (per 10–year)	0.342 (0.300, 0.389)	0.340 (0.301, 0.384)	0.879
Baseline BMI (per 5 kg/m^2)	1.480 (1.343, 1.631)	1.485 (1.356, 1.626)	0.868
Cigarettes/day (per 10 cigs/day)	0.976 (0.920, 1.034)	0.975 (0.922, 1.032)	0.938
Pack Years (per 20 pack year)	1.190 (1.085, 1.304)	1.192 (1.089, 1.305)	0.968
Study Site			
1	(Reference)		
2	2.028 (1.429, 2.878)	1.989 (1.444, 2.739)	0.837
3	1.422 (1.001, 2.019)	1.386 (1.003, 1.914)	0.847
4	1.811 (1.268, 2.588)	1.820 (1.319, 2.512)	0.815
5	2.671 (1.909, 3.738)	2.564 (1.885, 3.488)	0.838
6	1.950 (1.374, 2.770)	1.861 (1.348, 2.569)	0.847
7	0.908 (0.635, 1.297)	0.902 (0.652, 1.248)	0.825
8	1.724 (1.234, 2.409	1.667 (1.227, 2.264)	0.838
9	2.016 (1.425, 2.852)	1.956 (1.427, 2.680)	0.825
10	2.307 (1.585, 3.357)	2.261 (1.559, 3.198)	0.853

there were only 361 distinct values of the outcome. Table 4.7 shows odds ratio estimates of higher FEV1 and 95% confidence intervals obtained from the two methods. The odds ratios from the two models are very close. The variance ratios (VRs) shown in the last column indicate that the log-odds ratios obtained by CPMs with exchangeable working correlation are more efficient than those from CPMs with independence working correlation, particularly for time-varying covariates (visit and BMI change from baseline). The confidence interval for the interaction term did not cover 1, which means that rs12194741 was associated with lung function decline. BMI change from baseline, baseline age, and lifetime smoking status were negatively associated with FEV1 while baseline BMI and the average number of cigarettes smoked per day had positive associations with FEV1. For example, keeping other covariates constant, 5 kg/m^2 increase in BMI change from baseline was associated with a 33-35% decrease in the odds of having a higher FEV1 value.

Conditional quantities including means, medians, and probabilities that FEV1 less than or equal to 2L were derived from the models and are shown in Figure 4.5 of the Supplementary Material as a function of study visit and genotype.

In an additional analysis, we adjusted for baseline FEV1 in the models. Results are detailed in Table 4.11 in the Supplementary Material. The interaction term between study visit and genotype was still significant.

However, efficiency gains using the exchangeable working correlation compared to the independence working correlation were smaller in these analyses. The reason could be that most of the variation was explained by baseline FEV1 (log odds ratio = 8.61 and 8.47; Pearson's correlation between FEV1 and baseline FEV1 was 0.93). We did see that CPMs with exchangeable working correlation was more efficient in estimating time-varying covariates, with the VR for BMI change from baseline being 0.619, for example. The confidence intervals for conditional quantities were also much narrower after adjusting for the baseline FEV1 as shown in Figure 4.6 of the Supplementary Material.

4.6 Discussion

We extended CPMs, a class of ordinal regression models for cross-sectional responses, to analyze clustered continuous response variables. Only rank information is used in CPMs when estimating $\boldsymbol{\beta}$, and thus fitting such ordinal regression models can avoid transformation of response variables. In cross-sectional settings, CPMs have been used to fit different types of continuous response variables (Liu et al., 2017; Tian et al., 2020). To account for correlation between observations within each cluster, we estimate parameters in CPMs using GEE techniques. With estimated parameters, we can easily obtain CDFs, expectation and quantiles conditional on covariates to help better interpret regression results.

We proposed two feasible and computationally efficient approaches for fitting CPMs depending on the working correlation structure. With low within cluster correlation, CPMs with independence working correlation is able to provide unbiased estimations with proper confidence interval coverage rates without losing much efficiency. With high within cluster correlation, CPMs with exchangeable/AR1 working correlation can provide more efficient estimations. Our approaches work well under a variety of simulation settings. We have built an R package, **cpmgee**, for CPMs with independence, exchangeable and AR1 working correlation.

Our methods can fit fully continuous clustered data with independence working correlation, but might require binning or rounding if using exchangeable or AR1 working correlation structures. For futures research, we could consider extending CPMs to include weights. With weighted CPMs, we could fit fully continuous clustered data with more complex working correlation structures by choosing different weighting matrices. GEE methods assume observations are missing completely at random, which can be violated in practice. We could extend our methods under the less restrictive missingness assumption of missing at random in the future.

4.7 Supplementary Material

4.7.1 Asymptotic Properties of CPMs with Independence Working Correlation

Li et al. (2022b) has shown consistency and asymptotic normality for NPMLEs in CPMs in cross-sectional settings. The proof for CPMs with independence working correlation is very similar as the proof in Li et al. (2022b) with minor modifications to address for correlated responses and sandwich estimator for covariance. We use the same notation in Li's paper ($\boldsymbol{\gamma}$, $\boldsymbol{X}$, and G^{-1} in this paper are equivalent to A, Z, and G in Li et al. (2022b) respectively).

First, for clustered data, the log-likelihood $l_n(\beta, A)$ is now the summed log-likelihood from each observation within one cluster

$$l(\beta, A) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\{I(Y_{it} \leq L)\log G(A(L) - \beta^T Z_{it}) + I(Y_{it} > U)\log(1 - G(A(U) - \beta^T Z_{it})) +$$
$$I(L < Y_{it} \leq U)\log(G(A(Y_{it}) - \beta^T Z_{it}) - G(A(Y_{it}-) - \beta^T Z_{it}))\},$$

where N is the number of clusters and T_i is the number of observations in cluster i.

The boundedness of $\hat{A}(y)$ and $nA\{Y_i\}$, and consistency of parameters still hold for clustered data following the same proof in Li et al. (2022b). For proof of the asymptotic distribution, note that (A.8) and (A.9) are still valid for clustered data, where $\boldsymbol{S}_{11}, \boldsymbol{S}_{12}, \boldsymbol{S}_{21}$, and $\boldsymbol{S}_{22}$ are the second order differentiation operators based on the newly defined log-likelihood and $\boldsymbol{S}(Y, Z)$ is the first order differentiation operator derived from the pseudo log-likelihood.

From (A.10), the asymptotic variance for the parameters takes the sandwich form described in (4.10) since v and h are the inverse of the information operator. In the cross-sectional setting, we have shown the consistency of the variance, the inverse of the information matrix. The middle part of the sandwich estimator is also consistent because it is a function of the score function, which is differentiable with respect to β and A, so the estimator is also consistent based on the Glivenko-Cantelli Theorem. Therefore, the sandwich estimator for variance is also consistent.

4.7.2 CPMs with Independence Working Correlation and Ordinal GEE with Independence Working Correlation Structure

The marginal regression model used in GEE methods for ordinal response variables is the CPM. CPMs with independence correlation and GEE methods for ordinal response variables with independence working correlation both ignore the within cluster correlation. We would like to show the estimations for $\boldsymbol{\theta}$ and $V_{\boldsymbol{\theta}}$ from CPMs with independence working correlation and GEE methods for ordinal response variables with independence working correlation are equivalent. More specifically, we first show that the score equation in

CPMs is the equivalent to the estimating function in GEE methods when assuming independence working correlation, then we demonstrate the equivalence of the covariance estimator.

Before directly working with the likelihood of CPMs, we first introduce some new notations. Let $O_{itj} = I(Y_{it} = y_{(j)}) = Z_{itj} - Z_{it(j-1)}$ and $\pi_{itj} = E(O_{itj}|\boldsymbol{x}_{it})$. Then $\boldsymbol{O}_{it} = (O_{it1}, \ldots, O_{itJ})^T$ and $\boldsymbol{O}_{it} \sim \text{Multinomial}(1, \boldsymbol{\pi}_{it})$, which belongs to the exponential family. The probability mass function (PMF) is

$$\begin{aligned} P(\boldsymbol{O}_{it}|\boldsymbol{x}_{it}) &= \left(\prod_{j=1}^{J-1} \pi_{itj}^{O_{itj}}\right)\left(1 - \sum_{j=1}^{J-1} \pi_{itj}\right)^{\left(1-\sum_{j=1}^{J-1} O_{itj}\right)} \\ &= \exp\left\{\sum_{j=1}^{J-1} O_{itj}\log(\pi_{itj}) + \left(1 - \sum_{j=1}^{J-1} O_{itj}\right)\log\left(1 - \sum_{j=1}^{J-1}\pi_{itj}\right)\right\} \\ &= \exp\left\{\sum_{j=1}^{J-1} O_{itj}\log\left(\frac{\pi_{itj}}{1-\sum_{j=1}^{J-1}\pi_{itj}}\right) + \log\left(1 - \sum_{j=1}^{J-1}\pi_{itj}\right)\right\}. \end{aligned}$$

The log-likelihood is

$$l_O = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{J-1} O_{itj}\log\left(\frac{\pi_{itj}}{1-\sum_{j=1}^{J-1}\pi_{itj}}\right) + \log\left(1 - \sum_{j=1}^{J-1}\pi_{itj}\right) \tag{4.13}$$

The score equation of variables in the exponential family have a specific form (McCullagh and Nelder, 1983). Let $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_{it}^T, \ldots, \boldsymbol{\pi}_{iT_i}^T)^T$ and $\boldsymbol{O}_i = (\boldsymbol{O}_{i1}^T, \ldots, \boldsymbol{O}_{iT_i}^T)^T$. The score equation based on (4.13) is

$$U_O(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{S}_{Oi}^{-1}(\boldsymbol{O}_i - \boldsymbol{\pi}_i) = \boldsymbol{0}, \tag{4.14}$$

where $\boldsymbol{S}_{Oi}$ is a block diagonal matrix with $\text{Cov}(\boldsymbol{O}_{it}) = \text{diag}(\boldsymbol{\pi}_{it}) - \boldsymbol{\pi}_{it}\boldsymbol{\pi}_{it}^T$ on the diagonal, i.e. $\boldsymbol{S}_{Oi} = \text{diag}\{\text{Cov}(\boldsymbol{O}_{i1}), \ldots, \text{Cov}(\boldsymbol{O}_{iT_i})\}$.

In CPMs, we use cumulative indicators $Z_{itj} = \sum_{k=1}^{j} O_{itk}$ and cumulative probabilities $\mu_{itj} = \sum_{k=1}^{j}\pi_{itk}$. The underlying model is still multinomial and can be converted by a $(J-1)\times(J-1)$ matrix $\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ & \vdots & & & \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}$. The score function of CPMs can be derived by $\boldsymbol{Z}_i = \boldsymbol{L}\boldsymbol{O}_i$ and $\boldsymbol{\mu}_i = \boldsymbol{L}\boldsymbol{\pi}_i$ (McCullagh and Nelder, 1983):

$$U_Z(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{S}_{Zi}^{-1}(\boldsymbol{Z}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{4.15}$$

where $\boldsymbol{S}_{Zi} = \text{diag}\{\text{Cov}(\boldsymbol{Z}_{i1}), \ldots, \text{Cov}(\boldsymbol{Z}_{iT_i})\}$ and $\boldsymbol{S}_{Zi} = \boldsymbol{L}\boldsymbol{S}_{Oi}\boldsymbol{L}^T$.

Similarly, the information is

$$I_Z(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{S}_{Zi}^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right)^T. \tag{4.16}$$

Then the robust covariance of CPMs can be estimated based on (4.10)

$$\begin{aligned}\hat{V}_{\boldsymbol{\theta},\text{CPM}} = &\left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T\right]^{-1} \left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} (\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)\right] \\ &\left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T\right]^{-1}.\end{aligned} \tag{4.17}$$

For GEE methods, the independence working correlation indicates that $\boldsymbol{W}_i = \boldsymbol{S}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{S}_i^{\frac{1}{2}} = \boldsymbol{S}_i$, where $\boldsymbol{S}_i$ is a block diagonal matrix with $\text{Cov}(\boldsymbol{Z}_{it})$ be the diagonal elements. This means $\boldsymbol{S}_i = \boldsymbol{S}_{Zi}$. The estimating equation with independence working correlation is

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{S}_{Zi}^{-1} (\boldsymbol{Z}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \tag{4.18}$$

Now (4.15) and (4.18) are identical and thus solving the two equations would result in the same point estimations.

The covariance matrix in GEE methods assuming independence is estimated by

$$\begin{aligned}\hat{V}_{\boldsymbol{\theta},\text{GEE}} = &\left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)\right]^{-1} \left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} (\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \theta}\right)\right] \\ &\left[\sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)^T \hat{\boldsymbol{S}}_{Zi}^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\theta}}\right)\right]^{-1}.\end{aligned} \tag{4.19}$$

(4.17) and (4.19) are also identical. Therefore, we have shown that CPMs with independence working correlation is equivalent to GEE methods for ordinal response variables with independence working correlation.

4.7.3 Simulations

4.7.3.1 Complete Data

In an ideal situation, no value is missing. With complete data and the same time-varying covariate pattern across all subjects, each observation contributes approximately equally to the estimating equation, so the independence working correlation structure is as efficient as a more complex working correlation structure (Lipsitz et al., 1994).

We evaluated the performances of the two CPM methods with different association parameter α when we

Table 4.8: Simulation results for the complete data scenarios

Scenario	Method	Metric	β_X	β_T	$E(Y\|X=1,T=0.2)$	$Q(0.5\|X=1,T=0.2)$	$F(5\|X=1,T=0.2)$
$\alpha = 0.3$	GEE (correct)	Bias(%)	-0.051	0.073	-	-	-
		RMSE	0.037	0.058	-	-	-
		Coverage	0.942	0.949	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.072	0.207	-0.006	-0.031	-0.058
		RMSE	0.038	0.056	1.235	1.204	0.171
		Coverage	0.942	0.947	0.959	0.955	0.956
		RE	1.035	0.953	-	-	-
	CPM (ex)	Bias(%)	0.183	2.899	-0.338	-0.387	-0.402
		RMSE	0.038	0.065	1.219	1.188	0.170
		Coverage	0.943	0.912	0.953	0.945	0.948
		RE	1.075	1.002	-	-	-
$\alpha = 0.7$	GEE (correct)	Bias(%)	-0.006	0.063	-	-	-
		RMSE	0.050	0.038	-	-	-
		Coverage	0.950	0.945	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.157	0.300	0	-0.038	-0.122
		RMSE	0.051	0.043	1.233	1.200	0.170
		Coverage	0.945	0.950	0.953	0.956	0.950
		RE	1.041	1.251	-	-	-
	CPM (ex)	Bias(%)	0.277	2.955	-0.327	-0.394	-0.303
		RMSE	0.051	0.053	1.218	1.185	0.169
		Coverage	0.949	0.888	0.952	0.953	0.951
		RE	1.051	1.317	-	-	-

have complete data. The results are in Table 4.8. We do not expect and did not observe efficiency gain by using exchangeable working correlation with complete data. The CPM methods with independence working correlation had slightly better performance under this circumstance for its lower bias, more proper coverage rates, and similar RMSE. The CPM method was almost as efficient as the GEE method for continuous response variables with the correct transformation when the within cluster correlation is small.

4.7.3.2 Time Effects

We varied the coefficient for time, β_T, from 0 to 2 to investigate the performance under scenarios with different time effects. Results are in Table 4.9. When $\beta_T = 0$, the percent bias for both methods was ∞ because the true value (in the denominator) is 0, and the bias for the all methods was small. As the time effects increase, CPM methods were less efficient than the standard GEE method with the correct transformation, but they still had small bias and good coverage rates.

Table 4.9: Simulation results for different time effects

Scenario	Method	Metric	β_X	β_T	$E(Y\|X{=}1,T{=}0.2)$	$Q(0.5\|X{=}1,T{=}0.2)$	$F(5\|X{=}1,T{=}0.2)$
$\beta_T = 0$	GEE (correct)	Bias(%)	-0.01	Inf	-	-	-
		RMSE	0.710	0.710	-	-	-
		Coverage	0.953	0.944	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.125	Inf	-0.009	-0.079	-0.103
		RMSE	0.712	0.711	1.184	1.150	0.173
		Coverage	0.958	0.947	0.954	0.959	0.956
		RE	1.130	2.166	-	-	-
	CPM (ex)	Bias(%)	0.157	Inf	-0.121	-0.210	-0.144
		RMSE	0.712	0.709	1.180	1.145	0.174
		Coverage	0.959	0.940	0.945	0.956	0.956
		RE	1.046	1.180	-	-	-
$\beta_T = 0.5$	GEE (correct)	Bias(%)	-0.010	0.174	-	-	-
		RMSE	0.358	0.359	-	-	-
		Coverage	0.953	0.944	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.126	0.275	-0.009	-0.074	-0.089
		RMSE	0.361	0.363	1.208	1.174	0.172
		Coverage	0.957	0.944	0.955	0.957	0.957
		RE	1.130	2.189	-	-	-
	CPM (ex)	Bias(%)	0.180	3.111	-0.146	-0.236	-0.047
		RMSE	0.369	0.349	1.202	1.168	0.172
		Coverage	0.95	0.933	0.941	0.954	0.958
		RE	1.046	1.204	-	-	-
$\beta_T = 2$	GEE (correct)	Bias(%)	-0.020	0.047	-	-	-
		RMSE	0.710	0.711	-	-	-
		Coverage	0.953	0.944	-	-	-
		RE	baseline	baseline	-	-	-
	CPM (ind)	Bias(%)	0.132	0.279	-0.013	-0.086	-0.188
		RMSE	0.708	0.727	1.285	1.255	0.165
		Coverage	0.960	0.945	0.959	0.956	0.955
		RE	1.130	2.637	-	-	-
	CPM (ex)	Bias(%)	0.411	2.766	-0.266	-0.358	-0.147
		RMSE	0.706	0.751	1.272	1.242	0.165
		Coverage	0.957	0.888	0.951	0.947	0.949
		RE	1.061	1.728	-	-	-

Table 4.10: The odds ratios with 95% confidence intervals from standard GEE with exchangeable working correlation on log-transformed CD4:CD8 ratio.

Predictor	Odds Ratio	CI
Time (in years)	1.079	(1.061, 1.097)
Enrollment Year	0.996	(0.989 1.002)
Race		
African American	(Reference)	
Caucasian	1.000	(0.950, 1.052)
Hispanic	0.871	(0.740, 1.025)
Other	0.859	(0.736,1.004)
Baseline Age	0.986	(0.984, 0.988)
Sex		
Male	(Reference)	
Female	1.202	(1.187, 1.218)
Route		
Heterosexual	(Reference)	
Injection Drug Use	1.017	(0.927, 1.116)
MSM	0.946	(0.938, 0.954)
Other/Unknown	0.919	(0.750, 1.127)
HCV	0.946	(0.914, 0.980)
HBV	1.010	(0.853, 1.197)

4.7.4 Applications

4.7.4.1 CD4:CD8 Ratio

We analyzed the data using standard GEE methods with exchangeable working correlation after log-transforming the CD4:CD8 ratios. Results in Table 4.10 suggested that time, baseline age, sex, route, and HCV have significant association with the outcome. The conclusion and estimations were different from the results from CPM methods. Log-transformation might be a natural choice for such right-skewed responses, but not the optimal transformation for CD4:CD8 ratios.

4.7.4.2 The Lung Health Study

The distribution of FEV1 at the first follow-up visit is shown in Figure 4.4.

In Figure 4.5, we show the conditional mean and median of FEV1, and the conditional probability of FEV1 being less than or equal to 2L while fixing other covariates at median (continuous covariates) or mode (categorical covariates).

We adjusted for baseline FEV1 in the models and the results are in Table 4.11.

The conditional quantities after adjusting for baseline FEV1 are shown in Figure 4.6.

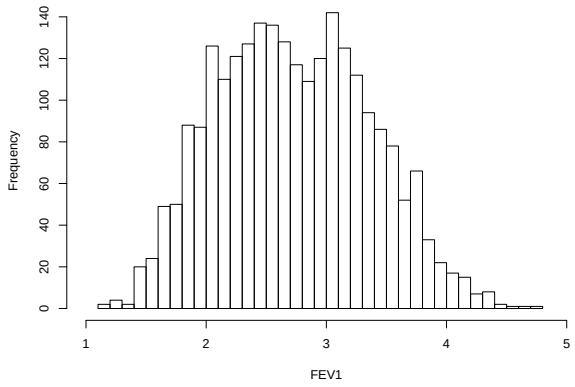

Figure 4.4: The histogram of the FEV1 measured at the first follow-up visit of participants in The Lung Health Study who were smokers for all 5 visits with at minimum 2 visits.

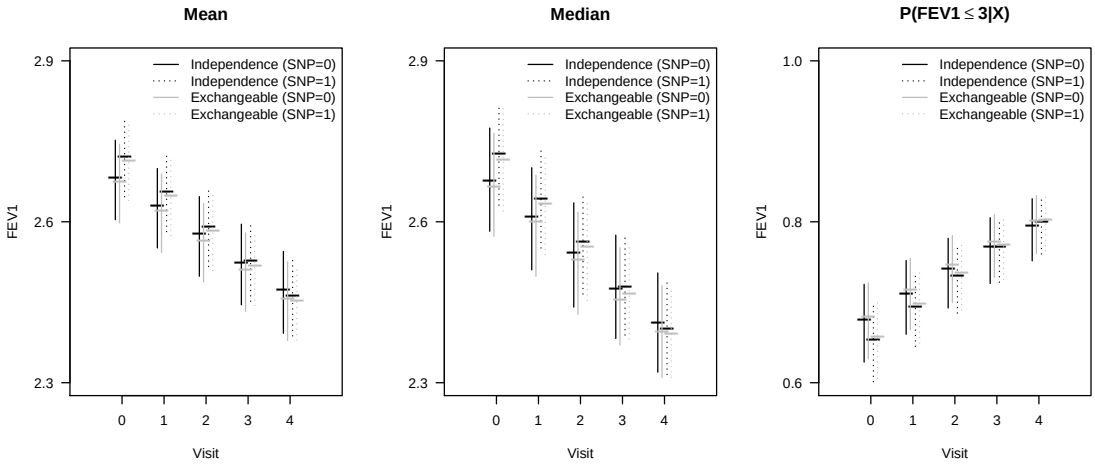

Figure 4.5: The estimated conditional mean, median of FEV1 and the conditional probability of FEV1 being less than or equal to 2 as functions of study visit while fixing other covariates at their medians (for continuous covariates) or modes (for categorical covariates) under the circumstances that rs12194741 is present (dotted lines) and is not present (solid lines).

Table 4.11: We show odds ratios estimates for higher FEV1 with 95% confidence intervals from CPMs with independence and exchangeable working correlation adjusting for the baseline FEV1. The last column shows the VRs calculated by the variances of the log-odds ratios from CPMs with exchangeable working correlation divided by the variances of the log-odds ratios from CPMs with independence working correlation.

Predictor	Independence	Exchangeable	VR
Visit:rs12194741	0.946 (0.909, 0.984)	0.947 (0.912, 0.983)	0.880
Visit	0.608 (0.919, 1.157)	0.607 (0.588, 0.626)	0.934
rs12194741	1.031 (0.919, 1.157)	1.003 (0.882, 1.140)	1.244
BMI Change (per 5 kg/m^2)	0.344 (0.289, 0.410)	0.336 (0.293, 0.386)	0.619
Baseline Age (per 10–year)	0.845 (0.761, 0.938)	0.853 (0.768, 0.946)	0.993
Baseline BMI (per 5 kg/m^2)	1.075 (0.999, 1.157)	1.081 (1.002, 1.116)	1.071
Cigarettes/day (per 10 cigs/day)	0.936 (0.891, 0.984)	0.941 (0.897, 0.988)	0.940
Pack Years (per 20 pack year)	0.960 (0.888, 1.039)	0.957 (0.885, 1.035)	1.000
Study Site			
1	(Reference)		
2	1.642 (1.251, 2.155)	1.634 (1.246, 2.143)	0.995
3	0.984 (0.755, 1.283)	0.982 (0.750, 1.286)	1.039
4	1.654 (1.274, 2.147)	1.632 (1.248, 2.134)	1.057
5	1.284 (0.980, 1.682)	1.250 (0.959, 1.630)	0.966
6	1.409 (1.079, 1.840)	1.403 (1.069, 1.842)	1.041
7	0.888 (0.680, 1.159)	0.891 (0.676, 1.173)	1.068
8	1.163 (0.895, 1.509)	1.147 (0.883, 1.491)	1.005
9	1.951 (1.487, 2.560)	1.918 (1.474, 2.496)	0.940
10	1.072 (0.781, 1.472)	1.059 (0.781, 1.436)	0.923
Baseline FEV1	5497.550 (4231.518, 7142.369)	4782.765 (3719.052, 6150.717)	0.924

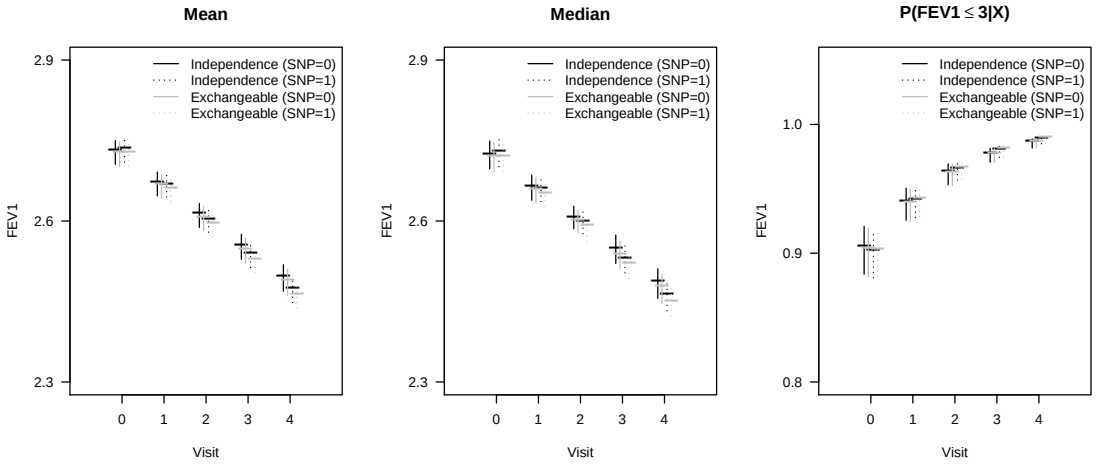

Figure 4.6: The estimated conditional mean, median of FEV1 and the conditional probability of FEV1 being less than or equal to 2 as functions of study visit while fixing other covariates (including the baseline FEV1) at their medians (for continuous covariates) or modes (for categorical covariates) under the circumstances that rs12194741 is present (dotted lines) and is not present (solid lines).

CHAPTER 5

Conclusion

5.1 Summary

Continuous data are commonly encountered in practice and often are too complicated to be simply modeled by linear regression. In this dissertation, we address skewed, censored, and clustered continuous response data with semiparametric CPMs. The goal of the research has been to present more robust and flexible approaches to analyze different types of continuous response data. .

In Chapter 2, we studied the similarities and differences of two transformation models, CPMs and MLTs. Both models assume that a response variable can be modeled linearly with errors following a specified distribution after an unspecified monotonic transformation. Therefore, neither of the models require transformation of response variables before modeling, but estimate the transformation as part of the modeling procedure. The two models mainly differ by the way that they estimate the transformation. CPMs regard each distinct response value as its own ordinal level, and estimate the transformation nonparametrically with a step function, while MLTs utilize parametric basis functions for the estimation of the transformation. They both had good and similar performance in most cases, even with complex transformation and skewed distributions. We note that they handle left censored response data differently. MLTs make assumptions on the distribution of censored data based on uncensored data, while CPMs treat censored values as the smallest ordinal level. This leads us to the next chapter focusing on censored continuous response data.

In Chapter 3, we proposed approaches to analyze continuous response variables subject to DLs based on CPMs. Most existing methods for DLs require dichotomization, single imputation, or distribution assumptions for values outside DLs. CPMs, an ordinal regression models, do not make such assumptions and use all available information. Continuous data subject to DLs effectively follow a mixture distribution of discrete and continuous data. We described our approaches in two scenarios, single DLs, where there is a single lower DL and/or upper DL, and multiple DLs, where data are collected from different sites/times with different DLs or no DL. With single DLs, CPMs, as ordinal regression models, make minimal assumptions on distributions outside DLs, and assign ordinal level to unobserved values. We extended CPMs to address multiple DLs by modifying the CPM likelihood to appropriately distribute probability mass. In addition, we proposed a new estimator for conditional quantiles from CPMs that is more interpretable with DLs. Our approaches had good performance even with small sample sizes and large censoring rates. The work is accompanied by a new R package, **multipleDL**.

In Chapter 4, we extended CPMs for clustered continuous data to avoid transformation of response variables. In previous chapters, we estimated parameters in CPMs based on the likelihood. For clustered continuous data, parameters are estimated by GEE to account for correlated observations within each cluster. To overcome computation limitations due to the large number of ordinal levels with continuous data, we proposed two feasible approaches. With independence working correlation, CPMs can be directly applied to obtain the point estimation, and then we use a robust estimator for covariance to correct for misspecification of correlation structure. To gain greater efficiency by specifying a more complex working correlation, we propose a one-step GEE estimator for CPMs using the framework of repolr for estimating the association parameter. Our approaches work well under a variety of simulation settings. We have built an R package, **cpmgee**, for CPMs to fit clustered continuous response variables with independence, uniform and AR1 working correlation.

We hope that our work provides new directions and tools for researchers to analyze skewed, censored, and clustered continuous response data.

5.2 Future Research

All CPM-based approaches described in this dissertation make an implicit parallelism assumption that $\boldsymbol{\beta}$ does not depend on j, the order of response values. In practice, this assumption might be violated. For example, the effects of at least on covariate might vary across different cut-off points. In the future, we might consider developing more flexible partial CPMs that allow for different relationships for different covariate levels. Partial CPMs can also be useful if addressing composite endpoints (e.g., death, heart disease, and blood pressure) longitudinally in a single analysis.

CPMs assume the error terms are independent and identically distributed. With correlated observations, this assumption is violated and we proposed approaches in Chapter 4 to deal with it by treating correlation as nuisance parameters with GEE. Another solution could be weighted CPMs to put different weights on observations.

GEE methods assume observations are missing completely at random (MCAR), which can be a strong assumption in practice. We could extend our methods under the less restrictive missingness assumption of missing at random (MAR).

References

Acion, L., Peterson, J. J., Temple, S., and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25(4):591–602.

Agresti, A. (2003). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, volume 656. John Wiley & Sons.

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A. C., Bonzini, M., Patterson Jr, D. G., Bertazzi, P. A., and Landi, M. T. (2005). Handling of dioxin measurement data in the presence of non-detectable values: overview of available methods and their application in the Seveso chloracne study. *Chemosphere*, 60(7):898–906.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

Castilho, J. L., Shepherd, B. E., Koethe, J., Turner, M., Bebawy, S., Logan, J., Rogers, W. B., Raffanti, S., and Sterling, T. R. (2016). CD4/CD8 ratio, age, and risk of serious non-communicable diseases in HIV-infected adults on antiretroviral therapy. *AIDS*, 30(6):899.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

da Silva, C. M., de Peder, L. D., Silva, E. S., Previdelli, I., Pereira, O. C. N., Teixeira, J. J. V., and Bertolini, D. A. (2018). Impact of HBV and HCV coinfection on CD4 cells among HIV-infected patients: a longitudinal retrospective study. *The Journal of Infection in Developing Countries*, 12(11):1009–1018.

De Neve, J., Thas, O., and Gerds, T. A. (2019). Semiparametric linear transformation models: Effect measures, estimators, and applications. *Statistics in Medicine*, 38(8):1484–1501.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Fedorov, V., Mannino, F., and Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1):50–61.

Fiévet, B. and Della Vedova, C. (2010). Dealing with non-detect values in time-series measurements of radionuclide concentration in the marine environment. *Journal of Environmental Radioactivity*, 101(1):1–7.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302.

Garland, M., Morris, J. S., Rosner, B. A., Stampfer, M. J., Spate, V. L., Baskett, C. J., Willett, W. C., and Hunter, D. J. (1993). Toenail trace element levels as biomarkers: reproducibility over a 6-year period. *Cancer Epidemiology and Prevention Biomarkers*, 2(5):493–497.

Genter, F. C. and Farewell, V. T. (1985). Goodness-of-link testing in ordinal regression models. *Canadian Journal of Statistics*, 13(1):37–44.

Gras, L., May, M., Ryder, L. P., Trickey, A., Helleberg, M., Obel, N., Thiebaut, R., Guest, J., Gill, J., Crane, H., et al. (2019). Determinants of restoration of CD4 and CD8 cell counts and their ratio in HIV-1–positive individuals with sustained virological suppression on antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes*, 80(3):292.

Hansel, N. N., Ruczinski, I., Rafaels, N., Sin, D. D., Daley, D., Malinina, A., Huang, L., Sandford, A., Murray, T., Kim, Y., et al. (2013). Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Human Genetics*, 132(1):79–90.

Harel, O., Perkins, N., and Schisterman, E. F. (2014). The use of multiple imputation for data subject to limits of detection. *Sri Lankan Journal of Applied Statistics*, 5(4):227.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, 26(16):3057–3077.

Harrell, F. (2020). rms: Regression modeling strategies. R package version 6.1.0.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harrell Jr, F. E. et al. (2016). rms: Regression modeling strategies. *R package version*, 5(2).

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91(435):1024–1036.

Helsel, D. R. (2011). *Statistics for Censored Environmental Data Using Minitab and R*, volume 77. John Wiley & Sons.

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.

Hothorn, T. (2018). *mlt: Most Likely Transformations*. R package version 1.0-4.

Hothorn, T. (2019). Transformation boosting machines. *Statistics and Computing*.

Hothorn, T. (2020). Most likely transformations: The mlt package. *Journal of Statistical Software*, 92:1–68.

Hothorn, T., Möst, L., and Bühlmann, P. (2018). Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134.

Hothorn, T. and Zeileis, A. (2017). Transformation forests. Technical report, arXiv 1701.02110, v2.

Huang, G.-H., Bandeen-Roche, K., and Rubin, G. S. (2002). Building marginal models for multiple ordinal measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(1):37–57.

Jiamsakul, A., Kariminia, A., Althoff, K. N., Cesar, C., Cortes, C. P., Davies, M.-A., Do, V. C., Eley, B., Gill, J., Kumarasamy, N., et al. (2017). HIV viral load suppression in adults and children receiving antiretroviral therapy–results from the IeDEA collaboration. *Journal of Acquired Immune Deficiency Syndromes*, 76(3):319.

Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, pages 945–953.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Koethe, J. R., Bian, A., Shintani, A. K., Boger, M. S., Mitchell, V. J., Erdem, H., and Hulgan, T. (2012). Serum leptin level mediates the association of body composition and serum C-reactive protein in HIV-infected persons on antiretroviral therapy. *AIDS Research and Human Retroviruses*, 28(6):552–557.

Koethe, J. R., Grome, H., Jenkins, C. A., Kalams, S. A., and Sterling, T. R. (2016). The metabolic and cardiovascular consequences of obesity in persons with HIV on long-term antiretroviral therapy. *AIDS*, 30(1):83.

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., and Rathouz, P. J. (2008). Sleep duration: how well do self-reports reflect objective measures? The CARDIA Sleep Study. *Epidemiology*, 19(6):838.

Li, C., Chen, G., and Shepherd, B. E. (2022a). Fitting semiparametric cumulative probability models for big data. *Submitted*.

Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480.

Li, C., Zeng, D., Tian, Y., and Shepherd, B. E. (2022b). A semiparametric transformation model for continuous outcomes. *In preparation*.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford University Press.

Lipsitz, S., Fitzmaurice, G., Sinha, D., Hevelone, N., Hu, J., and Nguyen, L. L. (2017). One-step generalized estimating equations with large cluster sizes. *Journal of Computational and Graphical Statistics*, 26(3):734–737.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13(11):1149–1163.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78(1):153–160.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Liu, Q., Shepherd, B. E., Li, C., and Harrell Jr, F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in Medicine*, 36(27):4316–4335.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*, 112(17):1691–1696.

Manuguerra, M. and Heller, G. Z. (2010). Ordinal regression models for continuous scales. *The International Journal of Biostatistics*, 6(1):14.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Routledge.

Pan, W., Wu, H., Luo, J., Deng, Z., Ge, C., Chen, C., Jiang, X., Yin, W.-J., Niu, G., Zhu, L., et al. (2017). Cs2AgBiBr6 single-crystal X-ray detectors with a low detection limit. *Nature Photonics*, 11(11):726–732.

Parsons, N. (2017). repolr: an R package for fitting proportional-odds models to repeated ordinal scores.

Parsons, N. R., Costa, M. L., Achten, J., and Stallard, N. (2009). Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Computational Statistics & Data Analysis*, 53(3):632–641.

Parsons, N. R., Edmondson, R., and Gilmour, S. (2006). A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):507–524.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-simulation and Computation*, 23(4):939–951.

Petoumenos, K., Choi, J. Y., Hoy, J., Kiertiburanakul, S., Ng, O. T., Boyd, M., Rajasuriar, R., and Law, M. (2017). CD4:CD8 ratio comparison between cohorts of HIV-positive Asians and Caucasians upon commencement of antiretroviral therapy. *Antiviral Therapy*, 22(8):659–668.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, pages 825–839.

Sauter, R., Huang, R., Ledergerber, B., Battegay, M., Bernasconi, E., Cavassini, M., Furrer, H., Hoffmann, M., Rougemont, M., Günthard, H. F., et al. (2016). CD4/CD8 ratio and CD8 counts predict CD4 response in HIV-1-infected drug naive and in patients on cART. *Medicine*, 95(42).

Schildcrout, J. S., Haneuse, S., Tao, R., Zelnick, L. R., Schisterman, E. F., Garbett, S. P., Mercaldo, N. D., Rathouz, P. J., and Heagerty, P. J. (2020). Two-phase, generalized case-control designs for the study of quantitative longitudinal outcomes. *American Journal of Epidemiology*, 189(2):81–90.

Serrano-Villar, S., Caruana, G., Zlotnik, A., Pérez-Molina, J. A., and Moreno, S. (2017). Effects of maraviroc versus efavirenz in combination with zidovudine-lamivudine on the CD4/CD8 ratio in treatment-naive HIV-infected individuals. *Antimicrobial Agents and Chemotherapy*, 61(12):e01763–17.

Shepherd, B. E., Li, C., and Liu, Q. (2016). Probability-scale residuals for continuous, discrete, and censored data. *Canadian Journal of Statistics*, 44(4):463–479.

Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, pages 592–607.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Steegen, K., Luchters, S., De Cabooter, N., Reynaerts, J., Mandaliya, K., Plum, J., Jaoko, W., Verhofstede, C., and Temmerman, M. (2007). Evaluation of two commercially available alternatives for HIV-1 viral load testing in resource-limited settings. *Journal of Virological Methods*, 146(1-2):178–187.

Tian, Y., Hothorn, T., Li, C., Harrell Jr, F. E., and Shepherd, B. E. (2020). An empirical comparison of two novel transformation models. *Statistics in Medicine*, 39(5):562–576.

Touloumis, A., Agresti, A., and Kateri, M. (2013). GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69(3):633–640.

Tsao, C.-H., Shiau, M.-Y., Chuang, P.-H., Chang, Y.-H., and Hwang, J. (2014). Interleukin-4 regulates lipid metabolism by inhibiting adipogenesis and promoting lipolysis. *Journal of Lipid Research*, 55(3):385–397.

Tukey, J. W. et al. (1957). On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, 28(3):602–632.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

Wing, S., Shy, C. M., Wood, J. L., Wolf, S., Cragle, D. L., and Frome, E. (1991). Mortality among workers at Oak Ridge National Laboratory: evidence of radiation effects in follow-up through 1984. *Journal of the American Medical Association*, 265(11):1397–1402.

Wu, L., Thompson, D. K., Li, G., Hurt, R. A., Tiedje, J. M., and Zhou, J. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology*, 67(12):5780–5790.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564.

Zhang, D., Fan, C., Zhang, J., and Zhang, C.-H. (2009). Nonparametric methods for measurements below detection limit. *Statistics in Medicine*, 28(4):700–715.