Rank-Based Semiparametric Methods: Covariate-Adjusted Spearman's Correlation with Probability-Scale Residuals and Cumulative Probability Models

By

Qi Liu

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

August, 2016

Nashville, Tennessee

Approved:

Bryan E. Shepherd, Ph.D.

Frank E. Harrel, Ph.D.

Jonathan S. Schildcrout, Ph.D.

David W. Haas, M.D.

ACKNOWLEDGMENTS

First of all, I would like to take this opportunity to thank all the faculty, staff, and graduate students in the Department of Biostatistics. I am extremely fortunate to be part of this big family. Studying at Vanderbilt University is one of the best decisions I have made in my life.

Especially, I would like to thank Dr. Bryan Shepherd for his excellent guidance and constant inspiration. As my advisor, he has taught and helped me more than I could ever give him credit for. He has shown me what a good statistician (and person) should be. I am also very grateful to all the other members of my dissertation committee, Drs. Frank Harrel, Jonathan Schildcrout, and David Haas. Each of them has given me valuable suggestions on my research and my professional development. Besides, I would like to thank Dr. Chun Li. Although he has left Vanderbilt and is not in my dissertation committee formally, I have received tremendous guidance from him and have benefited a lot from the weekly group meetings with him.

Special thanks go to Dr. Jeffrey Blume and Linda Wilson for making our graduate program happen and for making it a great place to stay.

Finally, I would like to thank my parents for always believing in me and encouraging me to do my best. Without their love, this journey would not have been possible.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

LIST OF TABLES v

LIST OF FIGURES vii

ABSTRACT x

Chapter

1 Introduction 1

1.1 Rank-Based Correlations with Covariate-Adjustment 1
1.2 Rank-Based Regression Models 2
1.3 Overview of the Dissertation 4

2 Covariate-Adjusted Spearman's Rank Correlation with Probability-Scale Residuals 6

2.1 Introduction 6
2.2 Covariate-Adjusted Spearman's Rank Correlations 9
2.2.1 Spearman's Rank Correlation and PSRs 9
2.2.2 Conditional Spearman's Rank Correlation 11
2.2.3 Partial Spearman's Rank Correlation 11
2.2.4 Conditional Partial Spearman's Rank Correlation 12
2.3 Estimation and Inference 13
2.3.1 Calculation of PSRs 13
2.3.2 Standard Error of Partial Estimators 14
2.3.3 Conditional Estimators 16
2.4 Numerical Illustrations and Simulations 17
2.5 Application Examples 21
2.5.1 Wage and Education 21
2.5.2 SCIP Survey Data 24
2.6 Discussion 27
2.7 Supplemental Materials 28
2.7.1 Derivation of Population Parameter of Our Partial Estimator 28
2.7.2 Derivation of Population Parameter of Our Conditional Partial Estimator 29

3 Modeling Continuous Outcomes Using Ordinal Regression with Cumulative Probabilities . . . 32

3.1 Introduction . . . 32
3.2 Cumulative Probability Models for Continuous Outcomes . . . 34
3.2.1 Latent Variable Motivation . . . 34
3.2.2 Nonparametric Maximum Likelihood Estimation . . . 37
3.2.3 An Illustration of Cumulative Probability Models . . . 41
3.2.4 Assumptions of Cumulative Probability models . . . 43
3.2.5 Model Diagnostics . . . 46
3.3 Simulation Studies . . . 47
3.3.1 Estimation with Proper Link Function Specification . . . 47
3.3.2 Estimation with Link Function Misspecification . . . 56
3.4 Application Examples . . . 62
3.4.1 CD4 Count . . . 62
3.4.2 Viral Load . . . 67
3.5 Discussion . . . 70
3.6 Supplemental Materials . . . 74
3.6.1 Performance of Cumulative Probability Models with an Automatic Link Function Selection Procedure . . . 74
3.6.2 Details of Simulation Results . . . 75
3.6.3 Cumulative Probability Models for Measurements Subject to Detection Limits . . . 93

4 PResiduals: An R Package for Residual Analysis Using Probability-Scale Residuals . . . 95

4.1 Introduction . . . 95
4.2 Review of Methods . . . 96
4.2.1 PSRs . . . 96
4.2.2 Test of Residual Correlation with PSRs . . . 97
4.2.3 Covariate-Adjusted Spearman's Rank Correlation with PSRs . . . 98
4.3 Analysis with the PResiduals Package . . . 100
4.3.1 Wage Data . . . 100
4.3.2 Calculation of PSRs . . . 101
4.3.3 Tests of Conditional Association . . . 114
4.3.4 Covariate-Adjusted Spearman's Rank Correlation with PSRs . . . 117
4.4 Summary . . . 123

5 Conclusions . . . 124

5.1 Summary . . . 124
5.2 Future Research . . . 124

REFERENCES . . . 126

LIST OF TABLES

Table Page

2.1 Simulation results: estimation of our partial Spearman's rank correlation with PSRs 20

2.2 Simulation results: the performance of our partial Spearman's rank correlation on testing covariate-adjusted associations 21

3.1 Commonly used link functions and their corresponding error distributions. . 35

3.2 Simulation results: the performance of cumulative probability models with an automatic link function selection procedure 76

3.3 Simulation results: the performance of cumulative probability models on estimating slopes and intercepts with proper link function specification . . . 77

3.4 Simulation results: the performance of cumulative probability models on estimating conditional CDFs with proper link function specification 78

3.5 Simulation results: the performance of cumulative probability models on estimating conditional means with proper link function specification 79

3.6 Simulation results: the performance of cumulative probability models on estimating conditional quantiles with proper link function specification . . . 80

3.7 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (a) – (d) with the sample size of 50 81

3.8 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (e) – (h) with the sample size of 50 82

3.9 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (a) – (d) with the sample size of 50 83

3.10 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (e) – (h) with the sample size of 50 84

3.11 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (a) – (d) with the sample size of 100 85

3.12 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (e) – (h) with the sample size of 100 86

3.13 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (a) – (d) with the sample size of 100 87

3.14 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (e) – (h) with the sample size of 100 88

3.15 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (a) – (d) with the sample size of 200 89

3.16 Simulation results: the performance of cumulative probability models on estimating conditional means for Scenarios (e) – (h) with the sample size of 200 90

3.17 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (a) – (d) with the sample size of 200 91

3.18 Simulation results: the performance of cumulative probability models on estimating conditional medians for Scenarios (e) – (h) with the sample size of 200 92

3.19 Simulation results: the performance of cumulative probability models for outcomes subject to detection limits 94

LIST OF FIGURES

Figure Page

2.1 Population parameters of partial and conditional Spearman's rank correlation in Scenarios I and III 18

2.2 Application results: the age-specific conditional Spearman's rank correlation between wage and education 23

2.3 Application results: the heatmap of our covariate-adjusted partial Spearman's correlation matrix for responses from the SCIP survey 26

3.1 Non-smooth transformation function $H(\cdot)$ for semiparametric transformation models 36

3.2 The sparse structure of score function and hessian matrix of cumulative probability models 39

3.3 Estimation of conditional CDFs and quantiles with cumulative probability models 40

3.4 The estimated conditional CDF from cumulative probability models compared with that from parametric and nonparametric models in a simple example 42

3.5 The assumption of parallelism of cumulative probability models 44

3.6 Simulation results: the performance of cumulative probability models on estimating proper transformations 48

3.7 Simulation results: the performance of cumulative probability models on estimating slopes and intercepts with proper link function specification . . . 49

3.8 Simulation results: the relative efficiency of properly specified cumulative probability models 52

3.9 Simulation results: the performance of cumulative probability models on estimating conditional CDFs with proper link function specification 53

3.10 Simulation results: the performance of cumulative probability models on estimating conditional means with proper link function specification 54

3.11 Simulation results: the performance of cumulative probability models on estimating conditional quantiles with proper link function specification . . . 55

3.12 The extent of violation to the parallel assumption with commonly used link functions for Scenarios (a) – (d) 58

3.13 The extent of violation to the parallel assumption with commonly used link functions for Scenarios (e) – (h) 59

3.14 Simulation results: the performance of cumulative probability models on estimating conditional means with commonly used link functions 60

3.15 Simulation results: the performance of cumulative probability models on estimating conditional medians with commonly used link functions 61

3.16 Application results for CD4 count: the estimated transformation and QQ plots for model diagnostics 63

3.17 Application results for CD4 count: residuals-by-predictor plots for model diagnostics 65

3.18 Application results for CD4 count: the estimated conditional mean, median, and probabilities 66

3.19 Application results for viral load: the estimated transformation and QQ-plots for model diagnostics 70

3.20 Application results for viral load: residuals-by-predictor plots for model diagnostics 71

3.21 Application results for viral load: the estimated conditional probabilities and quantiles 72

4.1 The `PResiduals` package: residual-by-predictor plots with PSRs from proportional odds models 103

4.2 The `PResiduals` package: PSRs from linear regression models with different assumptions 105

4.3 The `PResiduals` package: residual-by-predictor plots with PSRs from linear regression models 106

4.4 The `PResiduals` package: QQ-plots with PSRs from cumulative probability models with different link functions 108

4.5 The `PResiduals` package: residual-by-predictor plots with PSRs from cumulative probability models 109

4.6 The `PResiduals` package: QQ-plots with Cox-Snell-like PSRs from parametric survival models 111

4.7 The `PResiduals` package: residual-by-predictor plots with PSRs from parametric survival models 112

4.8 The `PResiduals` package: the QQ-plot with Cox-Snell-like PSRs and residual-by-predictor plots with standard PSRs for Cox proportional hazards models 114

4.9 The `PResiduals` package: modeling conditional Spearman's rank correlation with linear models 121

4.10 The `PResiduals` package: modeling conditional Spearman's rank correlation with kernel smoothing 122

ABSTRACT

In this dissertation, we develop semiparametric rank-based methods. These types of methods are particularly useful with skewed data, nonlinear relationships, and truncated measurements. Semiparametric rank-based methods can achieve a good balance between robustness and efficiency.

The first part of this dissertation develops new estimators for covariate-adjusted Spearman's rank correlation, both partial and conditional, using probability-scale residuals (PSRs). These estimators are consistent for natural extensions of the population parameter of Spearman's rank correlation in the presence of covariates and are general for both continuous and discrete variables. We evaluate their performance with simulations and illustrate their application in two examples. To preserve the rank-based nature of Spearman's correlation, we obtain PSRs from ordinal cumulative probability models for both discrete and continuous variables.

Cumulative probability models were first invented to handle discrete ordinal outcomes, and their potential utility for the analysis of continuous outcomes has been largely unrecognized. This motivates the second part of this dissertation: an in-depth study of the application of cumulative probability models to continuous outcomes. When applied to continuous outcomes, these models can be viewed as semiparametric transformation models. We present a latent variable motivation for these models; describe estimation, inference, assumptions, and model diagnostics; conduct extensive simulations to investigate the finite sample performance of these models with and without proper link function specification; and illustrate their application in an HIV study.

Finally, we developed an R package, `PResiduals`, to compute PSRs, to incorporate them into conditional tests of association, and to implement our covariate-adjusted Spearman's rank correlation. The third part of this dissertation contains a vignette for this package, in which we illustrate its usage with a publicly available dataset.

Chapter 1

Introduction

Rank-based statistical methods, such as Spearman's rank correlation and Wilcoxon's rank sum test, are frequently used in biomedical and social science research, especially for data with nonlinear relationships, skewed distributions, extreme values, and censored observations. These methods use order information but do not rely on parametric distribution assumptions, and therefore, are favored for their robustness. However, many classic rank-based tests do not easily handle covariates and in general, nonparametric methods are inefficient when there are covariates. To achieve a good balance between robustness and efficiency, we develop and study semiparametric rank-based methods in this dissertation. Specifically, we extend Spearman's rank correlation for covariate adjustment and investigate the use of ordinal cumulative probability models to continuous outcomes.

This chapter provides some background and briefly reviews the literature for covariate-adjusted rank correlations and rank-based semiparametric regression models. It also provides an overview of this dissertation.

1.1 Rank-Based Correlations with Covariate-Adjustment

Correlation coefficients are commonly used to summarize the degree of association between two variables. For well behaved continuous variables, a common choice is Pearson's correlation coefficient. When dealing with ordered categorical data, nonlinear relationships, skewed distributions, and extreme values, rank correlation coefficients such as Spearman's rho or Kendall's tau are preferred for their robustness. Spearman's rank correlation is simply Pearson's correlation between the ranks of observations (Spearman, 1904), whereas Kendall's rank correlation measures concordance and discordance among all possible pair combinations (Kendall, 1942).

In many applications, it is desirable to adjust correlation coefficients for the influence of other variables. There are generally two types of covariate-adjusted correlations in the literature: partial and conditional correlations.

The partial correlation removes the effect of covariates and summarizes the relationship with a single number. It was originally developed for Pearson's correlation with normal data (Fisher, 1924), and has been extended by plugging rank-based estimates into the formula for Pearson's correlation (Kendall, 1942). However, these rank-based partial correlations are ad hoc, do not correspond with sensible population parameters, and at least for the partial Kandall's correlation have theoretical problems that make them not useful (Kendall, 1942; Korn, 1984; Gripenberg, 1992).

The conditional correlation assesses the relationship at specific levels of covariates. Rank-based conditional correlations have been studied for continuous data with copulas (Nelsen, 2006) and have been extended to adjust for covariates (Gijbels et al., 2011). However, these methods cannot be directly extended to discrete data because copula functions are not uniquely defined for discrete data (Genest & Nešlehová, 2007; Nešlehová, 2007).

Covariate-adjusted correlations can be constructed with residuals. For example, the partial Pearson's correlation between X and Y adjusting for Z can be computed as the correlation between the observed-minus-expected residuals (OMER) resulting from linear regression models of X on Z and of Y on Z (Fisher, 1924). Our methods are similarly motivated. We construct partial and conditional covariate-adjusted Spearman's rank correlations using the correlation between probability-scale residuals (Li & Shepherd, 2012; Shepherd et al., in press) resulting from semiparametric transformation models of X on Z and of Y on Z.

1.2 Rank-Based Regression Models

Rank-based regression models for continuous variables are attractive due to their robustness. They only use the order information of response variables and are therefore in-

variant to any monotonic transformation of outcomes. This is particularly useful when the distributions of continuous responses are skewed and different transformations may give conflicting results.

An intuitive yet flawed rank-based regression method is simply replacing outcome variables with their ranks and then fitting them with linear regression models. This approach has been seen in practice occasionally for data with extreme values or skewed distributions (Tian et al., 2014). Although this approach is easy to implement, it lacks justification: ranks are nonparametric statistics and they vary in different datasets; modeling their conditional distribution does not make sense. Regression coefficients and predicted values using this approach are generally not interpretable.

Another type of rank-based regression is linear regression with an unspecified monotonic transformation of the response variable. For example, consider the linear transformation model $Y = H(\beta X + \varepsilon)$, where Y is a continuous response variable, X is the p-dimensional covariates, ε is the error term, and $H(\cdot)$ is an unspecified monotonic transformation function. This model is invariant to any monotonic transformation of Y since $H(\cdot)$ is unspecified. There is a large literature for linear transformation models. For example, when both $H(\cdot)$ and the distribution of error term ε are unspecified, the model is nonparametric. Several rank-based estimators for β and their asymptotic properties have been studied, including the maximum rank correlation estimator (Han, 1987), the monotone rank estimator (Cavanagh & Sherman, 1998), the partial rank estimator (Khan & Tamer, 2007) and a modified partial rank estimator (Song et al., 2007). If the distribution of error term ε is assumed, the model is semiparametric and has been studied extensively for censored data in survival analysis. With Gumbel minimum error distribution (the type I extreme value distribution), this model is the well-known proportional hazards model (Cox, 1972). For this semiparametric transformation model with censored outcomes, estimation procedures based on partial likelihood Cox (1975), rank-based estimation equations (Cheng et al., 1995; Chen et al., 2002), and nonparametric maximum likelihood estimation (NPMLE)

(Bennett, 1983; Murphy et al., 1997; Zeng & Lin, 2006) have been studied. The consistency, asymptotic normality, and asymptotic efficiency of the NPMLE have been established by Zeng & Lin (2006) for censored outcomes. This semiparametric transformation model has also been considered for uncensored continuous outcomes but the asymptotic properties of the NPMLE have not been developed (Zeng & Lin, 2006).

Ordinal regression models, specifically cumulative probability models, are also rank-based. They are closed related to the nonparametric rank-based test. Specifically, when there is only a single binary covariate, the score test of a cumulative probability model with the logit link is equivalent to the Wilcoxon rank sum test (McCullagh, 1980). Cumulative probability models also have natural connections with linear transformation models. For example, cumulative probability models for discrete ordinal variables can be motivated from linear transformation models where the transformation maps a latent continuous variable to the observed discrete response. Although cumulative probability models were initially invented for discrete ordinal outcomes, they can be used for continuous variables because continuous variables are also ordinal. However, this approach is typically not used in practice, we believe largely due to a lack of awareness.

1.3 Overview of the Dissertation

In Chapter 2, we propose new estimators for covariate-adjusted Spearman's rank correlation, both partial and conditional, using probability-scale residuals. Our estimators are consistent for natural extensions of the population parameter of Spearman's rank correlation in the presence of covariates and are general for both continuous and discrete variables. We describe estimation and inference, and highlight the use of cumulative probability models, which allow our method to preserve the rank-based nature of Spearman's correlation. We conduct simulations to evaluate the performance of our estimators and compare them with other popular measures of association, demonstrating their robustness and efficiency. We illustrate our method in two application examples.

In Chapter 3, we study the application of cumulative probability models for continuous outcomes. We present a latent variable motivation for these models, describe estimation and inference, and discuss model assumptions. Extensive simulations are performed to investigate the finite sample performance of these models with and without correct link function specification. We illustrate their application in an HIV study.

These methods will be of greatest use if they can be easily implemented using standard statistical software. We have created an R package `Presiduals` that computes probability-scale residuals for a wide range of statistical models, including the cumulative probability models studied in Chapter 3. This package also implements both the partial and the conditional covariate-adjusted Spearman's rank correlations developed in Chapter 2. In Chapter 4, we present the `PResiduals` package. A publicly available dataset is used to illustrate its usage in model diagnostics, tests of conditional associations, and covariate-adjustment for Spearman's rank correlation.

Chapter 5 concludes the dissertation with a discussion of future research.

Chapter 2

Covariate-Adjusted Spearman's Rank Correlation with Probability-Scale Residuals

In this chapter, we propose new estimators for covariate-adjusted Spearman's rank correlation, both partial and conditional, using probability-scale residuals (PSRs). Our partial estimator for Spearman's correlation between X and Y adjusted for Z is the correlation of PSRs from models of X on Z and of Y on Z, which is analogous to the partial Pearson's correlation derived as the correlation of observed-minus-expected residuals. Our conditional estimator is the conditional correlation of PSRs. Our estimators are consistent for natural extensions of the population parameter of Spearman's rank correlation in the presence of covariates and are general for both continuous and discrete variables. We describe estimation and inference, and highlight the use of ordinal cumulative probability models, which allow our method to preserve the rank-based nature of Spearman's correlation. We conduct simulations to evaluate the performance of our estimators and compare them with other popular measures of association, demonstrating their robustness and efficiency. We illustrate our method in two application examples.

2.1 Introduction

It is often of interest to summarize the degree of association between two variables using a single number. To this end, associations are frequently described using correlation coefficients, which, well over a century after their introduction, remain popular in practice. Although correlation coefficients have limitations (e.g., an inability to accurately describe non-monotonic relationships), their continued popularity is due in part to their simplicity and interpretability. For well behaved continuous variables, a common choice is Pearson's correlation coefficient. When dealing with ordered categorical data, nonlinear relationships, skewed distributions, and extreme values, rank correlation coefficients such

as Spearman's rho or Kendall's tau are preferred.

In many applications, it is desirable to adjust the correlation coefficients for the influence of other variables. For example, when quantifying the association between educational attainment and wage, investigators may want to adjust for age, gender, job class, and health condition, because those variables are associated with both educational attainment and wage, and could be potential confounding factors. As a second example, researchers may be interested in quickly and robustly assessing all pairwise associations between responses in a large survey while controlling for the influence of demographic factors. In general, there are two approaches to adjusting the correlation for covariates. One is to obtain a partial correlation, i.e., removing the effect of covariates and then summarizing the relationship with a single number. The other is to obtain conditional correlations, i.e., assessing the correlation at specific levels of the covariates.

The partial correlation was originally developed for the Pearson's correlation with normal data (Fisher, 1924). The partial Pearson's correlation coefficient between X and Y controlling for Z, denoted as $\rho_{XY \cdot Z}$, is the correlation between the residuals resulting from linear regression models of X on Z and of Y on Z. When Z is a single variable, it can be calculated with the formula

$$\rho_{XY \cdot Z} = (\rho_{XY} - \rho_{XZ}\rho_{YZ}) / \sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}, \tag{2.1}$$

where ρ_{AB} represents the Pearson's correlation between A and B. Partial Spearman's and partial Kendall's correlations have also been proposed with the same formula: substituting ρ_{AB} with corresponding rank correlations (Kendall, 1942). However, they have limitations. The partial Kendall's correlation calculated with this formula can be far from 0 even under conditional independence, and therefore, is generally not useful (Korn, 1984). The partial Spearman's correlation using this formula is an ad hoc procedure, has little theoretical justification, and does not correspond with a sensible population parameter (Kendall, 1942;

Gripenberg, 1992).

Conditional rank correlations have been studied for continuous data with copulas. For continuous variables, Spearman's rank correlation and Kendall's rank correlation can be expressed as functions of copulas (Nelsen, 2006). Gijbels et al. (2011) proposed a kernel-based method to estimate the conditional copula and the associated conditional Spearman's and Kendall's rank correlations. However, their approach cannot be directly extended to discrete data because rank correlations between discrete variables cannot be easily described with copulas (see the discussions in Genest & Nešlehová (2007); Nešlehová (2007)).

In this paper, we propose new estimators for covariate-adjusted Spearman's rank correlations, both partial and conditional, using probability-scale residuals (PSRs) (Li & Shepherd, 2012). Our partial estimator is the correlation of PSRs from models of X on Z and of Y on Z, which is analogous to the partial Pearson's correlation derived as the correlation of observed-minus-expected residuals. Our conditional estimator is the conditional correlation of PSRs, which can be used to capture changes in Spearman's correlations between different values of covariates. Since PSRs are widely defined for any orderable variable (Shepherd et al., in press), our estimators are quite general. In the absence of covariates, our partial estimator reduces to the usual sample Spearman's rank correlation; and in the presence of covariates, it averages the conditional Spearman's rank correlation across different values of covariates.

The paper is organized as follows. In Section 2.2, we review PSRs and illustrate their connection with Spearman's rank correlation. We then describe new estimators for conditional and partial Spearman's rank correlation using PSRs. In Section 2.3, we discuss estimation and inference, highlighting the use of semiparametric ordinal models. In Section 2.4, we conduct simulations to evaluate the performance of our estimators and compare them with other popular measures of association, demonstrating their robustness and efficiency. In Section 2.5, we illustrate our approach in two application examples: one looking

at the association between education and wage after controlling for other potentially confounding variables, and a second application estimating all pairwise correlations between responses in a large survey after adjusting for relevant demographic and community-level factors. Section 2.6 discusses the methods and future research, and Section 2.7 contains some technical details and proofs.

2.2 Covariate-Adjusted Spearman's Rank Correlations

2.2.1 Spearman's Rank Correlation and PSRs

Fundamentally, Spearman's rank correlation is a scale-invariant concordance measure (Kruskal, 1958). Its population parameter, γ_{XY}, can be interpreted as the scaled difference between the probability of concordance and the probability of discordance among (X,Y) and (X_0,Y_0), where X_0 and Y_0 have the same marginal distributions with X and Y, respectively; but $X_0 \perp Y_0$, and $(X_0,Y_0) \perp (X,Y)$ (Kruskal, 1958). That is, $\gamma_{XY} = c(P_c - P_d)$, where $P_c = P[(X-X_0)(Y-Y_0) > 0]$, $P_d = P[(X-X_0)(Y-Y_0) < 0]$, and c is a scaling factor so that $-1 \leq \gamma_{XY} \leq 1$. The scaling factor c is constant and equal to 3 for continuous X and Y (Kruskal, 1958; Nelsen, 2006). For non-continuous X and/or Y, the scaling factor c is usually not a constant but a function of the marginal distributions of the non-continuous variables (Nešlehová, 2007). We now express γ_{XY} in terms of a new type of residual: the probability-scale residual (PSR).

For an orderable random variable X from distribution F, the PSR of an observed value x is defined as

$$r(x,F^*) = E[\text{sign}(x,X^*)] = P(X^* < x) - P(X^* > x) = F^*(x-) + F^*(x) - 1,$$

where F^* is an assumed or fitted distribution of X, X^* is a random variable with distribution F^*, and $\text{sign}(a,b)$ is -1, 0, and 1 for $a < b$, $a = b$, and $a > b$, respectively (Li & Shepherd, 2012; Shepherd et al., in press). We use $X_{res} = r(X,F^*)$ to denote the corresponding random

variable. With properly specified models, $F^* \to F$ as $n \to \infty$; therefore, $X_{res} = r(X, F^*) \to r(X, F)$, and the properties of $r(X, F)$ are hence applicable to $r(X, F^*)$ in its asymptote. As shown in Shepherd et al. (in press), $E[r(X, F)] = 0$ and $\text{var}[r(X, F)] = 1/3$ if X is continuous or $\text{var}[r(X, F)] = (1 - \sum f_x^3)/3$ if X is discrete, where $f_x = P(X = x)$.

Note that the difference between the probability of concordance and the probability of discordance can be written as

$$\begin{aligned}
P_c - P_d &= P[(X - X_0)(Y - Y_0) > 0] - P[(X - X_0)(Y - Y_0) < 0] \\
&= E\left[\text{sign}(X - X_0)\text{sign}(Y - Y_0)\right] \\
&= E_{(X,Y)}\left\{E\left[\text{sign}(X - X_0)\text{sign}(Y - Y_0)|(X, Y)\right]\right\} \\
&= E_{(X,Y)}\left\{E\left[\text{sign}(X - X_0)|X\right] E\left[\text{sign}(Y - Y_0)|Y\right]\right\} \\
&= E_{(X,Y)}\left\{\left[P(X_0 < X|X) - P(X_0 > X|X)\right]\left[P(Y_0 < Y|Y) - P(Y_0 > Y|Y)\right]\right\} \\
&= E\left\{\left[F(X-) + F(X) - 1\right]\left[G(Y-) + G(Y) - 1\right]\right\} \\
&= E[r(X, F)r(Y, G)] \\
&= \text{cov}[r(X, F), r(Y, G)],
\end{aligned}$$

with the last equality holding because $E[r(X, F)] = E[r(Y, G)] = 0$. Therefore, $\gamma_{XY} = c(P_c - P_d)$ is bounded between -1 and 1 if and only if $c = \{\text{var}[r(X, F)]\text{var}[r(Y, G)]\}^{-1/2}$. Then,

$$\gamma_{XY} = \text{corr}[r(X, F), r(Y, G)]. \tag{2.2}$$

This expression suggests that Spearman's rank correlation can be estimated with PSRs, i.e., $\hat{\gamma}_{XY} = \sum_{i=1}^{n}(x_{i,res} - \bar{x}_{res})(y_{i,res} - \bar{y}_{res}) / \sqrt{\sum_{i=1}^{n}(x_{i,res} - \bar{x}_{res})^2 \sum_{i=1}^{n}(y_{i,res} - \bar{y}_{res})^2}$, where $x_{i,res} = r(x_i, F^*)$, $y_{i,res} = r(y_i, G^*)$, $\bar{x}_{res} = \sum_{i=1}^{n} x_{i,res}/n$, and $\bar{y}_{res} = \sum_{i=1}^{n} y_{i,res}/n$. In the absence of covariates, F^* and G^* are often estimated with empirical distribution functions. In that case, $x_{i,res}$ and $y_{i,res}$ are linear functions of the ranks of x_i and y_i (Shepherd et al., in press); therefore, the sample Spearman's rank correlation estimated with PSRs is indeed

equal to the usual sample Spearman's rank correlation.

2.2.2 Conditional Spearman's Rank Correlation

In the presence of covariates Z, the fitted distributions F^* and G^* will typically be conditional on covariates, denoted as $F^*_{X|Z}$ and $G^*_{Y|Z}$, respectively. Then, the PSRs of X and Y for subject i are $r(x_i, F^*_{X|Z=z_i})$ and $r(y_i, G^*_{Y|Z=z_i})$, respectively. If both models of X on Z and Y on Z are properly specified, the fitted distributions $F^*_{X|Z} \to F_{X|Z}$ and $G^*_{Y|Z} \to G_{Y|Z}$. Then, $r(X, F^*_{X|Z}) \to r(X, F_{X|Z})$ and $r(Y, G^*_{Y|Z}) \to r(Y, G_{Y|Z})$. Similarly, we have $E[r(X, F_{X|Z})|Z] = 0$ and $\text{var}[r(X, F_{X|Z})|Z] = 1/3$ for continuous X or $\text{var}[r(X, F_{X|Z})|Z] = (1 - \sum_x f^3_{x|Z})/3$ for discrete X, where $f_{x|Z} = P(X = x|Z)$.

We can define the population version of the conditional Spearman's rank correlation as $\gamma_{XY|Z} = c_Z(P_{c|Z} - P_{d|Z})$ with $P_{c|Z} = P[(X - X_0)(Y - Y_0) > 0|Z]$ and $P_{d|Z} = P[(X - X_0)(Y - Y_0) < 0|Z]$, where $X_0|Z \sim F_{X|Z}$, $Y_0|Z \sim G_{Y|Z}$, $X_0 \perp Y_0|Z$, and $(X_0, Y_0) \perp (X, Y)$. As shown in the unconditional case, $P_{c|Z} - P_{d|Z} = \text{cov}[r(X, F_{X|Z}), r(Y, G_{Y|Z})|Z]$. With the scaling factor $c_Z = \left\{\text{var}[r(X, F_{X|Z})|Z]\text{var}[r(Y, G_{Y|Z})|Z]\right\}^{-1/2}$, $\gamma_{XY|Z}$ can be expressed as the conditional correlation of PSRs:

$$\gamma_{XY|Z} = \text{corr}[r(X, F_{X|Z}), r(Y, G_{Y|Z})|Z] \tag{2.3}$$

When both X and Y are continuous, the scaling factor c_Z is constant and equal to 3; this expression is mathematically equivalent to the conditional Spearman's rank correlation defined in Gijbels et al. (2011) with conditional copulas. However, since PSRs are well defined and easily calculated for a wide variety of outcomes and models, our expression of $\gamma_{XY|Z}$ with PSRs has the advantages that it is general for any orderable variables.

2.2.3 Partial Spearman's Rank Correlation

Mimicking the derivation of the partial Pearson's correlation as the correlation of observed-minus-expected residuals, we propose a new partial Spearman's rank correlation as the cor-

relation of PSRs from the distributions of X given Z and of Y given Z. That is, $\gamma_{XY\cdot Z} = \text{corr}[r(X, F_{X|Z}), r(Y, G_{Y|Z})]$. Unlike the traditional ad hoc partial Spearman's rank correlation calculated using (2.1), our partial estimator corresponds to a meaningful population parameter. Specifically, $\gamma_{XY\cdot Z}$ can be interpreted as the rescaled average of the conditional concordance measure, since $\gamma_{XY\cdot Z} = c^* E_Z(P_{c|Z} - P_{d|Z})$ with the scaling factor $c^* = \left\{\text{var}[r(X, F_{X|Z})]\text{var}[r(Y, G_{Y|Z})]\right\}^{-1/2}$ so that $-1 \leq \gamma_{XY\cdot Z} \leq 1$ (see Section 2.7.1).

Note that $\gamma_{XY\cdot Z}$ is a weighted average of $\gamma_{XY|Z}$ (see Section 2.7.1). For continuous X and Y, the weights are constant and equal to 1; therefore, $\gamma_{XY\cdot Z} = E(\gamma_{XY|Z})$. For non-continuous X or Y, the weight is a function of covariates Z. For example, if both X and Y are discrete variables, $\gamma_{XY\cdot Z}$ can be expressed as $E(w_Z \gamma_{XY|Z})$ with the weight $w_Z = \sqrt{(1 - \sum_x f_{x|Z}^3)(1 - \sum_y g_{y|Z}^3)} / \sqrt{[1 - \sum_x E_Z(f_{x|Z}^3)][1 - \sum_y E_Z(g_{y|Z}^3)]}$, where $f_{x|Z} = P(X = x|Z)$ and $g_{y|Z} = P(Y = y|Z)$. Since the denominator of w_z is fixed for all levels of Z, larger weights are assigned to the levels of Z with which the discrete variables are less likely to have ties (i.e., have more categories and/or are more evenly distributed).

2.2.4 Conditional Partial Spearman's Rank Correlation

When covariates are multidimensional, it may be useful to condition the partial Spearman's rank correlation on one or a subset of covariates. For example, suppose Z can be divided into two (potentially multidimensional) components, i.e., $Z = (Z_1, Z_2)$. To describe the rank correlation between X and Y for a specific level of Z_1 while adjusting for the other covariates, we can define a conditional partial Spearman's rank correlation as $\gamma_{XY\cdot Z|Z_1} = \text{corr}[r(X, F_{X|Z}), r(Y, G_{Y|Z})|Z_1]$. It can be shown that $\gamma_{XY\cdot Z|Z_1} = c^*_{Z_1} E_{Z_2|Z_1}(P_{c|Z} - P_{d|Z}) = E_{Z_2|Z_1}(w^*_{Z_1} \gamma_{XY|Z})$ (see derivation and the expressions of $c^*_{Z_1}$ and $w^*_{Z_1}$ in Section 2.7.2). For continuous X and Y, it can be shown that $c^*_{Z_1} = 3$ and $w^*_{Z_1} = 1$; therefore, $\gamma_{XY\cdot Z|Z_1} = E_{Z_2|Z_1}(P_{c|Z} - P_{d|Z}) = E_{Z_2|Z_1}(\gamma_{XY|Z})$ and $\gamma_{XY\cdot Z} = E_{Z_1}(\gamma_{XY\cdot Z|Z_1})$. However, for non-continuous X or Y, $w^*_{Z_1}$ and $c^*_{Z_1}$ are usually not constant and this relationship is generally not guaranteed.

2.3 Estimation and Inference

2.3.1 Calculation of PSRs

To obtain our estimators, we first need to fit models for X on Z and for Y on Z, and compute the two sets of PSRs. PSRs are functions of the fitted distribution, therefore, they are well defined as long as the fitted distributions are estimable. Since Spearman's rank correlation is a nonparametric statistic, it is natural to consider obtaining the PSRs using nonparametric models. For example, given a dataset (x_i, y_i, z_i) for $i = 1, 2, ..., n$, a kernel estimator for the conditional distribution could be $\hat{F}_{X|Z=z}(x) = \sum_{i=1}^{n} w_i(z,h) I(x_i \leq x)$, where the kernel weight $w_i(z,h)$ is given by $K[(z_i - z)/h] / \sum_{i=1}^{n} K[(z_i - z)/h]$ with kernel function $K(\cdot)$ and bandwidth h (Gijbels et al., 2011). Similarly, $\hat{F}_{X|Z=z}(x-) = \sum_{i=1}^{n} w_{ni}(z,h) I(x_i < x)$. Then, the PSR for observation x_i can be calculated as $x_{i,res} = \hat{F}_{X|Z=z_i}(x_i) + \hat{F}_{X|Z=z_i}(x_i-) - 1$. The PSR for y_i can be calculated similarly.

Although such an approach is feasible, there are challenges to incorporate nonparametric models when applied to real data. One challenge of kernel based approaches is that fitted models can be highly dependent on the selected bandwidth. Additional challenges arise with multidimensional covariates due to the curse of dimensionality (Hastie et al., 2009). Nonparametric models are also typically less efficient. But, on the other hand, although parametric models are easier to fit, more efficient, and more convenient for obtaining PSRs, they are less robust. Given the robustness of Spearman's rank correlation, it seems that assessing the correlation of PSRs derived from parametric models is contrary to the nature of Spearman's rank correlation.

To achieve a good compromise between robustness and efficiency, we consider semiparametric models that only use the order information of the outcomes, e.g., the semiparametric transformation model $X = T(\beta Z + \varepsilon)$, where $T(\cdot)$ is an unspecified monotonic increasing transformation and ε is a random error term with a specified parametric distribution F_e (Zeng & Lin, 2007). Since the transformation $T(\cdot)$ is left unspecified and only

needs to be monotonic, this model only depends on the order of X. Note, this model can be written in the form of the ordinal cumulative probability model: $g[F_{X|Z}(x)] = \alpha(x) - \beta Z$, where $F_{X|Z}$ is the conditional distribution of X on Z, the link function $g(\cdot) = F_e^{-1}(\cdot)$, and the intercept $\alpha(x) = T^{-1}(x)$. Based on this fact, Harrell (2015) proposed an estimating procedure which maximizes the multinomial likelihood of the corresponding ordinal cumulative probability model and implemented this procedure as the function `orm()` in the `rms` package (Harrell, 2016). Specifically, the orm procedure treats a continuous response variable as an ordered categorical variable (each unique value as one category) and fits it with ordinal cumulative probability models. Estimates from the orm procedure are very similar to the nonparametric maximum likelihood estimators (Zeng & Lin, 2007), whose asymptotic properties have been well studied with right censored data (Murphy et al., 1997; Zeng & Lin, 2006, 2007); but in practice the orm procedure is much easier to implement. Another advantage of using the orm procedure is its wide applicability: it can be fit to any orderable outcome. For example, when the outcome is ordered categorical, the orm procedure with the logit link function is the commonly used proportional odds model; when the outcome is binary, it is the commonly used logistic regression model.

2.3.2 Standard Error of Partial Estimators

After obtaining the PSRs from models of X on Z and of Y on Z, we can obtain our partial estimator simply as the correlation of PSRs. Here, we focus our discussion on the estimation of its standard error. Li & Shepherd (2010) described two approaches, a bootstrap method and a large sample approximation, for obtaining the distribution of the correlation of PSRs in the special case where X and Y are both ordered categorical variables. Similar approaches can be applied with more general X and Y.

The bootstrap method is easy to implement: sample with replacement from (X,Y,Z), re-fit models, compute PSRs, calculate their correlation, and repeat multiple times. The resulting bootstrap distribution can be used to construct confidence intervals and to test

the null hypothesis that $\gamma_{XY \cdot Z} = 0$. Note that when PSRs are obtained from nonparametric models such as kernel smoothers, if the bandwidth is not pre-specified but chosen based on the data, the bandwidth selection needs to be incorporated into each bootstrap replication to account for this extra variability.

When both sets of PSRs are obtained from parametric or semiparametric models that can be written in the form of estimating equations, it may be more computationally efficient to obtain standard error estimates based on large sample approximation using M-estimation techniques. Briefly, let $\Psi_x(\cdot)$ denote estimating equations for the model of X on Z with parameter θ_x, and $\Psi_y(\cdot)$ denote estimating equations for the model of Y on Z with parameter θ_y. $\Psi_x(\cdot)$ and $\Psi_y(\cdot)$ can be stacked together with the components necessary for computing the correlation of PSRs, resulting in the following estimating function:

$$\Psi(X_i, Y_i, Z_i; \theta) = \begin{cases} \Psi_x(X_i, Z_i; \theta_x) \\ \Psi_y(Y_i, Z_i; \theta_y) \\ Y_{i,res} - \theta_1 \\ X_{i,res} - \theta_2 \\ Y_{i,res} X_{i,res} - \theta_3 \\ Y_{i,res}^2 - \theta_4 \\ X_{i,res}^2 - \theta_5 \end{cases}$$

where $\theta = (\theta_x, \theta_y, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, with $\theta_1 = E(Y_{i,res})$, $\theta_2 = E(X_{i,res})$, $\theta_3 = E(Y_{i,res} X_{i,res})$, $\theta_4 = E(Y_{i,res}^2)$ and $\theta_5 = E(X_{i,res}^2)$, and $\sum_{i=1}^n \Psi(X_i, Y_i, Z_i; \hat{\theta}) = 0$. Under the usual conditions, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0, V(\theta)]$, where $V(\theta) = A(\theta)^{-1} B(\theta) [A(\theta)^{-1}]', A(\theta) = E[-\partial \Psi_i(\theta)/\partial \theta]$, and $B(\theta) = E[\Psi_i(\theta) \Psi_i(\theta)']$. Since $\hat{\gamma}_{XY \cdot Z} = (\hat{\theta}_3 - \hat{\theta}_1 \hat{\theta}_2) / \sqrt{(\hat{\theta}_4 - \hat{\theta}_1^2)(\hat{\theta}_5 - \hat{\theta}_2^2)}$, the delta-method can be employed to obtain the large sample distribution of $\hat{\gamma}_{XY \cdot Z}$. In practice, estimating the large sample distribution of the Fisher's transformation of $\hat{\gamma}_{XY \cdot Z}$, i.e., $\log[(1 +$

$\hat{\gamma}_{XY\cdot Z})/(1-\hat{\gamma}_{XY\cdot Z})]/2$, typically results in more rapid convergence to normality, and is therefore preferable for constructing confidence intervals and test statistics based on large sample approximations. In addition, some parameters may have known values and can be removed from the estimating procedure. For example, $\theta_1 = \theta_2 = 0$ if F^* and G^* are properly specified, and $\theta_4 = \theta_5 = 1/3$ if X and Y are continuous and F^* and G^* are properly specified. In our experience, however, estimation of these parameters even when the truth is known does not have much impact on resulting confidence intervals and test statistics, so we generally include them in the estimating equations.

2.3.3 Conditional Estimators

To obtain our conditional estimator, we need to estimate the correlation of PSRs conditional on the value of Z. If Z is a categorical variable with sufficient numbers in each category, it is natural to compute the correlation of PSRs within each level of Z. For continuous Z, smoothing is needed and can be achieved nonparametrically or parametrically.

For example, we can estimate the conditional correlation of PSRs nonparametrically with a kernel smoothing approach, such as,

$$\hat{\gamma}_{XY|Z}(z) = \frac{\sum x_{i,res}y_{i,res}w_i(z,h) - \sum x_{i,res}w_i(z,h)\sum y_{i,res}w_i(z,h)}{\sqrt{\sum x_{i,res}^2 w_i(z,h) - [\sum x_{i,res}w_i(z,h)]^2}\sqrt{\sum y_{i,res}^2 w_i(z,h) - [\sum y_{i,res}w_i(z,h)]^2}},$$

where $w_i(z,h)$ are weights which sum to 1, e.g., $w_i(z,h) = K[(z_i - z)/h]/\sum_{i=1}^{n} K[(z_i - z)/h]$ with kernel function $K(\cdot)$ and bandwidth h.

Alternatively, we can estimate $\gamma_{XY|Z}$ using parametric models. For example, since under properly specified models, both $E(X_{res}|Z)$ and $E(Y_{res}|Z)$ converge to 0, $\hat{\gamma}_{XY|Z}$ can be approximated with $\hat{E}(X_{res}Y_{res}|Z)/\sqrt{\hat{E}(X_{res}^2|Z)\hat{E}(Y_{res}^2|Z)}$. To obtain $\hat{E}(X_{res}Y_{res}|Z)$, $\hat{E}(X_{res}^2|Z)$, and $\hat{E}(Y_{res}^2|Z)$, one might fit linear regression models of $X_{res}Y_{res}$ on Z, X_{res}^2 on Z, and Y_{res}^2 on Z and transform Z with spline functions to allow flexible modeling. Specifically, when X and/or Y are continuous variables, $E(X_{res}^2|Z)$ and/or $E(Y_{res}^2|Z)$ converge to a constant equal

to 1/3 under properly specified models; plugging in 1/3 or empirical estimates ($\sum X_{res}^2/n$ and/or $\sum Y_{res}^2/n$) could further simplify the estimation procedure.

Depending on the parametric or nonparametric methods used, standard errors and confidence intervals can be estimated using the bootstrap or large sample approximation techniques similar to those described in Section 2.3.2.

2.4 Numerical Illustrations and Simulations

We conducted simulations to evaluate the finite sample performance of our partial estimator. Let $Z \sim N(0,1)$ and $\binom{X_1}{Y_1}|Z \sim N\left[\binom{\alpha_0+\alpha_1 Z}{\beta_0+\beta_1 Z}, \begin{pmatrix} 1 & \rho_{XY|Z} \\ \rho_{XY|Z} & 1 \end{pmatrix}\right]$ with $\alpha_0 = \beta_0 = 0$, $\alpha_1 = 1$, and $\beta_1 = -1$. We consider four scenarios: (I) $X = X_1$ and $Y = Y_1$; (II) $X = X_1$ and $Y = \exp(Y_1)$; (III) $Y = Y_1$ and X is generated by discretizing X_1 with cut-off values as the 20^{th}, 40^{th}, 60^{th}, 80^{th} percentiles of the standard normal distribution; (IV) $Y = \exp(Y_1)$ and X is the discretized version of X_1 described in (III).

We first discuss the population values of our estimators in these four scenarios. In Scenario I, since (X,Y) conditional on Z is normally distributed, $\gamma_{XY|Z}$ has a closed-form relationship with $\rho_{XY|Z}$, i.e., $\gamma_{XY|Z} = 6\arcsin(\rho_{XY|Z}/2)/\pi$ (Fisher, 1924). Because of this relationship, $\gamma_{XY|Z}$ is constant if we set $\rho_{XY|Z}$ constant for different values of Z, as done throughout this section. Therefore, our partial estimator $\gamma_{XY\cdot Z} = E(\gamma_{XY|Z}) = \gamma_{XY|Z} = 6\arcsin(\rho_{XY|Z}/2)/\pi$. Figure 2.1 (left panel) plots $\gamma_{XY\cdot Z}$ as a function of $\rho_{XY|Z}$ ranging from -1 and 1. For the purpose of comparison, the traditional partial Pearson's, Spearman's, and Kendall's correlations obtained by plugging corresponding parameters into (2.1) are also plotted and denoted as $\rho^*_{XY\cdot Z}$, $\gamma^*_{XY\cdot Z}$, and $\tau^*_{XY\cdot Z}$, respectively. As shown in Figure 2.1, $\rho^*_{XY\cdot Z} = \rho_{XY|Z}$; $\tau^*_{XY\cdot Z}$ has poor performance since its value departs from 0 even under conditional independence ($\rho_{XY|Z} = 0$); and our partial estimator $\gamma_{XY\cdot Z}$ is a better approximation to $\rho_{XY|Z}$ than the traditional partial Spearman's correlation $\gamma^*_{XY\cdot Z}$, especially in the upper tail. Scenario II is similar to Scenario I except that the exponential transformation of Y generates extreme values and nonlinearity. All rank based correlations, including our

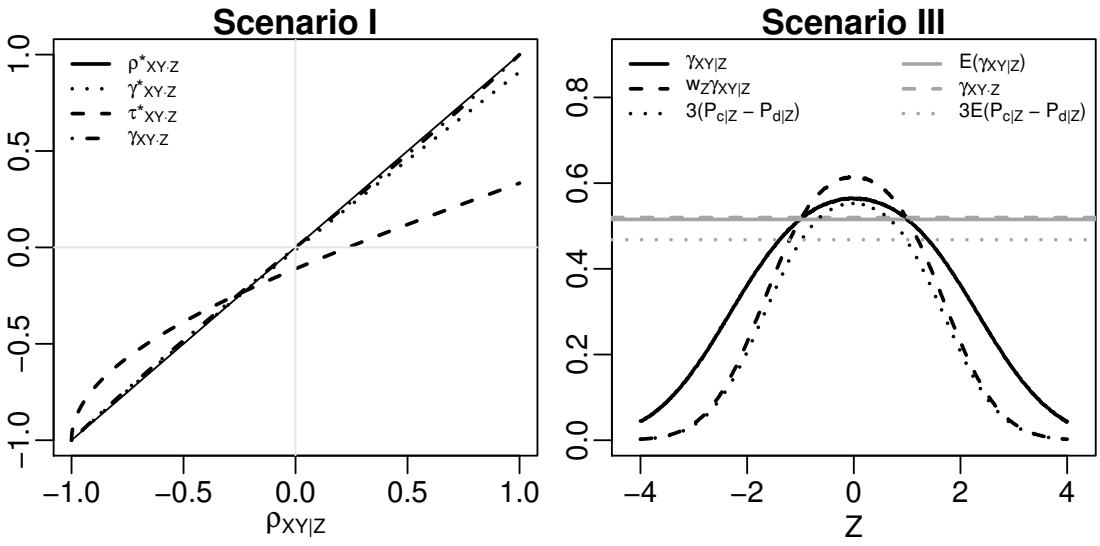

Figure 2.1: Parameters in Scenarios I and III. $\gamma_{XY|Z}$ and $\gamma_{XY\cdot Z}$ are the population parameters of our conditional and partial estimators, respectively; $\rho^*_{XY\cdot Z}$, $\gamma^*_{XY\cdot Z}$, and $\tau^*_{XY\cdot Z}$ are traditional partial Pearson's, Spearman's, and Kendall's correlations based on (1), respectively; $(P_{c|Z} - P_{d|Z})$ is the difference between the probability of concordance and the probability of discordance conditional on Z.

partial estimator, are unchanged, whereas partial Pearson's correlation cannot accurately summarize the covariate-adjusted association.

In Scenario III, X is discrete and has, when $Z = 0$, five evenly distributed categories. As Z departs from 0, the conditional distribution of X becomes skewed, therefore, it is more likely to have ties. We would expect the conditional correlation between X and Y to become weaker as Z departs from 0. Figure 2.1 (right panel) plots $\gamma_{XY|Z}$ as a function of Z, showing this trend. Our partial estimator $\gamma_{XY\cdot Z}$ differs, although only slightly in this example, from $E(\gamma_{XY|Z})$. Since $\gamma_{XY\cdot Z} = E(w_Z\gamma_{XY|Z}) = c^*E(P_{c|Z} - P_{d|Z})$, we also plot $w_Z\gamma_{XY|Z}$ and $3(P_{c|Z} - P_{d|Z})$ as references. As shown in Figure 2.1, the weight, w_Z, is bigger when there are less likely to be ties in X (i.e., Z is closer to 0); and the scaling factor c^* is larger than 3. Again, because it is a rank based correlation, our partial Spearman's correlation is identical in Scenarios III and IV.

We next evaluate the performance of estimating our partial estimators with PSRs derived from linear regression models (LM), kernel smoothers (kernel), and orm procedures

in finite samples ($n = 200$) in these four scenarios. In this set of simulations, we set $\rho_{XY|Z} = 0.6$ so that $\gamma_{XY \cdot Z} \approx 0.582$ in Scenarios I and II, and $\gamma_{XY \cdot Z} \approx 0.520$ in Scenarios III and IV, respectively. With linear regression models, we computed two sets of PSRs: 1) assuming normality of the error distribution, and 2) empirically, assuming a constant variance of the error distribution (Shepherd et al., in press). With kernel smoothers, we used a Gaussian kernel and chose the bandwidth based on Silverman's rule of thumb (Wand & Jones, 1995). When applying orm procedures, we used both the properly specified link function (probit) and misspecified link functions (logit, loglog, and cloglog). All simulations used Fisher's transformations and large sample approximations to compute variance and confidence intervals for our partial estimator. The results based on 10,000 simulation replications are shown in Table 2.1.

In summary, our partial estimators using PSRs from the orm procedure had minimal bias, good coverage, and low mean squared error (MSE) across all four simulation scenarios. In Scenario I, estimators using the orm procedure properly specified with the probit link performed similarly to fully parametric estimators correctly assuming normality. However, because of their invariance to the exponential transformation of Y, in Scenarios II and IV, estimators using the orm procedure easily out-performed those using PSRs from linear models. Surprisingly, our estimators using PSRs from the orm procedure were very robust to link function misspecification, with only slight increases in bias. The bias and MSE of our estimators using the orm procedures were also generally smaller than those using kernel smoothers. Other bandwidth selection algorithms may improve the performance of kernel smoothers, but may also increase computational complexity.

In the second set of simulations, we investigated the performance of our partial estimators using PSRs from the orm procedure for testing covariate-adjusted association, and compared them with tests based on three traditional partial correlation coefficients. We set $\rho_{XY|Z} = 0$ under the null hypothesis (H_0) and $\rho_{XY|Z} = 0.2$ under the alternative hypothesis (H_A). For tests based on the traditional partial correlation coefficients, p-values were

Table 2.1: Simulation results for evaluating our partial estimator with PSRs derived from linear models (LM), kernel estimation (kernel), and the orm procedures with n=200 and 10,000 simulation replicates

Scenarios			truth	est	% bias	est.se	emp.se	MSE	CP
I									
	LM								
		normality	0.582	0.582	0.01	0.048	0.049	0.0024	0.946
		empirically	0.582	0.580	-0.35	0.048	0.049	0.0024	0.945
	kernel								
		Silverman	0.582	0.552	-5.10	—	0.050	0.0034	—
	orm								
		probit	0.582	0.577	-0.92	0.050	0.049	0.0025	0.950
		logit	0.582	0.573	-1.59	0.050	0.050	0.0026	0.948
		loglog	0.582	0.565	-2.85	0.050	0.049	0.0027	0.942
		cloglog	0.582	0.565	-2.88	0.050	0.049	0.0027	0.941
II									
	LM								
		normality	0.582	0.394	-32.33	0.058	0.063	0.0393	0.057
		empirically	0.582	0.386	-33.68	0.060	0.063	0.0424	0.051
	kernel								
		Silverman	0.582	0.552	-5.10	—	0.050	0.0034	—
	orm								
		probit	0.582	0.577	-0.92	0.050	0.049	0.0025	0.950
		logit	0.582	0.573	-1.59	0.050	0.050	0.0026	0.948
		loglog	0.582	0.565	-2.85	0.050	0.049	0.0027	0.942
		cloglog	0.582	0.565	-2.88	0.050	0.049	0.0027	0.941
III									
	LM								
		normality	0.520	0.499	-3.86	0.054	0.055	0.0034	0.931
		empirically	0.520	0.500	-3.72	0.054	0.055	0.0034	0.930
	kernel								
		Silverman	0.520	0.499	-3.92	—	0.053	0.0032	—
	orm								
		probit	0.520	0.517	-0.52	0.053	0.053	0.0028	0.945
		logit	0.520	0.514	-1.02	0.053	0.053	0.0029	0.945
		loglog	0.520	0.505	-2.84	0.053	0.053	0.0030	0.943
		cloglog	0.520	0.504	-2.90	0.053	0.053	0.0030	0.939
IV									
	LM								
		normality	0.520	0.361	-30.52	0.054	0.056	0.0283	0.144
		empirically	0.520	0.382	-26.50	0.053	0.055	0.0219	0.226
	kernel								
		Silverman	0.520	0.499	-3.92	—	0.053	0.0032	—
	orm								
		probit	0.520	0.517	-0.52	0.053	0.053	0.0028	0.945
		logit	0.520	0.514	-1.02	0.053	0.053	0.0029	0.945
		loglog	0.520	0.505	-2.84	0.053	0.053	0.0030	0.943
		cloglog	0.520	0.504	-2.90	0.053	0.053	0.0030	0.939

est is the mean of the point estimates.

est.se is the mean of the standard error estimates.

emp.se is the standard deviation of the point estimates

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

Table 2.2: Type I error rate and power (%) for testing covariate-adjusted association using our partial estimator and the traditional partial correlation coefficients with n=200 and 10,000 simulation replicates

Scenarios		**our partial estimator**				**traditional partial coefficients**		
		probit	logit	loglog	cloglog	Pearson	Spearman	Kendall
I								
	H_0	4.96	4.97	5.11	5.04	5.01	6.36	67.42
	H_1	77.47	76.62	76.44	76.86	81.89	69.17	1.77
II								
	H_0	4.96	4.97	5.11	5.04	5.20	6.36	67.33
	H_1	77.47	76.62	76.44	76.86	33.79	69.16	1.77
III								
	H_0	5.12	5.12	5.29	5.29	5.16	6.64	72.04
	H_1	67.19	66.81	64.79	65.02	68.42	65.56	4.49
IV								
	H_0	5.12	5.12	5.29	5.29	5.16	6.64	72.04
	H_1	67.19	66.81	64.79	65.02	26.85	65.56	4.49

obtained based on large sample approximations using the R package `ppcor` (Kim, 2015). Type I error rate and power are reported in Table 2.2. Consistent with the observation in Figure 2.1, tests based on partial Kendall's correlations had poor performance: high type I error rate and almost no power. Compared with the partial Pearson's correlation, our estimators were slightly less efficient when the relationships were linear or approximately linear (Scenarios I and III), but much more robust in the presence of nonlinearity and extreme values (Scenarios II and IV). Also, our estimators had better performance than the traditional ad hoc partial Spearman's correlation: type I error rate closer to 5% and generally higher power.

2.5 Application Examples

2.5.1 Wage and Education

We first illustrate our estimators by assessing the association between wage and education using a dataset of 3,000 male workers in the mid-Atlantic region of the United States from 2003 to 2009. This dataset is available in the R package `ISLR` (James et al., 2013). Since education was collected as an ordered categorical variable with 5 levels and the distribution of wage is right-skewed, Spearman's rank correlation is preferred for its robustness.

Also, since wage and education are associated with other factors, e.g., job class and age, it is desirable to adjust Spearman's rank correlation for these factors.

Using our approach, we computed PSRs for education (X_{res}) and for wage (Y_{res}) using the orm procedure with the logit link function and including job class, age (transformed with restricted cubic splines using 5 knots), race, health condition, marital status, and calender year as covariates. Our partial estimator was 0.44 with 95% confidence interval (CI) (0.41, 0.47), which was lower than the unadjusted Spearman's rank correlation 0.50 (95% CI: 0.47, 0.53). These confidence intervals and all others in this section, were based on large sample approximations with Fisher's transformation; bootstrap confidence intervals were very similar and are not shown. To explore whether the association between wage and education varied for different job classes and age groups, we estimated the correlation of PSRs conditional on job class and conditional on age using the approaches described in Section 3.3. Specifically, we fitted linear regression models of $X_{res}Y_{res}$, X_{res}^2, and Y_{res}^2 including job class or age (transformed with restricted cubic splines using 5 knots) as covariates, estimated the conditional expectations of $X_{res}Y_{res}$, X_{res}^2, and Y_{res}^2, and then plugged in these estimates to obtain $\hat{\gamma}_{XY|Z} = \hat{E}(X_{res}Y_{res}|Z)/\sqrt{\hat{E}(X_{res}^2|Z)\hat{E}(Y_{res}^2|Z)}$. We found that after adjusting for other covariates, the Spearman's rank correlation between age and education was significantly higher in the information job class than in the industrial class: 0.48 (95% CI: 0.44, 0.52) vs. 0.41 (95% CI: 0.36, 0.45); the p-value for the difference was 0.02. The age-specific conditional estimates and 95% pointwise confidence intervals, plotted in Figure 2.2, suggested that after adjusting for other factors, the Spearman's rank correlation between education and wage was weaker among those who were younger (< 30 years). We repeated the analysis with the nonparametric approaches, i.e., obtaining the kernel weighted conditional correlation of PSRs for age and calculating the correlation of PSRs stratified by job class, and found very similar results (see Figure 2.2).

Additional analyses were performed including implementing the orm procedure with other link functions, and results were very similar to those reported here. For the purpose

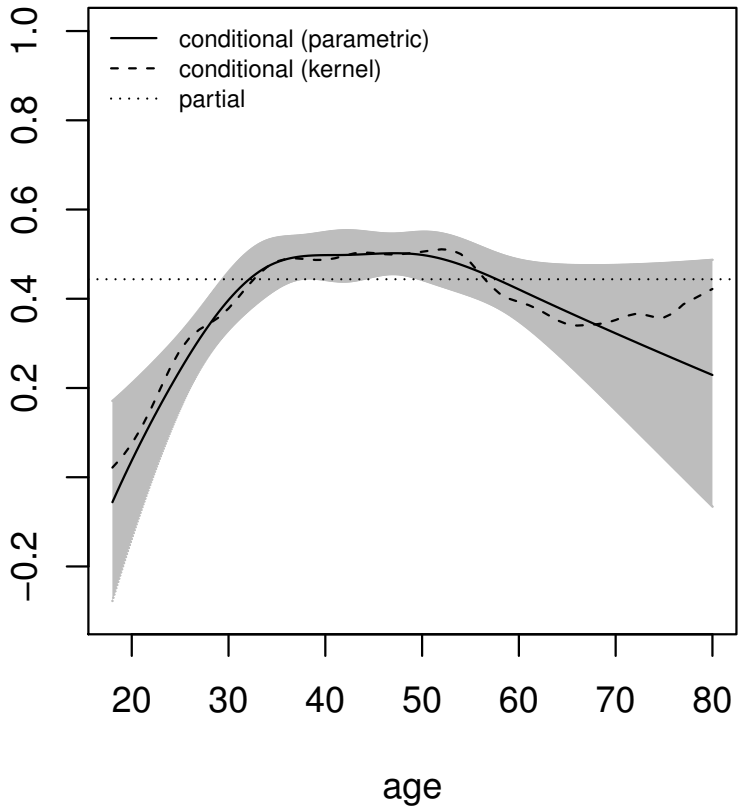

Figure 2.2: The age-specific conditional Spearman's rank correlation between wage and education modeled with parametric and nonparametric (kernel-based) approaches. The shaded region is the pointwise 95% confidence intervals from the parametric approach. For the nonparametric approach, we used a Gaussian kernel and selected the bandwidth based on Silverman's rule of thumb (Wand & Jones, 1995). Note, these estimates are also adjusted for job class, race, health condition, marital status, and calender year, since we also included them in the models as covariates when computing PSRs.

of comparison, we also fitted a linear regression model on log wage, including education as a categorical variable, all other covariates, and also interaction terms between education and job class, and between education and age. This linear regression model suggested both interaction terms were significant, which was consistent with the results using our approach.

2.5.2 SCIP Survey Data

In a second example, we use our partial estimator as a quick and robust tool to summarize a large number of covariate-adjusted associations. We illustrate our method with the endline survey data of the Strengthening Communities through Integrated Programming (SCIP) project. This World Vision project was funded by the U.S. Government with the aim to improve the health and livelihood of children, women, and families in Mozambique. In this survey, 3,892 female heads of household were asked to give opinions on their overall quality of life, health care, nutrition, education, and other aspects of livelihood. Many of the survey questions asked the participants to rank their opinions on an ordinal scale. The investigators were interested in the correlations among the participants' responses to different questions while adjusting for relevant demographic factors, such as the age, primary language, marital status, religion, urban/rural region, and district. The purpose of such an analysis is largely exploratory – to provide a quick view of the pairwise correlation between the responses to all survey questions. This information can then be used to focus on particular sets of questions for further study.

We included all 171 questions with orderable responses from 13 modules of the questionnaire, including 54 questions with binary responses (e.g, Yes or No), 106 questions with ordinal responses (e.g., strongly disagree, disagree, neutral, agree, strongly agree), and 11 questions with continuous responses. Missing values were common in this dataset. We performed simple imputation with the median or mode for the relatively few missing values of demographic factors ($\leq 5\%$). We found that a lot of missing values were not due

to nonresponse from participants, but rather due to the design of the questionnaire. For example, some questions were designed to follow up an earlier question, and they were only asked if the participants gave certain answers to the earlier questions. Because of this design, and also because the investigators were interested in a quick initial pass, we decided to focus our analysis on participants who actually responded to the survey questions, i.e., we did not impute any missing responses.

We fit a total of 171 models using the orm procedure with the logit link function for the survey questions with the demographic factors listed above included as covariates, and then obtained the PSRs. The pairwise correlation of PSRs were computed among participants who responded to both questions. The heatmap of our covariate-adjusted partial Spearman's correlation matrix is plotted in Figure 2.3. As we expected, responses to questions within the same module tended to be correlated since those questions were usually related to the same topics. But the strength of the correlation could be quite different even within the same module. To facilitate the visualization of the results, we developed a web application (https://scip.shinyapps.io/scip_app) to allow investigators to zoom into any specific area in the heatmap and check the detailed information about the questionnaire and responses. The web application includes 95% confidence intervals based on large sample approximations and Fisher's transformation, and also compares results with the unadjusted Spearman's rank correlation. Although their patterns were very similar, we found some interesting changes after adjusting for demographic factors. For example, the unadjusted Spearman's rank correlation among the 3,782 pairs of responses to questions "Do you have a separate room which is used as a kitchen? (1) No (2) Yes" and "How many years of education have you completed?" was 0.2 (95% CI: 0.17, 0.23), whereas the corresponding correlation of PSRs was 0.08 (95% CI: 0.05, 0.12), suggesting that the strength of association between the responses to these questions might be partly due to demographic factors. Similar results were obtained when fitting the orm models with other link functions, again demonstrating the robustness of our approach.

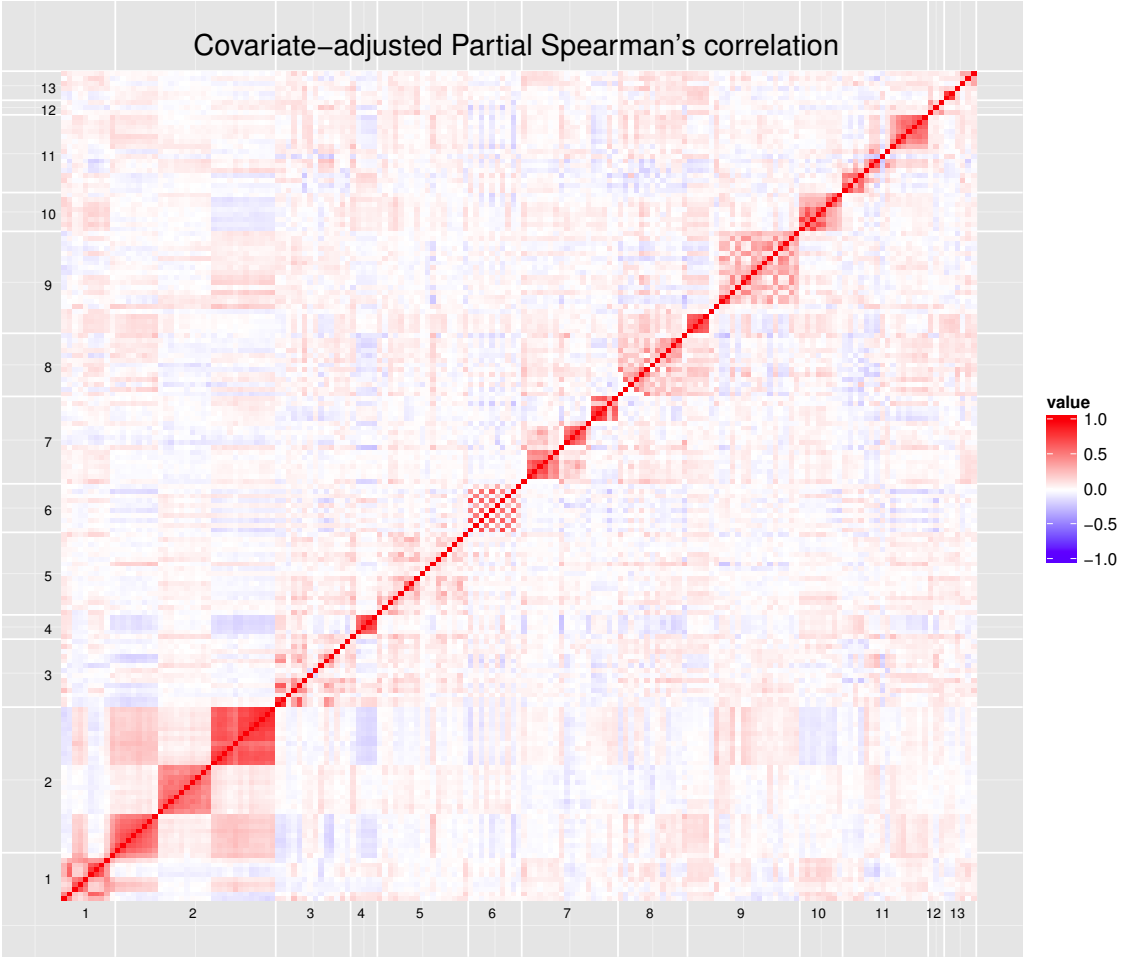

Figure 2.3: The heatmap of our partial estimators for pairwise Spearman's rank correlation adjusting for demographic factors for responses to 171 questions from 13 modules of the survey labeled as 1: overall quality of life, 2: mental health, 3: income, 4: food and nutrition, 5: material goods, 6: transportation, 7: health care, 8: voluntary counseling and testing (VCT) services, 9: HIV prevention, 10: social support, 11: community service, 12: education test result, and 13: perception of education. An interactive figure of results is at https://scip.shinyapps.io/scip_app.

2.6 Discussion

In this work, we express the population parameter of Spearman's rank correlation in terms of residuals, which connects this commonly used nonparametric statistic to a variety of regression models. Our methods therefore permit the adjustment of Spearman's rank correlation for multidimensional covariates. Our framework is very general, applicable to any orderable variables modeled with estimable fitted distributions. As a specific application with the orm procedure, our proposed covariate-adjusted Spearman's rank correlation preserves the rank-based nature of Spearman's correlation while allowing flexible modeling of covariates. The wide applicability, robustness, and computational simplicity of our estimators make them very useful, particularly when dealing with big data, as hinted at in our survey example.

For implementation of our covariate-adjusted Spearman's correlation, we favor using PSRs from the orm procedure, as they appear to be quite robust and efficient. It should be noted that the validity of the large sample distribution of our estimators when using the orm procedure relies on the asymptotic normality of the orm estimators themselves, which has never formally been proved. More generally, the asymptotic properties of the nonparametric maximum likelihood estimator of semiparametric transformation models with uncensored data have not been fully developed and are quite challenging to derive (Zeng & Lin, 2007; Zeng, Kosorok, & Lin, personal communication). From fairly extensive simulations (not shown), the orm estimators appear to be consistent and asymptotically normal, and all of our simulation results of the covariate-adjusted Spearman's correlation using the orm procedure have been well behaved (e.g., confidence intervals based on large sample distributions covering at the nominal level) and quite robust to misspecification of the link function. Further study of the orm procedure is certainly warranted.

Since the PSRs are widely defined, our framework has the potential to be extended to more complicated settings, such as longitudinal data in which the observations are not independent and censored outcomes in which fitted distributions are not completely determined.

We are studying extensions in these settings.

2.7 Supplemental Materials

2.7.1 Derivation of Population Parameter of Our Partial Estimator

Here we derive the population parameter of our partial estimator.

Since $E[r(X,F_{X|Z})|Z] = E[r(Y,G_{Y|Z})|Z] = 0$, $E[r(X,F_{X|Z})] = E_Z\{E[r(X,F_{X|Z})|Z]\} = 0$ and $E[r(Y,G_{Y|Z})] = E_Z\{E[r(X,G_{Y|Z})|Z]\} = 0$. Then, we have

$$
\begin{aligned}
\gamma_{XY\cdot Z} =& \text{corr}[r(X,F_{X|Z}), r(Y,G_{Y|Z})] \\
&= E[r(X,F_{X|Z})r(Y,G_{Y|Z})]/\sqrt{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]} \\
&= E_Z\{E[r(X,F_{X|Z})r(Y,G_{Y|Z})|Z]\}/\sqrt{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]} \\
&= E_Z\{\text{cov}[r(X,F_{X|Z}), r(Y,G_{Y|Z})|Z]\}/\sqrt{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]} \\
&= c^* E_Z(P_{c|Z} - P_{d|Z}) \\
&= \frac{E_Z\left\{\sqrt{\text{var}[r(X,F_{X|Z})|Z]\text{var}[r(Y,G_{Y|Z})|Z]}\,\text{corr}[r(X,F_{X|Z}), r(Y,G_{Y|Z})|Z]\right\}}{\sqrt{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]}} \\
&= E_Z\left\{\sqrt{\frac{\text{var}[r(X,F_{X|Z})|Z]\text{var}[r(Y,G_{Y|Z})|Z]}{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]}}\gamma_{XY|Z}\right\} \\
&= E_Z(w_Z\gamma_{XY|Z}),
\end{aligned}
$$

where $c^* = \left\{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]\right\}^{-1/2}$ and $w_Z = \sqrt{\frac{\text{var}[r(X,F_{X|Z})|Z]\text{var}[r(Y,G_{Y|Z})|Z]}{\text{var}[r(X,F_{X|Z})]\text{var}[r(Y,G_{Y|Z})]}}$.

Note, the weight w_Z can be written as $w_Z = \sqrt{a_{X|Z}a_{Y|Z}}$, where $a_{X|Z} = \frac{\text{var}[r(X,F_{X|Z})|Z]}{\text{var}[r(X,F_{X|Z})]}$ and $a_{Y|Z} = \frac{\text{var}[r(Y,G_{Y|Z})|Z]}{\text{var}[r(Y,G_{Y|Z})]}$. Since $E[r(X,F_{X|Z})|Z] = 0$, we have

$$\text{var}[r(X,F_{X|Z})] = E[\text{var}[r(X,F_{X|Z})]|Z].$$

The denominator of $a_{X|Z}$ is the expectation of its numerator, and therefore $a_{X|Z}$ is a nor-

malized function with $E(a_{X|Z}) = 1$. Similar is $a_{Y|Z}$. This normalization effectively gives a high weight to Z with a relatively high $\text{var}[r(X, F_{X|Z})|Z]$ or $\text{var}[r(Y, G_{Y|Z})|Z]$.

Specifically, for continuous X and Y, $\text{var}[r(X, F_{X|Z})] = \text{var}[r(Y, G_{Y|Z})] = 1/3$. Therefore, th scaling factor $c^* = 3$ and the weight $w_Z = 1$. Then,

$$\gamma_{XY \cdot Z} = 3E_Z(P_{c|Z} - P_{d|Z}) = E_Z(\gamma_{XY|Z})$$

2.7.2 Derivation of Population Parameter of Our Conditional Partial Estimator

Here we derive the population parameter of our conditional partial Spearman's rank correlation $\gamma_{XY \cdot Z|Z_1} = \text{corr}[r(X, F_{X|Z}), r(Y, G_{Y|Z})|Z_1]$.

Note,

$$\begin{aligned}
E[r(X, F_{X|Z})|Z_1] &= \int_{(X,Z_2)} r(X, F_{X|Z}) f(X, Z|Z_1) \mathrm{d}(X, Z_2) \\
&= \int_{(X,Z_2)} r(X, F_{X|Z}) f(X|Z) f(Z) / f(Z_1) \mathrm{d}(X, Z_2) \\
&= \int_{(X,Z_2)} r(X, F_{X|Z}) f(X|Z) f(Z_2|Z_1) \mathrm{d}(X, Z_2) \\
&= \int_{Z_2} \left[\int_X r(X, F_{X|Z}) f(X|Z) \mathrm{d}X \right] f(Z_2|Z_1) \mathrm{d}Z_2 \\
&= \int_{Z_2} E[r(X, F_{X|Z})|Z] f(Z_2|Z_1) \mathrm{d}Z_2 \\
&= 0.
\end{aligned}$$

This holds because $E[r(X, F_{X|Z})|Z]$ =0 for properly specified models. Similarly, we also

have $E[r(Y,G_{Y|Z})|Z_1]=0$. Therefore,

$$
\begin{aligned}
&\operatorname{cov}[r(X,F_{X|Z}),r(Y,G_{Y|Z})|Z_1]\\
=&E[r(X,F_{X|Z})r(Y,G_{Y|Z})|Z_1]\\
=&\int_{(X,Y,Z_2)} r(X,F_{X|Z})r(Y,G_{Y|Z})f(X,Y,Z_2|Z_1)\mathrm{d}(X,Y,Z_2)\\
=&\int_{(X,Y,Z_2)} r(X,F_{X|Z})r(Y,G_{Y|Z})f(X,Y|Z)f(Z)/f(Z_1)\mathrm{d}(X,Y,Z_2)\\
=&\int_{Z_2}\left[\int_{(X,Y)} r(X,F_{X|Z})r(Y,G_{Y|Z})f(X,Y|Z)\mathrm{d}(X,Y)\right]f(Z_2|Z_1)\mathrm{d}Z_2\\
=&\int_{Z_2} E\left[r(X,F_{X|Z})r(Y,G_{Y|Z})|Z\right]f(Z_2|Z_1)\mathrm{d}Z_2\\
=&\int_{Z_2} \operatorname{cov}[r(X,F_{X|Z}),r(Y,G_{Y|Z})|Z]f(Z_2|Z_1)\mathrm{d}Z_2\\
=&\int_{Z_2} (P_{c|Z}-P_{d|Z})f(Z_2|Z_1)\mathrm{d}Z_2\\
=&E_{Z_2|Z_1}(P_{c|Z}-P_{d|Z})\\
=&E_{Z_2|Z_1}(c_Z^{-1}\gamma_{XY|Z})\text{ , where } c_Z=\left\{\operatorname{var}[r(X,F_{X|Z})|Z]\operatorname{var}[r(Y,G_{Y|Z})|Z]\right\}^{-1/2}.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\gamma_{XY\cdot Z|Z_1}=&\frac{\operatorname{cov}[r(X,F_{X|Z}),r(Y,G_{Y|Z})|Z_1]}{\sqrt{\operatorname{var}[r(X,F_{X|Z})|Z_1]\operatorname{var}[r(Y,G_{Y|Z})|Z_1]}}\\
=&c^*_{Z_1}E_{Z_2|Z_1}(P_{c|Z}-P_{d|Z})\\
=&E_{Z_2|Z_1}(w^*_{Z_1}\gamma_{XY|Z}),
\end{aligned}
$$

where $c^*_{Z_1}=\left\{\operatorname{var}[r(X,F_{X|Z})|Z_1]\operatorname{var}[r(Y,G_{Y|Z})|Z_1]\right\}^{-1/2}$ and

$$
w^*_{Z_1}=\sqrt{\frac{\operatorname{var}[r(X,F_{X|Z})|Z]\operatorname{var}[r(Y,G_{Y|Z})|Z]}{\operatorname{var}[r(X,F_{X|Z})|Z_1]\operatorname{var}[r(Y,G_{Y|Z})|Z_1]}}.
$$

When both X and Y are continuous, since

$$\begin{aligned}
\mathrm{var}[r(X,F_{X|Z})|Z_1] =& E[r(X,F_{X|Z})^2|Z_1] \\
=& \int_{(X,Z_2)} r^2(X,F_{X|Z}) f(X,Z_2|Z_1)\mathrm{d}(X,Z_2) \\
=& \int_{(X,Z_2)} r^2(X,F_{X|Z}) f(X|Z) f(Z)/f(Z_1)\mathrm{d}(X,Z_2) \\
=& \int_{(X,Z_2)} r^2(X,F_{X|Z}) f(X|Z) f(Z_2|Z_1)\mathrm{d}(X,Z_2) \\
=& \int_{Z_2} \left[\int_X r^2(X,F_{X|Z}) f(X|Z)\mathrm{d}X \right] f(Z_2|Z_1)\mathrm{d}Z_2 \\
=& \int_{Z_2} E[r(X,F_{X|Z})|Z]^2 f(Z_2|Z_1)\mathrm{d}Z_2 \\
=& \int_{Z_2} \mathrm{var}[r(X,F_{X|Z})|Z] f(Z_2|Z_1)\mathrm{d}Z_2 \\
=& \int_{Z_2} 1/3 f(Z_2|Z_1)\mathrm{d}Z_2 \\
=& 1/3,
\end{aligned}$$

and similarly, $\mathrm{var}[r(Y,G_{Y|Z})|Z_1] = 1/3$, we have $c^*_{Z_1} = 3$ and $w^*_{Z_1} = 1$. That is,

$$\gamma_{XY\cdot Z|Z_1} = E_{Z_2|Z_1}(\gamma_{XY|Z}) = 3E_{Z_2|Z_1}(P_{c|Z} - P_{d|Z})$$

Then, it is easy to verify $\gamma_{XY\cdot Z} = E_{Z_1}(\gamma_{XY\cdot Z|Z_1})$. However, when one or both of X and Y are discrete, since $w^*_{Z_1}$ and $c^*_{Z_1}$ are usually not constant, this relationship generally does not hold.

Chapter 3

Modeling Continuous Outcomes Using Ordinal Regression with Cumulative Probabilities

In this chapter, we study the application of a widely used ordinal regression model, the cumulative probability model, for continuous outcomes. Cumulative probability models applied to continuous outcomes can be thought of as semiparametric transformation models: a linear relationship is assumed between covariates and a latent variable whose distribution is implied by the choice of link function, and the observed continuous response variable arises from an unspecified monotonic transformation of the latent variable. We describe estimation and inference, and discuss model assumptions. Extensive simulations are performed to investigate the finite sample performance of these models with and without correct link function specification. We find that properly specified cumulative probability models generally have good finite sample performance with moderate sample sizes, but that bias may occur when the sample size is small. Cumulative probability models are fairly robust to minor or moderate link function misspecification in our simulations. We illustrate their application in a study of the treatment of HIV. CD4 cell count and viral load 6 months after the initiation of antiretroviral therapy are modeled using cumulative probability models; both variables typically require transformations and viral load has a large proportion of measurements below a detection limit.

3.1 Introduction

Continuous data are also ordinal, and ordinal regression models can be fit to continuous outcomes (Sall, 1991; Harrell, 2015). The first ordinal regression model was developed by Walker & Duncan (1967) as an extension of logistic regression to ordered categorical data. This class of models was later studied by McCullagh (1980) and referred to as proportional odds models for the logit link and proportional hazards models for the complementary

log-log link. To distinguish this class of models from other ordinal regression models (e.g., continuation ratio and adjacent-categories models), these models have been referred to as cumulative link models (Agresti, 2010). However, this nomenclature is problematic because probabilities, not link functions, are added. Hence, we refer to this class of models as cumulative probability models.

The use of ordinal cumulative probability models for continuous outcomes has many attractive features. First, ordinal regression models are robust because they only incorporate the order information of response variables and are therefore invariant to any monotonic transformation of outcomes. This is particularly useful when the distributions of continuous responses are skewed and different transformations may give conflicting results. Second, cumulative probability models directly model the conditional cumulative distribution function (CDF), from which other components of the conditional distribution (e.g., expectation and quantiles) can be easily derived. Therefore, one can examine various aspects of the conditional distribution from a single cumulative probability model. Finally, cumulative probability models can handle any orderable response, including those with mixed types of continuous and discrete distributions. This may be particularly useful when dealing with detection limits, e.g., measurements censored at an assay detection limit resulting in a mixture of an undetectable category and detectable quantities.

Although the idea of using ordinal cumulative probability models for continuous outcomes is attractive and has been around for a while, we have not seen it used very often in practice. This may be in part due to computing limitations, as most software that fit cumulative probability models are currently designed for ordered categorical outcomes with relatively small numbers of potential response values. This need not be the case anymore, with modern computing power and improved algorithms (Sall, 1991), implemented in existing software (Harrell, 2016). However, we believe the primary reason for limited use of these models with continuous outcomes is a lack of awareness. Cumulative probability models were first invented to handle discrete ordinal outcomes, and their potential utility

for the analysis of continuous outcomes has been largely unrecognized. Furthermore, we are unaware of an in-depth study of the use of cumulative probability models for continuous outcomes.

The goal of this manuscript is to describe and study the application of cumulative probability models to continuous outcomes. In Section 3.2 we describe details of the approach including the motivation, connections with semiparametric transformation models, estimation, inference, assumptions and model diagnostics. In Section 3.3, we investigate the finite sample performance of cumulative probability models with and without proper link function specification through simulations. In Section 3.4, we illustrate their application to an HIV study modelling CD4 cell count and viral load 6 months after the initiation of antiretroviral therapy. Both variables typically require transformations and viral load has a large proportion of measurements below a detection limit. Section 3.5 contains a discussion and Section 3.6 contains additional simulation results.

3.2 Cumulative Probability Models for Continuous Outcomes

3.2.1 Latent Variable Motivation

For a discrete ordinal response variable Y with K categories and covariates X, a cumulative probability model can be written as

$$G[P(Y \leq j|X)] = \alpha_j - \beta X \text{ for } j = 1, ..., K-1, \tag{3.1}$$

where G is a link function (McCullagh, 1980; Agresti, 2010). It is well known that the cumulative probability model (3.1) can be motivated by assuming there is an underlying continuous latent variable Y^* from a parametric linear model $Y^* = \beta X + \varepsilon$, where $\varepsilon \sim F_\varepsilon \equiv G^{-1}$, and then assuming the observed response variable Y is generated by discretizing the latent variable Y^*, i.e., $Y = j$ if and only if $\alpha_{j-1} < Y^* \leq \alpha_j$, where $-\infty \equiv \alpha_0 < \alpha_1 < ... < \alpha_{K-1} < \alpha_K \equiv +\infty$. Hence, the cumulative probability model (3.1) can be motivated

Table 3.1: Commonly used link functions and their corresponding error distributions. $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Name	Link Function	Error Distribution	CDF
logit	$\log[y/(1-y)]$	logistic	$\exp(y)/[1+\exp(y)]$
probit	$\Phi^{-1}(y)$	normal	$\Phi(y)$
loglog	$-\log[-\log(y)]$	extreme value type II (Gumbel Maximum)	$\exp[-\exp(-y)]$
cloglog	$\log[-\log(1-y)]$	extreme value type I (Gumbel Minimum)	$1-\exp[-\exp(y)]$
cauchit	$\tan[\pi(y-1/2)]$	Cauchy	$\tan^{-1}(y)/\pi+1/2$

from a linear transformation model: $Y = H(\beta X + \varepsilon)$, where $\varepsilon \sim G^{-1}$ and H is a left-continuous non-decreasing step function mapping the latent variable Y^* to the observable discrete response Y, e.g., $H(y^*) = j$ if $\alpha_{j-1} < y^* \leq \alpha_j$. Table 3.1 summarizes commonly used link functions and their corresponding error distributions.

Cumulative probability models for continuous response variables can be similarly motivated. Consider an underlying latent variable $Y^* = \beta X + \varepsilon$, where $\varepsilon \sim G^{-1}$ and the observed response variable Y results from a monotonic increasing mapping, $H(Y^*)$. That is, consider a linear transformation model

$$Y = H(\beta X + \varepsilon), \tag{3.2}$$

where the transformation function $H(\cdot)$ is a monotonic increasing function. Since

$$\begin{aligned} P(Y \leq y|X) &= P[H(\beta X + \varepsilon) \leq y|X] \\ &= P[\beta X + \varepsilon \leq H^{-1}(y)|X] \\ &= P[\varepsilon \leq H^{-1}(y) - \beta X|X] \\ &= G^{-1}[H^{-1}(y) - \beta X], \end{aligned}$$

we can rewrite the model in the form of cumulative probability models

$$G[P(Y \leq y|X)] = \alpha(y) - \beta X, \tag{3.3}$$

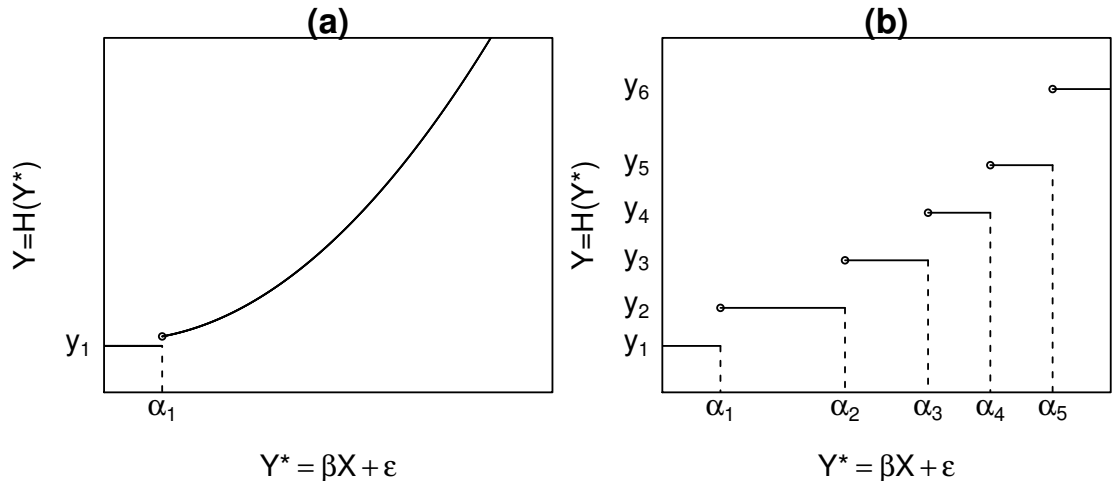

Figure 3.1: (a): $H(\cdot)$ is not strictly increasing. Define $H^{-1}(y) = \sup\{x : H(x) \leq y\}$. In this example, $H^{-1}(y_1) = \alpha_1$. This could be used to illustrate the detection limit issue. When the true underlying continuous quantity Y^* is below the detection limit α_1, the observed measurement Y collapses to the "undetectable" (the lowest) category, which results in a mixed type of continuous and discrete distributions for Y. (b): A special case where $H(\cdot)$ is step function: $H^{-1}(y_i) = \alpha_i$. Y can only take on limited number of possible values and is a discrete ordinal variable.

where $\alpha(y) = H^{-1}(y)$. The above argument only needs $H(\cdot)$ to be a monotonic increasing function and does not require a specific form for $H(\cdot)$. That is, we do not assume a linear relationship between response Y and covariates, but only require linearity after some unspecified monotonic transformation of Y. The above derivation also works when $H(\cdot)$ is not strictly increasing. If $H(\cdot)$ is constant on some interval as shown in Figure 3.1(a), the observed response variable Y has the mixed types of continuous and discrete distributions, and we can define $H^{-1}(y) = sup\{x : H(x) \leq y\}$. In the extreme case where $H(\cdot)$ is a step function as shown in Figure 3.1(b), the observed response Y is truly a discrete ordinal variable with a fixed number of categories. Therefore, cumulative probability models motivated from the transformation of latent variables are applicable to continuous, discrete, and mixed types of ordinal variables.

Unlike other commonly used continuous regression models which only focus on one aspect of the conditional distribution, e.g., the conditional mean for linear regression models or the conditional quantiles for quantile regression models, cumulative probability mod-

els model the entire conditional distribution, $F_{Y|X}(y) = G^{-1}[\alpha(y) - \beta X]$. Specifically, $F_{Y|X=0}(y) = G^{-1}[\alpha(y)]$, or $\alpha(y) = G[F_{Y|X=0}(y)]$; therefore, the intercept $\alpha(y)$ can be interpreted as the link function transformed outcome CDF at baseline $X = 0$. Since $\alpha(y) = H^{-1}(y)$, it can also be interpreted as the transformation needed for Y to best fit a parametric linear regression model with the error term $\varepsilon \sim G^{-1}$. The association between covariates X and the response variable Y is captured by the slope parameter, $-\beta$. Depending on the link function chosen, the slope, $-\beta$, may have a nice interpretation, e.g., as a log-odds ratio if G is the logit link function or as a log-hazard ratio if G is the cloglog link function. Because cumulative probability models directly model the conditional CDF, other components of the distribution, such as the conditional expectation or conditional quantiles, can be readily derived. In this sense, cumulative probability models can provide more details of the conditional distributions with fewer assumptions compared with other commonly used regression models for continuous responses.

3.2.2 Nonparametric Maximum Likelihood Estimation

The latent variable motivation relates cumulative probability models to semiparametric transformation models. Zeng & Lin (2007) studied these type of semiparametric transformation models with censored data. They proposed a nonparametric maximum likelihood estimator (NPMLE) for this model and established its consistency, asymptotic normality, and asymptotic efficiency. NPMLE can be viewed as a generalization of the maximum likelihood estimator for semiparametric or nonparametric models, in which the likelihood function is modified by replacing the functional parameter by an empirical function with jumps only at the observed data. For example, in Zeng & Lin (2007), the likelihood function was first constructed based on the counting process and then modified by approximating the baseline cumulative intensity function using a step function with jumps at the observed failure times.

For cumulative probability models with general continuous outcomes, we consider a

similar approach but with a different parameterization. For a given dataset of size n with the continuous response Y and p-dimensional covariate X, we order the observations so that $y_1 < y_2 < \cdots < y_n$ (assuming no ties), and then approximate $\alpha(y)$ using a step function with steps at each observed value y_i and denote them as $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, where $\alpha_1 < \alpha_2 \cdots < \alpha_n$. Since the conditional probability density function for observation i can be written as $f(y_i|X = x_i) = \lim_{\Delta y \to 0} \frac{F_{Y|X=x_i}(y_i) - F_{Y|X=x_i}(y_i - \Delta y)}{\Delta y} = \lim_{\Delta y \to 0} \frac{G^{-1}[\alpha(y_i) - \beta x] - G^{-1}[\alpha(y_i - \Delta y) - \beta x]}{\Delta y}$, we approximate the contribution of observation i to the likelihood with $f^*(y_i|x_i) \propto [G^{-1}(\alpha_i - \beta X_i) - G^{-1}(\alpha_{i-1} - \beta X_i)]$. Therefore, the NPMLEs can be obtained by maximizing

$$\begin{aligned} L^*(\beta, \boldsymbol{\alpha}) &= \prod_{i=1}^{n} f^*(y_i|x_i) \\ &\propto \prod_{i=1}^{n} [G^{-1}(\alpha_i - \beta x_i) - G^{-1}(\alpha_{i-1} - \beta x_i)]. \end{aligned}$$

Since α_0 and α_n are only present in the first and the last term of L^*, respectively, and also since G^{-1} is a monotonic increasing function, L^* is maximized when $\hat{\alpha}_0 = -\infty$ and $\hat{\alpha}_n = +\infty$. Plugging in $\hat{\alpha}_0$ and $\hat{\alpha}_n$, L^* can be simplified as

$$L^*(\beta, \boldsymbol{\alpha}) \propto [G^{-1}(\alpha_1 - \beta x_1)][G^{-1}(\alpha_2 - \beta x_2) - G^{-1}(\alpha_1 - \beta x_2)] \cdots [1 - G^{-1}(\alpha_{n-1} - \beta x_n)], \tag{3.4}$$

with a total of $n-1$ intercepts (assuming no ties) and p slopes to estimate.

Note that $L^*(\beta, \boldsymbol{\alpha})$ has the same structure as the multinomial likelihood of the cumulative probability model for a discrete variable with only one observation in each category. Therefore, maximizing $L^*(\beta, \boldsymbol{\alpha})$ can be easily achieved by treating Y as a discrete variable (each value of observed Y as one category) and fitting the discrete cumulative probability model using standard statistical software. Although this approach is convenient in practice, computational challenges arise with large sample sizes because the Newton-Raphson algorithm typically used for maximization requires inverting the Hessian matrix whose dimensions ($n-1+p$ by $n-1+p$) increase with the sample size. However, as

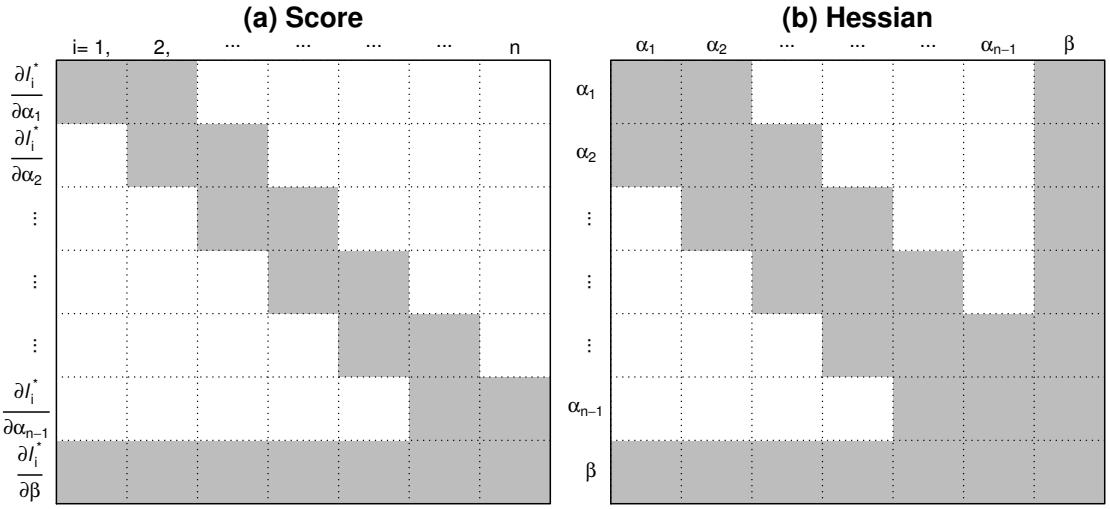

Figure 3.2: (a) Each observation's contribution to the score function assuming observations are ordered by the value of y, i.e., $y_1 < y_2 < \cdots < y_n$. White region indicates zero and grey region indicates none-zero values. (b) The bordered tridiagonal structure of Hessian matrix of $\log(L^*)$ with respect to intercepts and slopes.

shown in Figure 3.2, due to the special structure of the score functions, the Hessian matrix of $\log(L^*)$ with respect to the intercepts has a block tridiagonal structure. That is, large regions of the Hessian are 0. With this structure, the matrix inversion can be solved efficiently through Cholesky decomposition (Sall, 1991). Taking advantage of these facts, Harrell (2016) implemented a computationally efficient algorithm to obtain the NPMLE with the `orm` function in the `rms` R package. Note, the `orm` function uses a slightly different notation, i.e., $G[1-F_{Y|X}(y)] = \alpha_{orm}(y) + \beta_{orm}X$. For symmetric error distributions such as normal and logistic distributions, whose corresponding link functions have the property $G(1-x) = G(x)$, the regression coefficients from the `orm` function differ with those from Formula (3.3) only by sign. But for the cloglog and loglog link functions, this property does not hold. Instead, Formula (3.3) with the cloglog link function corresponds to the `orm` function with the loglog link function and vice versa. In this paper, we use the `orm` function to fit all cumulative probability models for continuous responses, but we use the notation from (3.3) to be consistent with the literature for cumulative probability models.

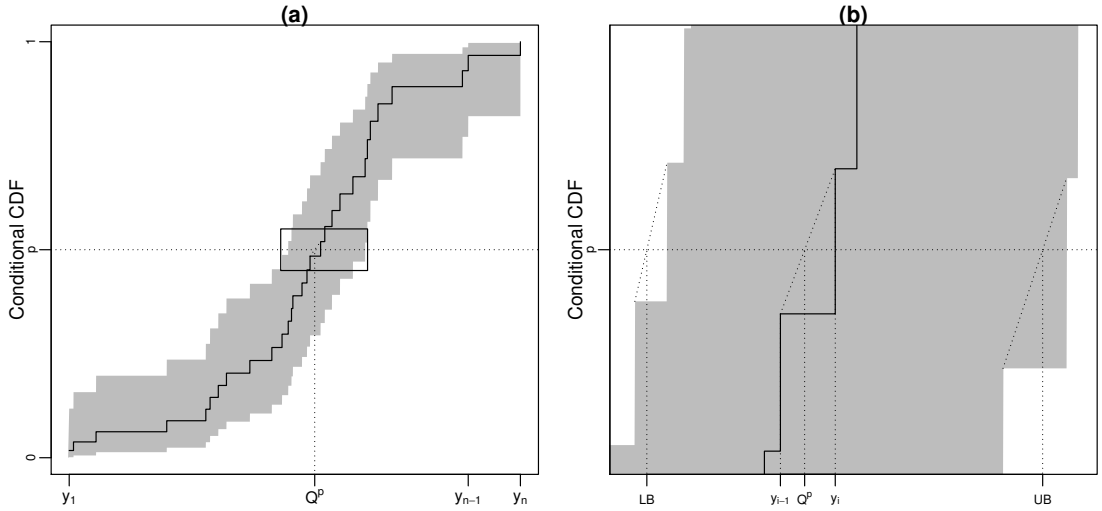

Figure 3.3: (a): An example of the estimated conditional CDF and its pointwise confidence intervals. (b): An example of the estimated p^{th} quantile and its confidence intervals. The p^{th} quantile of the conditional distribution, denoted as Q^p, and its confidence interval (LB, UB) could be obtained from the linear interpolation of y_{i-1} and y_i, where $y_i = \inf\{y : \hat{F}_{Y|X}(y) \geq p\}$, based on the estimated conditional CDF (or its pointwise confidence intervals).

The variance and covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ can be obtained by treating L^* as a parametric likelihood and inverting the observed information matrix for all these parameters. Again, this can be achieved in a computationally efficient way due to the bordered tridiagonal structure of the Hessian matrix. With the NPMLE of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and their variance-covariance estimators, the conditional CDF evaluated at the observed values can be calculated directly, and their standard error estimates can be obtained by the delta method. Given the order constraint of the cumulative probabilities, it is natural to use a step function with jumps at the observed values to approximate the entire conditional CDF. Figure 3.3(a) shows an example of the step function estimator for the entire conditional distribution and its pointwise 95% confidence interval. With the estimated CDF, other properties of the conditional distribution can be easily derived. For example, we can estimate the conditional mean as $\hat{E}(Y|X) = \sum_{i=1}^{n} y_i \hat{f}(y_i|X) = \sum_{i=1}^{n} y_i [G^{-1}(\hat{\alpha}_i - \hat{\boldsymbol{\beta}}X) - G^{-1}(\hat{\alpha}_{i-1} - \hat{\boldsymbol{\beta}}X)]$ and obtain its standard error estimate by the delta method. Also, as illustrated in Figure

3.3(b), the conditional quantiles and their confidence intervals can be obtained from linear interpolation of the inverse of the conditional CDF and its pointwise confidence intervals, respectively.

It should be noted that there is no general theory for the asymptotic properties of NPMLEs. In the paper of Zeng & Lin (2007), the authors only prove the consistency, asymptotic normality, and asymptotic efficiency for censored data and their proofs rely on the boundedness of the estimator of $H(\cdot)$. Based on personal communication, these authors have been working on establishing the asymptotic properties of NPMLEs for semiparametric transformation models with uncensored data. However, for a general continuous response, both the true value of $H(\cdot)$ and its estimator could be unbounded. This imposes tremendous technical difficulties in the proofs. In this manuscript we do not attempt to fill this gap in the theory. Instead, we perform extensive simulations (Section 3.3 and additional simulations in Section 3.6) that suggest that with proper model specification (i.e., correctly specifying the link function and the mean model, but leaving $H(\cdot)$ unspecified), the NPMLE procedure just described results in consistent, asymptotically normal estimators with well-approximated variance.

3.2.3 An Illustration of Cumulative Probability Models

We illustrate cumulative probability models and compare them with nonparametric and parametric models in the following simple example. Consider a single binary covariate $X \sim Bernoulli(p)$, where $p = 0.5$, the latent response $Y^* = \beta X + \varepsilon$, where $\beta = 1$ and $\varepsilon \sim N(0,1)$, and the observable response $Y = H(Y^*)$. For simplicity, we set $H(y) = y$, that is, no transformation is needed. We order the observations so that $y_1 < y_2 < ... < y_n$. In this setting, since the covariate is binary, both nonparametric and parametric models can be easily applied to estimate the conditional CDF. Nonparametrically, we can compute the empirical CDFs for the subgroup $X = 0$ and the subgroup $X = 1$, respectively. Parametrically, we can fit a linear regression model to estimate the conditional mean and variance,

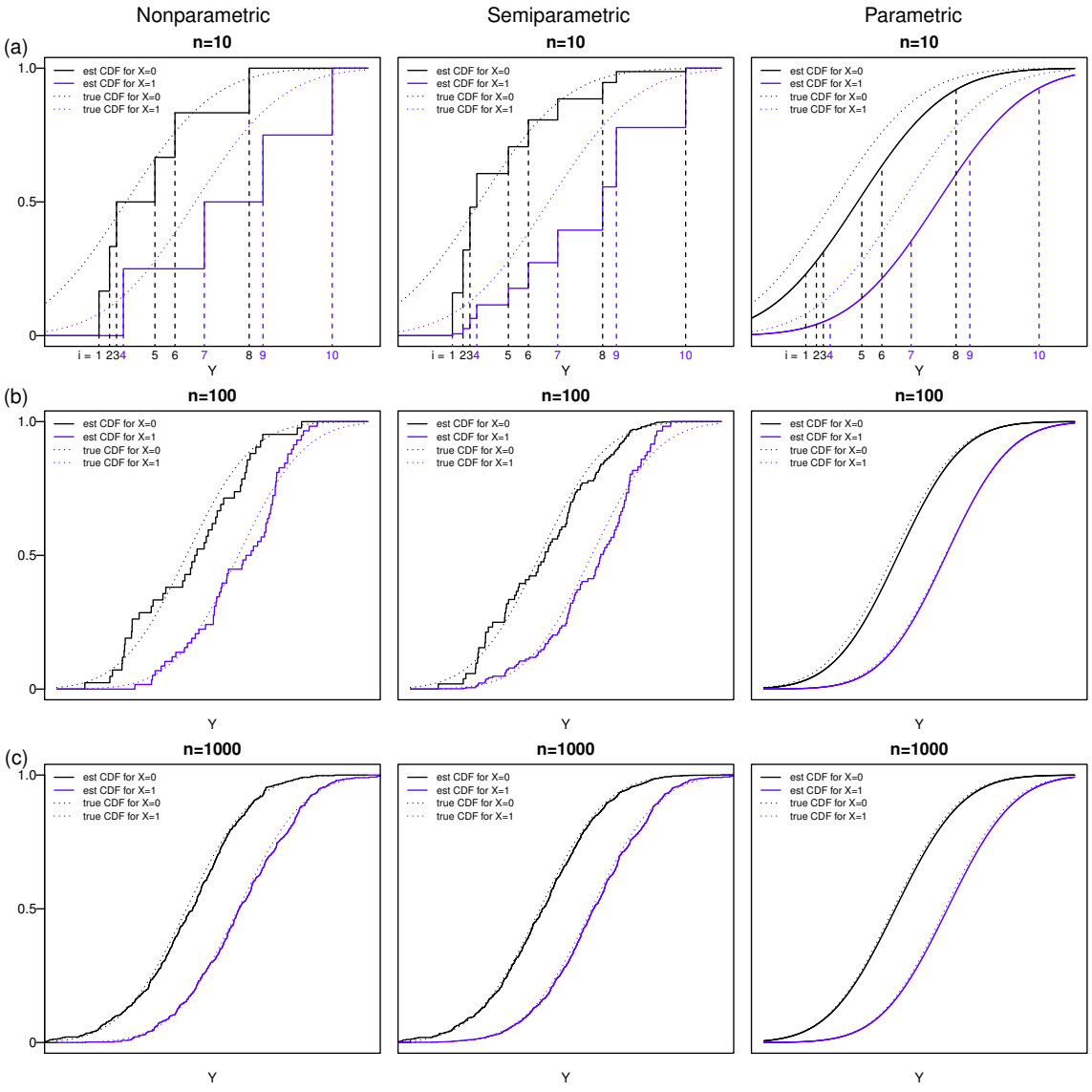

Figure 3.4: Estimation of conditional CDF from cumulative probability models compared with parametric and nonparametric models in a simple example: (a) with the sample size of 10, (b) with the sample size of 100, and (c) with the sample size of 1000.

and then calculate the conditional CDF using the cumulative probability function of the normal distribution. Figure 3.4 shows the estimation of conditional CDFs with different approaches when the sample size is 10, 100 and 1000, respectively.

The nonparametric approach only uses information from each subgroup to estimate its conditional CDF, i.e., the empirical CDF is a step function with jumps only at the observed values within each subgroup, e.g., 6 jumps for $X = 0$ and 4 jumps for $X = 1$ in the specific example with the sample size of 10 in Figure 3.4(a). The parametric approach (the normal linear regression model) pools information from all observations and also uses the assumption of a normal error distribution to provide smooth estimates for the conditional CDFs. The cumulative probability model (semiparametric) provides something in between, i.e., the estimates are step functions but with jumps at observed values from both subgroups, e.g., 10 jumps for both $X = 0$ and $X = 1$ as shown in the middle panel of Figure 3.4(a). The nonparametric approach does not make any assumptions about the conditional distributions; therefore, it is the most robust estimation procedure. But it is not efficient because it only uses information within each subgroup and it is not easily extended to continuous or multivariate X. The parametric approach assumes normality for the conditional distribution. It is most efficient if the assumption is properly specified, but not robust. The cumulative probability model does not make full parametric assumptions about the conditional distributions (i.e., it only assumes that after some unspecified transformation, the data are normal), but still pools information by assuming a shared shape for the conditional distributions. Therefore, the cumulative probability model provides a compromise between efficiency and robustness.

3.2.4 Assumptions of Cumulative Probability models

Cumulative probability models have an assumption of parallelism. That is, the difference between link function transformed conditional CDFs for different values of covariates is constant, i.e., $G^{-1}[F_{Y|X=x_2}(y)] - G^{-1}[F_{Y|X=x_1}(y)] = \beta(x_2 - x_1)$, which is free of y. Specif-

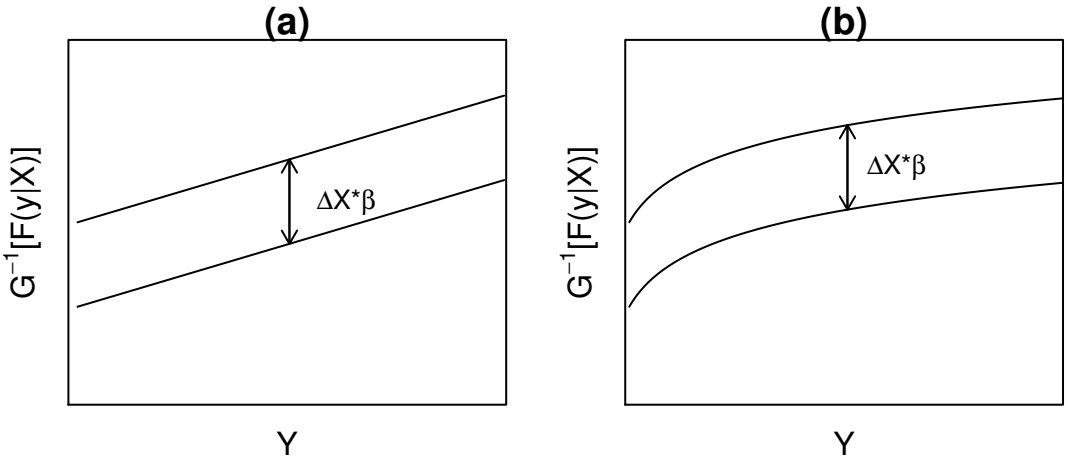

Figure 3.5: (a): The parallelism assumption is implicitly assumed in normal linear regression models. But additionally, these models assume linearity for the link function (probit) transformed conditional CDF as a linear function of y. (b): The parallelism assumption is also assumed in other parametric transformation models where both the error distribution and the transformation are specified. In this example, we set $\varepsilon \sim G^{-1}$ and $H(y) = \exp(y)$. But in addition to the parallelism assumption, these parametric transformation models also assume a specific smooth shape for the link function transformed conditional CDF as a function of y. A monotonic transformation of Y can be viewed as keeping the y-axis unchanged but changing the scale of the x-axis in this plot. It could change the shape of the curve but would keep distance between two curves constant (parallel). Cumulative probability models do not specify the transformation. Therefore, their link transformed conditional CDF could be any monotonic function of Y as long as the parallelism assumption holds. Adapted from Harrell (2015).

ically, it assumes proportional odds with the logit link and proportional hazards with the cloglog link. Cumulative probability models assume a parametric error distribution on the transformed latent variable scale, i.e., $\varepsilon \sim G^{-1}$, but leave the monotonic transformation $H(\cdot)$ unspecified; therefore, they can be viewed as semiparametric transformation models.

Figure 3.5 illustrates the different assumptions between cumulative probability models and parametric transformation models. A parametric transformation model is written in the form $Y = H(\beta X + \varepsilon)$, where $H(\cdot)$ and the distribution of ε are specified. Note that a normal linear model is a parametric transformation model with $H(y) = y$ and $\varepsilon \sim$ N(0,1). The parallelism assumption is also implicitly assumed in these parametric linear transformation models. In addition, however, parametric transformation models implicitly assume a specific form for the link function transformed conditional CDF, $G^{-1}[F(y|X)]$. In contrast, cumulative probability models do not assume the form of $G^{-1}[F(y|X)]$, but only require a constant distance between different levels of covariates (parallelism).

Unlike least squares regressions, which mainly make assumptions on the mean and variance of the error distribution, cumulative probability models require the specification of a link function, which corresponds to specifying the CDF of the error distribution on the transformed scale. Although the commonly used link functions represent various types of error distributions, e.g., the probit and logit link functions correspond to bell-shaped and symmetric error distributions with different tail densities, whereas cloglog and loglog link functions represent error distributions skewed in opposite directions, in practice there is no guarantee that the latent error has exactly the same CDF as the inverse of the link function. Therefore, it is of interest to study the robustness of cumulative probability models with link function misspecification, especially when the misspecification is only moderate, e.g., using the probit link function when the true latent error distribution is the t-distribution or using the cloglog/loglog link functions when the true error distribution is skewed in the same direction but does not have exactly the same shape as the specified link function. We examine these misspecification through simulations in Section 3.3.

3.2.5 Model Diagnostics

As with discrete variables, the link function of cumulative probability models for continuous responses should ideally be pre-specified based on preliminary scientific knowledge and convenience of interpretation. This may be challenging in application since it requires the specification of the error distribution on the unknown transformed scale. Model diagnostics, especially for the choice of link functions, is important. A Goodness-of-link test has been developed by Genter & Farewell (1985) to discriminate the model fit between probit, cloglog, and loglog link functions for discrete variables. The main idea is that these three link functions can be considered as special cases of a generalized log-gamma link function with an extra parameter. Then, a likelihood ratio test comparing the full model (using the log-gamma link) with the reduced model (using the probit, cloglog, or loglog link) could provide information about the goodness of fit. To avoid the computational burden of fitting the full model, Genter & Farewell (1985) proposed to compare the log-likelihoods of probit, cloglog, and loglog models directly as a conservative approximation to the formal likelihood ratio test. Specifically, they claim that if twice the difference of two log-likelihoods exceeds the appropriate percentile of a chi-square distribution with 1 degree of freedom, one can infer that the link with the smaller likelihood is inappropriate. This approach is also applicable to continuous responses. However, an automated link function selection procedure based on the largest log-likelihood should be used with caution because it seems to share the problems of step-wise selection procedures (Huberty, 1989; Shepherd, 2008) according to our simulations (see details in Section 3.6.1), i.e., type I error rates are inflated, standard error estimates are too small, and the confidence intervals are too narrow.

The model fit of cumulative probability models can also be examined graphically with residuals. As described in Section 3.2.1, the intercept $\alpha(y)$ can be interpreted as the best transformation needed to fit the parametric linear model. Therefore, we can transform the observed y according to $\hat{\alpha}(y)$ and explore the model fit with residuals on the transformed scale. Alternatively, we can use probability-scale residuals, which were originally pro-

posed for discrete ordinal variables (Li & Shepherd, 2012; Shepherd et al., in press). The probability-scale residuals are functions of fitted conditional CDFs. They are uniformly distributed with proper model specification of continuous outcomes, and are therefore particularly useful in this setting. We illustrate the application of this new residual with details in Section 3.4.

3.3 Simulation Studies

3.3.1 Estimation with Proper Link Function Specification

In this section, we first conduct simulations to evaluate the finite sample performance of cumulative probability models for continuous response with proper link function specification. We generate data from $Y = \exp(\beta_1 X_1 + \beta_2 X_2 + \varepsilon)$, where $X_1 \sim \text{Bernoulli}(0.5)$, $X_2 \sim N(0,1)$, $\beta_1 = 1$ and $\beta_2 = -0.5$, and with two error distributions: (i) $\varepsilon \sim N(0,1)$ and (ii) ε generated from the extreme value distribution (type I) with location parameter 0 and scale parameter 1. The corresponding properly specified link functions for these two error distributions are probit and cloglog, respectively. We conducted simulations for different sample sizes with $n = 25, 50, 100, 200, 500$, and 1000. For each sample size, simulations were replicated $10,000$ times. For the purpose of better visualization, we summarize the results with Figures. More details of simulation results can be found in Section 3.6.

As discussed in Section 3.2.1, when the link function is properly specified, the intercept $\alpha(y)$ of cumulative probability models corresponds to the proper transformation needed for the parametric linear model, which is $\log(y)$ in this case. With NPMLE, $\alpha(y)$ is approximated using a step function with jumps at observed values of Y. To illustrate this point, we plot $\hat{\alpha}(y)$ vs. the proper transformation $\log(y)$ from the first simulation replicate with the sample size of 100 in Figure 3.6. The average of the step function estimates $\hat{\alpha}(y)$ over all $10,000$ simulation replicates is also plotted. According to Figure 3.6, in our simulations with sample size of 100, the NPMLE $\hat{\alpha}(y)$ shows little bias compared with the true

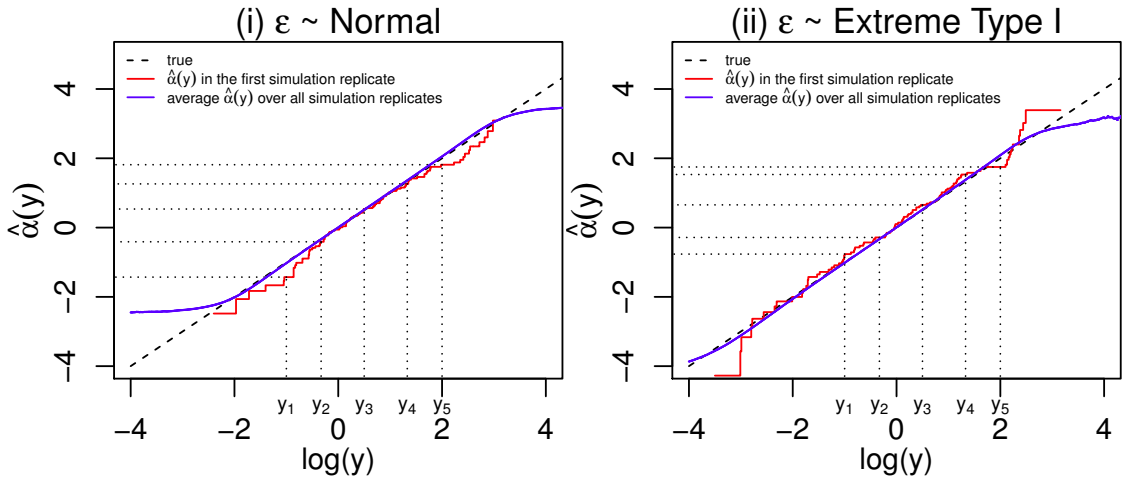

Figure 3.6: The intercept $\hat{\alpha}(y)$ can be interpreted as the proper transformation needed to fit the parametric linear model. In one simulation replication, not all possible values of Y would be observed, therefore, $\alpha(y)$ is approximated using a step function with jumps at observed values of Y. However, the average of the step function estimates over all simulation replicates is smooth and very close to the true transformation except in the tail regions where little information is observed.

transformation except in the tail regions.

Figure 3.7 summarizes the performance for estimating the regression coefficients, including slopes β_1, β_2, and the intercept $\alpha(y)$ with different sample sizes. For the purpose of evaluation, we examine the values of $\hat{\alpha}(y)$ at $y_1 = 0.368, y_2 = 0.719, y_3 = 1.649, y_4 = 3.781,$ and $y_5 = 7.389$ (shown in Figure 3.6). At those values, the marginal cumulative probabilities of Y with the error distribution (i) are close to $0.1, 0.25, 0.5, 0.75, 0.9$, and are close to $0.23, 0.39, 0.63, 0.84, 0.95$ with the error distribution (ii), respectively. The log transformation (the true proper transformation) at those values are $\log(y_1) = -1, \log(y_2) = -0.33, \log(y_3) = 0.5, \log(y_4) = 1.33,$ and $\log(y_5) = 2$. In simulation replicates, those exact values of Y are not observed, therefore, $\hat{\alpha}_(y_j)$ is not directly estimated from the model but approximated from the step function with $\max_{y \le y_j}\{\hat{\alpha}(y)\}$.

Our simulations demonstrate that when the link function is properly specified, the NPMLE has good performance for estimating the regression coefficients with moderate to large sample sizes (e.g., $n \geq 50$), i.e., bias is small and the standard error estimators agree

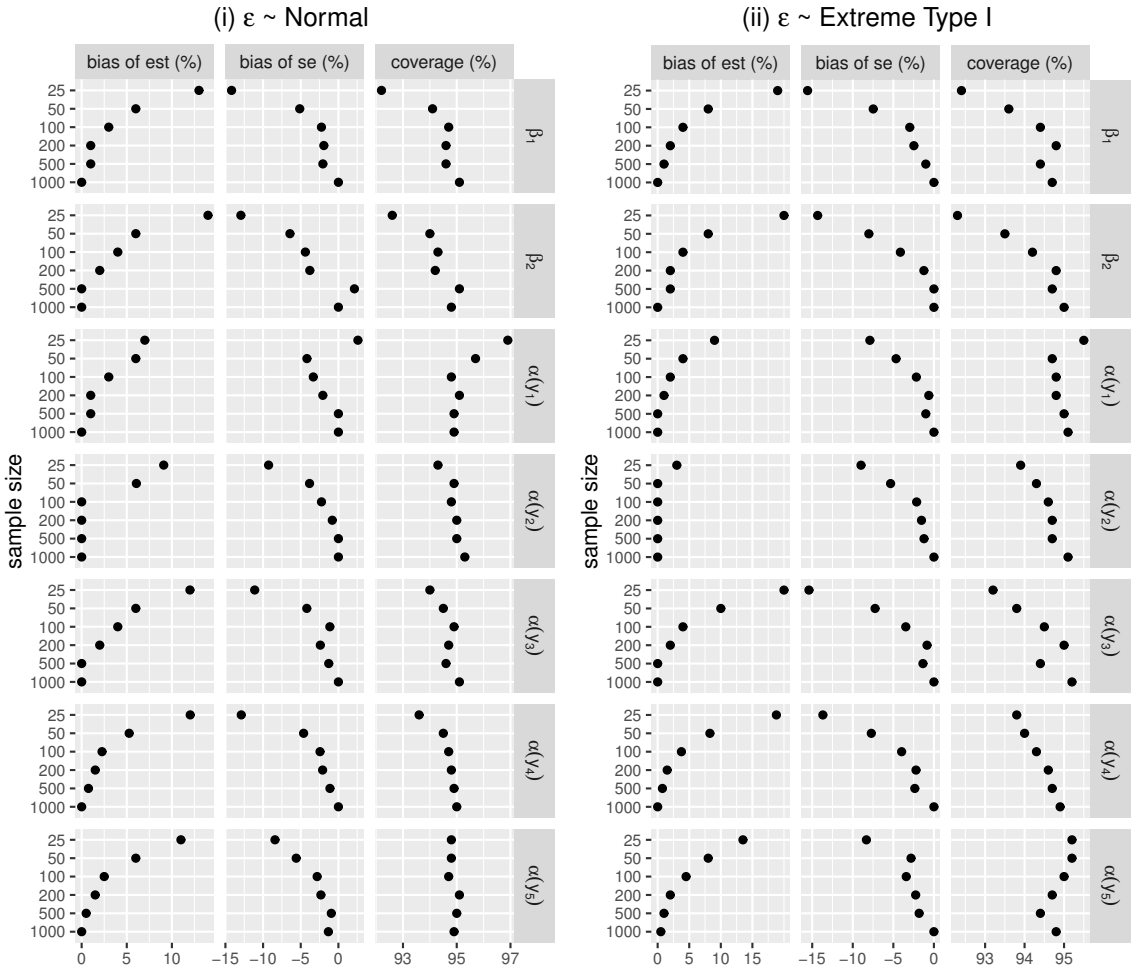

Figure 3.7: The performance of cumulative probability models on estimating the slopes β_1 and β_2 and the transformation at $y_1 = 0.368$, $y_2 = 0.719$, $y_3 = 1.649$, $y_4 = 3.781$, and $y_5 = 7.389$ with properly specified link functions. The results are based on 10,000 simulation replicates for each sample size. We report the percent bias (%) of the point estimate, the percent bias (%) of the standard error estimate, and the coverage probability of 95% confidence intervals. The percent bias of the point estimate is calculated as the mean of point estimates in 10,000 simulation replicates minus the true value and then divided by the true value. The percent bias of the standard error estimate is calculated as the mean of standard error estimates in 10,000 simulation replicates minus the standard deviation of point estimates in 10,000 simulation replicates, and then divided by the standard deviation of point estimates in 10,000 simulation replicates.

well with the empirical standard errors. Although we do observe some bias (up to 20%) when the sample size is small (e.g., $n = 25$), the bias in estimating regression coefficients and their standard errors decreases quickly with increasing sample size. For example, in our two simulation settings, the bias for regression coefficients is less than 8% with $n = 50$, less than 5% with the sample size of 100, and less than 2% with the sample size of 200. When the sample size gets to $1,000$, the average of estimates are almost identical to the true parameters.

For the purpose of comparison, we investigated the relative efficiency of the properly specified cumulative probability models compared with other commonly used parametric or semiparametric models with different sample sizes. For error distribution (i), we compare the cumulative probability model with the linear regression model after a Box-Cox transformation (Box & Cox, 1964). The latter can be viewed as a parametric transformation method, since the Box-Cox transformation with parameter λ can be written as

$$B(y;\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$$

The usual practice of Box-Cox transformation is to first find the $\hat{\lambda}$ which gives the highest profile likelihood for the transformed linear regression model, and then to treat $\hat{\lambda}$ as known (which is potentially problematic because it ignores uncertainty in its estimation) and transform the response y to $y^* = B(y;\hat{\lambda})$. Finally, linear regression is performed on the transformed response y^*. In our simulations, we performed the Box-Cox transformation using the `boxcox` function in the `MASS` R package. We used its default option in which λ is estimated by a grid search algorithm searching in the range of -2 to 2 with increments of 0.1.

In our simulation setting, the true proper transformation is a special case of the Box-Cox transformation with parameter $\lambda = 0$. With the latent scale error $\varepsilon \sim N(0,1)$, the

Box-Cox transformation model is a properly specified parametric transformation model. Figure 3.8(a) shows the relative efficiency of the properly specified cumulative probability model compared with the Box-Cox transformation model on estimating the slopes β_1 and β_2. Since the estimates from the cumulative probability models show small bias when the sample size is small, we estimate the relative efficiency with the ratio of mean square errors (MSE) instead of the ratio of variances. As expected, the properly specified cumulative probability model is generally less efficient than its parametric counterpart. However, the relative efficiency increases with the sample size. Specifically, in our simulation setting, the relative efficiency is over 80% when $n = 100$, and over 90% when $n = 1,000$, indicating only small efficiency loss when using properly specified cumulative probability models with moderate or large sample sizes.

For error distribution (ii), we compare the properly specified cumulative probability model with the Cox proportional hazards model. Note, in this scenario, both models assume proportional hazards and are properly specified semiparametric models. However, their estimation procedures are different. The Cox proportional hazards model maximizes the partial likelihood which is only a function of slopes, whereas the cumulative probability model maximizes the full likelihood L^* which is a function of both intercepts and slopes. Usually, the Cox proportional hazards model is applied to right censored data, but it can also be fitted to uncensored data with negative values (e.g., assume the response variable Y is a centered time without censoring). Figure 3.8(b) shows the relative efficiency of the cumulative probability and Cox proportional hazards models on estimating the slopes β_1 and β_2, again, measured with the MSE ratio. In our simulations, the cumulative probability model with the cloglog link function has similar performance with the Cox proportional hazards model when the sample size is moderate or large. For example, the relative efficiency is over 90% when $n = 100$ and very close to 1 when $n = 1,000$. However, for small sample size, such as $n = 25$, the estimates from the cumulative probability model seem to be less efficient than those from the Cox proportional hazards model.

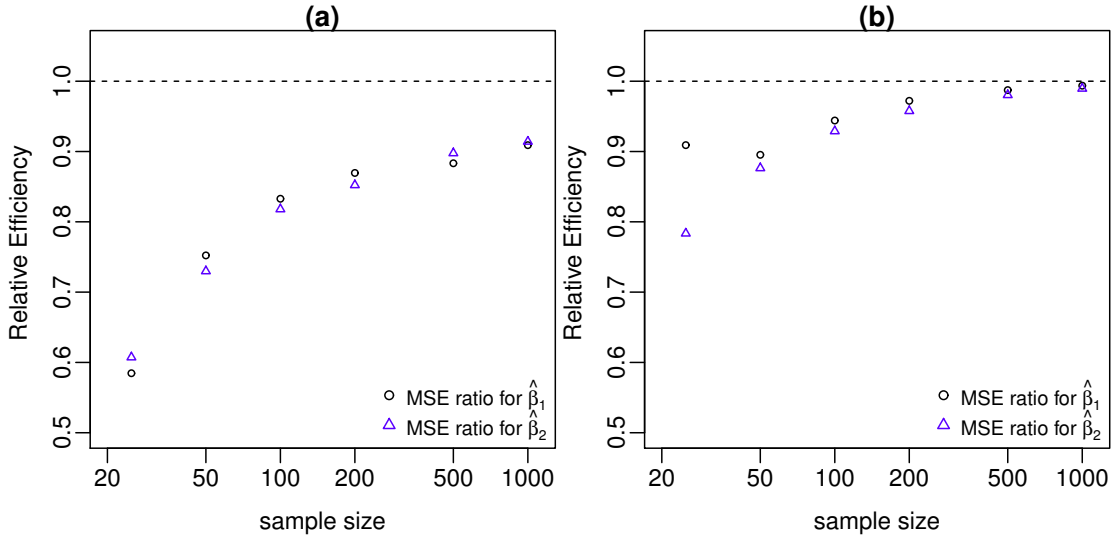

Figure 3.8: (a): The relative efficiency of properly specified cumulative probability model (using the probit link function) compared with properly specified Box-Cox transformation models in Scenario (i) $\varepsilon \sim N(0,1)$. For the Box-Cox transformation models, $\hat{\lambda}$ is estimated as the value with the highest profile likelihood by a grid search algorithm searching in the range of -2 to 2 with increments of 0.1. As in the common practice,$\hat{\lambda}$ is then treated as known and the uncertainty of estimating λ is ignored. (b): The relative efficiency of properly specified cumulative probability model (using the cloglog link function) compared with Cox proportional hazard model in Scenario (ii) when ε is generated from type I extreme value distribution.

As described in Section 3.2.2, the conditional CDF and other properties of the conditional distribution can be derived with the regression coefficients from cumulative probability models. Figure 3.9 summarizes the performance on estimating conditional CDFs when the link function is properly specified. Figures 3.10 and 3.11 summarize the performance on estimating conditional means and conditional quantiles using the properly specified cumulative probability models, respectively. Similar patterns are observed as in estimating the regression coefficients. Generally, the cumulative probability models have good performance for estimating different aspects of the conditional distribution with moderate or large sample sizes (e.g., $n \geq 50$ in our simulation settings). But when the sample size is relatively small (e.g., $n = 25$), there may be substantial bias both in estimating the parameters and their standard errors.

It is worth pointing out that although the cumulative probability model is usually less

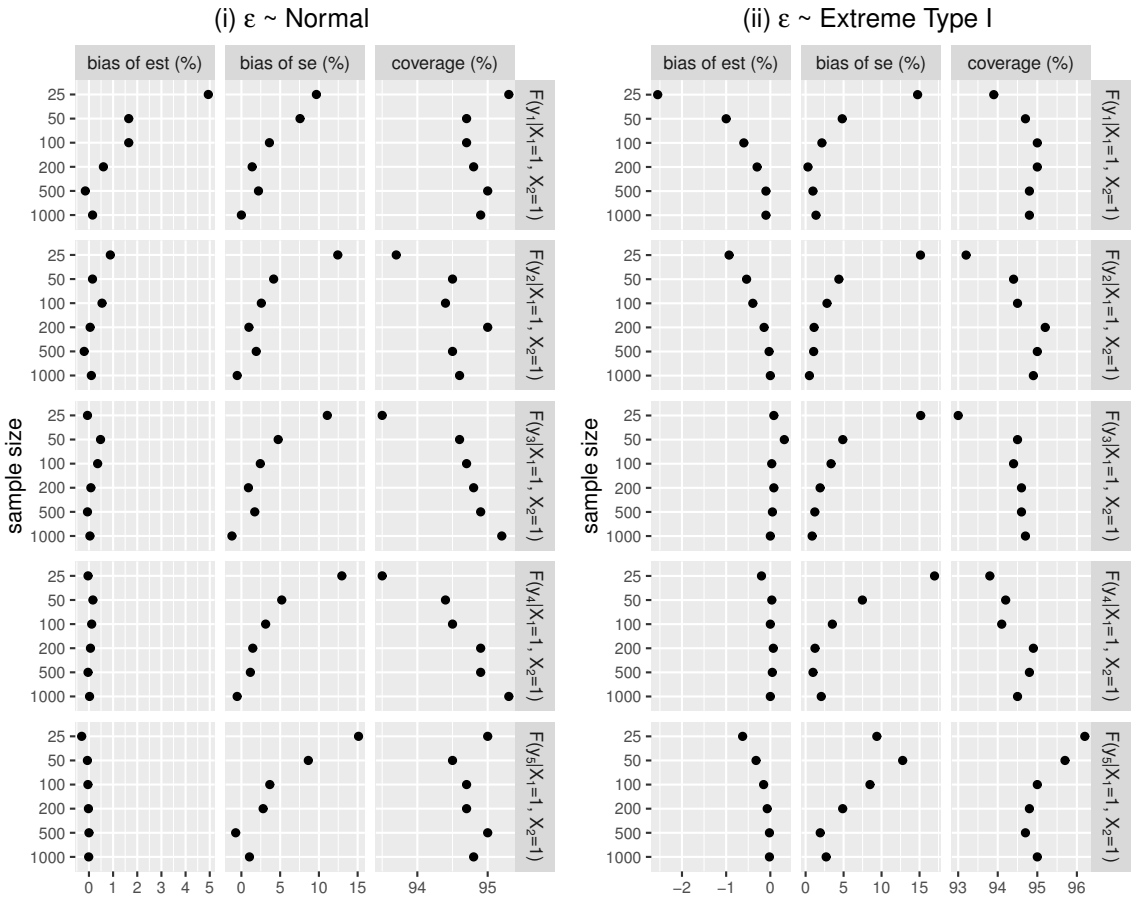

Figure 3.9: The performance of cumulative probability models on estimating conditional CDF evaluated at $y_1 = 0.368$, $y_2 = 0.719$, $y_3 = 1.649$, $y_4 = 3.781$, and $y_5 = 7.389$ with properly specified link functions. The results are based on 10,000 simulation replicates for each sample size.

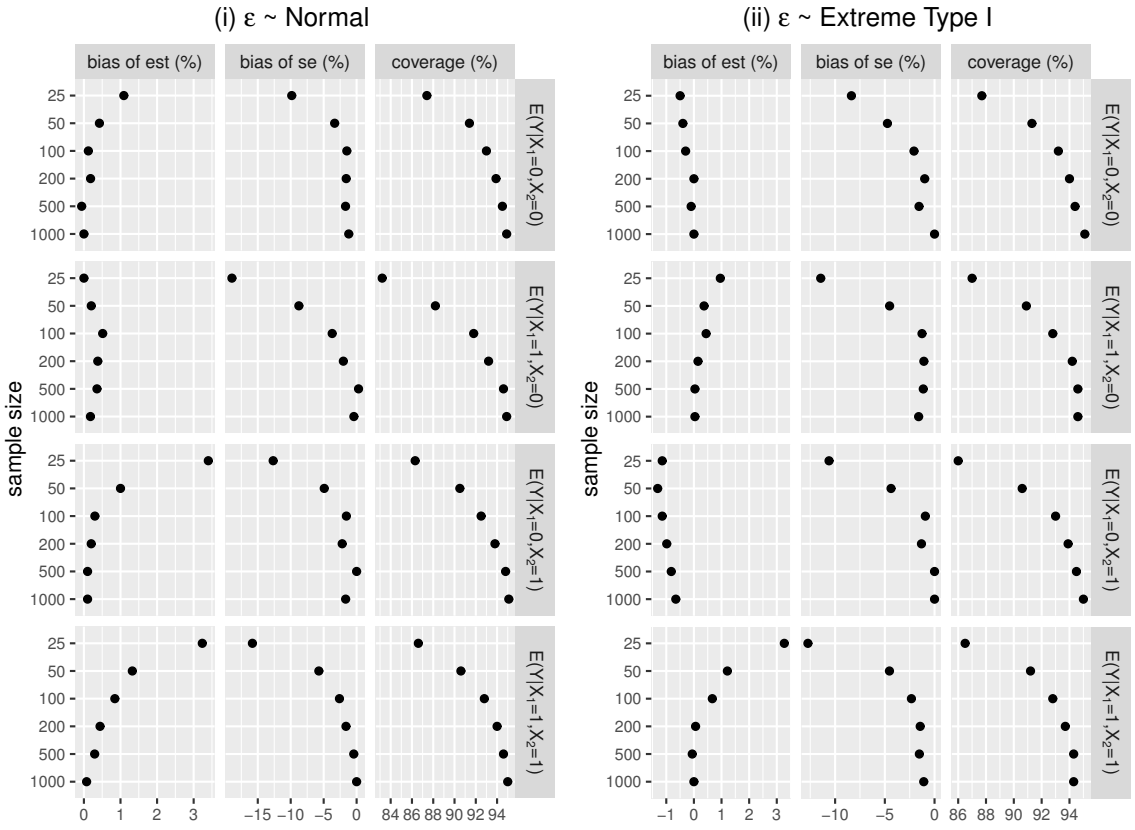

Figure 3.10: The performance of cumulative probability models on estimating conditional means with properly specified link functions. The results are based on 10,000 simulation replicates for each sample size.

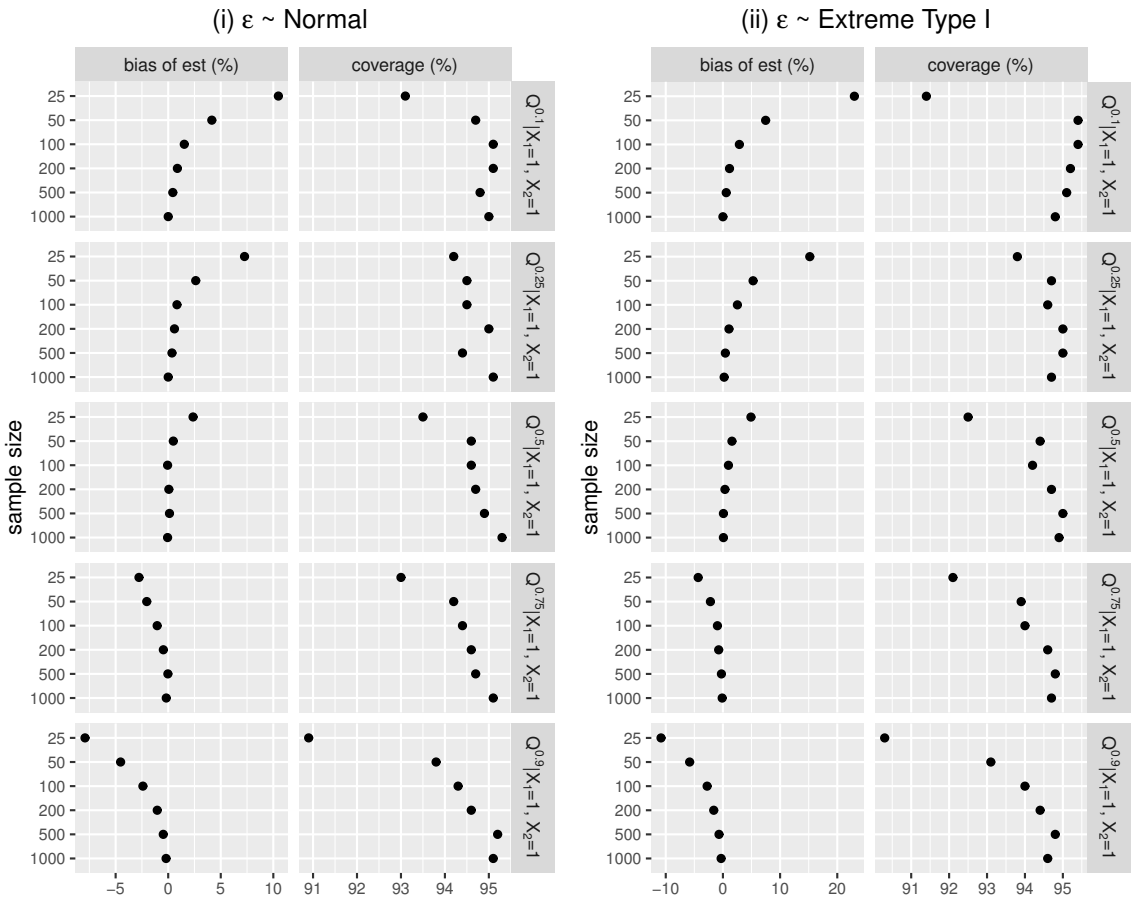

Figure 3.11: The performance of cumulative probability models on estimating conditional 10^{th},25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles with properly specified link functions. The results are based on 10,000 simulation replicates for each sample size.

efficient than the properly specified parametric models, it may still have some advantage when estimating some properties that are easier to derive from the conditional CDF. For example, with the error distribution (i) where the Box-Cox transformation model is properly specified, although the cumulative probability model is generally less efficient for estimating the slope parameters, it has good performance for estimating conditional mean with moderate to large sample sizes when the link function is properly specified (shown in Figure 3.10). However, with the regression coefficients from the Box-Cox transformation model, we can only consistently estimate the conditional mean on the transformed scale, i.e., $\hat{Y}^*|X \to E[B(Y,\lambda)|X]$. Performing a simple back transformation, $B^{-1}(\hat{Y}^*,\hat{\lambda})$, usually does not yield a consistent estimator for the conditional mean on the original scale because $B^{-1}(\hat{Y}^*,\hat{\lambda}) \to B^{-1}\{E[B(Y,\lambda)|X],\lambda\}$, which is well known from Jensen's inequality to not equal $E(Y|X)$.

3.3.2 Estimation with Link Function Misspecification

We now study the performance of cumulative probability models with link function misspecification. We generate data from $Y = H(\beta_1 X_1 + \beta_2 X_2 + \varepsilon)$, where $X_1 \sim \text{Bernoulli}(0.5)$, $X_2 \sim N(0,1)$, $\beta_1 = 1$ and $\beta_2 = -0.5$. For simplicity, we set $H(y) = y$, that is, no transformation is needed. The error term ε is generated from: (a) the standard normal distribution, (b) the standard logistic distribution, (c) the Type I extreme value distribution, (d) the Type II extreme value distribution, (e) the t distribution with 5 degrees of freedom, (f) the uniform distribution with range from -5 to 5, (g) the standardized beta distribution with parameters $\alpha = 5$ and $\beta = 2$ (standardized by subtracting the mean and then dividing by the standard deviation), and (h) the standardized beta distribution with parameters $\alpha = 2$ and $\beta = 5$. Figures 3.12 and 3.13 show the probability density functions (PDFs) of these error distributions and the extent of violation to the parallel assumption when probit, logit, cloglog, and loglog link functions are used, respectively. The proper link functions for (a) – (d) are probit, logit, cloglog, and loglog, respectively, whereas, there are no proper link functions

for (e) – (h), i.e., no perfect parallelism in Figure 3.13. However, there are some similarities among error distributions (e) – (h) and (a) – (d) in terms of shape and skewness. We are interested in the robustness of cumulative probability models in these settings.

We fitted cumulative probability models with probit, logit, cloglog, and loglog link functions for each scenario with sample sizes of 50, 100, and 200, respectively. For each sample size, we repeated the analysis 10,000 times. With different link functions, the regression coefficients are not in the same scale, and therefore are not directly comparable. Instead, we compared the estimated conditional means and conditional medians for $(X_1 = 0, X_2 = 0)$, $(X_1 = 1, X_2 = 0)$, $(X_1 = 0, X_2 = 1)$, and $(X_1 = 1, X_2 = 1)$. For the purpose of comparison, we also obtained the estimates for these conditional means and medians from linear regression and median regression models. The efficiency of the cumulative probability model on estimating conditional means and medians is compared with the properly specified linear regression and median regression models, respectively. The relative efficiency is measured with the MSE ratio.

Figures 3.14 and 3.15 summarize the performance for estimating conditional means and medians for sample sizes of 100, respectively. More details of these simulation results are reported as supplemental materials in Section 3.6.2. We also conduct simulations for sample sizes of 50 and 200. The results are similar and we report also them in Section 3.6.2. In summary, we find with moderate or large sample size (e.g., $n \geq 50$), the cumulative probability models with properly specified link functions have good performance for estimating conditional means and medians, i.e., the bias is small and the coverage probability of 95% confidence intervals is close to 0.95. It is worth pointing out that cumulative probability models with properly specified link functions seem to be more efficient than median regression, i.e., MSE ratios are generally greater than 1. Cumulative probability models with properly specified link functions may also be more efficient than linear regression when the error distribution is skewed, e.g., error distribution (c) with the cloglog link function and error distribution (d) with the loglog link function. The cumulative probability models

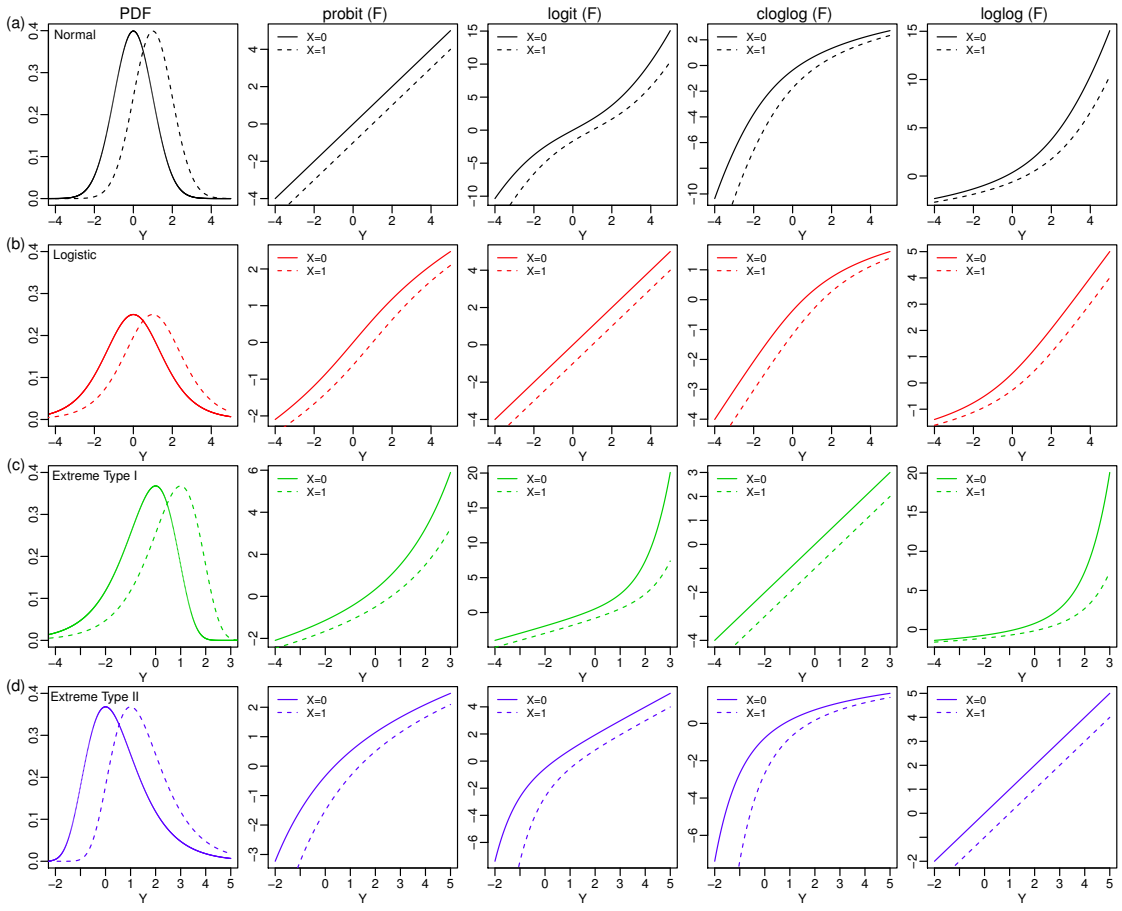

Figure 3.12: The probability density functions of error distributions for (a) – (d) and the extent of violation to the parallel assumption with commonly used link functions.

seem to have reasonable performance under minor or moderate link function misspecification, e.g., error distribution (a) with the logit link function, error distribution (b) with the probit link function, error distribution (e) with the logit or probit link function, error distribution (g) with the cloglog link function, and error distribution (h) with the loglog link function. However, with severe link function misspecification, i.e., the error distributions have totally different shapes or are skewed in opposite directions, the cumulative probability models may have poor performance, e.g., error distributions (a) and (b) with the cloglog or loglog link functions, error distributions (c) and (g) with the loglog link function, and error distributions (d) and (h) with the loglog link function.

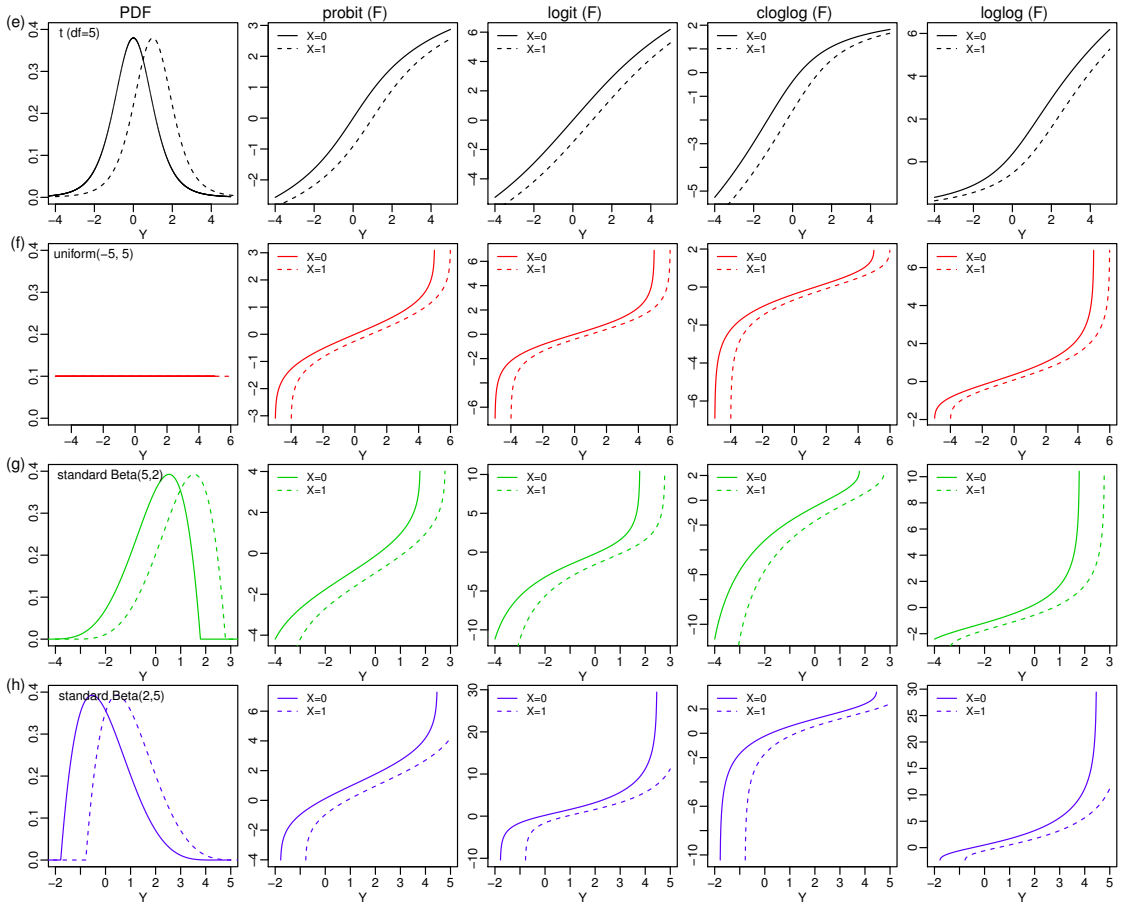

Figure 3.13: The probability density functions of error distributions for (e) – (h) and the extent of violation to the parallel assumption with commonly used link functions.

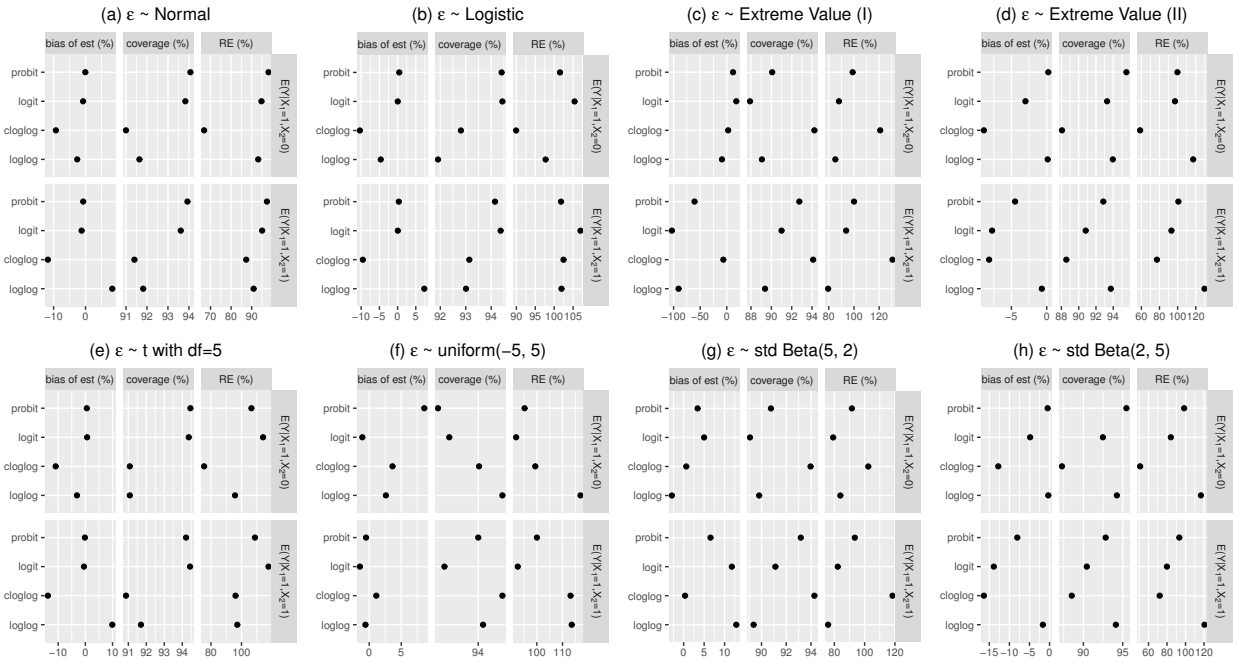

Figure 3.14: The performance of cumulative probability models on estimating conditional means with commonly used link functions link functions. We report the percent bias (%) of the point estimate, the coverage probability of 95% confidence intervals, and the relative efficiency (RE). The percent bias of the point estimate is calculated as the mean of point estimates in 10,000 simulation replicates minus the true value and then divided by the true value. The relative efficiency is compared with properly specified linear regression measured with MSE ratio. The sample size is 100 and the results are based on 10,000 simulation replicates.

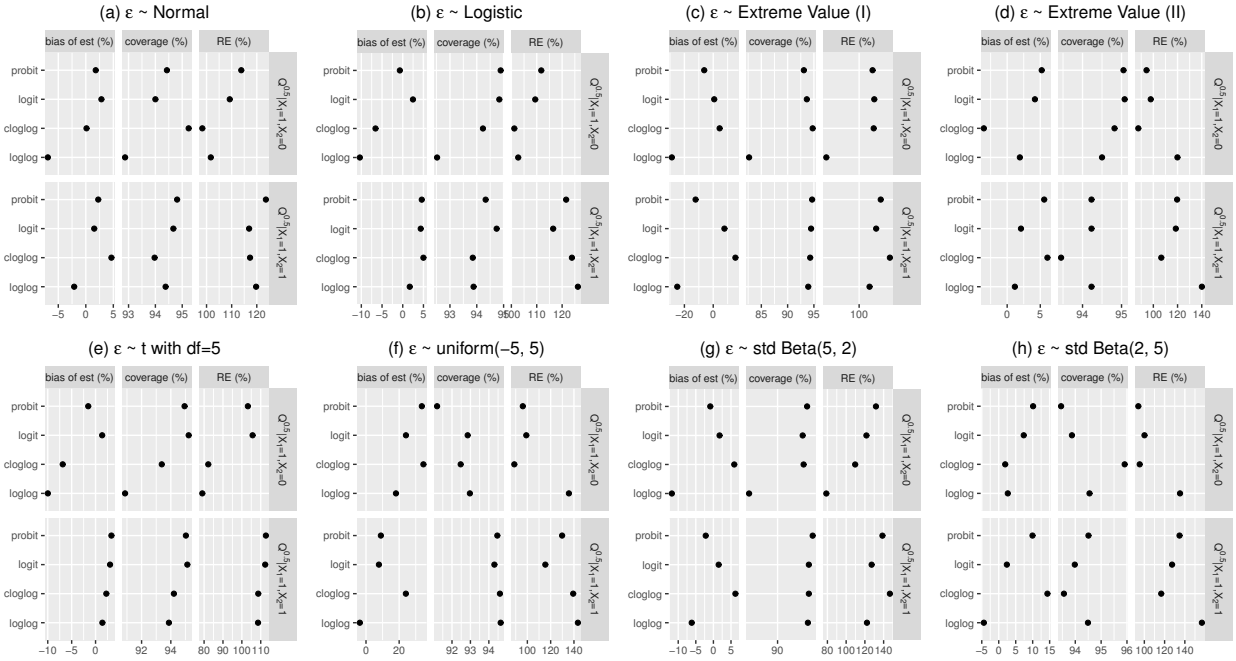

Figure 3.15: The performance of cumulative probability models on estimating medians with commonly used link functions. We report the percent bias (%) of the point estimate, the coverage probability of 95% confidence intervals, and the relative efficiency (RE). The percent bias of the point estimate is calculated as the mean of point estimates in 10,000 simulation replicates minus the true value and then divided by the true value. The relative efficiency is compared with properly specified median regression measured with MSE ratio. The sample size is 100 and the results are based on 10,000 simulation replicates.

3.4 Application Examples

We illustrate the application of cumulative probability models for continuous outcomes using a dataset of 4,776 HIV-infected persons starting antiretroviral therapy (ART) in Latin America (McGowan et al., 2007). We are interested in modeling CD4 count and viral load 6 months after the initiation of ART using patients' demographics and baseline covariates. CD4 count and viral load are important measures of an HIV-infected patient's immune system function and control of the virus. Their distributions are often skewed. When modeling them with linear regression models, transformations are often applied, e.g., square root transformation for CD4 count and log transformation for viral load, although other transformations are sometimes used and different transformations may yield conflicting results. With cumulative probability models, the proper transformation can be estimated semiparametrically. In addition, measurements of viral load are often censored at assay detection limits, especially when patients are on ART. To deal with this issue, common practice is to categorize the viral load (e.g., "undetectable" vs. 'detectable") or to impute particular values for those measurements below the detection limit (e.g., if detection limit is 400 copies/mL, then to record all measurements below the detection limit as 399 copies/mL). However, these strategies either ignore the information of viral load above the detection limit or make assumptions for viral load below the detection limit. We believe that cumulative probability models are particularly useful in these settings.

3.4.1 CD4 Count

We fit cumulative probability models for CD4 count 6 months after ART initiation with the probit, logit, cloglog, and loglog link functions, including age, gender, treatment class, study site, probable infection route, year of ART initiation, baseline nadir CD4, baseline viral load, and baseline AIDS status as covariates. The baseline nadir CD4 was square-root transformed, and then modeled with restricted cubic splines using 5 knots. The baseline

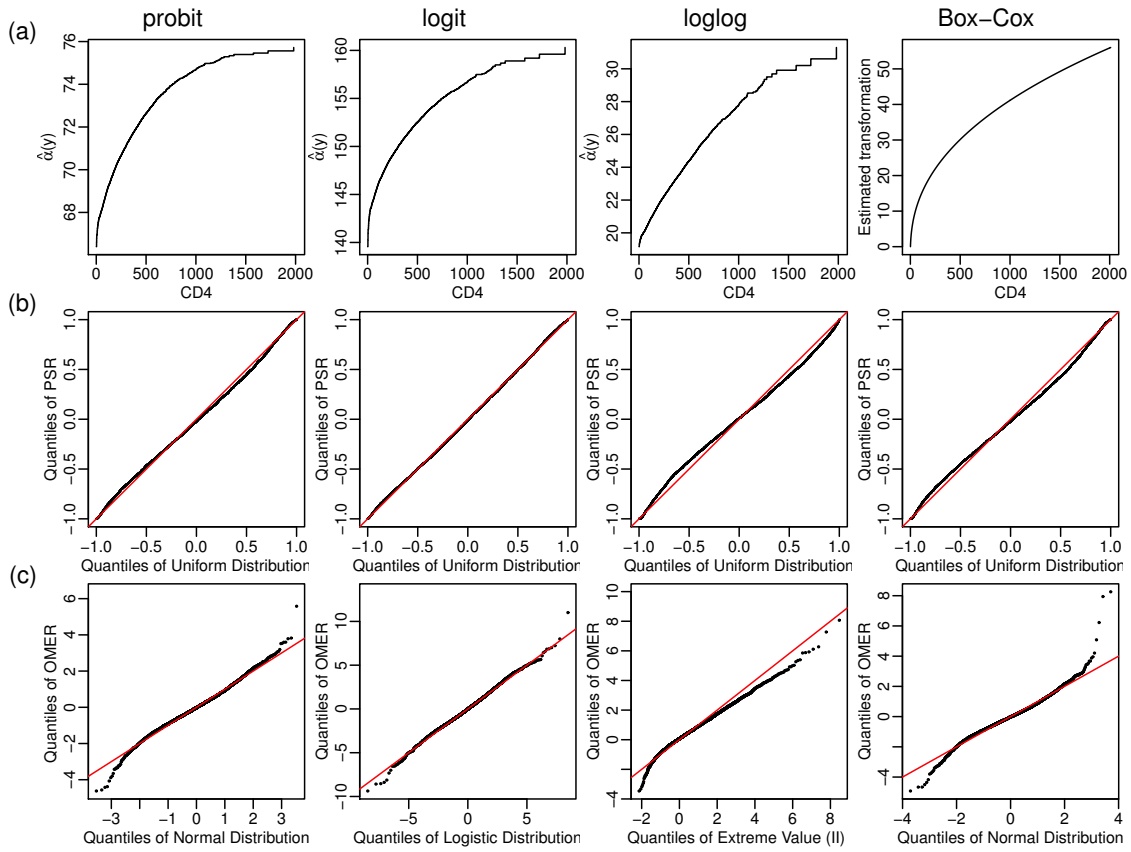

Figure 3.16: (a): the estimated intercepts $\hat{\alpha}(y)$ from the cumulative probability models using the probit, logit, and loglog link functions, which can be interpreted as semiparametric estimates of the best transformation for the 6-month CD4 count. For purpose of comparison, we also plot the estimated Box-Cox transformation, (b): QQ-plots of probability-scale residuals (PSRs). (c): QQ-plots of observed-minus-expected residuals (OMERs), removing the residual for the observation with the largest value of 6-month CD4 count.

viral load was log transformed, and then modeled with restricted cubic splines using 5 knots. All other continuous predictors were transformed with restricted cubic splines using 5 knots directly.

Figure 3.16(a) plots the estimated intercepts $\hat{\alpha}(y)$ resulting from the cumulative probability models, which can be interpreted as semiparametric estimates of the best transformation for the 6-month CD4 count. For purpose of comparison, we also plot the estimated Box-Cox transformation, which was estimated to be $(y^{0.421}-1)/0.421$, close to the commonly used square-root transformation.

The cumulative probability model with the cloglog link function did not converge after

12 iterations, suggesting poor model fit. The log likelihoods for models using the probit, logit, and loglog link functions were -28796.59, -28709.87, and -29067.54, respectively, suggesting better model fit using the symmetric probit and logit link functions. To further assess the goodness-of-fit with different link functions, we computed the probability-scale residuals (PSRs), defined as $P(Y^* < y) - P(Y^* > y)$, where y is the observed value and Y^* is a random variable from the fitted distribution (Li & Shepherd, 2012; Shepherd et al., in press). PSRs of a continuous outcome under the properly specified model are uniformly distributed with range from -1 to 1. Therefore, the QQ-plot of PSRs vs. the uniform distribution can be used to assess the overall model fit (Figure 3.16(b)). We can also assess the model fit using the observed-minus-expected residuals (OMERs) on the transformed scale. Specifically, since the cumulative probability model (3.3) can be interpreted as the semi-parametric transformation model $Y = H(\beta X + \varepsilon)$ and the intercept $\alpha(y) = H^{-1}(y)$, we can compute the OMERs on the transformed scale as $\hat{\varepsilon}_i = \hat{\alpha}(y_i) - \hat{\beta} x_i$. However, as discussed in Section 3.2.1, the NPMLE of $\alpha(y)$ is an unbounded step function with $\hat{\alpha}(y_{max}) = +\infty$; therefore, the OMER for the observation with the largest value of y is also unbounded. Figure 3.16(c) shows the QQ plot of OMERs vs. the error distributions corresponding to the specific link functions, removing the residual for the observation with the largest CD4 count. Both (b) and (c) of Figure 3.16 suggest better model fit using the logit link function. This is consistent with the fact that the cumulative probability model with the logit link function has the highest log likelihood among all link functions considered. The QQ-plots of PSRs and OMERs from the linear regression model with Box-Cox transformation are also shown in Figure 3.16, suggesting after the Box-Cox transformation, the error distribution, although fairly symmetric, is not normally distributed, especially in the tail regions.

PSRs and OMERs can also be used in residual-by-predictor plots to detect the lack of fit for models. For example, in Figure 3.17, we compare the residual-by-predictor plots using both PSRs and OMERs from the cumulative probability models including and not including the baseline nadir CD4 in the model. The smoothed curves show a clear pattern

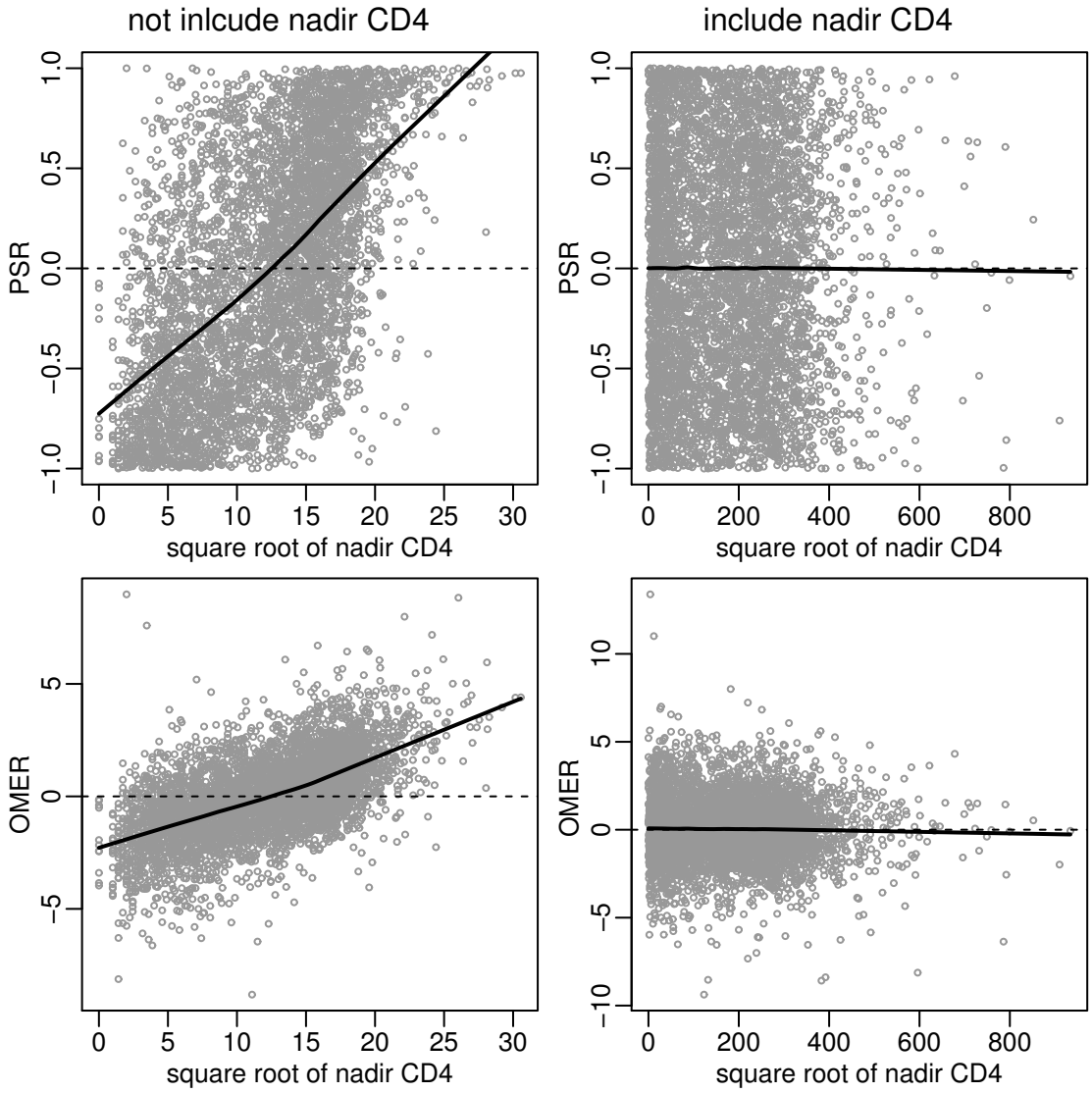

Figure 3.17: Residual-by-predictor plots using PSRs (top panel) and OMERs on the transformed scale (bottom panel) from cumulative probability models (using the logit link function) including and not including baseline nadir CD4 count in the models. Smoothed curves using Friedman's super smoother are added.

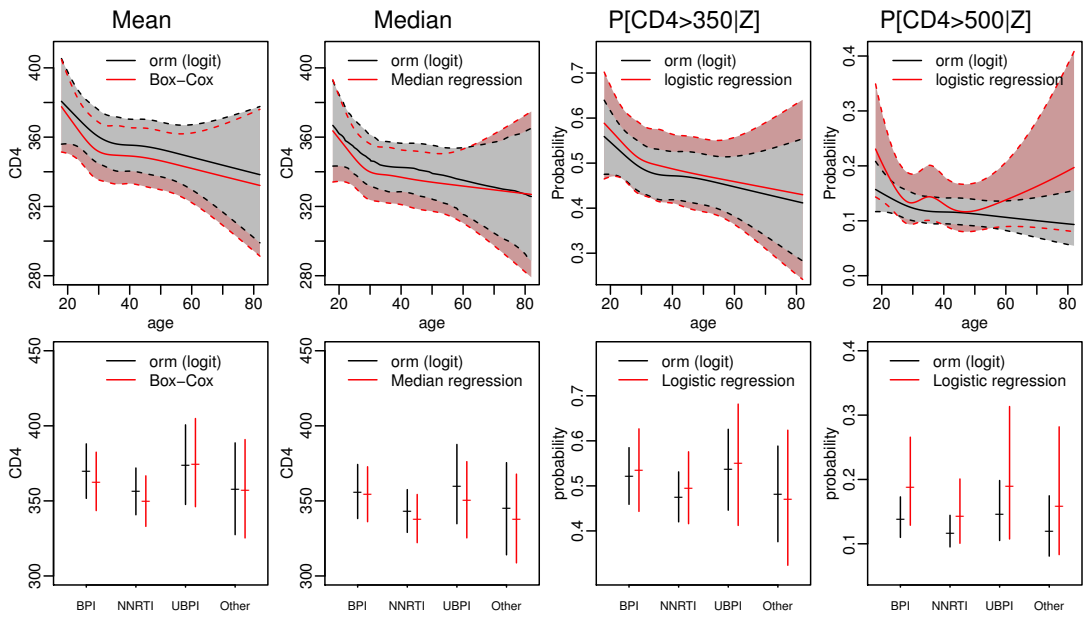

Figure 3.18: The estimated mean, median, and the probabilities of CD4 being greater than 500 cells/uL and CD4 being greater than 350 cells/uL as functions of age (top panel) or treatment class (bottom panel, BPI: boosted protease inhibitors, NNRTI: non-nucleoside reverse transcriptase inhibitors, and UBPI: unboosted protease inhibitors) from the cumulative probability model with the logit link function fixing other predictors at their medians (for continuous variables) or modes (for categorical variables). For purpose of comparison, we also plot the conditional means from linear regression models with Box-Cox transformation, the conditional medians from median regression models, and conditional probabilities from logistic regression models using the dichotomized CD4 count as outcomes. The shaded regions are point-wise 95% confidence intervals.

of a positive relationship between residuals and the baseline nadir CD4 when it is not included (left panel of Figure 3.17). The relationship disappears when the baseline nadir CD4 is included (right panel of Figure 3.17).

As described in Section 3.2.2, one can easily obtain different aspects of conditional distributions using cumulative probability models for continuous outcomes. This is particularly useful when modeling CD4 count. For example, we might want to summarize the central tendency of the conditional distribution with medians instead of means since the distribution is skewed. Besides the central tendency, we may also be interested in the probabilities of CD4 count below or above some commonly used thresholds. Figure 3.18 plots the estimated means, medians, and the probabilities of CD4 count being above 350 or 500 cells/uL as functions of age or treatment class fixing other predictors at their medians (for

continuous variables) or modes (for categorical variables). For purpose of comparison, we also obtain the conditional means through a back transformation from the linear regression model with the Box-Cox transformation, the conditional medians from a median regression model, and the conditional probabilities of CD4 being above 350 or 500 cells/uL from corresponding logistic regression models using dichotomized CD4 count as outcomes. The right sides of these models (i.e., the predictor variables and their transformations) are the same as those in the cumulative probability models.

The estimated conditional means and medians from the different models were generally similar with comparable 95% confidence intervals except that some point estimates for conditional means from the back transformation of linear models were slightly lower than those from the cumulative probability model. This trend is generally consistent with the direction of Jensen's inequality, i.e., $E[g(Y|X)] \leq g[E(Y|X)]$, where $g(y) = (y^{0.421} - 1)/0.421$. The cumulative probability models were more efficient at estimating the conditional probabilities of CD4 count being above 350 or 500 cells/uL than the corresponding logistic regression models as evidenced by their narrow 95% confidence intervals. It is interesting to note that the two models give similar point estimates for the probabilities of CD4 count being above 350 cells/uL but slightly different point estimates for the probabilities of CD4 count being above 500 cells/uL. The estimates from the logistic regression models tend to be larger and perhaps more susceptible to over-fitting given that relatively few patients (16.5%) had CD4 greater than 500 cells/uL after 6 months.

In summary, rather than fitting three separate models, some of which require transforming the outcome, we were able to obtain similar and likely less biased and more efficient estimates by fitting a single cumulative probability model.

3.4.2 Viral Load

In this dataset, 85% of patients had viral load below the detection limit (400 copies/mL) 6 months after ART initiation. For those measurements above the detection limit, their

distribution was highly skewed, ranging from 400 to 7,800,000 copies/mL with a median of 1,300 copies/mL. Due to the large proportion of undetectable viral loads, common practice is to dichotomize the viral load into two categories ("undetectable" and "detectable") and then fit logistic regression models, which ignore the numerical information in the detectable measurements. To make full use of ordinal information of viral load, we fit cumulative probability models for the 6-month viral load measures using the probit, logit, cloglog, and loglog link functions. We included the same covariates with similar transformations as in the CD4 models except that we used 4 knots when transforming the continuous variables using restricted cubic splines because of concerns of over-fitting due to the large proportion of undetectable viral loads.

The log likelihoods for models using the probit, logit, cloglog, and loglog link functions were -5213.69, -5185.26, -5245.03, and -5162.70, respectively, suggesting better model fit with the skewed loglog link function, followed by the symmetric logit and probit link functions, then the cloglog link function which is skewed in the opposite direction as the loglog link function. Figure 3.19 displays the estimated transformations and the QQ-plots of PSRs and OMERs. Note, since the distribution of 6-month viral load is a mixture of discrete and continuous distributions, PSRs are not uniformly distributed, even if the model is properly specified. OMERs on the transformed scale similarly suffer. QQ-plots, in this setting, are not useful for assessing model fit. In this example, although the model with the loglog link function had a slightly higher log likelihood, we also consider the logit probability model for purpose of convenient interpretation, particularly for comparisons with logistic regression models on the dichotomized viral load outcome.

Although PSRs are generally not uniformly distributed for variables with mixed types of discrete and continuous distributions, they have expectation 0 under properly specified models, and therefore can still be used in residual-by-predictor plots (Shepherd et al., in press). Figure 3.20 plots the PSRs vs. age from cumulative probability models including and not including age. When age is not included in the models, PSRs show weak negative

association with age; whereas when age is included in the model, the association disappears, suggesting that age should be included. The results using the loglog and the logit link functions are generally similar.

Figure 3.21 plots the estimated probabilities of 6-month viral load being detectable and being greater than 1,000 copies/uL, and the estimated 95^{th} percentiles as functions of age or treatment class fixing other predictors at their medians or modes using both the loglog and the logit link functions. The results using these two different link functions were very similar. For purpose of comparison, we also obtain the estimated probabilities from logistic regression models using dichotomized viral loads as the outcomes. We also obtain estimates of the 95^{th} percentiles from two separate quantile regression models: one imputing values below the detection limit as the detection limit (i.e., 400 copies/uL) and the other imputing values below the detection limit as 0. Cumulative probability models and logistic regression models gave very similar results when estimating the probabilities of 6-month viral load being detectable: similar point estimates and comparable 95% confidence intervals. Results were more different when estimating the probabilities of 6-month viral load being greater than 1,000 copies/mL: the point estimates from the cumulative probability models were generally smaller with narrower 95% confidence intervals than those from logistic regression models, suggesting that the cumulative probability models incorporated information from all levels of detectable viral loads. The estimates for the 95^{th} percentiles from the quantile regression were very unstable: the point estimates varied with values imputed for undetectable viral load and their 95% confidence intervals were very wide (with negative values for the lower bounds, results not shown). In contrast, cumulative probability models did not require any imputation and they gave sensible point estimates and 95% confidence intervals.

We performed a limited simulation to compare the performance of cumulative probability models, logistic regression models with dichotomized outcomes, and linear regression models with imputed values under various undetectable proportions (see details in Section

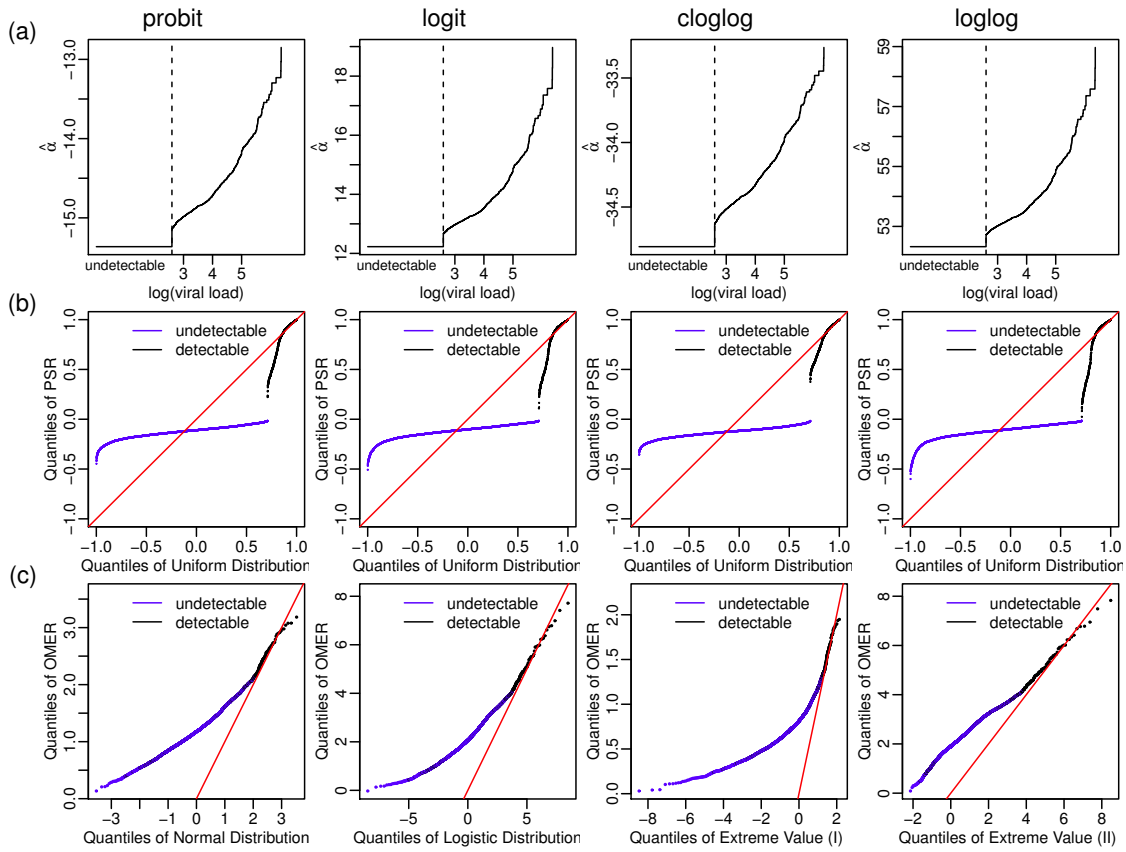

Figure 3.19: (a): the estimated intercepts $\hat{\alpha}(y)$ resulting from the cumulative probability models using the probit, logit, cloglog, and loglog link functions, which can be interpreted as semiparametric estimates of the best transformation for the 6-month viral load. (b): QQ-plots of probability-scale residuals (PSRs). (c): QQ-plots of observed-minus-expected residuals (OMERs), removing the residual for the observation with the largest value of viral load.

3.6.3). We found that cumulative probability models with properly specified link functions were generally more robust than the other two approaches. We also saw that gains in efficiency, particularly when compared with logistic regression models, were minimal when the proportion of undetectable measurements was large, e.g. $\geq 75\%$, as was the case in our actual data example.

3.5 Discussion

In this paper, we have studied cumulative probability models for continuous outcomes. By relating them to semiparametric transformation models, we have shown that the esti-

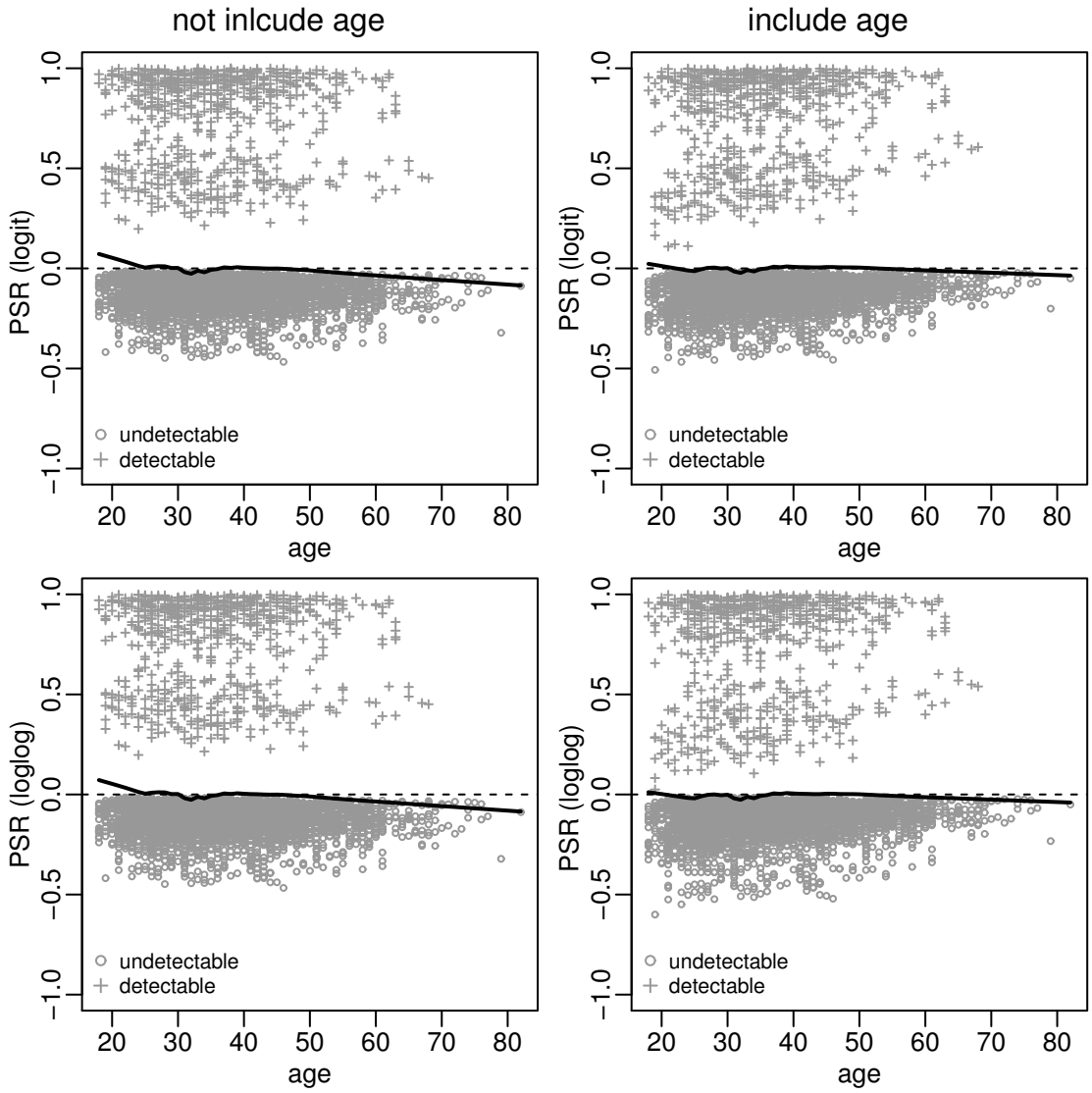

Figure 3.20: Residual-by-predictor plots using PSRs from cumulative probability models using the logit link function (top panel) and the loglog link function (bottom panel). Smoothed curves using Friedman's super smoother are added.

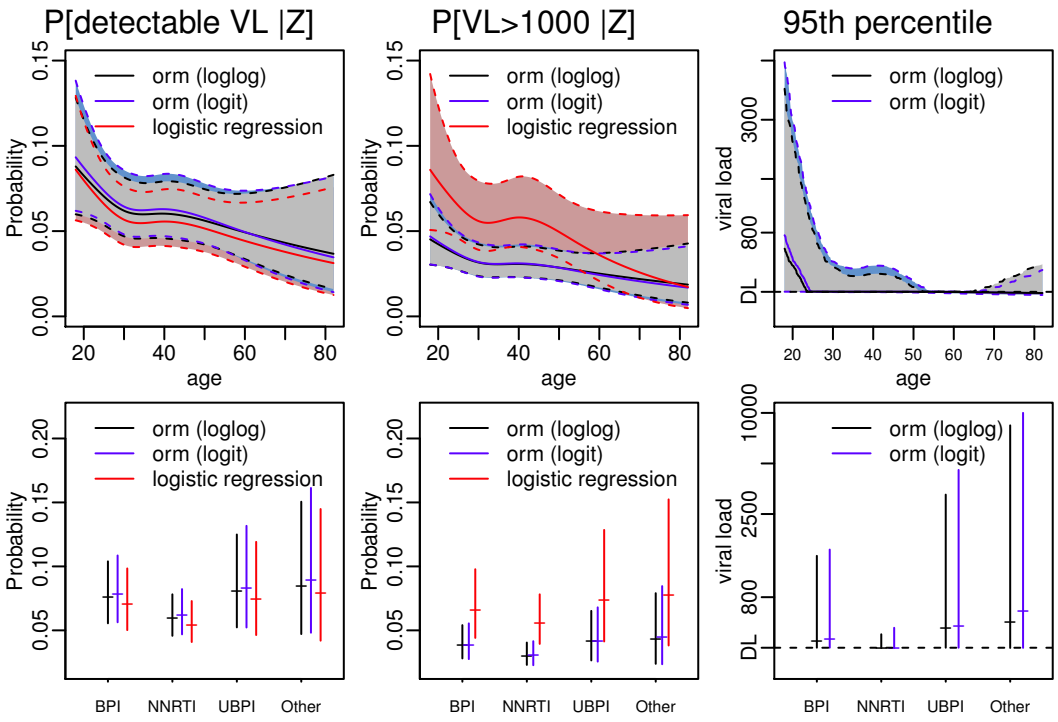

Figure 3.21: The probabilities of 6-month viral load being detectable ($\geq$ 400 copies/mL) and being greater than 1000 copies/mL, and the 95^{th} percentiles as functions of age (top panel) or treatment class (bottom panel), estimated using the cumulative probability model with the loglog and logit link functions, fixing other predictors at their medians (for continuous variables) or modes (for categorical variables). The shaded regions are the point-wise 95% confidence intervals. For purpose of comparison, we also show the estimates of the conditional probabilities from logistic regression models using the dichotomized viral load as outcomes. We also estimated conditional percentiles from quantile regression models by imputing the measurements below the detection limit to be the detection limit or 0. However, since the estimates from quantile regression models were very unstable with very wide 95% confidence intervals crossing 0, we did not plot the results.

mated intercepts can be viewed as the estimated semiparametric transformation. Therefore, these models are particularly useful when flexible transformations are needed for the response variables. Another suitable application of these models is for measurements with detection limits, since they only require outcomes to be orderable, but do not require assigning specific values to those under the detection limit. Various aspects of the conditional distribution can be easily derived from the regression coefficients, giving a full picture of the conditional distribution.

Our simulation studies show that properly specified cumulative probability models have good finite sample performance with moderate or relatively large sample sizes, but that some bias may occur when the sample size is small. In addition, cumulative probability models seem to be fairly robust to minor or moderate link function misspecification according to our simulations. These results are comforting given that the asymptotic properties of the NPMLE of these models have not been formally developed and are quite challenging (Zeng & Lin, 2007; Zeng, Kosorok, & Lin, personal communication). We have focused on the application side of these models and we hope our study provides additional motivation and insight to this challenging theoretical problem.

Using cumulative probability models for continuous variables is motivated from a simple and intuitive idea: continuous variables are also ordinal and can be modeled ordinally. This idea can be naturally extended to other ordinal regression models, such as the continuation-ratio models and adjacent-categories models, and to more complicated settings, such as longitudinal data in which the observations are not independent. We are studying extensions in these settings.

3.6 Supplemental Materials

3.6.1 Performance of Cumulative Probability Models with an Automatic Link Function Selection Procedure

With cumulative probability models, the primary assumptions are made through the choice of link function. The goodness-of-link test proposed by Genter & Farewell (1985) suggests using the log likelihood to discriminate the model fit using the probit, loglog, and cloglog link functions. In a real application when the link function cannot be pre-specified based on preliminary scientific knowledge, one would hope to automate the link function selection. We conducted simulations to investigate the performance of cumulative probability models with an automatic link function selection procedure, in which three cumulative probability models are fitted separately using the probit, loglog, and cloglog links, and the model with the largest log likelihood is selected. We generated data from $Y = H(\beta X + \varepsilon)$, where $X \sim N(0,1)$, $\beta = 0$ under the null hypothesis H_0, and $\beta = 0.1$ under the alternative hypothesis H_1. For simplicity, we set $H(y) = y$, that is, no transformation was needed. The error term ε was generated from: (a) the standard normal distribution, (b) the standard logistic distribution, (c) the Type I extreme value distribution, and (d) the Type II extreme value distribution. The proper link functions in these scenarios are probit, logit, cloglog, and loglog, respectively. Table 3.2 reports the type I error rate and power of cumulative probability models with such a link function selection procedure compared with those with a pre-specified link function. For models with the link function selection procedure, we also report the proportion of times that the chosen link functions were probit, loglog, and cloglog during the 10,000 simulation replicates. Simulations were repeated for sample sizes of 25, 50, 100, 200, 500, and 1000, respectively.

The link function selection based on the log likelihood seems not to perform well in our simulation setting. It generally favors the skewed link functions (cloglog and loglog) under the null no matter what the true error distribution was. Under the alternative hypothesis, it

still had difficulty choosing the proper link function, e.g., the proper link function was only favored with large sample sizes, e.g., $n = 1000$. Cumulative probability models with such a link function selection procedure have inflated type I error rates. It is interesting to note that with moderate or large sample sizes, e.g., $n \geq 50$, the type I error rates are close to 5% for any pre-specified link function even for those link functions that do not correspond to the true error distributions. This is because there is only one covariate in our simulation setting and under the null hypothesis ($\beta = 0$), there is always a transformation $H^{-1}(\cdot)$ such that $H^{-1}(Y)$ has the distribution the link function specifies (Zeng & Lin, 2007). That is, in our simulation setting there is no link function misspecification under the null. However, this is generally not the case when there are multiple covariates in the model.

3.6.2 Details of Simulation Results

We report the detailed simulation results of Section 3.3.1 in Table 3.3 – Table 3.6. These results have been summarized in Figures 3.7 – 3.11.

Here, we report the detailed simulation results of Section 3.3.2. The results for sample size of 100 have been summarized in Figures 3.14 and 3.15. We also conducted simulations for the same setting with the sample sizes of 50 and 200 to investigate the finite sample performance. Tables 3.7 - 3.10 summarize the results for $n = 50$. Tables 3.11 - 3.14 summarize the results for $n = 100$. Tables 3.15 - 3.18 summarize the results for $n = 200$. The results from these additional simulations are generally similar with what we report in Section 3.3.2.

Table 3.2: Type I error rate and power of cumulative probability models with an automatic link function selection procedure (selecting the link function with the highest likelihood among probit, cloglog, and loglog) compared with those with a pre-specified link function. The numbers in the parentheses are the proportions of chosen link function being probit, cloglog, or loglog during 10,000 simulation replicates.

True error distribution		link function selection	probit	cloglog	loglog
Normal					
$n = 25$	H_0	0.118 (9.78%, 45.02%, 45.20%)	0.068	0.068	0.074
	H_1	0.153 (10.86%, 44.67%, 44.47%)	0.096	0.098	0.095
$n = 50$	H_0	0.102 (12.52%, 43.51%, 43.97%)	0.057	0.060	0.060
	H_1	0.176 (13.95%, 42.89%, 43.16%)	0.113	0.106	0.112
$n = 100$	H_0	0.099 (13.20%, 44.04%, 42.76%)	0.054	0.058	0.057
	H_1	0.247 (18.54%, 40.91%, 40.55%)	0.173	0.156	0.161
$n = 200$	H_0	0.096 (14.83%, 43.31%, 41.86%)	0.052	0.053	0.056
	H_1	0.373 (24.16%, 37.31%, 38.53%)	0.294	0.261	0.251
$n = 500$	H_0	0.091 (15.10%, 41.58%, 43.32%)	0.048	0.051	0.055
	H_1	0.678 (37.19%, 31.02%, 31.79%)	0.607	0.522	0.526
$n = 1000$	H_0	0.099 (15.62%, 42.44%, 41.94%)	0.058	0.057	0.055
	H_1	0.910 (50.02%, 24.51%, 25.47%)	0.880	0.808	0.814
Logistic					
$n = 25$	H_0	0.116 (10.15%, 45.87%, 43.98%)	0.066	0.072	0.069
	H_1	0.126 (10.75%, 45.48%, 43.77%)	0.076	0.080	0.078
$n = 50$	H_0	0.106 (12.02%, 43.35%, 44.63%)	0.057	0.064	0.061
	H_1	0.129 (13.37%, 42.53%, 44.10%)	0.077	0.077	0.077
$n = 100$	H_0	0.100 (13.15%, 42.90%, 43.95%)	0.054	0.056	0.057
	H_1	0.142 (15.73%, 41.76%, 42.51%)	0.089	0.083	0.087
$n = 200$	H_0	0.099 (14.78%, 42.44%, 42.78%)	0.054	0.055	0.055
	H_1	0.194 (19.45%, 40.16%, 40.39%)	0.137	0.121	0.119
$n = 500$	H_0	0.091 (14.87%, 42.83%, 42.30%)	0.050	0.050	0.050
	H_1	0.317 (26.45%, 36.87%, 36.68%)	0.244	0.205	0.204
$n = 1000$	H_0	0.095 (15.60%, 42.45%, 41.95%)	0.052	0.054	0.051
	H_1	0.506 (35.05%, 32.52%, 32.43%)	0.428	0.350	0.358
Extreme Type I					
$n = 25$	H_0	0.116 (10.15%, 45.87%, 43.98%)	0.066	0.072	0.069
	H_1	0.147 (10.32%, 47.74%, 41.94%)	0.089	0.100	0.085
$n = 50$	H_0	0.106 (12.02%, 43.35%, 44.63%)	0.057	0.064	0.061
	H_1	0.165 (13.41%, 47.42%, 39.17%)	0.104	0.119	0.088
$n = 100$	H_0	0.100 (13.15%, 42.90%, 43.95%)	0.054	0.056	0.057
	H_1	0.227 (16.23%, 50.30%, 33.47%)	0.152	0.171	0.109
$n = 200$	H_0	0.099 (14.78%, 42.44%, 42.78%)	0.054	0.055	0.055
	H_1	0.352 (19.37%, 55.72%, 24.91%)	0.256	0.296	0.166
$n = 500$	H_0	0.091 (14.87%, 42.83%, 42.30%)	0.050	0.050	0.050
	H_1	0.652 (21.80%, 66.99%, 11.21%)	0.525	0.606	0.312
$n = 1000$	H_0	0.095 (15.60%, 42.45%, 41.95%)	0.052	0.054	0.051
	H_1	0.903 (21.17%, 75.20%, 3.63%)	0.819	0.886	0.534
Extreme Type II					
$n = 25$	H_0	0.116 (10.15%, 45.87%, 43.98%)	0.066	0.072	0.069
	H_1	0.145 (10.72%, 43.29%, 45.99%)	0.088	0.086	0.096
$n = 50$	H_0	0.106 (12.02%, 43.35%, 44.63%)	0.057	0.064	0.061
	H_1	0.166 (13.71%, 38.10%, 48.19%)	0.106	0.088	0.118
$n = 100$	H_0	0.100 (13.15%, 42.90%, 43.95%)	0.054	0.056	0.057
	H_1	0.224 (15.54%, 32.69%, 51.77%)	0.148	0.103	0.174
$n = 200$	H_0	0.099 (14.78%, 42.44%, 42.78%)	0.054	0.055	0.055
	H_1	0.356 (18.58%, 23.89%, 57.53%)	0.254	0.163	0.305
$n = 500$	H_0	0.091 (14.87%, 42.83%, 42.30%)	0.050	0.050	0.050
	H_1	0.652 (21.52%, 11.24%, 67.24%)	0.524	0.311	0.609
$n = 1000$	H_0	0.095 (15.60%, 42.45%, 41.95%)	0.052	0.054	0.051
	H_1	0.900 (21.07%, 3.72%, 75.21%)	0.818	0.534	0.882

Table 3.3: The performance of cumulative probability models on estimating the slopes β_1 and β_2 and the intercepts at $y_1 = 0.368$, $y_2 = 0.719$, $y_3 = 1.649$, $y_4 = 3.781$, and $y_5 = 7.389$ with properly specified link functions. The results are based on 10,000 simulation replicates for each sample size.

			(i) $\varepsilon \sim$ Normal				**(ii) $\varepsilon \sim$ Extreme Type I**			
		true	est	est.se	emp.se	CP	est	est.se	emp.se	CP
$n = 25$										
	β_1	1.00	1.13	0.472	0.550	0.922	1.19	0.515	0.610	0.924
	β_2	-0.50	-0.57	0.242	0.278	0.926	-0.60	0.269	0.314	0.923
	$\alpha(y_1)$	-1.00	-1.07	0.436	0.425	0.969	-1.09	0.490	0.532	0.955
	$\alpha(y_2)$	-0.33	-0.36	0.372	0.410	0.943	-0.34	0.395	0.434	0.939
	$\alpha(y_3)$	0.50	0.56	0.376	0.423	0.940	0.60	0.368	0.435	0.932
	$\alpha(y_4)$	1.33	1.49	0.439	0.504	0.936	1.58	0.454	0.526	0.938
	$\alpha(y_5)$	2.00	2.22	0.533	0.582	0.948	2.27	0.573	0.625	0.952
$n = 50$										
	β_1	1.00	1.06	0.314	0.331	0.941	1.08	0.334	0.361	0.936
	β_2	-0.50	-0.53	0.160	0.171	0.940	-0.54	0.172	0.187	0.935
	$\alpha(y_1)$	-1.00	-1.06	0.298	0.311	0.957	-1.04	0.327	0.343	0.947
	$\alpha(y_2)$	-0.33	-0.35	0.251	0.261	0.949	-0.33	0.266	0.281	0.943
	$\alpha(y_3)$	0.50	0.53	0.251	0.262	0.945	0.55	0.243	0.262	0.938
	$\alpha(y_4)$	1.33	1.40	0.289	0.303	0.945	1.44	0.287	0.311	0.940
	$\alpha(y_5)$	2.00	2.12	0.354	0.375	0.948	2.16	0.379	0.390	0.952
$n = 100$										
	β_1	1.00	1.03	0.217	0.222	0.947	1.04	0.228	0.235	0.944
	β_2	-0.50	-0.52	0.109	0.114	0.943	-0.52	0.116	0.121	0.942
	$\alpha(y_1)$	-1.00	-1.03	0.203	0.210	0.948	-1.02	0.226	0.231	0.948
	$\alpha(y_2)$	-0.33	-0.33	0.174	0.178	0.948	-0.33	0.184	0.188	0.946
	$\alpha(y_3)$	0.50	0.52	0.174	0.176	0.949	0.52	0.167	0.173	0.945
	$\alpha(y_4)$	1.33	1.36	0.200	0.205	0.947	1.38	0.193	0.201	0.943
	$\alpha(y_5)$	2.00	2.05	0.241	0.248	0.947	2.09	0.255	0.264	0.950
$n = 200$										
	β_1	1.00	1.01	0.152	0.155	0.946	1.02	0.158	0.162	0.948
	β_2	-0.50	-0.51	0.076	0.079	0.942	-0.51	0.080	0.081	0.948
	$\alpha(y_1)$	-1.00	-1.01	0.142	0.145	0.951	-1.01	0.158	0.159	0.948
	$\alpha(y_2)$	-0.33	-0.33	0.122	0.123	0.950	-0.33	0.129	0.131	0.947
	$\alpha(y_3)$	0.50	0.51	0.122	0.125	0.947	0.51	0.117	0.118	0.950
	$\alpha(y_4)$	1.33	1.35	0.140	0.143	0.948	1.35	0.133	0.136	0.946
	$\alpha(y_5)$	2.00	2.03	0.168	0.172	0.951	2.04	0.174	0.178	0.947
$n = 500$										
	β_1	1.00	1.01	0.095	0.097	0.946	1.01	0.099	0.100	0.944
	β_2	-0.50	-0.50	0.048	0.047	0.951	-0.51	0.050	0.050	0.947
	$\alpha(y_1)$	-1.00	-1.01	0.089	0.089	0.949	-1.00	0.099	0.100	0.950
	$\alpha(y_2)$	-0.33	-0.33	0.077	0.077	0.950	-0.33	0.081	0.082	0.947
	$\alpha(y_3)$	0.50	0.50	0.077	0.078	0.946	0.50	0.073	0.074	0.944
	$\alpha(y_4)$	1.33	1.34	0.088	0.089	0.949	1.34	0.083	0.085	0.947
	$\alpha(y_5)$	2.00	2.01	0.105	0.106	0.950	2.02	0.108	0.110	0.944
$n = 1000$										
	β_1	1.00	1.00	0.067	0.067	0.951	1.00	0.070	0.070	0.947
	β_2	-0.50	-0.50	0.034	0.034	0.948	-0.50	0.035	0.035	0.950
	$\alpha(y_1)$	-1.00	-1.00	0.063	0.063	0.949	-1.00	0.070	0.070	0.951
	$\alpha(y_2)$	-0.33	-0.33	0.054	0.054	0.953	-0.33	0.057	0.057	0.951
	$\alpha(y_3)$	0.50	0.50	0.054	0.054	0.951	0.50	0.052	0.052	0.952
	$\alpha(y_4)$	1.33	1.33	0.062	0.062	0.950	1.33	0.059	0.059	0.949
	$\alpha(y_5)$	2.00	2.00	0.074	0.075	0.949	2.01	0.076	0.076	0.948

est is the mean of the point estimates.

est.se is the mean of the standard error estimates.

emp.se is the standard deviation of the point estimates

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

Table 3.4: The performance of cumulative probability models on estimating the conditional CDFs evaluated at $y_1 = 0.368$, $y_2 = 0.719$, $y_3 = 1.649$, $y_4 = 3.781$, and $y_5 = 7.389$. The results are based on 10,000 simulation replicates for each sample size.

	(i) $\varepsilon \sim$ Normal					**(ii) $\varepsilon \sim$ Extreme Type I**				
	true	est	est.se	emp.se	CP	true	est	est.se	emp.se	CP
$n = 25$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0701	0.0680	0.0620	0.953	0.2000	0.1949	0.1074	0.0936	0.939
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2051	0.1275	0.1134	0.937	0.3534	0.3501	0.1492	0.1296	0.932
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.4997	0.1694	0.1525	0.935	0.6321	0.6326	0.1710	0.1485	0.930
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7964	0.1264	0.1119	0.935	0.8991	0.8973	0.1006	0.0860	0.938
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9304	0.0693	0.0602	0.950	0.9887	0.9825	0.0350	0.0320	0.962
$n = 50$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0679	0.0455	0.0423	0.947	0.2000	0.1980	0.0696	0.0664	0.947
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2036	0.0851	0.0817	0.945	0.3534	0.3515	0.0954	0.0914	0.944
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.5024	0.1148	0.1096	0.946	0.6321	0.6341	0.1115	0.1063	0.945
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7980	0.0849	0.0807	0.944	0.8991	0.8994	0.0676	0.0629	0.942
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9326	0.0453	0.0417	0.945	0.9887	0.9855	0.0203	0.0180	0.957
$n = 100$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0679	0.0315	0.0304	0.947	0.2000	0.1988	0.0477	0.0467	0.950
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2044	0.0598	0.0583	0.944	0.3534	0.3520	0.0658	0.0640	0.945
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.5018	0.0797	0.0778	0.947	0.6321	0.6323	0.0772	0.0747	0.944
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7976	0.0593	0.0575	0.945	0.8991	0.8991	0.0471	0.0455	0.941
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9328	0.0311	0.0300	0.947	0.9887	0.9872	0.0128	0.0118	0.950
$n = 200$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0672	0.0219	0.0216	0.948	0.2000	0.1994	0.0331	0.0330	0.950
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2034	0.0416	0.0412	0.950	0.3534	0.3529	0.0455	0.0450	0.952
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.5004	0.0555	0.0550	0.948	0.6321	0.6326	0.0535	0.0525	0.946
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7972	0.0415	0.0409	0.949	0.8991	0.8997	0.0328	0.0324	0.949
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9330	0.0220	0.0214	0.947	0.9887	0.9880	0.0086	0.0082	0.948
$n = 500$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0667	0.0139	0.0136	0.950	0.2000	0.1998	0.0210	0.0208	0.948
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2029	0.0266	0.0261	0.945	0.3534	0.3533	0.0287	0.0284	0.950
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.4997	0.0354	0.0348	0.949	0.6321	0.6324	0.0335	0.0331	0.946
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7964	0.0262	0.0259	0.949	0.8991	0.8995	0.0208	0.0206	0.948
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9332	0.0135	0.0136	0.950	0.9887	0.9885	0.0053	0.0052	0.947
$n = 1000$										
$F(y_1\|X_1 = 1, X_2 = 1)$	0.0668	0.0669	0.0097	0.0097	0.949	0.2000	0.1998	0.0149	0.0147	0.948
$F(y_2\|X_1 = 1, X_2 = 1)$	0.2033	0.2035	0.0184	0.0185	0.946	0.3534	0.3534	0.0202	0.0201	0.949
$F(y_3\|X_1 = 1, X_2 = 1)$	0.5000	0.5002	0.0243	0.0246	0.952	0.6321	0.6321	0.0236	0.0234	0.947
$F(y_4\|X_1 = 1, X_2 = 1)$	0.7967	0.7969	0.0182	0.0183	0.953	0.8991	0.8991	0.0149	0.0146	0.945
$F(y_5\|X_1 = 1, X_2 = 1)$	0.9332	0.9331	0.0097	0.0096	0.948	0.9887	0.9885	0.0038	0.0037	0.950

est is the mean of the point estimates.

est.se is the mean of the standard error estimates.

emp.se is the standard deviation of the point estimates

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

Table 3.5: The performance of cumulative probability models on estimating conditional means. The results are based on 10,000 simulation replicates for each sample size.

	(i) $\varepsilon \sim$ **Normal**					(ii) $\varepsilon \sim$ **Extreme Type I**				
	true	est	est.se	emp.se	CP	true	est	est.se	emp.se	CP
$n = 25$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.668	0.531	0.589	0.874	1.00	0.995	0.285	0.311	0.877
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.480	1.358	1.675	0.832	2.72	2.746	0.758	0.856	0.870
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.034	0.393	0.450	0.863	0.61	0.603	0.219	0.245	0.860
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.808	1.043	1.239	0.866	1.65	1.704	0.603	0.691	0.865
$n = 50$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.657	0.377	0.390	0.914	1.00	0.996	0.201	0.211	0.913
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.489	1.020	1.118	0.882	2.72	2.730	0.549	0.575	0.909
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.010	0.271	0.285	0.905	0.61	0.602	0.153	0.160	0.906
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.756	0.740	0.785	0.906	1.65	1.670	0.421	0.441	0.912
$n = 100$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.652	0.265	0.269	0.930	1.00	0.997	0.142	0.145	0.932
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.503	0.753	0.782	0.918	2.72	2.732	0.391	0.396	0.928
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.003	0.190	0.193	0.925	0.61	0.603	0.107	0.108	0.930
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.743	0.524	0.538	0.928	1.65	1.661	0.294	0.301	0.928
$n = 200$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.653	0.187	0.190	0.939	1.00	1.000	0.100	0.101	0.940
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.497	0.535	0.546	0.932	2.72	2.724	0.276	0.279	0.942
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.002	0.134	0.137	0.938	0.61	0.604	0.075	0.076	0.939
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.732	0.366	0.372	0.940	1.65	1.651	0.205	0.208	0.937
$n = 500$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.649	0.117	0.119	0.945	1.00	0.999	0.063	0.064	0.944
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.496	0.341	0.340	0.946	2.72	2.721	0.174	0.176	0.946
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.001	0.084	0.084	0.948	0.61	0.605	0.048	0.048	0.945
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.728	0.230	0.231	0.946	1.65	1.649	0.129	0.131	0.943
$n = 1000$										
$E(Y\|X_1 = 0, X_2 = 0)$	1.65	1.650	0.083	0.084	0.949	1.00	1.000	0.045	0.045	0.951
$E(Y\|X_1 = 1, X_2 = 0)$	4.48	4.488	0.241	0.242	0.949	2.72	2.721	0.123	0.125	0.946
$E(Y\|X_1 = 0, X_2 = 1)$	1.00	1.001	0.059	0.060	0.951	0.61	0.606	0.034	0.034	0.950
$E(Y\|X_1 = 1, X_2 = 1)$	2.72	2.722	0.162	0.162	0.950	1.65	1.650	0.091	0.092	0.943

est is the mean of the point estimates.

est.se is the mean of the standard error estimates.

emp.se is the standard deviation of the point estimates

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

Table 3.6: The performance of cumulative probability models on estimating the conditional 10^{th},25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles. The results are based on 10,000 simulation replicates for each sample size.

		(i) $\varepsilon \sim$ Normal				**(ii) $\varepsilon \sim$ Extreme Type I**			
		true	est	emp.se	CP	true	est	emp.se	CP
$n = 25$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.506	0.2657	0.931	0.174	0.214	0.1639	0.914
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.901	0.4165	0.942	0.474	0.546	0.3156	0.938
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.688	0.7428	0.935	1.143	1.199	0.5686	0.925
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.146	1.4068	0.930	2.286	2.187	0.9519	0.921
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.469	2.6814	0.909	3.796	3.385	1.4476	0.903
$n = 50$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.477	0.1623	0.947	0.174	0.187	0.0904	0.954
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.862	0.2602	0.945	0.474	0.499	0.1872	0.947
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.657	0.4863	0.946	1.143	1.161	0.3615	0.944
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.170	0.9467	0.942	2.286	2.236	0.6473	0.939
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.670	1.8859	0.938	3.796	3.575	1.0389	0.931
$n = 100$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.465	0.1083	0.951	0.174	0.179	0.0590	0.954
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.847	0.1773	0.945	0.474	0.486	0.1236	0.946
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.648	0.3305	0.946	1.143	1.154	0.2509	0.942
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.202	0.6600	0.944	2.286	2.264	0.4546	0.940
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.796	1.3193	0.943	3.796	3.691	0.7401	0.940
$n = 200$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.462	0.0738	0.951	0.174	0.176	0.0409	0.952
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.845	0.1221	0.950	0.474	0.479	0.0839	0.950
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.650	0.2311	0.947	1.143	1.147	0.1725	0.947
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.221	0.4636	0.946	2.286	2.269	0.3215	0.946
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.877	0.9432	0.946	3.796	3.735	0.5288	0.944
$n = 500$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.460	0.0469	0.948	0.174	0.175	0.0257	0.951
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.843	0.0783	0.944	0.474	0.476	0.0529	0.950
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.651	0.1465	0.949	1.143	1.144	0.1081	0.950
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.235	0.2942	0.947	2.286	2.280	0.2022	0.948
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.911	0.5867	0.952	3.796	3.770	0.3385	0.948
$n = 1000$									
	$Q^{0.10}\|X_1 = 1, X_2 = 1$	0.458	0.458	0.0325	0.950	0.174	0.174	0.0181	0.948
	$Q^{0.25}\|X_1 = 1, X_2 = 1$	0.840	0.840	0.0536	0.951	0.474	0.475	0.0371	0.947
	$Q^{0.50}\|X_1 = 1, X_2 = 1$	1.649	1.648	0.1009	0.953	1.143	1.144	0.0759	0.949
	$Q^{0.75}\|X_1 = 1, X_2 = 1$	3.236	3.230	0.2037	0.951	2.286	2.283	0.1432	0.947
	$Q^{0.90}\|X_1 = 1, X_2 = 1$	5.939	5.927	0.4167	0.951	3.796	3.784	0.2420	0.946

est is the mean of the point estimates.

emp.se is the standard deviation of the point estimates

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

Table 3.7: The performance of cumulative probability models on estimating conditional means with the sample size of 50. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.938	0.979	0.01	0.931	0.950	0.02	0.910	0.940	0.08	0.928	0.808
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.934	0.978	0.99	0.930	0.956	0.92	0.924	0.810	0.98	0.908	0.935
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.50	0.930	0.967	-0.48	0.924	0.925	-0.38	0.857	0.792	-0.42	0.918	0.743
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.934	0.962	0.49	0.928	0.944	0.44	0.908	0.898	0.53	0.917	0.915
(b) $\varepsilon \sim$ Logistic													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.937	1.012	0.00	0.938	1.043	0.04	0.912	0.989	0.09	0.930	0.962
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.936	1.014	1.00	0.935	1.049	0.90	0.932	0.969	0.96	0.909	0.985
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.51	0.938	0.982	-0.50	0.936	1.017	-0.36	0.887	0.935	-0.40	0.928	0.853
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.938	1.013	0.50	0.939	1.055	0.45	0.922	1.030	0.52	0.922	0.997
(c) $\varepsilon \sim$ Extreme Type I													
$E(Y\|X_1=0,X_2=0)$	-0.58	-0.58	0.950	0.997	-0.53	0.933	1.000	-0.58	0.931	1.155	-0.45	0.901	0.718
$E(Y\|X_1=1,X_2=0)$	0.42	0.47	0.896	1.032	0.49	0.880	0.967	0.44	0.929	1.187	0.39	0.880	0.870
$E(Y\|X_1=0,X_2=1)$	-1.08	-1.18	0.953	0.744	-1.11	0.948	0.857	-1.10	0.930	1.060	-0.98	0.909	0.578
$E(Y\|X_1=1,X_2=1)$	-0.08	-0.04	0.922	1.030	0.00	0.908	1.009	-0.07	0.933	1.282	-0.02	0.890	0.826
(d) $\varepsilon \sim$ Extreme Type II													
$E(Y\|X_1=0,X_2=0)$	0.58	0.53	0.893	1.022	0.50	0.878	0.949	0.61	0.879	0.871	0.56	0.926	1.172
$E(Y\|X_1=1,X_2=0)$	1.58	1.58	0.945	0.997	1.53	0.930	0.996	1.44	0.899	0.720	1.58	0.926	1.149
$E(Y\|X_1=0,X_2=1)$	0.08	0.07	0.896	1.193	0.06	0.891	1.122	0.27	0.817	0.725	0.06	0.923	1.273
$E(Y\|X_1=1,X_2=1)$	1.08	1.03	0.925	1.014	1.00	0.908	0.979	0.99	0.881	0.816	1.06	0.929	1.225

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.8: The performance of cumulative probability models on estimating conditional means with the sample size of 50 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$E(Y\|X_1=0,X_2=0)$	0	-0.01	0.937	1.054	-0.01	0.938	1.123	0.02	0.905	0.979	0.09	0.925	0.902
$E(Y\|X_1=1,X_2=0)$	1	1.01	0.942	1.058	1.01	0.940	1.133	0.91	0.931	0.919	0.98	0.905	0.973
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.52	0.938	0.935	-0.52	0.939	1.024	-0.37	0.855	0.828	-0.43	0.921	0.720
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.936	1.071	0.50	0.936	1.145	0.44	0.910	0.994	0.54	0.911	0.991
(f) $\varepsilon \sim$ uniform(-5, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.08	0.927	0.962	0.00	0.924	0.913	-0.03	0.928	1.132	-0.04	0.926	0.964
$E(Y\|X_1=1,X_2=0)$	1	1.07	0.926	0.959	0.99	0.925	0.910	1.03	0.924	0.958	1.02	0.929	1.132
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.65	0.912	0.960	-0.49	0.917	0.913	-0.54	0.925	1.200	-0.55	0.920	0.974
$E(Y\|X_1=1,X_2=1)$	0.5	0.49	0.930	0.984	0.48	0.926	0.920	0.50	0.931	1.073	0.49	0.928	1.071
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.948	0.968	0.04	0.932	0.888	0.00	0.934	1.133	0.11	0.910	0.674
$E(Y\|X_1=1,X_2=0)$	1	1.03	0.905	0.940	1.05	0.894	0.841	1.01	0.932	1.023	0.97	0.891	0.842
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.55	0.943	0.840	-0.49	0.929	0.836	-0.50	0.926	1.080	-0.39	0.905	0.597
$E(Y\|X_1=1,X_2=1)$	0.5	0.53	0.928	0.945	0.55	0.914	0.864	0.50	0.934	1.148	0.55	0.902	0.789
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.03	0.912	0.943	-0.04	0.895	0.838	0.03	0.893	0.848	-0.01	0.930	1.010
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.946	0.973	0.96	0.929	0.887	0.89	0.907	0.662	1.00	0.931	1.127
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.49	0.901	1.069	-0.49	0.894	0.942	-0.33	0.808	0.654	-0.50	0.922	1.098
$E(Y\|X_1=1,X_2=1)$	0.5	0.46	0.929	0.946	0.44	0.910	0.858	0.42	0.890	0.769	0.49	0.934	1.147

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.9: The performance of cumulative probability models on estimating conditional medians with the sample size of 50. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.03	0.945	1.148	0.02	0.946	1.140	0.11	0.923	0.957	0.03	0.948	0.996
$Q^{0.5}\|X_1=1,X_2=0$	1	1.04	0.943	1.095	1.04	0.943	1.069	1.04	0.953	0.940	0.95	0.935	1.089
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.47	0.945	1.265	-0.48	0.947	1.255	-0.30	0.878	0.853	-0.49	0.947	0.976
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.53	0.946	1.214	0.52	0.946	1.158	0.55	0.936	1.123	0.49	0.938	1.157
(b) $\varepsilon \sim$ Logistic													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.07	0.942	1.088	0.04	0.946	1.095	0.17	0.919	0.972	0.11	0.932	0.969
$Q^{0.5}\|X_1=1,X_2=0$	1	1.02	0.948	1.115	1.04	0.948	1.088	0.98	0.948	1.046	0.92	0.937	1.110
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.40	0.942	1.171	-0.45	0.948	1.201	-0.21	0.882	0.924	-0.36	0.931	0.935
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.54	0.945	1.226	0.54	0.948	1.172	0.56	0.942	1.233	0.51	0.940	1.222
(c) $\varepsilon \sim$ Extreme Type I													
$Q^{0.5}\|X_1=0,X_2=0$	-0.37	-0.39	0.956	1.090	-0.37	0.956	1.095	-0.34	0.946	1.251	-0.30	0.941	0.866
$Q^{0.5}\|X_1=1,X_2=0$	0.63	0.61	0.933	1.207	0.65	0.936	1.183	0.68	0.947	1.168	0.47	0.894	0.832
$Q^{0.5}\|X_1=0,X_2=1$	-0.87	-0.96	0.956	0.939	-0.95	0.954	0.972	-0.86	0.947	1.300	-0.79	0.939	0.689
$Q^{0.5}\|X_1=1,X_2=1$	0.13	0.12	0.948	1.282	0.15	0.948	1.223	0.17	0.946	1.389	0.09	0.937	1.097
(d) $\varepsilon \sim$ Extreme Type II													
$Q^{0.5}\|X_1=0,X_2=0$	0.37	0.46	0.921	1.108	0.42	0.931	1.159	0.60	0.854	0.681	0.38	0.945	1.222
$Q^{0.5}\|X_1=1,X_2=0$	1.37	1.46	0.950	0.941	1.45	0.950	0.967	1.37	0.954	0.877	1.41	0.939	1.163
$Q^{0.5}\|X_1=0,X_2=1$	-0.13	0.02	0.892	1.166	-0.02	0.918	1.276	0.27	0.692	0.504	-0.11	0.941	1.434
$Q^{0.5}\|X_1=1,X_2=1$	0.87	0.94	0.944	1.195	0.91	0.946	1.182	0.95	0.938	1.044	0.89	0.944	1.350

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

Table 3.10: The performance of cumulative probability models on estimating conditional medians with the sample size of 50 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.06	0.939	0.985	0.04	0.946	1.050	0.15	0.909	0.793	0.09	0.925	0.826
$Q^{0.5}\|X_1=1,X_2=0$	1	1.01	0.949	1.034	1.03	0.953	1.044	0.98	0.950	0.904	0.92	0.932	0.931
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.41	0.939	1.002	-0.45	0.950	1.107	-0.23	0.848	0.639	-0.39	0.928	0.751
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.53	0.948	1.134	0.53	0.950	1.124	0.54	0.939	1.056	0.50	0.936	1.096
(f) $\varepsilon \sim$ uniform(-5, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	0	-0.18	0.938	1.124	-0.09	0.942	1.065	-0.02	0.942	1.372	-0.21	0.940	1.017
$Q^{0.5}\|X_1=1,X_2=0$	1	1.36	0.923	1.006	1.28	0.933	0.996	1.39	0.936	0.902	1.20	0.930	1.307
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.91	0.919	1.204	-0.75	0.934	1.177	-0.64	0.936	1.722	-0.92	0.931	1.032
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.58	0.946	1.253	0.57	0.945	1.134	0.68	0.948	1.305	0.51	0.944	1.329
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$Q^{0.5}\|X_1=0,X_2=0$	0.13	0.09	0.952	1.146	0.11	0.951	1.137	0.15	0.945	1.357	0.14	0.951	0.964
$Q^{0.5}\|X_1=1,X_2=0$	1.13	1.14	0.938	1.317	1.16	0.932	1.214	1.21	0.942	1.144	1.01	0.916	1.002
$Q^{0.5}\|X_1=0,X_2=1$	-0.37	-0.49	0.942	0.994	-0.46	0.947	1.038	-0.37	0.952	1.464	-0.41	0.956	0.794
$Q^{0.5}\|X_1=1,X_2=1$	0.63	0.63	0.945	1.367	0.65	0.943	1.250	0.69	0.946	1.435	0.59	0.938	1.154
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	-0.13	-0.07	0.927	1.251	-0.09	0.933	1.234	0.05	0.876	0.801	-0.14	0.945	1.307
$Q^{0.5}\|X_1=1,X_2=0$	0.87	0.98	0.940	0.923	0.96	0.944	0.963	0.93	0.958	0.850	0.92	0.937	1.241
$Q^{0.5}\|X_1=0,X_2=1$	-0.63	-0.52	0.904	1.397	-0.55	0.917	1.392	-0.31	0.750	0.619	-0.64	0.935	1.592
$Q^{0.5}\|X_1=1,X_2=1$	0.37	0.43	0.940	1.294	0.40	0.940	1.237	0.45	0.935	1.083	0.37	0.943	1.500

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

Table 3.11: The performance of cumulative probability models on estimating conditional means with the sample size of 100. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.941	0.982	0.00	0.940	0.955	0.02	0.918	0.930	0.09	0.914	0.685
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.941	0.982	0.99	0.938	0.948	0.91	0.910	0.669	0.97	0.916	0.932
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.50	0.944	0.973	-0.48	0.935	0.926	-0.37	0.818	0.626	-0.41	0.918	0.654
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.939	0.974	0.49	0.936	0.951	0.44	0.914	0.873	0.54	0.918	0.909
(b) $\varepsilon \sim$ Logistic													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.945	1.016	0.00	0.945	1.054	0.05	0.918	0.987	0.11	0.926	0.886
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.944	1.015	1.00	0.944	1.054	0.90	0.928	0.899	0.95	0.919	0.977
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.51	0.945	0.984	-0.50	0.944	1.026	-0.34	0.869	0.821	-0.38	0.926	0.777
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.942	1.018	0.50	0.944	1.070	0.45	0.931	1.025	0.54	0.930	1.019
(c) $\varepsilon \sim$ Extreme Type I													
$E(Y\|X_1=0,X_2=0)$	-0.58	-0.58	0.955	0.999	-0.53	0.932	0.969	-0.58	0.942	1.177	-0.43	0.884	0.576
$E(Y\|X_1=1,X_2=0)$	0.42	0.47	0.901	0.989	0.50	0.879	0.881	0.43	0.942	1.208	0.38	0.891	0.852
$E(Y\|X_1=0,X_2=1)$	-1.08	-1.19	0.958	0.665	-1.11	0.955	0.860	-1.09	0.943	1.089	-0.96	0.912	0.507
$E(Y\|X_1=1,X_2=1)$	-0.08	-0.03	0.928	1.000	0.00	0.910	0.938	-0.07	0.941	1.306	-0.01	0.894	0.795
(d) $\varepsilon \sim$ Extreme Type II													
$E(Y\|X_1=0,X_2=0)$	0.58	0.53	0.896	0.990	0.50	0.878	0.884	0.61	0.892	0.864	0.57	0.940	1.189
$E(Y\|X_1=1,X_2=0)$	1.58	1.58	0.955	0.997	1.53	0.933	0.971	1.44	0.881	0.586	1.58	0.940	1.171
$E(Y\|X_1=0,X_2=1)$	0.08	0.07	0.910	1.194	0.05	0.903	1.106	0.28	0.739	0.526	0.07	0.937	1.264
$E(Y\|X_1=1,X_2=1)$	1.08	1.03	0.928	1.008	0.99	0.908	0.931	0.99	0.886	0.772	1.07	0.937	1.297

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.12: The performance of cumulative probability models on estimating conditional means with the sample size of 100 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$E(Y\|X_1=0,X_2=0)$	0	-0.01	0.942	1.060	-0.01	0.942	1.128	0.03	0.905	0.957	0.10	0.913	0.781
$E(Y\|X_1=1,X_2=0)$	1	1.01	0.944	1.065	1.01	0.944	1.141	0.89	0.911	0.755	0.97	0.911	0.957
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.53	0.945	0.918	-0.53	0.947	1.028	-0.35	0.806	0.676	-0.43	0.925	0.656
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.942	1.088	0.50	0.944	1.177	0.43	0.909	0.961	0.55	0.917	0.973
(f) $\varepsilon \sim$ uniform(-5, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.08	0.935	0.957	0.01	0.938	0.922	-0.03	0.941	1.183	-0.03	0.939	0.994
$E(Y\|X_1=1,X_2=0)$	1	1.09	0.935	0.952	0.99	0.936	0.919	1.04	0.940	0.994	1.03	0.943	1.170
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.66	0.922	0.912	-0.47	0.928	0.910	-0.54	0.940	1.255	-0.56	0.929	0.984
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.940	1.000	0.49	0.936	0.926	0.51	0.943	1.131	0.50	0.941	1.136
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.954	0.984	0.04	0.929	0.849	0.00	0.939	1.158	0.12	0.877	0.521
$E(Y\|X_1=1,X_2=0)$	1	1.03	0.908	0.916	1.05	0.891	0.792	1.01	0.940	1.024	0.97	0.898	0.840
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.56	0.945	0.812	-0.49	0.939	0.835	-0.49	0.932	1.106	-0.38	0.897	0.511
$E(Y\|X_1=1,X_2=1)$	0.5	0.53	0.932	0.935	0.56	0.911	0.822	0.50	0.943	1.184	0.56	0.894	0.759
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.03	0.908	0.921	-0.05	0.895	0.795	0.03	0.896	0.832	0.00	0.936	1.018
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.955	0.985	0.95	0.925	0.841	0.87	0.873	0.509	1.00	0.943	1.166
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.49	0.909	1.072	-0.50	0.901	0.938	-0.32	0.712	0.475	-0.50	0.936	1.112
$E(Y\|X_1=1,X_2=1)$	0.5	0.46	0.928	0.933	0.43	0.904	0.798	0.42	0.885	0.720	0.49	0.941	1.205

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.13: The performance of cumulative probability models on estimating conditional medians with the sample size of 100. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.01	0.951	1.130	0.00	0.949	1.108	0.10	0.908	0.861	0.03	0.949	0.943
$Q^{0.5}\|X_1=1,X_2=0$	1	1.02	0.944	1.138	1.03	0.940	1.092	1.00	0.952	0.983	0.93	0.929	1.017
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.49	0.945	1.226	-0.50	0.945	1.215	-0.30	0.809	0.616	-0.48	0.951	0.912
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.51	0.948	1.236	0.51	0.947	1.169	0.52	0.940	1.173	0.49	0.944	1.197
(b) $\varepsilon \sim$ Logistic													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.05	0.945	1.092	0.02	0.949	1.112	0.15	0.909	0.894	0.12	0.924	0.922
$Q^{0.5}\|X_1=1,X_2=0$	1	0.99	0.949	1.118	1.02	0.948	1.094	0.93	0.942	1.012	0.90	0.925	1.027
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.42	0.941	1.121	-0.48	0.949	1.176	-0.21	0.820	0.700	-0.33	0.910	0.805
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.52	0.943	1.216	0.52	0.947	1.165	0.52	0.938	1.239	0.51	0.939	1.262
(c) $\varepsilon \sim$ Extreme Type I													
$Q^{0.5}\|X_1=0,X_2=0$	-0.37	-0.40	0.958	1.068	-0.38	0.957	1.080	-0.35	0.950	1.257	-0.28	0.929	0.808
$Q^{0.5}\|X_1=1,X_2=0$	0.63	0.59	0.931	1.163	0.64	0.936	1.184	0.66	0.948	1.176	0.45	0.827	0.602
$Q^{0.5}\|X_1=0,X_2=1$	-0.87	-0.98	0.952	0.814	-0.97	0.949	0.847	-0.86	0.953	1.264	-0.75	0.923	0.603
$Q^{0.5}\|X_1=1,X_2=1$	0.13	0.12	0.947	1.262	0.14	0.945	1.206	0.15	0.943	1.373	0.10	0.939	1.126
(d) $\varepsilon \sim$ Extreme Type II													
$Q^{0.5}\|X_1=0,X_2=0$	0.37	0.44	0.920	1.054	0.40	0.934	1.156	0.58	0.782	0.516	0.37	0.948	1.188
$Q^{0.5}\|X_1=1,X_2=0$	1.37	1.44	0.950	0.944	1.42	0.951	0.978	1.32	0.948	0.876	1.39	0.945	1.199
$Q^{0.5}\|X_1=0,X_2=1$	-0.13	0.00	0.873	1.003	-0.05	0.909	1.185	0.27	0.465	0.295	-0.13	0.944	1.341
$Q^{0.5}\|X_1=1,X_2=1$	0.87	0.91	0.942	1.196	0.88	0.942	1.186	0.92	0.934	1.066	0.88	0.942	1.401

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

Table 3.14: The performance of cumulative probability models on estimating conditional medians with the sample size of 100 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.05	0.941	0.964	0.02	0.951	1.046	0.13	0.882	0.677	0.10	0.913	0.734
$Q^{0.5}\|X_1=1,X_2=0$	1	0.98	0.949	1.032	1.01	0.952	1.057	0.93	0.934	0.822	0.90	0.909	0.791
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.42	0.933	0.955	-0.48	0.948	1.109	-0.23	0.725	0.428	-0.36	0.900	0.619
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.52	0.950	1.127	0.52	0.951	1.124	0.51	0.942	1.087	0.51	0.939	1.086
(f) $\varepsilon \sim$ uniform(-5, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	0	-0.24	0.928	1.075	-0.14	0.940	1.068	-0.08	0.941	1.445	-0.24	0.936	0.992
$Q^{0.5}\|X_1=1,X_2=0$	1	1.34	0.912	0.961	1.24	0.928	0.991	1.35	0.925	0.888	1.18	0.930	1.357
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.99	0.882	0.960	-0.81	0.920	1.074	-0.72	0.927	1.683	-0.98	0.907	0.862
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.54	0.945	1.298	0.54	0.943	1.154	0.62	0.946	1.393	0.48	0.946	1.434
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$Q^{0.5}\|X_1=0,X_2=0$	0.13	0.08	0.948	1.099	0.10	0.949	1.122	0.14	0.948	1.423	0.15	0.955	0.974
$Q^{0.5}\|X_1=1,X_2=0$	1.13	1.12	0.937	1.318	1.15	0.932	1.216	1.20	0.933	1.096	1.00	0.864	0.788
$Q^{0.5}\|X_1=0,X_2=1$	-0.37	-0.51	0.918	0.819	-0.47	0.925	0.904	-0.38	0.944	1.470	-0.38	0.955	0.796
$Q^{0.5}\|X_1=1,X_2=1$	0.63	0.62	0.944	1.387	0.64	0.939	1.272	0.67	0.939	1.467	0.59	0.938	1.222
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	-0.13	-0.09	0.924	1.231	-0.12	0.929	1.225	0.04	0.818	0.644	-0.16	0.941	1.248
$Q^{0.5}\|X_1=1,X_2=0$	0.87	0.95	0.934	0.940	0.93	0.939	1.001	0.88	0.959	0.954	0.89	0.946	1.352
$Q^{0.5}\|X_1=0,X_2=1$	-0.63	-0.54	0.898	1.292	-0.58	0.913	1.342	-0.31	0.550	0.391	-0.67	0.938	1.493
$Q^{0.5}\|X_1=1,X_2=1$	0.37	0.40	0.945	1.347	0.38	0.940	1.274	0.42	0.936	1.167	0.35	0.945	1.568

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

Table 3.15: The performance of cumulative probability models on estimating conditional means with the sample size of 200. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.945	0.992	0.01	0.940	0.957	0.03	0.911	0.903	0.10	0.846	0.491
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.950	0.989	1.00	0.945	0.960	0.90	0.857	0.490	0.97	0.916	0.894
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.50	0.944	0.980	-0.48	0.934	0.920	-0.36	0.706	0.433	-0.40	0.879	0.517
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.944	0.981	0.50	0.943	0.959	0.44	0.898	0.797	0.55	0.910	0.832
(b) $\varepsilon \sim$ Logistic													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.948	1.018	0.00	0.949	1.063	0.05	0.914	0.940	0.11	0.911	0.752
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.946	1.020	1.00	0.947	1.057	0.89	0.910	0.756	0.95	0.911	0.946
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.51	0.948	0.985	-0.50	0.948	1.042	-0.32	0.806	0.630	-0.37	0.916	0.686
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.947	1.022	0.50	0.947	1.077	0.45	0.927	1.006	0.54	0.933	1.008
(c) $\varepsilon \sim$ Extreme Type I													
$E(Y\|X_1=0,X_2=0)$	-0.58	-0.58	0.961	1.005	-0.53	0.935	0.909	-0.58	0.944	1.195	-0.43	0.831	0.413
$E(Y\|X_1=1,X_2=0)$	0.42	0.47	0.887	0.930	0.50	0.853	0.764	0.43	0.945	1.224	0.38	0.882	0.811
$E(Y\|X_1=0,X_2=1)$	-1.08	-1.20	0.947	0.550	-1.11	0.961	0.862	-1.08	0.945	1.093	-0.95	0.898	0.434
$E(Y\|X_1=1,X_2=1)$	-0.08	-0.03	0.928	0.966	0.00	0.897	0.836	-0.08	0.947	1.342	0.00	0.880	0.720
(d) $\varepsilon \sim$ Extreme Type II													
$E(Y\|X_1=0,X_2=0)$	0.58	0.53	0.891	0.932	0.50	0.856	0.769	0.62	0.880	0.803	0.57	0.947	1.221
$E(Y\|X_1=1,X_2=0)$	1.58	1.58	0.960	1.007	1.53	0.929	0.910	1.43	0.833	0.415	1.58	0.946	1.188
$E(Y\|X_1=0,X_2=1)$	0.08	0.07	0.910	1.198	0.05	0.901	1.097	0.30	0.552	0.332	0.07	0.945	1.317
$E(Y\|X_1=1,X_2=1)$	1.08	1.03	0.923	0.961	0.99	0.890	0.819	0.99	0.871	0.686	1.07	0.943	1.323

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.16: The performance of cumulative probability models on estimating conditional means with the sample size of 200 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$E(Y\|X_1=0,X_2=0)$	0	-0.01	0.948	1.060	-0.01	0.947	1.133	0.03	0.912	0.929	0.11	0.868	0.563
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.951	1.072	1.01	0.950	1.144	0.88	0.860	0.551	0.96	0.909	0.927
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.53	0.949	0.908	-0.52	0.949	1.034	-0.33	0.715	0.464	-0.41	0.909	0.545
$E(Y\|X_1=1,X_2=1)$	0.5	0.50	0.948	1.095	0.50	0.949	1.179	0.43	0.902	0.866	0.56	0.906	0.895
(f) $\varepsilon \sim$ uniform(-5, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.08	0.934	0.928	0.02	0.943	0.923	-0.02	0.950	1.202	-0.03	0.944	1.007
$E(Y\|X_1=1,X_2=0)$	1	1.09	0.934	0.922	0.99	0.939	0.920	1.03	0.941	1.005	1.03	0.947	1.192
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.67	0.910	0.822	-0.47	0.934	0.903	-0.54	0.942	1.256	-0.56	0.936	0.984
$E(Y\|X_1=1,X_2=1)$	0.5	0.49	0.948	1.014	0.49	0.941	0.931	0.50	0.950	1.179	0.50	0.950	1.170
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$E(Y\|X_1=0,X_2=0)$	0	0.00	0.961	0.989	0.05	0.921	0.778	0.00	0.948	1.189	0.13	0.794	0.348
$E(Y\|X_1=1,X_2=0)$	1	1.03	0.910	0.892	1.05	0.887	0.733	1.00	0.944	1.038	0.96	0.893	0.790
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.56	0.943	0.734	-0.49	0.948	0.831	-0.49	0.940	1.110	-0.37	0.857	0.379
$E(Y\|X_1=1,X_2=1)$	0.5	0.53	0.934	0.929	0.55	0.910	0.765	0.49	0.948	1.224	0.56	0.887	0.696
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$E(Y\|X_1=0,X_2=0)$	0	-0.03	0.910	0.888	-0.05	0.889	0.734	0.03	0.898	0.790	0.00	0.945	1.049
$E(Y\|X_1=1,X_2=0)$	1	1.00	0.960	0.989	0.96	0.923	0.781	0.87	0.807	0.357	1.00	0.944	1.178
$E(Y\|X_1=0,X_2=1)$	-0.5	-0.49	0.922	1.080	-0.50	0.915	0.938	-0.31	0.533	0.291	-0.50	0.944	1.140
$E(Y\|X_1=1,X_2=1)$	0.5	0.47	0.932	0.914	0.44	0.900	0.723	0.42	0.878	0.652	0.50	0.948	1.237

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified linear regression measured with MSE ratios.

Table 3.17: The performance of cumulative probability models on estimating conditional medians with the sample size of 200. The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(a) $\varepsilon \sim$ Normal													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.01	0.947	1.123	0.00	0.944	1.110	0.09	0.868	0.706	0.04	0.938	0.884
$Q^{0.5}\|X_1=1,X_2=0$	1	1.01	0.948	1.114	1.02	0.944	1.065	0.98	0.952	0.942	0.92	0.896	0.807
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.49	0.949	1.200	-0.51	0.944	1.173	-0.30	0.652	0.392	-0.47	0.946	0.857
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.51	0.948	1.236	0.50	0.945	1.165	0.51	0.944	1.191	0.49	0.942	1.198
(b) $\varepsilon \sim$ Logistic													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.05	0.945	1.043	0.01	0.953	1.089	0.15	0.871	0.722	0.12	0.902	0.765
$Q^{0.5}\|X_1=1,X_2=0$	1	0.98	0.949	1.102	1.01	0.948	1.089	0.90	0.920	0.870	0.88	0.895	0.837
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.42	0.934	1.046	-0.49	0.949	1.173	-0.21	0.690	0.452	-0.31	0.868	0.620
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.51	0.947	1.206	0.51	0.949	1.153	0.51	0.945	1.259	0.51	0.939	1.258
(c) $\varepsilon \sim$ Extreme Type I													
$Q^{0.5}\|X_1=0,X_2=0$	-0.37	-0.40	0.956	1.024	-0.39	0.954	1.053	-0.36	0.948	1.249	-0.27	0.905	0.659
$Q^{0.5}\|X_1=1,X_2=0$	0.63	0.59	0.921	1.049	0.63	0.935	1.157	0.65	0.949	1.155	0.44	0.671	0.377
$Q^{0.5}\|X_1=0,X_2=1$	-0.87	-0.98	0.932	0.703	-0.97	0.927	0.732	-0.86	0.949	1.256	-0.72	0.881	0.475
$Q^{0.5}\|X_1=1,X_2=1$	0.13	0.11	0.946	1.218	0.14	0.946	1.179	0.14	0.948	1.371	0.11	0.938	1.114
(d) $\varepsilon \sim$ Extreme Type II													
$Q^{0.5}\|X_1=0,X_2=0$	0.37	0.43	0.906	0.939	0.39	0.932	1.134	0.58	0.620	0.333	0.37	0.947	1.177
$Q^{0.5}\|X_1=1,X_2=0$	1.37	1.42	0.946	0.923	1.41	0.948	0.970	1.29	0.924	0.741	1.38	0.947	1.237
$Q^{0.5}\|X_1=0,X_2=1$	-0.13	0.00	0.824	0.795	-0.06	0.900	1.109	0.28	0.178	0.160	-0.13	0.948	1.342
$Q^{0.5}\|X_1=1,X_2=1$	0.87	0.90	0.942	1.186	0.87	0.946	1.187	0.90	0.937	1.098	0.87	0.947	1.402

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

Table 3.18: The performance of cumulative probability models on estimating conditional medians with the sample size of 200 (continued). The results are based on 10,000 simulation replicates.

		probit			**logit**			**cloglog**			**loglog**		
	true	est	CP	RE	est	CP	RE	est	CP	RE	est	CP	RE
(e) $\varepsilon \sim$ t with df=5													
$Q^{0.5}\|X_1=0,X_2=0$	0	0.04	0.938	0.910	0.01	0.949	1.030	0.13	0.822	0.510	0.11	0.870	0.541
$Q^{0.5}\|X_1=1,X_2=0$	1	0.97	0.945	0.985	1.00	0.951	1.065	0.90	0.888	0.610	0.89	0.853	0.583
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-0.42	0.926	0.847	-0.48	0.949	1.080	-0.22	0.512	0.250	-0.33	0.829	0.421
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.51	0.950	1.146	0.51	0.950	1.139	0.50	0.941	1.109	0.51	0.940	1.101
(f) $\varepsilon \sim$ uniform(-5, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	0	-0.26	0.906	0.937	-0.16	0.931	1.025	-0.11	0.938	1.442	-0.26	0.917	0.894
$Q^{0.5}\|X_1=1,X_2=0$	1	1.32	0.884	0.829	1.22	0.919	0.934	1.31	0.904	0.793	1.17	0.927	1.312
$Q^{0.5}\|X_1=0,X_2=1$	-0.5	-1.03	0.813	0.689	-0.84	0.898	0.922	-0.75	0.906	1.484	-1.01	0.855	0.662
$Q^{0.5}\|X_1=1,X_2=1$	0.5	0.52	0.950	1.330	0.52	0.948	1.167	0.58	0.953	1.464	0.47	0.950	1.489
(g) $\varepsilon \sim$ standardized Beta(5, 2)													
$Q^{0.5}\|X_1=0,X_2=0$	0.13	0.07	0.935	0.954	0.10	0.944	1.055	0.14	0.952	1.409	0.15	0.956	0.951
$Q^{0.5}\|X_1=1,X_2=0$	1.13	1.11	0.935	1.298	1.14	0.934	1.236	1.19	0.927	1.082	0.99	0.734	0.519
$Q^{0.5}\|X_1=0,X_2=1$	-0.37	-0.52	0.866	0.585	-0.48	0.892	0.707	-0.38	0.948	1.450	-0.37	0.959	0.789
$Q^{0.5}\|X_1=1,X_2=1$	0.63	0.61	0.947	1.345	0.63	0.945	1.250	0.66	0.948	1.510	0.59	0.938	1.179
(h) $\varepsilon \sim$ standardized Beta(2, 5)													
$Q^{0.5}\|X_1=0,X_2=0$	-0.13	-0.10	0.925	1.190	-0.13	0.932	1.216	0.03	0.699	0.448	-0.17	0.937	1.140
$Q^{0.5}\|X_1=1,X_2=0$	0.87	0.95	0.918	0.829	0.92	0.930	0.937	0.87	0.956	0.970	0.88	0.945	1.379
$Q^{0.5}\|X_1=0,X_2=1$	-0.63	-0.55	0.884	1.155	-0.59	0.911	1.304	-0.31	0.276	0.219	-0.68	0.931	1.366
$Q^{0.5}\|X_1=1,X_2=1$	0.37	0.40	0.943	1.321	0.37	0.940	1.262	0.41	0.937	1.150	0.35	0.944	1.571

est is the mean of the point estimates.

CP is the coverage probability of 95% confidence intervals in the 10,000 simulation replicates

RE is the relative efficiency compared with properly specified median regression measured with MSE ratios.

3.6.3 Cumulative Probability Models for Measurements Subject to Detection Limits

We conducted simulations to investigate the performance of cumulative probability models for handling measurements subject to detection limits. We generated data with sample size of 100 from $Y^* = \alpha + \beta X + \varepsilon$, where $X \sim N(0,1)$, $\varepsilon \sim N(0,1)$, $\alpha = 3$, $\beta = 0$ under the null hypothesis (H_0), and $\beta = 0.25$ under the alternative hypothesis (H_1). The outcome variable Y was then generated by left censoring Y^* at the detection limit (DL). That is, we set Y as undetectable if $Y^* < \text{DL}$ and set $Y = Y^*$ if $Y^* \geq \text{DL}$. We changed the values of DL to vary the proportion of undetectable measurements. Specifically, we set DL as 1.64, 2.28, 3.0, 3.72, and 4.36 so that the marginal proportion of undetectable measurements were 10%, 25%, 50%, 75%, and 90%, respectively.

We fitted the properly specified cumulative probability model using the probit link function. For purpose of comparison, we also fitted the logistic regression using dichotomized outcomes (undetectable vs. detectable) and two separate linear regression models: one imputing values below the detection limit as DL and the other imputing values below the detection limit as 0. To compare results with the logistic regression models, we also fitted cumulative probability models with the logit link function.

Table 3.19 summarizes the type I error rate and power of these approaches under various proportions of undetectable measurements. We found that the performances of cumulative probability models with the probit and logit link functions were generally similar, and they were generally more robust than the other two approaches. However, the gains in efficiency varied with the proportion of undetectable measurements. For example, when the proportion of undetectable measurements was small, e.g., $\leq 25\%$, cumulative probability models did not gain much efficiency compared with the linear regression models, but they were much more efficient than the logistic regression models; when the proportion of undetectable measurements was large, e.g., $\geq 75\%$, all three models had similar power.

Table 3.19: Type I error rate and power of cumulative probability models for outcomes subject to detection limit (DL) compared with logistic regression models (dichotomizing the outcome into two categories: detectable and undetectable) and linear regression models (imputing with the detection limit or 0). The sample size is 100 and we repeat the simulation 10,000 times for each scenario.

% below DL		**orm**		**logistic regression**	**linear regression**	
		probit	logit	binary outcome	impute DL	impute 0
10%						
	H_0	0.053	0.053	0.042	0.050	0.050
	H_1	0.689	0.671	0.263	0.674	0.662
25%						
	H_0	0.052	0.053	0.045	0.050	0.050
	H_1	0.673	0.654	0.406	0.646	0.625
50%						
	H_0	0.052	0.049	0.044	0.049	0.050
	H_1	0.606	0.586	0.478	0.549	0.570
75%						
	H_0	0.051	0.049	0.045	0.049	0.050
	H_1	0.464	0.446	0.412	0.387	0.450
90%						
	H_0	0.042	0.041	0.041	0.050	0.047
	H_1	0.279	0.271	0.261	0.237	0.288

Chapter 4

PResiduals: An R Package for Residual Analysis Using Probability-Scale Residuals

We have created an R package `Presiduals` that computes probability-scale residuals for a wide range of statistical models, including cumulative probability models we studied in Chapter 3. This package also implements both the partial and the conditional covariate-adjusted Spearman's rank correlations developed in Chapter 2. In this chapter, we present the `PResiduals` package. A publicly available dataset is used to illustrate its usage in model diagnostics, tests of conditional associations, and covariate-adjustment for Spearman's rank correlation.

4.1 Introduction

We have developed a new type of residual, the probability-scale residual (PSR), defined as $r(y,F^*) = F^*(y-) + F^*(y) - 1$, where y is the observed value and F^* is a fitted distribution (Li & Shepherd, 2010, 2012; Shepherd et al., in press). This residual is on the probability scale ranging from -1 to 1. It is well defined for a wide variety of outcome types and models, including some settings where other popular residuals are not applicable. Under properly specified models, the PSR has expectation 0, and it can, therefore, be used for model diagnostics. In addition, PSRs can be used to test for conditional associations (Li & Shepherd, 2010) and to construct covariate-adjusted Spearman's rank correlation (Liu, Shepherd, Wanga & Li, 2016). These methods are applicable to any orderable variables. They use order information but do not require assigning scores to ordered categorical variables or transforming continuous outcomes, and therefore, can achieve a good balance between robustness and efficiency.

The R package, `PResiduals`, has been developed to facilitate residual analyses using PSRs. The purpose of this vignette is to provide an introduction to the `PResiduals` pack-

age. We organize this paper as follows. In Section 4.2, we provide a brief review of PSRs and related methods. In Section 4.3, we illustrate the main functions in `PResiduals` with examples. Section 4.4 contains a summary.

4.2 Review of Methods

4.2.1 PSRs

A residual can be viewed as a contrast between the observed value and its fitted distribution. For example, the commonly used observed-minus-expected residual (OMER) can be written as $y-\hat{y}=E(y-Y^*)$, where y is the observed value, Y^* is a random variable from the fitted distribution F^*, and the contrast function is the difference. The PSR can be written similarly with a more general contrast function $\text{sign}(y,Y^*)$, where $\text{sign}(a,b)$ is -1, 0, and 1 for $a<b$, $a=b$, and $a>b$. Specifically, $r(y,F^*)=E[\text{sign}(y,Y^*)]=P(Y^*<y)-P(Y^*>y)=F^*(y-)+F^*(y)-1$. The PSR was originally proposed for ordered categorical variables where the difference between categories is not well defined (Li & Shepherd, 2010, 2012). Later, it was extended to other types of orderable variables, including continuous, discrete, and censored outcomes (Shepherd et al., in press).

With continuous outcomes, the PSR is $2F^*(y)-1$ (Shepherd et al., in press). If the model is properly specified, i.e., $F^*\to F$, then $r(Y,F^*)\to 2F(Y)-1$. Note that $F(Y)$ is the probability integral transformation and it is uniformly distributed from 0 to 1. Therefore, if the PSR is from the properly specified model, it will be approximately uniformly distributed from -1 to 1 with expectation 0 and constant variance $1/3$. A quantile-quantile (QQ) plot of PSRs versus the theoretical quantiles of the uniform distribution can be used to assess the overall model fit. In addition, PSRs can also be used in residual-by-predictor plots to detect lack of fit for specific predictors.

With discrete outcomes, the PSR is $2F^*(y)-f^*(y)-1$, where f^* is the probability mass function of the fitted distribution (Li & Shepherd, 2012; Shepherd et al., in press). In the

extreme case where Y is binary, the PSR reduces to $y - P(Y^* = 1)$, which is the OMER or unscaled Pearson residual. Although the PSR still has the expectation 0 under the properly specified model, it is not uniformly distributed due to the discreteness. Therefore, residual-by-predictor plots still provide information for the fit of specific predictors, but QQ-plots with PSRs are generally not useful.

With right censored outcomes, we denote T as the time to event and C as the time to censoring. Rather than directly observing T we only observe $Y = \min(T, C)$ and $\Delta = I(T \leq C)$. The above formula for the PSR can only be applied to non-censored observations. If censored, the failure time is unknown but it occurs after the censoring time y. Therefore, we define the PSR as its conditional expectation given that $t > y$, i.e., $E[r(T^*, F^*)|T^* > y] = F^*(y)$ (Shepherd et al., in press). Formally, the PSR for censored outcomes is defined in terms of y and δ, the observed values of Y and Δ as $r(y, F^*, \delta) = F^*(y) - \delta[1 - F^*(y-)]$. Note that with this definition, the PSR for censored observations is always non-negative. But under the properly specified model and $T \perp C$, it still has expectation 0. Therefore, the PSR can be used for model diagnostics for censored outcomes (Shepherd et al., in press).

4.2.2 Test of Residual Correlation with PSRs

The PSR was initially proposed as a component of test statistics for testing conditional association between two ordered categorical variables X and Y while adjusting for covariates Z, referred to as COBOT (conditional ordinal by ordinal tests) in Li & Shepherd (2010). Traditional regression approaches treat the ordinal predictor as either categorical or numerical, whereas the former ignores the order information and the latter makes linear assumptions. The basic idea of COBOT is to obtain conditional distributions of X and Y from models of X on Z and of Y on Z, and then to determine whether these conditional distributions are independent.

Three test statistics were proposed based on this idea. The first test statistic (T1) compares the observed joint distribution between X and Y with their expected distribution under

the null of conditional independence. If X and Y are independent conditional on Z, their joint distribution given Z is expected to follow the product of the conditional distribution of X and Y given Z. Therefore, we can test the difference between the observed and expected distributions; specifically, this is achieved by computing Goodman and Kruskal's gamma for the observed and expected joint distributions and taking their difference. The second test statistic (T2) is based on the residuals (PSRs). Specifically, it computes PSRs from models of X on Z and of Y on Z and tests the null of no residual correlation. The third test statistic (T3) evaluates the concordance-discordance of data draw from the joint fitted distribution of X and Y under conditional independence with those drawn from the empirical joint distributions, which can be written as the covariance of PSRs. P-values are computed based on large sample theory using M-estimation procedures. More details of these test statistics are given in Li & Shepherd (2010).

Note that the test statistic of T2 is analogous to the partial Pearson's correlation where the same procedure is performed with linear regression using the OMER. Since the PSR is well defined and on the same scale across various outcome types and models, we can generalize T2 of COBOT to other settings: using the correlation of PSRs, we can test for conditional associations between any orderable X and Y, including continuous, binary, ordered categorical, or count outcomes.

4.2.3 Covariate-Adjusted Spearman's Rank Correlation with PSRs

When there are no covariates, the PSR is a linear transformation of ranks and the correlation of PSRs is simply Spearman's rank correlation (Li & Shepherd, 2012; Shepherd et al., in press). Formally, the population parameter of Spearman's rank correlation can be expressed as the correlation of PSRs (Liu, Shepherd, Wanga & Li, 2016). With covariates, the PSR can be viewed as a linear transformation of adjusted ranks. This motivates us to use PSRs to construct covariate-adjusted rank correlations (Liu, Shepherd, Wanga & Li, 2016).

There are generally two types of covariate-adjusted correlations. One is the partial correlation, i.e., removing the effect of covariates and summarizing the relationship with a single number. The other is the conditional correlation, i.e., assessing the correlation at specific levels of the covariates. We have proposed estimators for both partial and conditional Spearman's rank correlations: our partial estimator is the correlation of PSRs and our conditional estimator is the conditional correlation of PSRs (Liu, Shepherd, Wanga & Li, 2016).

To obtain those estimators, we first need to fit models of X on Z and of Y on Z, and then compute PSRs from both models. Although the PSR is well defined and can be easily computed from many parametric or nonparametric models, to achieve a good balance between robustness and efficiency we favor fitting rank-based semiparametric models. Specifically, we considered the cumulative probability models (Liu, Shepherd, Wanga & Li, 2016). This class of models was originally developed for discrete ordinal data (McCullagh, 1980; Agresti, 2010), but can be applied to continuous data (Sall, 1991; Harrell, 2015; Liu, Shepherd, Li & Harrell, 2016). Since the model fit only uses the order information of X and Y, using PSRs from this type of models can preserve the rank-based nature of Spearman's rank correlation.

After obtaining PSRs from models of X on Z and of Y on Z, our partial estimator can be obtained simply as the correlation of PSRs. M-estimation techniques can be used to obtain its standard error (Liu, Shepherd, Wanga & Li, 2016). Since the correlation coefficient is bounded between -1 and 1, Fisher's transformation can be used to obtain better convergence. Technical details can be found in Liu, Shepherd, Wanga & Li (2016).

To obtain the conditional estimator for Spearman's rank correlation, we need to model the conditional correlation between PSRs. If Z is a categorical variable with sufficient numbers in each category, we can do a stratified analysis, i.e., compute the correlation of PSRs within each level of Z. If Z is continuous, smoothing is needed and can be achieved nonparametrically or parametrically. We have described a nonparametric approach based

on kernel weighting and a parametric approach using linear regression in Liu, Shepherd, Wanga & Li (2016).

4.3 Analysis with the PResiduals Package

4.3.1 Wage Data

Throughout this section, we repeatedly use a publicly available dataset, the wage data, as an example to illustrate the usage of key functions in the `PResiduals` package. This dataset can be obtained from the R package `ISLR` (James et al., 2013). It contains annual wages (in thousands of dollars) and other information for 3,000 male workers in the mid-Atlantic region of the United States from 2003 and 2009. With this dataset, we can build regression models for wages and study their relationship with other variables.

```
library(ISLR)
data(Wage)
head(Wage)
       year age     sex           maritl     race       education
231655 2006  18 1. Male 1. Never Married 1. White    1. < HS Grad
86582  2004  24 1. Male 1. Never Married 1. White 4. College Grad
161300 2003  45 1. Male       2. Married 1. White 3. Some College
155159 2003  43 1. Male       2. Married 3. Asian 4. College Grad
11443  2005  50 1. Male      4. Divorced 1. White      2. HS Grad
376662 2008  54 1. Male       2. Married 1. White 4. College Grad
                  region       jobclass          health health_ins
231655 2. Middle Atlantic  1. Industrial      1. <=Good      2. No
86582  2. Middle Atlantic 2. Information 2. >=Very Good      2. No
161300 2. Middle Atlantic  1. Industrial      1. <=Good     1. Yes
155159 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
11443  2. Middle Atlantic 2. Information      1. <=Good     1. Yes
376662 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
        logwage      wage
231655 4.318063  75.04315
86582  4.255273  70.47602
161300 4.875061 130.98218
155159 5.041393 154.68529
11443  4.318063  75.04315
376662 4.845098 127.11574
```

4.3.2 Calculation of PSRs

We first illustrate how to obtain PSRs from various models. The function `presid()` is implemented to compute PSRs. Its usage is very similar to the function `residuals()` from the `stats` library. Specifically, it takes a model object and returns a numerical vector containing PSRs in the order of original observations in the data set. Currently supported model objects include `lm` and `glm` (Poisson, binomial, and gaussian families) in the `stats` library; `polr` and `glm.nb` in the `MASS` library; `ols`, `Glm`, `lrm`, `orm`, `psm`, and `cph` in the `rms` library; and `survreg` (Weibull, exponential, gaussian, logistic, and lognormal distributions) and `coxph` in the `survival` library. Hence, using the function `presid()`, we can easily obtain PSRs from proportional odds models (more generally cumulative probability models), linear regression models, generalized linear regression models (such as Poisson and negative binomial models), parametric survival models, and Cox proportional hazards models. We now illustrate the calculation of PSRs from some of these models and their application in model diagnostics with the wage data.

We start with ordinal regression models for ordered categorical variables, for which the PSR was originally created. Specifically, we model the ordered categorical variable `education`, which has 5 levels, with a proportional odds model. We include `age`, `race`, `jobclass`, `maritl` (marital status), `health` (health status), and `year` (calender year) as covariates with `age` transformed using restricted cubic splines to account for a potential nonlinear relationship. PSRs are natural residuals for this type of models and can be obtained as functions of regression coefficients directly. In R, proportional odds models can be fitted using the function `polr()` from the `MASS` package or the function `orm()` from the `rms` package. The following chunk of code illustrates the usage of these two functions along with `presid()`. When using `orm()`, we need to set the arguments `x=TRUE` and `y=TRUE` so that the expanded design matrix and the values of the response variable are returned; this is a convention of the `rms` package. In this specific example, the PSRs obtained using these two functions are slightly different at the sixth digit after the decimal. This is

because `polr()` and `orm()` use different fitting procedures and yield slightly different regression coefficients.

```
library(PResiduals)
library(MASS)
po.polr <- polr(education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage)
PSR.po.polr <- presid(po.polr)

library(rms)
po.orm <- orm(education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage, x = TRUE, y = TRUE)
PSR.po.orm <- presid(po.orm)

summary(cbind(PSR.po.polr, PSR.po.orm))
  PSR.po.polr           PSR.po.orm
 Min.   :-0.9882886   Min.   :-0.988289
 1st Qu.:-0.4510896   1st Qu.:-0.451093
 Median :-0.0072629   Median :-0.007264
 Mean   : 0.0000001   Mean   : 0.000000
 3rd Qu.: 0.5012186   3rd Qu.: 0.501223
 Max.   : 0.9716579   Max.   : 0.971659
```

Figure 4.1 shows the application of PSRs in residual-by-predictor plots. Specifically, in the left panel of Figure 4.1, we include both linear and nonlinear terms by transforming `age` using restricted cubic splines with 5 knots, whereas in the right panel, we only include the linear term. The smoothed curve shows a nonlinear relationship between PSRs and `age` when only including the linear term, suggesting a better fit when both linear and nonlinear terms are included.

```
par(mfrow = c(1, 2))  #### residual-by-predictor plots
plot(Wage$age, PSR.po.orm, cex = 0.3, xlab = "age", ylab = "PSRs",
    col = gray(0.6), main = "(a) include both linear and
    nonlinear terms", cex.main = 0.8)
lines(lowess(Wage$age, PSR.po.orm), lwd = 3)
abline(h = 0, lty = 2)

po.t.orm <- orm(education ~ age + race + jobclass + maritl +
    health +  year, data = Wage, x = TRUE, y = TRUE)
PSR.po.t.orm <- presid(po.t.orm)
plot(Wage$age, PSR.po.t.orm, cex = 0.3, xlab = "age", ylab = "PSRs",
    col = gray(0.6), main = "(b) only include the linear term",
    cex.main = 0.8)
lines(lowess(Wage$age, PSR.po.t.orm), lwd = 3)
abline(h = 0, lty = 2)
```

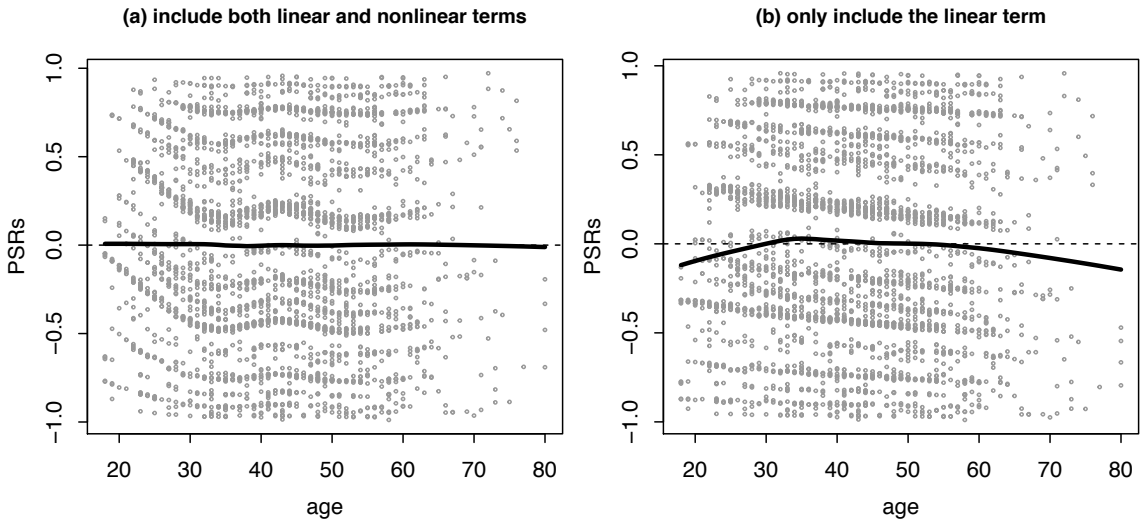

Figure 4.1: Residual-by-predictor plots with PSRs from proportional odds models. (a): PSRs are from the model including both linear and nonlinear terms. (b): PSRs are from the model only including the linear term.

Note that proportional odds models are special cases of a general class of ordinal regression models. This class of models has been referred to as cumulative link models in some literature (Agresti, 2010), but we refer to them as cumulative probability models because probabilities, not link functions, are added. Proportional odds models are cumulative probability models with the logit link (McCullagh, 1980). Other commonly used link functions include the probit link, the loglog link, and the complementary loglog link. Cumulative probability models with different link functions can be fitted by specifying the `method` argument in `polr()` or the `family` argument in `orm()`, and PSRs can be similarly obtained with `presid()`.

Next, we consider linear regression models. For linear regression models, PSRs can be obtained by assuming normality for the error distribution. For example, the PSR for observed value y_i can be computed as $2\Phi(y_i - \hat{y}_i) - 1$, where $\hat{y}_i$ is the fitted value and $\Phi()$ is the cumulative distribution function (CDF) of the standard normal distribution. This is the default in `presid()` for linear model objects (`lm`, `ols` and `Glm`). But the normality

assumption may be not necessary since it is well known that linear regression models are fairly robust to nonnormal errors as long as they are not highly skewed. In some application, we may be willing to only assume homoscedasticity instead of normality. In that case, PSRs can be obtained by empirically ranking the observed-minus-expected residuals (OMER). Specifically, if we denote the OMER for observation i as $\hat{\varepsilon}_i = y_i - \hat{y}_i$, the corresponding empirical PSR would be $\sum_{j=1}^{n} I(\hat{\varepsilon}_j < \hat{\varepsilon}_i)/n - \sum_{j=1}^{n} I(\hat{\varepsilon}_j > \hat{\varepsilon}_i)/n$ (Shepherd et al., in press). This can be obtained with `presid()` by setting the argument `emp=TRUE`. In the wage example, consider a linear regression model of `logwage` on `education`, `age`, `race`, `jobclass`, `maritl`, `health`, and `year`, where we apply the log transformation to `wage` due to its skewed distribution. The following chunk of code illustrates how to use `presid()` to obtain PSRs from linear regression models under different assumptions.

```
lm.1 <- lm(logwage ~ education + rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage)

library(PResiduals)
PSR.lm.1.normal <- presid(lm.1)  # default, normality assumed
PSR.lm.1.emp <- presid(lm.1, emp = TRUE)  #normality not assumed
OMER.lm.1 <- residuals(lm.1)  # observe-minus-expected residuals

summary(cbind(OMER.lm.1, PSR.lm.1.normal, PSR.lm.1.emp))
   OMER.lm.1       PSR.lm.1.normal     PSR.lm.1.emp
 Min.   :-1.7070   Min.   :-1.00000   Min.   :-0.9997
 1st Qu.:-0.1551   1st Qu.:-0.41148   1st Qu.:-0.4998
 Median : 0.0138   Median : 0.03841   Median : 0.0000
 Mean   : 0.0000   Mean   : 0.01264   Mean   : 0.0000
 3rd Qu.: 0.1657   3rd Qu.: 0.43675   3rd Qu.: 0.4998
 Max.   : 1.1556   Max.   : 0.99994   Max.   : 0.9997
```

Figure 4.2 shows the PSRs under different assumptions and their relationships with OMERs. Since the PSRs of continuous responses are uniformly distributed over $(-1, 1)$ under properly specified models, the quantile-quantile (QQ) plot of the empirical quantiles of the PSRs versus theoretical quantiles of uniform(-1, 1) can be used to assess the overall model fit (Shepherd et al., 2016). A QQ plot of PSRs from linear regression assuming normality is also plotted in Figure 4.2, suggesting the normal linear assumption for `logwage` may not be ideal. Note, the PSR under the assumption of homoscedasticity is obtained

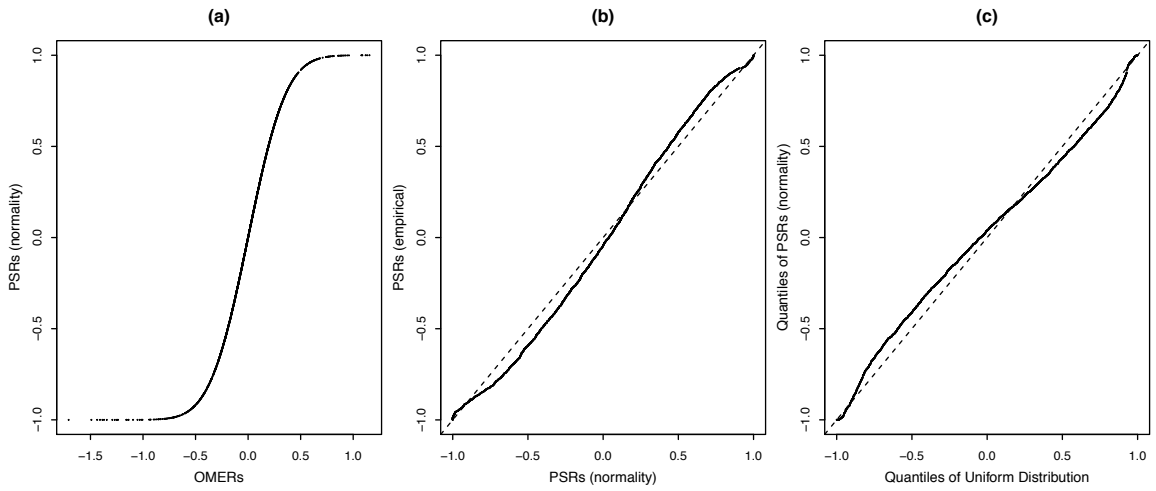

Figure 4.2: PSRs from linear regression models. (a): PSRs assuming normality are compared with OMERs. (b): empirical PSRs are compared with PSRs assuming normality. (c): QQ plot with PSRs under the assumption of normality.

by empirically ranking the OMERs, therefore, it is uniformly distributed by construction and its QQ-plot does not provide useful information about the model fit. However, this empirical PSR can still be used in residual-by-predictor plots to detect lack of fit for specific predictors. For example, in Figure 4.3, we compare the residual-by-predictor plots using the empirical PSRs from linear regression models including both linear and nonlinear terms for `age` (transformed using restricted cubic splines) and not including nonlinear terms. Again, the smoothed curves show a clear nonlinear pattern, suggesting lack of fit when only including the linear term.

```
par(mfrow = c(1, 3))
plot(OMER.lm.1, PSR.lm.1.normal, cex = 0.1, xlab = "OMERs", ylab =
    "PSRs (normality)", main = "(a)")
plot(PSR.lm.1.normal, PSR.lm.1.emp, cex = 0.1, xlab = "PSRs (normality)",
    ylab = "PSRs (empirical)", main = "(b)")
abline(0, 1, lty = 2)
qqplot(qunif(ppoints(length(PSR.lm.1.normal)), -1, 1), PSR.lm.1.normal,
    xlab = "Quantiles of Uniform Distribution",
    ylab = "Quantiles of PSRs (normality)",
    main = "(c)", cex = 0.1)  ## QQ plot of PSRs assuming normality
abline(a = 0, b = 1, lty = 2)
```

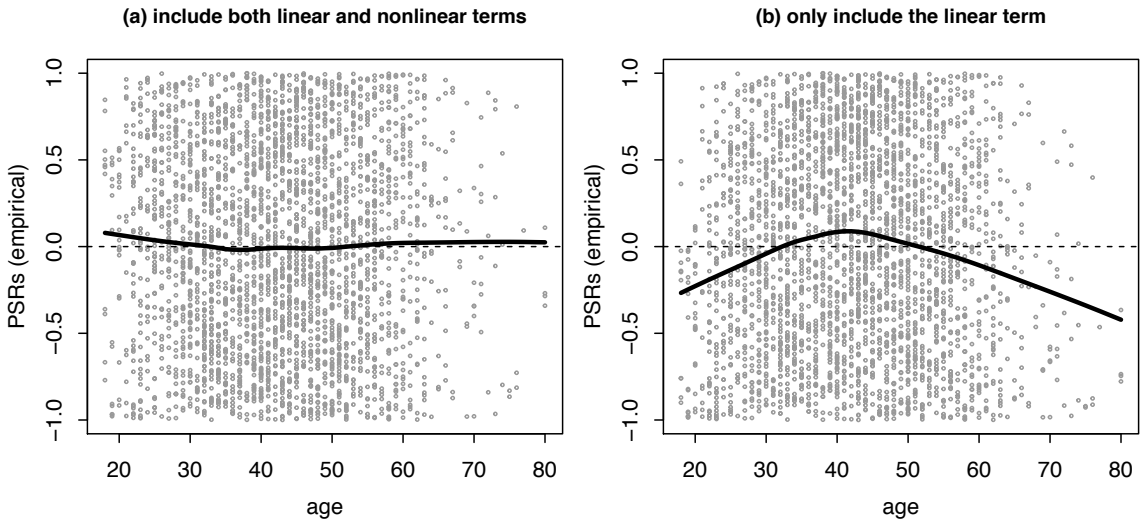

Figure 4.3: Residual-by-predictor plots using PSRs from linear regression models. (a): PSRs are from the model including both linear and nonlinear terms. (b): PSRs are from the model only including the linear term.

```
par(mfrow = c(1, 2))  ## residual-by-predictor plot
plot(Wage$age, PSR.lm.1.emp, cex = 0.3, xlab = "age",
    ylab = "PSRs (empirical)",  col = gray(0.6),
    main = "(a) include both linear and nonlinear terms",
    cex.main = 0.8)
lines(lowess(Wage$age, PSR.lm.1.emp), lwd = 3)
abline(h = 0, lty = 2)

lm.1.t <- ols(logwage ~ education + age + race + jobclass + maritl +
    health + year, data = Wage, x = TRUE, y = TRUE)
PSR.lm.1.t.emp <- presid(lm.1.t, emp = TRUE)
plot(Wage$age, PSR.lm.1.t.emp, cex = 0.3, xlab = "age",
    ylab = "PSRs (empirical)", col = gray(0.6),
    main = "(b) only include the linear term", cex.main = 0.8)
lines(lowess(Wage$age, PSR.lm.1.t.emp), lwd = 3)
abline(h = 0, lty = 2)
```

Although the log transformation is commonly used for right-skewed data, it may not be optimal. Different transformations may give conflicting results. A more robust analysis would estimate the transformation semiparametrically. Specifically, we have studied a semiparametric transformation model, which can be viewed as a natural extension of ordinal cumulative probability models to continuous responses (Sall, 1991; Harrell, 2015; Liu,

Shepherd, Li & Harrell, 2016). The `orm()` function in the `rms` package can be used to fit cumulative probability models for continuous responses. We now illustrate its usage and the calculation of PSRs with `presid()` using the wage data. Again, we need to set the arguments `x=TRUE` and `y=TRUE` when calling `orm()`.

```
library(rms)
cpm.logit <- orm(wage ~ education + rcs(age, 5) + race + jobclass +
    maritl + health + year, data = Wage, x = TRUE, y = TRUE)
cpm.cloglog <- orm(wage ~ education + rcs(age, 5) + race + jobclass +
    maritl + health + year, data = Wage, x = TRUE, y = TRUE,
    family = cloglog)

library(PResiduals)
PSR.cpm.logit <- presid(cpm.logit)
PSR.cpm.cloglog <- presid(cpm.cloglog)

summary(cbind(PSR.cpm.logit, PSR.cpm.cloglog))
 PSR.cpm.logit      PSR.cpm.cloglog
 Min.   :-0.99991   Min.   :-1.00000
 1st Qu.:-0.51619   1st Qu.:-0.42772
 Median : 0.01926   Median : 0.03212
 Mean   : 0.00000   Mean   : 0.00906
 3rd Qu.: 0.50798   3rd Qu.: 0.45051
 Max.   : 0.99976   Max.   : 0.99945
```

PSRs from cumulative probability models can also be used in QQ-plots and residual-by-predictor plots to assess model fit. Figure 4.4 shows QQ-plots of PSRs from cumulative probability models with the logit link and the cloglog link, suggesting better model fit with the logit link. The residual-by-predictor plots in Figure 4.5 show a similar nonlinear relationship between wages and age as seen in the linear regression models.

```
par(mfrow = c(1, 2))  ## QQ plots of PSRs
qqplot(qunif(ppoints(length(PSR.cpm.logit)), -1, 1), PSR.cpm.logit,
    cex = 0.1, xlab = "Quantiles of Uniform Distribution",
    ylab = "Quantiles of PSRs", main = "(a) logit")
abline(0, 1, col = 2, lty = 2)

qqplot(qunif(ppoints(length(PSR.cpm.cloglog)), -1, 1), PSR.cpm.cloglog,
    cex = 0.1, xlab = "Quantiles of Uniform Distribution",
    ylab = "Quantiles of PSRs", main = "(b) cloglog")
abline(0, 1, col = 2, lty = 2)
```

```
par(mfrow = c(1, 2))  ## residual-by-predictor plot
plot(Wage$age, PSR.cpm.logit, cex = 0.3, xlab = "age",
```

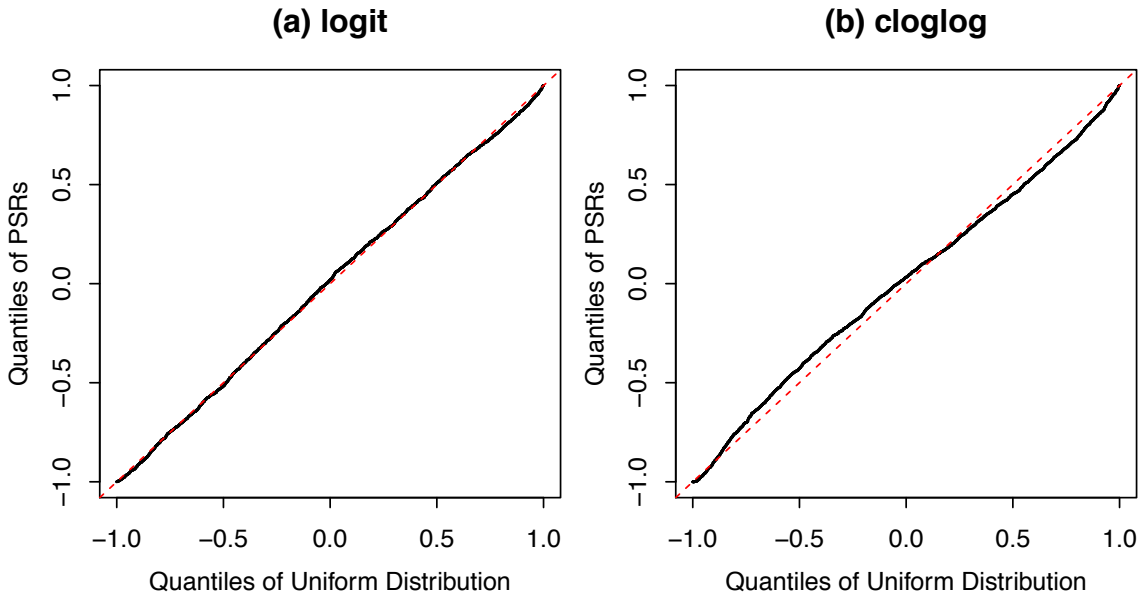

Figure 4.4: QQ-plots with PSRs from cumulative probability models with different link functions. (a) PSRs are from the model using the logit link. (b) PSRs are from the model using the cloglog link function.

```
    ylab = "PSRs (logit)", col = gray(0.6), cex.main = 0.8,
    main = "(a) include both linear and nonlinear terms")
lines(lowess(Wage$age, PSR.cpm.logit), lwd = 3)
abline(h = 0, lty = 2)

cpm.t.logit <- orm(wage ~ education + age + race + jobclass + maritl +
    health + year, data = Wage, x = TRUE, y = TRUE)
PSR.cpm.t.logit <- presid(cpm.t.logit)
plot(Wage$age, PSR.cpm.t.logit, cex = 0.3, xlab = "age",
    ylab = "PSRs (logit)", col = gray(0.6), cex.main = 0.8,
    main = "(b) only include the linear term")
lines(lowess(Wage$age, PSR.cpm.t.logit), lwd = 3)
abline(h = 0, lty = 2)
```

To illustrate PSRs for censored outcomes with the wage data, we artificially create a censoring indicator δ with the probability of being censored equal to 0.2. If $\delta = 0$, we pretend that the worker was not willing to share their exact wage and only reported a lower bound, i.e., the true wage is higher than the reported wage; whereas for workers with $\delta = 1$, we assume that they reported the exact value of their wages. In other words, we pretend that the wage data are right censored.

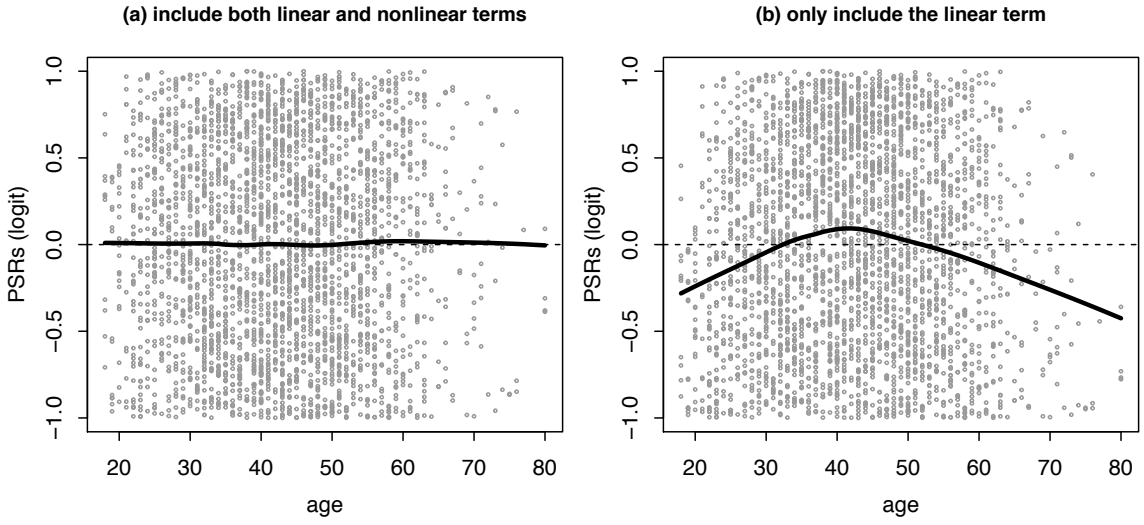

Figure 4.5: Residual-by-predictor plots using PSRs from cumulative probability models with the logit link function. (a): PSRs are from the model including both linear and nonlinear terms. (b): PSRs are from the model only including the linear term.

```
set.seed(1)
Wage$delta <- sample(c(0, 1), dim(Wage)[1], replace = TRUE, c(0.2, 0.8))
```

Survival models can be used to model right censored data. We first illustrate how to obtain PSRs from parametric survival models. Specifically, we use the `survreg()` function in the `survival` package to fit three parametric survival models, assuming the response distribution is Weibull, logistic, or Gaussian.

```
library(survival)
psm.1 <- survreg(Surv(wage, delta) ~ education + rcs(age, 5) + race +
    jobclass + maritl + health + year, dist = "weibull", data = Wage)
psm.2 <- survreg(Surv(wage, delta) ~ education + rcs(age, 5) + race +
    jobclass + maritl + health + year, dist = "logistic", data = Wage)
psm.3 <- survreg(Surv(wage, delta) ~ education + rcs(age, 5) + race +
    jobclass + maritl + health + year, dist = "gaussian", data = Wage)

library(PResiduals)
PSR.psm.1 <- presid(psm.1)
PSR.psm.2 <- presid(psm.2)
PSR.psm.3 <- presid(psm.3)

summary(cbind(PSR.psm.1, PSR.psm.2, PSR.psm.3))
   PSR.psm.1          PSR.psm.2          PSR.psm.3
 Min.   :-0.99335   Min.   :-0.99570   Min.   :-0.99899
```

```
 1st Qu.:-0.41191   1st Qu.:-0.45755   1st Qu.:-0.44846
 Median :-0.02840   Median : 0.03148   Median :-0.04532
 Mean   :-0.01622   Mean   : 0.00000   Mean   :-0.03946
 3rd Qu.: 0.34825   3rd Qu.: 0.41572   3rd Qu.: 0.32982
 Max.   : 1.00000   Max.   : 0.99997   Max.   : 1.00000
```

PSRs for censored outcomes are generally not uniformly distributed even when the model is properly specified. To assess the overall model fit, we have considered a modified version of the PSR, referred to as a Cox-Snell-like PSR (Shepherd et al., in press). This residual is simply the PSR evaluated at the observed value (ignoring censoring). It can be written as a one-to-one transformation of the Cox-Snell residual. Similar to the Cox-Snell residual which corresponds to a censored exponential(1) distribution, this modified PSR corresponds to a censored uniform distribution from -1 to 1 under the properly specified model. By comparing its Kaplan-Meier estimate with the uniform distribution, we can assess the goodness of fit. The following chunk of code shows the calculation of Cox-Snell-like PSRs. Note that this modified version of the PSR generally does not have expectation 0. Figure 4.6 shows QQ-plots of Cox-Snell-like PSRs based on the Kaplan-Meier estimates, suggesting better model fit when assuming the censored outcomes follow a logistic distribution.

```
library(PResiduals)
PSR.cox_snell.psm.1 <- presid(psm.1, type = "Cox-Snell-like")
PSR.cox_snell.psm.2 <- presid(psm.2, type = "Cox-Snell-like")
PSR.cox_snell.psm.3 <- presid(psm.3, type = "Cox-Snell-like")
summary(cbind(PSR.cox_snell.psm.1, PSR.cox_snell.psm.2,
              PSR.cox_snell.psm.3))
 PSR.cox_snell.psm.1 PSR.cox_snell.psm.2 PSR.cox_snell.psm.3
 Min.   :-0.9967     Min.   :-0.9957     Min.   :-0.9990
 1st Qu.:-0.4877     1st Qu.:-0.5487     1st Qu.:-0.5258
 Median :-0.1831     Median :-0.1731     Median :-0.2106
 Mean   :-0.1311     Mean   :-0.1133     Mean   :-0.1572
 3rd Qu.: 0.1684     3rd Qu.: 0.2728     3rd Qu.: 0.1567
 Max.   : 1.0000     Max.   : 1.0000     Max.   : 1.0000
```

```
par(mfrow = c(1, 3))  ## QQ plot with Cox-Snell-like PSRs

PSR.cox_snell.fit.1 <- survfit(Surv(PSR.cox_snell.psm.1, Wage$delta) ~
    1)
PSR.cox_snell.fit.2 <- survfit(Surv(PSR.cox_snell.psm.2, Wage$delta) ~
```

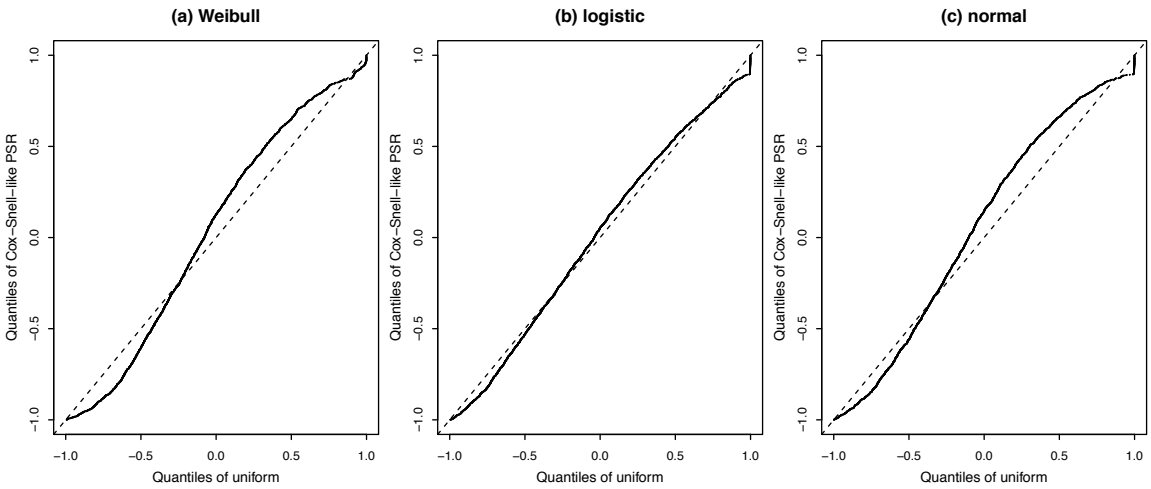

Figure 4.6: QQ-plots of Cox-Snell-like PSRs from parametric survival models with different distribution functions. (a): PSRs are from the model assuming the Weilbull distribution. (b): PSRs are from the model assuming the logistic distribution. (c): PSRs are from the model assuming the normal distribution

```
    1)
PSR.cox_snell.fit.3 <- survfit(Surv(PSR.cox_snell.psm.3, Wage$delta) ~
    1)

plot(summary(PSR.cox_snell.fit.1)$time,
    qunif(1 - summary(PSR.cox_snell.fit.1)$surv, -1, 1),  cex = 0.1,
    ylab = "Quantiles of Cox-Snell-like PSRs",  main = "(a) Weibull",
    xlab = "Quantiles of uniform", xlim = c(-1, 1), ylim = c(-1, 1))
abline(0, 1, lty = 2)

plot(summary(PSR.cox_snell.fit.2)$time,
    qunif(1 - summary(PSR.cox_snell.fit.2)$surv, -1, 1), cex = 0.1,
    ylab = "Quantiles of Cox-Snell-like PSRs", main = "(b) logistic"
    xlab = "Quantiles of uniform",  xlim = c(-1, 1), ylim = c(-1, 1))
abline(0, 1, lty = 2)

plot(summary(PSR.cox_snell.fit.3)$time,
    qunif(1 - summary(PSR.cox_snell.fit.3)$surv,  -1, 1),  cex = 0.1,
    ylab = "Quantiles of Cox-Snell-like PSRs",  main = "(c) normal"
    xlab = "Quantiles of uniform",  xlim = c(-1, 1), ylim = c(-1, 1))
abline(0, 1, lty = 2)
```

The standard PSR for censored data has expectation 0 under properly specified model and independent censoring; therefore, it can be used in residual-by-predictor plots (Shepherd et al., in press). Figure 4.7 plots PSRs from parametric survival models assuming the logistic distribution with and without the nonlinear terms for `age`, again suggesting a

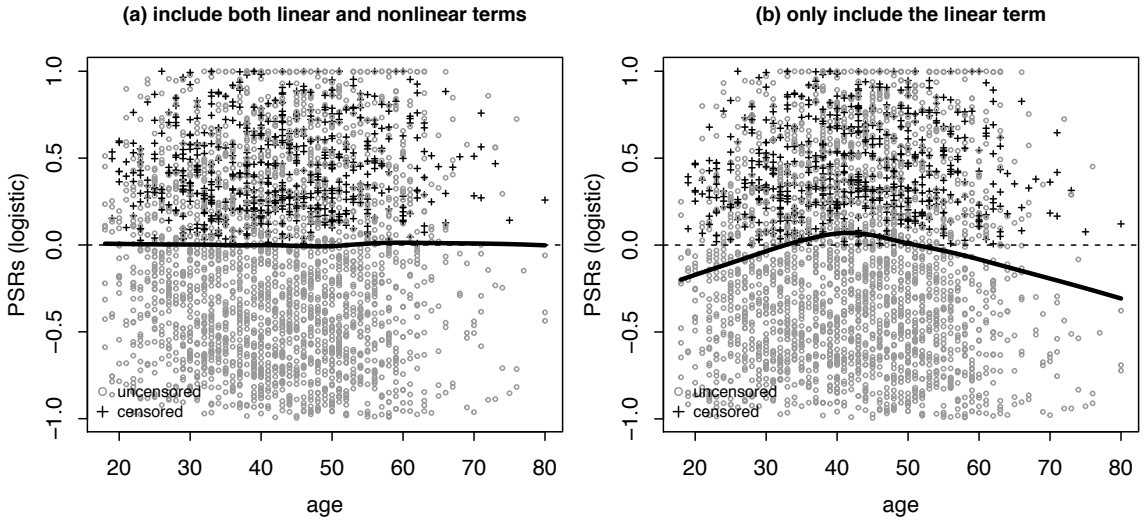

Figure 4.7: Residual-by-predictor plots using PSRs from parametric survival models assuming the logistic distribution. (a): PSRs are from the model including both linear and nonlinear terms. (b): PSRs are from the model only including the linear term.

better fit when including the nonlinear terms. For the purpose of illustration, we highlight the PSRs of censored observations, showing that they are always non-negative.

```
par(mfrow = c(1, 2))  ## residual-by-predictor plot with PSRs (standard)
plot(Wage$age, PSR.psm.2, cex = 0.4, xlab = "age",
    ylab = "PSRs (logistic)", col = ifelse(Wage$delta == 1,gray(0.6), 1),
    pch = ifelse(Wage$delta == 1, 1, 3), cex.main = 0.8,
    main = "(a) include both linear and nonlinear terms")
lines(lowess(Wage$age, PSR.psm.2), lwd = 3)
legend("bottomleft", legend = c("uncensored", "censored"), bty = "n",
    col = c(gray(0.6), 1), pch = c(1, 3), cex = 0.7)
abline(h = 0, lty = 2)

psm.2.t <- survreg(Surv(wage, delta) ~ education + age + race + jobclass
    + maritl + health + year, dist = "logistic", data = Wage)
PSR.psm.2.t <- presid(psm.2.t)
plot(Wage$age, PSR.psm.2.t, cex = 0.4, xlab = "age",
    ylab = "PSRs (logistic)", col = ifelse(Wage$delta == 1,gray(0.6), 1),
    pch = ifelse(Wage$delta == 1, 1, 3), cex.main = 0.8,
    main = "(b) only include the linear term")
legend("bottomleft", legend = c("uncensored", "censored"), bty = "n",
    col = c(gray(0.6), 1), pch = c(1, 3), cex = 0.7)
lines(lowess(Wage$age, PSR.psm.2.t), lwd = 3)
abline(h = 0, lty = 2)
```

We can also fit semiparametric survival models, e.g., the widely used proportional haz-

ards model, for censored wage data. The PSR and the Cox-Snell-like PSR can be obtained using the following chunk of code. Figure 4.8 shows the QQ plot of Cox-Snell-like PSRs and residual-by-predictor plots using PSRs from Cox proportional hazards models. The results are generally similar with those in the parametric survival models.

```
library(survival)
coxph.1 <- coxph(Surv(wage, delta) ~ education + rcs(age, 5) + race +
    jobclass + maritl + health + year, data = Wage)

library(PResiduals)
PSR.coxph.1 <- presid(coxph.1)  ##  standarad PSR
PSR.cox_snell.coxph <- presid(coxph.1, type = "Cox-Snell-like")
```

```
par(mfrow = c(1, 3))  ## PSRs from proportional hazards model

PSR.cox_snell.coxph.fit <- survfit(Surv(PSR.cox_snell.coxph, Wage$delta)
    ~ 1)
plot(summary(PSR.cox_snell.coxph.fit)$time,
    qunif(1 - summary(PSR.cox_snell.coxph.fit)$surv, -1, 1), cex = 0.1,
    ylab = "Quantiles of Cox-Snell-like PSRs", xlim = c(-1, 1),
    ylim = c(-1, 1), xlab = "Quantiles of uniform", main = "(a) Cox PH")
abline(0, 1, lty = 2)

plot(Wage$age, PSR.coxph.1, cex = 0.4, xlab = "age",
    ylab = "PSRs (Cox)", col = ifelse(Wage$delta == 1, gray(0.6), 1),
    pch = ifelse(Wage$delta == 1, 1, 3),
    main = "(b) include both linear and nonlinear terms")
legend("bottomleft", legend = c("uncensored", "censored"), bty = "n",
    col = c(gray(0.6), 1), pch = c(1, 3), cex = 0.7)
lines(lowess(Wage$age, PSR.coxph.1), lwd = 3)
abline(h = 0, lty = 2)

coxph.1.t <- coxph(Surv(wage, delta) ~ education + age + race + jobclass
    + maritl + health + year, data = Wage)
PSR.coxph.1.t <- presid(coxph.1.t)
plot(Wage$age, PSR.coxph.1.t, cex = 0.4, xlab = "age",
    ylab = "PSRs (Cox)", col = ifelse(Wage$delta == 1, gray(0.6), 1),
    pch = ifelse(Wage$delta == 1, 1, 3),
    main = "(c) only include the linear term")
legend("bottomleft", legend = c("uncensored", "censored"), bty = "n",
    col = c(gray(0.6), 1), pch = c(1, 3), cex = 0.7)
lines(lowess(Wage$age, PSR.coxph.1.t), lwd = 3)
abline(h = 0, lty = 2)
```

PSRs can also be computed for other types of data and models, for example, Poisson or negative binomial models for the count data. The usage of the `presid()` function for

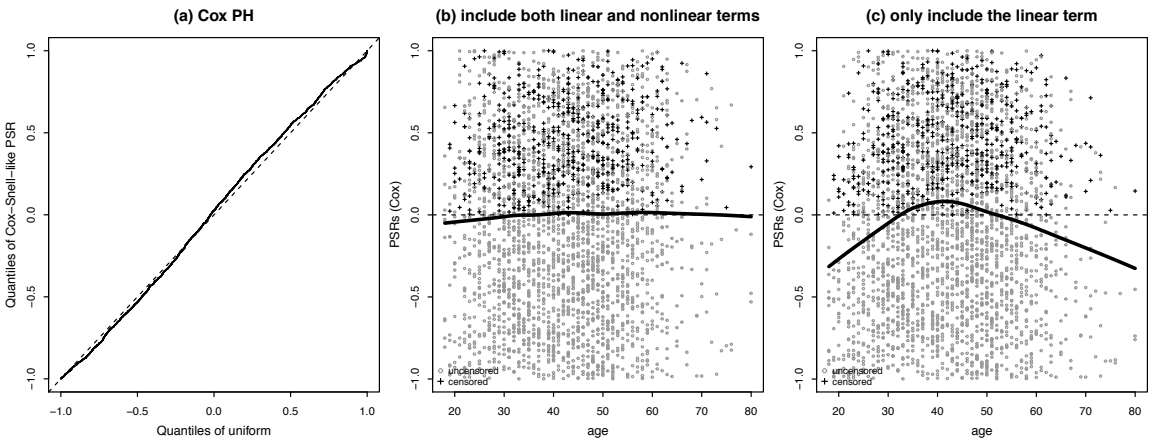

Figure 4.8: PSRs from Cox proportional hazards models. (a): QQ plot using Cox-Snell-like PSRs. (b): residual-by-predictor plot using PSRs from the model including both linear and nonlinear terms. (c) residual-by-predictor plot using PSRs from the model only including the linear term.

these models is similar with what we described above. We refer readers to the manual and the help file of `presid()` for more details and examples.

4.3.3 Tests of Conditional Association

In the previous section, we described the calculation of PSRs using the function `presid()` and illustrated their usage in model diagnostics. In this section, we focus on inference. Specifically, we describe how to use the `PResiduals` package to perform tests of conditional association.

Assume that we want to examine the association between `wage` and `education` while adjusting for a few potential confounders, such as `age`, `race`, `jobclass`, `maritl`, `health`, and `year`. One may consider fitting the linear regression or cumulative probability models we built earlier and then examining regression coefficients. In this example, both the linear regression model and the cumulative probability model suggest positive associations between `wage` and `education` after adjusting for other covariates (results not shown). Formal tests show significant associations between `wage` and `education` after adjusting for other covariates (results not shown).

However, in both regression models, the ordinal predictor `education` is coded as a categorical variable and the order information is ignored. To use the order information, we may consider assigning scores, e.g., the approximate years of education to different education levels, but that would force an assumption of linearity. This is a common issue for ordered categorical predictors in regression models and also one of the motivations for developing COBOT and PSRs.

The COBOT approach has been implemented in the `PResiduals` package as the `cobot()` function. To illustrate its usage with the wage data, we create an ordered categorical variable for wage, referred to as `wage.level`, by discretizing `wage` into five categories. Note, this is simply for the purpose of illustration and we do not recommend categorizing continuous variables in real data analyses because it does not use information efficiently and may lead to biased results (Royston et al., 2006). The `cobot()` function takes a formula object in the form of $X|Y \sim Z$, where X and Y are the ordinal variables whose relationship we are interested in, and Z designates the covariates we want to adjust for. Note that Z could be multidimensional covariates with transformations. By default, `cobot()` fits proportional odds models for both X on Z and Y on Z. Cumulative probability models with other link functions can be specified with the arguments `link.x` and `link.y`. The `cobot()` function reports three test statistics proposed in Li & Shepherd (2010) and their standard errors, p-values, and confidence intervals. The second statistic, T2, is the correlation of PSRs. Fisher's transformation is used by default to compute p-values and confidence intervals for T2. In this example, we find a strong positive association between education and the discretized wage with highly significant p-values.

```
Wage$wage.level <- cut(Wage$wage, breaks = c(0, quantile(Wage$wage,
    c(0.2, 0.4, 0.6, 0.8)), Inf))
summary(Wage$wage.level)
   (0,81.3] (81.3,97.5]  (97.5,114]   (114,135]   (135,Inf]
        661         548         635         558         598
cobot(wage.level | education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage)
                                est      stderr             p
Gamma(Obs) - Gamma(Exp)  0.3873366 0.015024858 1.497331e-146
Correlation of Residuals 0.4455342 0.015584921 4.729171e-134
```

```
Covariance of Residuals  0.1367667 0.004861128 3.680273e-174
                          lower CI  upper CI
Gamma(Obs) - Gamma(Exp)  0.3574999 0.4163833
Correlation of Residuals 0.4144760 0.4755558
Covariance of Residuals  0.1272267 0.1462814
Confidence Interval: 95%
Number of Observations: 3000
```

Since PSRs are well defined for a wide variety of outcomes, the COBOT approach based on PSRs can be extended to other types of X and Y as long as they are orderable. For example, in the `PResiduals` package, we have implemented `cocobot()` for an ordinal X and a continuous Y, `countbot()` for an ordinal X and a count variable Y, and a wrapper function `megabot()` for any orderable X and Y. The usage of `megabot()` is very similar with `cobot()` and is illustrated in the following chunk of code. Flexible modeling choices are available for both X on Z and Y on Z, and can be specified with the arguments `fit.x` and `fit.y`. Currently supported fitting procedures include `ordinal` (ordinal cumulative probability models fitted with `polr()`), `lm` (linear regression models assuming normality), `lm.emp` (linear regression models assuming homoscedasticity), `orm` (continuous or discrete cumulative probability models fitted with `orm()`), `poisson` (Poisson models for count data), and `nb` (negative binomial models for count data). If cumulative probability models are used (with either `polr()` or `orm()`), the default link function is the logit function and other link functions can be specified with arguments `link.x` and `link.y`. We give a few examples, using PSRs for `education` obtained from ordinal cumulative probability models fitted with either `polr()` or `orm()` and PSRs for `wage` obtained from either linear regression models or cumulative probability models. Note that when cumulative probability models are used for both models of X on Z and of Y on Z, the test results only use rank information of X and Y and are therefore invariant to any monotonic transformations of X or Y. Results are very similar across different models.

```
megabot(logwage | education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage, fit.x = "lm.emp", fit.y = "ordinal")
               est     stderr             p  lower CI  upper CI
cor PSRs 0.4403039 0.01585574 1.412076e-127 0.4087066 0.4708471
```

```
Confidence Interval: 95%
Number of Observations: 3000
Fisher Transform: TRUE
```

```
megabot(logwage | education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage, fit.x = "lm.emp", fit.y = "ordinal",
    link.y = "cloglog")
               est     stderr            p  lower CI  upper CI
cor PSRs 0.4409901 0.01562993 1.67254e-131 0.4098487 0.4711046
Confidence Interval: 95%
Number of Observations: 3000
Fisher Transform: TRUE
```

```
megabot(wage | education ~ rcs(age, 5) + race + jobclass + maritl +
    health + year, data = Wage, fit.x = "orm", fit.y = "orm")
               est     stderr             p  lower CI  upper CI
cor PSRs 0.4428448 0.01564295 5.103498e-132 0.4116738 0.4729808
Confidence Interval: 95%
Number of Observations: 3000
Fisher Transform: TRUE
```

4.3.4 Covariate-Adjusted Spearman's Rank Correlation with PSRs

As discussed in Section 4.2, PSRs can be used to construct partial and conditional Spearman's rank correlation adjusting for covariates (Liu, Shepherd, Wanga & Li, 2016). The test statistics in our tests for conditional association implemented in `megabot()` are actually partial Spearman's correlations.

We have implemented the function `partial.Spearman()` to obtain the partial Spearman's correlation where cumulative probability models are set as the default modeling method for both discrete and continuous ordinal X and Y (fitted with `orm()`). The following chunk of code illustrates its usage with different link functions in the example of `wage` and `education`.

```
partial.Spearman(wage | education ~ rcs(age, 5) + race + jobclass +
    maritl + health + year, data = Wage, link.x = "logit",
    link.y = "logit")
                      est     stderr             p  lower CI
partial Spearman 0.4428448 0.01564295 5.103498e-132 0.4116738
                 upper CI
partial Spearman 0.4729808
```

```
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

```
partial.Spearman(wage | education ~ rcs(age, 5) + race + jobclass +
    maritl + health + year, data = Wage, link.x = "probit",
    link.y = "probit")
                      est    stderr            p  lower CI
partial Spearman 0.4448799 0.0156437 8.34341e-133 0.4137038
                  upper CI
partial Spearman 0.4750138
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

```
partial.Spearman(wage | education ~ rcs(age, 5) + race + jobclass +
    maritl + health + year, data = Wage, link.x = "cloglog",
    link.y = "cloglog")
                      est     stderr             p  lower CI
partial Spearman 0.4580628 0.01555897 2.233473e-139 0.4270347
                  upper CI
partial Spearman 0.4880135
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

The result with the logit link function shows that after adjusting for other covariates, the partial Spearman's rank correlation between `wage` and `education` is 0.44 with 95% confidence interval (CI) (0.41, 0.47). This is lower than the unadjusted Spearman's rank correlation 0.50 (95% CI: 0.47, 0.53), suggesting part of the association between `wage` and `education` can be explained by their association with other covariates. Note that the point estimates and their confidence intervals are very similar with different link functions. (We did not report the result with the loglog link function because the cumulative probability model did not converge.). This is consistent with our simulations in Liu, Shepherd, Wanga & Li (2016) where we found that the partial Spearman's rank correlation using PSRs from `orm()` is robust to link function misspecification.

When covariates are multidimensional, as in the wage example, it may be useful to condition the correlation of PSRs on a single covariate. For example, we may be interested in whether Spearman's correlation varies for different job classes or ages while still

adjusting for other covariates. The function `conditional.partial.Spearman()` can be used to obtain the partial Spearman's correlation conditional on a specific covariate, denoted as Z_1. The usage of `conditional.partial.Spearman()` is very similar to `megabot()` and `partial.Spearman()`. It takes a formula object in the form of $X|Y \sim Z$ to specify the models of X on Z and of Y on Z. The fitting procedures can be specified with arguments `fit.x` and `fit.y` with the default as cumulative probability models with the logit link function. The covariate Z_1 is specified by the argument `conditional.by`. Different methods have been implemented to model the conditional correlation of PSRs and can be specified using the argument `conditional.method`. For categorical covariates such as `jobclass`, the conditional correlation of PSRs can be obtained by stratification, that is, we compute the correlation of PSRs within each category of `jobclass`. This can be achieved by setting `conditional.method= "stratification"`. For example,

```
conditional.partial.Spearman(education | wage ~ rcs(age, 5) + race +
    jobclass + maritl + health + year, conditional.by = "jobclass",
    conditional.method = "stratification",
    data = Wage)
Partial Spearman's correlation conditional by: jobclass
Conditional method: stratification
Number of levels of jobclass : 2
        jobclass       est     stderr            p  lower.CI
1  1. Industrial 0.4079285 0.02287611 4.085476e-56 0.3621315
2 2. Information 0.4782682 0.02107400 5.666486e-81 0.4359197
   upper.CI
1 0.4517609
2 0.5185035
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

If the stratification method is used, `conditional.partial.Spearman()` reports the point estimates, standard error estimates, p-values, and 95% confidence intervals for each category. In this example, after adjusting for other factors, Spearman's rank correlation between `wage` and `education` is higher in the information job class than that in the industrial class: 0.48 (95%CI: 0.44, 0.52) vs. 0.41 (95% CI: 0.36, 0.45).

For continuous variables such as `age`, two options are available for `conditional.method`: one is `lm`, which fits linear regression models for $X_{res}Y_{res}$ on Z_1, X_{res}^2 on Z_1, and Y_{res}^2 on Z_1 and then estimates the conditional correlation of PSRs using the fitted values, and the other is `kernel`, which estimates the conditional correlation of PSRs nonparametrically with kernel smoothing, allowing the user to input bandwidth parameters. If these methods are specified, `conditional.partial.Spearman()` prints the results for the first few observations and the results can be directly plotted using the function `plot()`.

```
conditional.lm <- conditional.partial.Spearman(wage | education ~
    rcs(age, 5) + race + jobclass + maritl + health + year,
    conditional.by = "age", conditional.method = "lm",
    conditional.formula = " ~ rcs(age,5)", data = Wage)

conditional.lm
Partial Spearman's correlation conditional by: age
Conditional method: lm
Conditional Formula: ~ rcs(age, 5)
  age        est     stderr            p   lower.CI  upper.CI
1  18 -0.0595070 0.11620834 6.094477e-01 -0.2804321 0.1674055
2  24  0.2014775 0.05041056 1.012077e-04  0.1009438 0.2979382
3  45  0.4983515 0.02531541 2.447728e-59  0.4471233 0.5463211
4  43  0.4958329 0.02878779 4.819362e-46  0.4373494 0.5501400
5  50  0.4982091 0.02759435 3.273183e-50  0.4422148 0.5503349
6  54  0.4767692 0.03082396 1.144778e-38  0.4141486 0.5348978
...
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

```
plot(conditional.lm)
```

```
conditional.kernel <- conditional.partial.Spearman(wage | education ~
    rcs(age, 5) + race + jobclass + maritl + health + year,
    conditional.by = "age", conditional.method = "kernel",
    kernel.bandwidth = "silverman", data = Wage)
conditional.kernel
Partial Spearman's correlation conditional by: age
Conditional method: kernel
kernel function: normal
kernel bandwidth: 2.467
     age        est
[1,]  18 0.01784734
[2,]  24 0.24475183
[3,]  45 0.50106153
[4,]  43 0.49954376
```

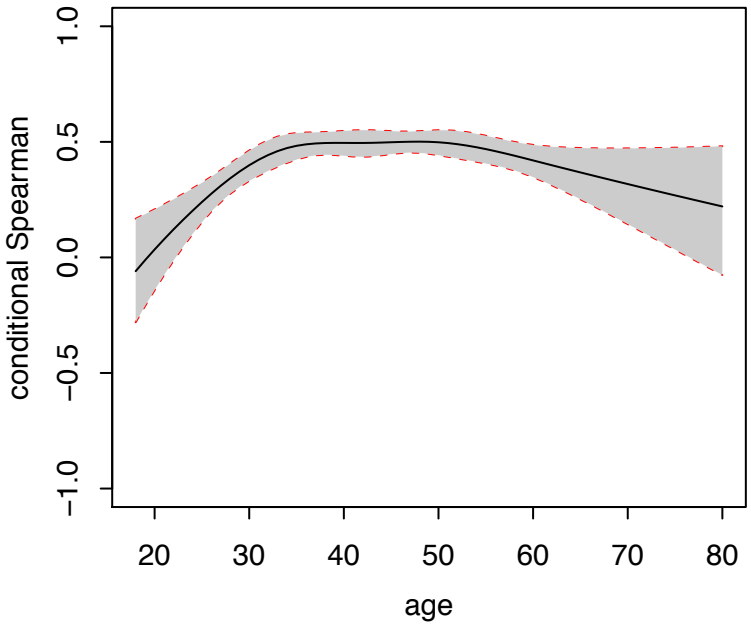

Figure 4.9: The age-specific conditional Spearman's rank correlation between wage and education. The conditional correlation of PSRs is modeled parametrically using linear regression models. The shaded regions are the point-wise 95% confidence intervals.

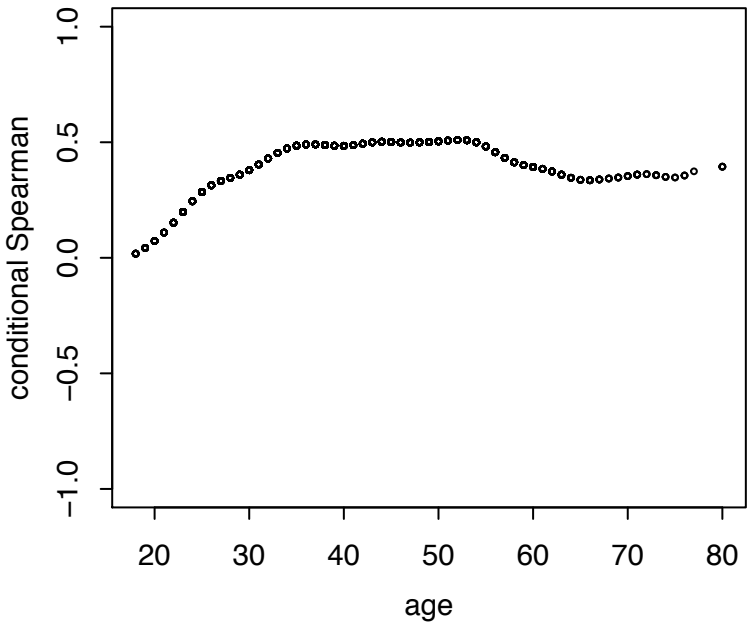

Figure 4.10: The age-specific conditional Spearman's rank correlation between wage and education. The conditional correlation of PSRs is modeled nonparametrically using kernel smoothing.

```
[5,]  50 0.50399806
[6,]  54 0.49989385
...
Fisher Transform: TRUE
Confidence Interval: 95%
Number of Observations: 3000
```

```
plot(conditional.kernel, cex = 0.4)
```

For the `lm` methods, `conditional.partial.Spearman()` reports standard error estimates and point-wise confidence intervals, obtained by M-estimation methods with Fisher's transformation. For the `kernel` methods, only the point estimates are returned. The standard error estimates can be obtained using bootstrap methods by users themselves. Figures 4.9 and 4.10 show that results from the two methods are similar, both suggesting that after adjusting for other factors, Spearman's rank correlation between wage and education is weaker among those who are younger (< 30 years).

4.4 Summary

The `PResiduals` package provides user-friendly functions for residual analysis with probability-scale residuals. This vignette illustrates its usage with examples. We hope users find it useful for model diagnostics and for assessing covariate-adjusted associations.

Chapter 5

Conclusions

5.1 Summary

In this dissertation, we have presented a general framework to construct covariate-adjusted Spearman's rank correlation. With the application of cumulative probability models, our estimators are rank-based and allow flexible modeling of covariates, achieving a good balance between efficiency and robustness. The wide applicability, robustness, and computational simplicity of our estimators make them very useful, particularly when dealing with big data.

In addition, we investigated the application of cumulative probability models to continuous outcomes. Our extensive simulations show that this approach has good finite sample performance and is fairly robust to minor or moderate link function misspecification. These results are important and will help promote the usage of this robust modeling strategy.

Finally, we developed the R package `PResiduals` to compute PSRs, to incorporate them into conditional tests of association, and to implement our covariate-adjusted Spearman's rank correlation.

5.2 Future Research

Since our framework is very general, there is much room for future work. Here we outline a few potential directions.

First, since PSRs are widely defined, we believe that our covariate-adjusted Spearman's rank correlation can be extended to more complicated settings, such as longitudinal data in which the observations are not independent and censored outcomes in which fitted distributions are not completely determined.

Second, we have shown good finite sample performance of semiparametric cumulative probability models for continuous outcomes through simulations. The asymptotic proper-

ties of this procedure have not been formally developed and warrant further study. Extensions of these models to correlated continuous outcomes may be possible, and would be a valuable statistical contribution. Finally, the idea of applying ordinal regression models to continuous outcomes is not limited to cumulative probability models. Other ordinal regression models, such as the continuous-ratio models and adjacent-categories models, could be similarly investigated.

REFERENCES

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

Bennett, S. (1983), Analysis of survival data by the proportional odds model, *Statistics in medicine* **2**(2), 273–277.

Box, G. E. P. & Cox, D. R. (1964), An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)* 211–252.

Cavanagh, C. & Sherman, R. P. (1998), Rank estimators for monotonic index models, *Journal of Econometrics* **84**(2), 351–381.

Chen, K., Jin, Z. & Ying, Z. (2002), Semiparametric analysis of transformation models with censored data, *Biometrika* **89**(3), 659–668.

Cheng, S., Wei, L. & Ying, Z. (1995), Analysis of transformation models with censored data, *Biometrika* **82**(4), 835–845.

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

Cox, D. R. (1975), Partial likelihood, *Biometrika* **62**(2), 269–276.

Fisher, R. A. (1924), The distribution of the partial correlation coefficient., *Metron* **3**, 329–332.

Genest, C. & Nešlehová, J. (2007), A primer on copulas for count data, *Astin Bulletin* **37**(2), 475–515.

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

Gijbels, I., Veraverbeke, N. & Omelka, M. (2011), Conditional copulas, association measures and their applications, *Computational Statistics and Data Analysis* **55**(5), 1919–1932.

Gripenberg, G. (1992), Confidence intervals for partial rank correlations, *Journal of the American Statistical Association* **87**(418), 546–551.

Han, A. K. (1987), A non-parametric analysis of transformations, *Journal of Econometrics* **35**(2), 191–209.

Harrell, F. E. (2015), *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2 edn, Springer.

Harrell, F. E. (2016), *rms: Regression Modeling Strategies*. R package version 4.5-0.
URL: *http://CRAN.R-project.org/package=rms*

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Huberty, C. J. (1989), Problems with stepwise methods—better alternatives, *Advances in social science methodology* **1**, 43–70.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *ISLR: Data for An Introduction to Statistical Learning with Applications in R*. R package version 1.0.
URL: *http://CRAN.R-project.org/package=ISLR*

Kendall, M. G. (1942), Partial rank correlation, *Biometrika* **32**(3–4), 277–283.

Khan, S. & Tamer, E. (2007), Partial rank estimation of duration models with general forms of censoring, *Journal of Econometrics* **136**(1), 251–280.

Kim, S. (2015), *ppcor: Partial and Semi-partial (Part) correlation*. R package version 1.1.
URL: *http://CRAN.R-project.org/package=ppcor*

Korn, E. L. (1984), The ranges of limiting values of some partial correlations under conditional independence, *The American Statistician* **38**(1), 61–62.

Kruskal, W. H. (1958), Ordinal measures of association, *Journal of the American Statistical Association* **53**(284), 814–861.

Li, C. & Shepherd, B. E. (2010), Test of association between two ordinal variables while adjusting for covariates, *Journal of the American Statistical Association* **105**(490), 612–620.

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

Liu, Q., Shepherd, B. E., Li, C. & Harrell, F. E. (2016), Modeling continuous outcomes using ordinal regression with cumulative probabilities.

Liu, Q., Shepherd, B. E., Wanga, V. & Li, C. (2016), Covariate-adjusted spearman's rank correlation with probability-scale residuals.

McCullagh, P. (1980), Regression models for ordinal data, *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(2), 109–142.

McGowan, C. C., Cahn, P., Gotuzzo, E., Padgett, D., Pape, J. W., Wolff, M., Schechter, M. & Masys, D. R. (2007), Cohort Profile: Caribbean, Central and South American Network for HIV research (CCASAnet) collaboration within the International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme, *International Journal of Epidemiology* **36**(5), 969–976.

Murphy, S. A., Rossini, A. J. & van der Vaart, A. W. (1997), Maximum likelihood estimation in the proportional odds model, *Journal of the American Statistical Association* **92**(439), 968–976.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer Science & Business Media.

Nešlehová, J. (2007), On rank correlation measures for non-continuous random variables, *Journal of Multivariate Analysis* **98**(3), 544–567.

Royston, P., Altman, D. G. & Sauerbrei, W. (2006), Dichotomizing continuous predictors in multiple regression: a bad idea, *Statistics in Medicine* **25**(1), 127–141.

Sall, J. (1991), A monotone regression smoother based on ordinal cumulative logistic regression, *ASA Proceedings of Statistical Computing Section* 276–137.

Shepherd, B. E. (2008), The cost of checking proportional hazards, *Statistics in Medicine* **27**(8), 1248–1260.

Shepherd, B. E., Li, C. & Liu, Q. (in press), Probability-scale residuals for continuous, discrete, and censored data, *Canadian Journal of Statistics* .

Song, X., Ma, S., Huang, J. & Zhou, X.-H. (2007), A semiparametric approach for the nonparametric transformation survival model with multiple covariates, *Biostatistics* **8**(2), 197–211.

Spearman, C. (1904), The proof and measurement of association between two things, *The American Journal of Psychology* **15**(1), 72–101.

Tian, C., Wan, T., Ying, L. & Xin, T. (2014), Rank regression: an alternative regression approach for data with outliers, *Shanghai Archives of Psychiatry* **26**(5), 310–315.

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

Wand, M. & Jones, M. (1995), *Kernel Smoothing*, Chapman&Hall, London.

Zeng, D. & Lin, D. (2006), Maximum likelihood estimation in semiparametric transformation models for counting processes, *Biometrika* **93**(3), 627–640.

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