ABSTRACT

ZHANG, MIN. Semiparametric Methods for Analysis of Randomized Clinical Trials and Arbitrarily Censored Time-to-event Data. (Under the direction of Dr. Marie Davidian and Dr. Anastasios A. Tsiatis.)

This dissertation includes two parts. In part one, using the theory of semiparametrics, we develop a general approach to improving efficiency of inferences in randomized clinical trials using auxiliary covariates. In part two, we study "smooth" semiparametric regression analysis for arbitrarily censored time-to-event data.

The primary goal of a randomized clinical trial is to make comparisons among two or more treatments. For example, in a two-arm trial with continuous response, the focus may be on the difference in treatment means; with more than two treatments, the comparison may be based on pairwise differences. With binary outcomes, pairwise odds-ratios or log-odds ratios may be used. In general, comparisons may be based on meaningful parameters in a relevant statistical model. Standard analyses for estimation and testing in this context typically are based on the data collected on response and treatment assignment only. In many trials, auxiliary baseline covariate information may also be available, and it is of interest to exploit these data to improve the efficiency of inferences. Taking a semiparametric theory perspective, we propose a broadly-applicable approach to adjustment for auxiliary covariates to achieve more efficient estimators and tests for treatment parameters in the analysis of randomized clinical trials. Simulations and applications demonstrate the performance of the methods.

A general framework for regression analysis of time-to-event data subject to arbitrary patterns of censoring is proposed. The approach is relevant when the analyst is willing to assume that distributions governing model components that are ordinarily left unspecified in popular semiparametric regression models, such as the baseline hazard function in the proportional hazards model, have densities satisfying mild "smoothness" conditions. Densities are approximated by a truncated series expansion that, for fixed degree of truncation, results in a "parametric" representation, which makes likelihood-based inference coupled with adaptive choice of the degree of truncation, and hence flexibility of the model, computationally and conceptually straightforward with data subject to any pattern of censoring. The formulation allows popular models, such as the proportional hazards, proportional odds, and accelerated failure time models, to be placed in a common framework; provides a principled basis for choosing among them; and renders useful extensions of the models straightforward. The utility and performance of the methods are demonstrated via simulations and by application to data from time-to-event studies.

Semiparametric Methods for Analysis of Randomized Clinical Trials and Arbitrarily Censored Time-to-event Data

by

Min Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2008

Approved By:

Dr. Daowen Zhang	Dr. Wenbin Lu
Dr. Marie Davidian Chair of Advisory Committee	Dr. Anastasios A. Tsiatis Co-Chair of Advisory Comittee

DEDICATION

To My Loving Family

BIOGRAPHY

Zhang, Min was born in Hunan Province, China to parents Zhang, Decheng and Xia, Shuyun in 1979. She has a cherishable childhood, full of love from her big family. Before she studied statistics, she received her B.S. in Environmental Sciences from Peking (Beijing) University, China, in 2001, and her M.A. in Ecology from Duke University, USA, in 2004. She made a difficult and important decision to switch her major, and she entered the Department of Statistics at North Carolina State University (NCSU) as a graduate student in 2004. While at NCSU, she received valuable guidance from her advisors Dr. Davidian, Marie and Dr. Tsiatis, Anastasios (Butch). She married Su, Fei in July, 2006. She will complete her Ph.D. in May, 2008.

ACKNOWLEDGMENTS

I am grateful to many persons, who have made important influence in my education and also my life. Words are never my strength, especially in occasions like this. Gratitude is in my heart.

My deepest gratitude goes to my advisors Marie Davidian and Anastasios (Butch), Tsiatis. I am so fortunate and honored to have advisors as talented and caring as you. If I have ever achieved anything in the last four years, they are yours. The only thing I have done is that, hopefully, I did not disappoint you. If I will ever achieve anything in the future, they won't happen without what you have given me.

I am especially grateful to Professor William Swallow. You believed me and opened the door to me when four years ago I came to talk to you about my determination and potential to study statistics. Not many people wanted to do that. I will always remember and appreciate your kindness.

I would also like to thank all faculty and staff members in our department. A very important person to me is Professor Dennis Boos, who wrote recommendation letters for me when I was applying for academic positions. I am thankful to Professors Daowen Zhang and Wenbin Lu for serving as my Ph.D. Dissertation members. I am thankful to Professor Cavell Brownie for her supervision when I worked as her research assistant. I am thankful to Professor Sastry Pantula for leading such an excellent department, which is my mother department wherever I go. Also I would like to thank my supervisor at Duke Clinical Research Institute, Karen Peiper, who gave me a lot of support in my work and in my job searching.

Finally, I want to acknowledge my family. You are the meaning of everything that I have and work for. To my parents, sisters and brother, grandparents, and in-laws, thank you for giving me endless love and support. To my husband, Fei Su, thank you for loving me, and always believing in me even when I myself doubted.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this dissertation, we study two topics. First, taking a semiparametric theory perspective, in Chapters 2 and 3 we develop theory and present a framework for improving efficiency of inferences in randomized clinical trials using auxiliary baseline covariates. Second, in Chapters 4 and 5 we study regression analysis of arbitrarily censored time-to-event data using a "smooth" semiparametric approach. Separate sets of notations are used for these two topics.

1.1 Introduction (I)

In randomized clinical trials, the primary objective is to compare two or more treatments on the basis of an outcome of interest. Along with treatment assignment and outcome, baseline auxiliary covariates may be recorded on each subject, including demographical and physiological characteristics, prior medical history, and baseline measures of the outcome. For example, the international Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy (PURSUIT) study (Harrington, 1998) in subjects with acute coronary syndromes compared the anti-coagulant Integrilin plus heparin and aspirin to heparin and aspirin alone (control) on the basis of the binary endpoint death or myocardial infarction at 30 days. Similarly, AIDS Clinical Trials Group (ACTG) 175 (Hammer et al., 1996) randomized HIV-infected subjects to four antiretroviral regimens with equal proba-

bilities, and an objective was to compare measures of immunological status under the three newer treatments to those under standard zidovudine (ZDV) monotherapy. In both studies, in addition to the endpoint, substantial auxiliary baseline information was collected.

Ordinarily, the primary analysis is based only on the data on outcome and treatment assignment. However, if some of the auxiliary covariates are associated with outcome, precision may be improved by "adjusting" for these relationships (e.g., Pocock et al., 2002), and there is an extensive literature on such covariate adjustment (e.g., Senn, 1989; Hauck, Anderson, and Marcus, 1998; Koch et al., 1998; Tangen and Koch, 1999; Lesaffre and Senn, 2003; Grouin, Day, and Lewis, 2004). Much of this work focuses on inference on the difference of two means and/or on adjustment via a regression model for mean outcome as a function of treatment assignment and covariates. In the special case of the difference of two treatment means, Tsiatis et al. (2007) proposed an adjustment method that follows from application of the theory of semiparametrics (e.g., van der Laan and Robins, 2003; Tsiatis, 2006) by Leon, Tsiatis, and Davidian (2003) to the related problem of "pretest-posttest" analysis, from which the form of the "optimal" (most precise) estimator for the treatment mean difference, adjusting for covariates, emerges readily. This approach separates estimation of the treatment difference from the adjustment, which may lessen concerns over bias that could result under regression-based adjustment because of the ability to inspect treatment effect estimates obtained simultaneously with different combinations of covariates and "to focus on the covariate model that best accentuates the estimate" (Pocock et al., 2002, p. 2925).

In Chapters 2 and 3, we expand on this idea by developing a broad framework for covariate adjustment in settings with two or more treatments and general outcome summary measures (e.g., log-odds ratios) by appealing to the theory of semiparametrics. The resulting methods seek to use the available data as efficiently as possible while making as few assumptions as possible. In Section 2.1, we present a semiparametric model framework involving parameters relevant to making general treatment comparisons. Using the theory of semiparametrics, we derive the class of estimating functions for these parameters in Section 2.2 and in Section 2.3 demonstrate how

these results lead to practical estimators. This development suggests a general approach to adjusting any test statistic for making treatment comparisons to increase efficiency, described in Section 2.4. Detailed theoretical proof is given in Section 2.5. Performance of the proposed methods is evaluated in simulation studies in Section 3.1 and is shown in representative applications in Section 3.2.

1.2 Introduction (II)

Regression analysis of censored time-to-event data is of central interest in health sciences research, and the most widely used approaches are based on semiparametric models. While representing some feature of the relationship between time-to-event and covariates by a parametric form, these models leave other aspects of their distribution unspecified.

Among such models, Cox's proportional hazards model (PH) (Cox, 1972) is unquestionably the most popular and is used almost by default in practice when data are right-censored, owing to straightforward, widely-available implementation. The hazard given covariates is represented as a parametric form modifying multiplicatively an unspecified baseline hazard function. This proportional hazards assumption is often not checked; however, effects of prognostic covariates often do not exhibit proportional hazards (e.g., Gray, 2000). Accordingly, there is considerable interest in alternative semiparametric regression models.

The accelerated failure time model (AFT) (Kalbfleisch and Prentice, 2002, sec. 2.2.3), in contrast to the PH model, where survival time and covariate effects are modeled indirectly through the hazard, represents the logarithm of event time directly by a parametric function of covariates plus a deviation with unspecified distribution, lending it practical appeal. However, this and similar models are used infrequently, likely due to computational challenges that undoubtedly dictate lack of commercially-available software. Although the iterative fitting method of Buckley and James (1979) (see also, e.g., Lin and Wei, 1992) for right-censored data is simple to program, it can exhibit problematic behavior, such as oscillation between two "solutions" (Jin, Lin, and Ying, 2006). Competing approaches based on rank tests (e.g., Tsiatis, 1990; Wei,

Ying, and Lin, 1990; Jin et al., 2003) may also admit multiple solutions (or have no solutions at all), can be computationally intensive (Lin and Geyer, 1992), and/or can involve rather complicated estimation of sampling variance.

The proportional odds (PO) model (Murphy, Rossini, and van der Vaart, 1997; Yang and Prentice, 1999) instead represents the logarithm of the ratio of the odds of survival given covariates to the baseline odds as a parametric function of covariates, where the associated baseline survival function is left unspecified. Despite its pleasing interpretation, the PO model is rarely used, again likely due to difficulty of implementation.

Hence, although the regression parameters in all of these models have intuitive interpretations, and although one model may be more suitable for representing the data than another, only the PH model is widely used. The PH and PO models are special cases of the linear transformation model (Cheng, Wei, and Ying, 1995; Chen, Jin, and Ying, 2002); Cheng, Wei, and Ying (1997) and Scharfstein, Gilbert, and Tsiatis (1998) also discuss a general class of models that includes both. The AFT and PH models are cases of the "extended" hazards model of Chen and Jewell (2001), including the "accelerated hazards" model of Chen and Wang (2000). However, there is no accessible framework that includes all three models and, indeed, further competitors, in which selection among them may be conveniently placed.

Moreover, the majority of developments for semiparametric time-to-event regression have been for right-censored (independently given covariates) event times. Fitting the PH model is straightforward under these conditions, but with interval censoring, specialized methods are required (Finkelstein, 1986; Satten, Datta, and Williamson, 1998; Goetghebeur and Ryan, 2000; Pan, 2000; Betensky et al., 2002), as they are for alternative models (e.g., Betensky, Rabinowitz, and Tsiatis, 2001; Sun, 2006). This require the analyst to seek out specialized, distinct techniques for different censoring patterns, even for the familiar PH model.

In Chapters 4 and 5, we propose a general framework for semiparametric regression analysis of censored time-to-event data that (i) provides a foundation for selection among competing models; (ii) unifies handling of different patterns of censoring, obviating the need for specialized techniques; and (iii) is computationally

tractable regardless of model or censoring pattern. To achieve simultaneously (i)–(iii), we sacrifice a bit of generality relative to traditional semiparametric methods by making the unrestrictive assumption that the distribution associated with unspecified model components has a density satisfying mild "smoothness" assumptions. Indeed, large sample theory for traditional methods requires similar assumptions (e.g., Ritov, 1990; Tsiatis, 1990; Jin et al., 2006). We assume that densities lie in a broad class whose elements may be approximated by the "SemiNonParametric" (SNP) density estimator of Gallant and Nychka (1987), tailored to provide an excellent approximation to virtually any plausible survival density. Many authors have used smoothing techniques in time-to-event regression (e.g., Kooperberg and Clarkson, 1997; Joly, Commenges, and Letenneur, 1998; Cai and Betensky, 2003; Komárek, Lesaffre, and Hilton, 2005). Our SNP approach endows likelihood-based inference for any of these models with "parametric-like" features and a virtually closed-form objective function under arbitrary censoring patterns that admits tractable implementation with standard optimization software. Competing models may be placed in a unified likelihood-based framework, providing a convenient, defensible basis for choosing among them via standard model selection techniques.

In Section 4.1, we review the SNP representation and describe its use in approximating any plausible survival density. We discuss SNP-based semiparametric time-to-event regression analysis with arbitrary censoring in Section 4.2. In Section 4.3, we discuss extension of the representation to more complex models, and in Section 4.4, we present the details. Simulation studies in Section 5.1 demonstrate performance. In Section 5.2, we apply the methods to two well-known data sets.

Chapter 2

Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

2.1 Semiparametric Model Framework

Denote the data from a k-arm randomized trial, $k \geq 2$, as (Y_i, X_i, Z_i), $i = 1, \ldots, n$, independent and identically distributed (iid) across i, where, for subject i, Y_i is outcome; X_i is the vector of all available auxiliary baseline covariates; and $Z_i = g$ indicates assignment to treatment group g with known randomization probabilities $P(Z = g) = \pi_g$, $g = 1, \ldots, k$, $\sum_{g=1}^{k} \pi_g = 1$. Randomization ensures that $Z \perp\!\!\!\perp X$, where "$\perp\!\!\!\perp$" means "independent of."

Let β denote a vector of parameters involved in making treatment comparisons under a specified statistical model. For example, in a two-arm trial, for a continuous real-valued response Y, a natural basis for comparison is the difference in means for each treatment, $E(Y \mid Z = 2) - E(Y \mid Z = 1)$, represented directly as β_2 in the model

$$E(Y \mid Z) = \beta_1 + \beta_2 I(Z = 2), \quad \beta_1 = E(Y \mid Z = 1), \quad \beta = (\beta_1, \beta_2)^T. \tag{2.1}$$

In a three-arm trial, we may consider the model

$$E(Y \mid Z) = \beta_1 I(Z = 1) + \beta_2 I(Z = 2) + \beta_3 I(Z = 3), \quad \beta = (\beta_1, \beta_2, \beta_3)^T. \tag{2.2}$$

In contrast to (2.1), we have parameterized (2.2) equivalently in terms of the three treatment means rather than differences relative to a reference treatment, and treatment comparisons may be based on pairwise contrasts among elements of β. For binary outcome $Y = 0$ or 1, where $Y = 1$ indicates the event of interest, we may consider for a k-arm trial

$$\text{logit}\{E(Y \mid Z)\} = \text{logit}\{P(Y = 1|Z)\} = \beta_1 + \beta_2 I(Z = 2) + \cdots + \beta_k I(Z = k), \quad (2.3)$$

where $\text{logit}(p) = \log\{p/(1-p)\}$; $\beta = (\beta_1, \ldots, \beta_k)^T$; and the log-odds ratio for treatment g relative to treatment 1 is β_g, $g = 2, \ldots, k$.

If Y_i is a vector of continuous longitudinal responses Y_{ij}, $j = 1, \ldots, m_i$, at times $t_{i1}, \ldots, t_{im_i}$, response-time profiles in a two-arm trial might be described by the simple linear mixed model

$$Y_{ij} = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}, \quad (b_{0i}, b_{1i})^T \overset{iid}{\sim} \mathcal{N}(0, D), \quad e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2), \quad (2.4)$$

where $\beta = (\beta_1, \beta_2)^T$, and β_2 is the difference in mean slope between the two treatments; extension to $k > 2$ treatment groups is straightforward. Alternatively, instead of considering the fully parametric model (2.4), one might make no assumption beyond

$$E(Y_{ij} \mid Z_i) = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij}, \quad j = 1, \ldots, m_i, \quad (2.5)$$

leaving remaining features of the distribution of Y given Z unspecified. For binary Y_{ij}, the marginal model $\text{logit}\{E(Y_{ij} \mid Z_i)\} = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij}$ might be adopted.

In all of (2.1)–(2.5), β $(p \times 1)$ is a parameter involved in making treatment comparisons in a model describing aspects of the conditional distribution of Y given Z and is of central interest. In addition to β, models like (2.4) and (2.5) depend on a vector of parameters γ, say; e.g., in (2.4), $\gamma = \{\alpha, \sigma_e^2, \text{vech}(D)^T\}^T$; and $\gamma = \alpha$ in (2.5). In general, we define $\theta = (\beta^T, \gamma^T)^T$ $(r \times 1)$, recognizing that models like (2.1)–(2.3) do not involve an additional γ, so that $\theta = \beta$.

For these and similar models, consistent, asymptotically normal estimators for θ, and hence for β and functions of its elements reflecting treatment comparisons, based

on the data (Y_i, Z_i), $i = 1, \ldots, n$, only and thus "unadjusted" for covariates, are readily available. Unadjusted, large-sample tests of null hypotheses of "no treatment effects" are also well-established. The difference of sample means is the obvious such estimator for β_2 in (2.1) and is efficient (i.e., has smallest asymptotic variance) among estimators depending only on these data, and a test of $H_0 : \beta_2 = 0$ may be based on the usual t statistic. Similarly, the maximum likelihood estimator (MLE) for β_2 in (2.4) and associated tests may be obtained from standard mixed model software. For $k > 2$, pairwise and global comparisons are possible; e.g., in (2.2), the sample means are efficient estimators for each element of β, and a test of $H_0 : \beta_1 = \beta_2 = \beta_3$ may be based on the corresponding two-degree-of-freedom Wald statistic.

As noted in Section 1.1, the standard approach in practice for covariate adjustment, thus using all of (Y_i, X_i, Z_i), $i = 1, \ldots, n$, is based on regression models for mean outcome as a function of X and Z. E.g., for $k = 2$ and continuous Y, a popular such estimator for β_2 in (2.1) is the ordinary least squares (OLS) estimator for ϕ in the analysis of covariance model

$$E(Y \mid X, Z) = \alpha_0 + \alpha_1^T X + \phi I(Z = 2); \tag{2.6}$$

extension to $k > 2$ treatments is immediate. See Tsiatis et al. (2007, Section 3) for discussion of related estimators for β_2 in the particular case of (2.1). If (2.6) is the correct model for $E(Y \mid X, Z)$, then ϕ and β_2 in (2.1) coincide, and, moreover, the OLS estimator for ϕ in (2.6) is a consistent estimator for β_2 that is generally more precise than the usual unadjusted estimator, even if (2.6) is not correct (e.g., Yang and Tsiatis, 2001). For binary Y, covariate adjustment is often carried out based on the logistic regression model

$$\text{logit}\{E(Y \mid X, Z)\} = \text{logit}\{P(Y = 1) \mid X, Z)\} = \alpha_0 + \alpha_1^T X + \phi I(Z = 2), \tag{2.7}$$

where the MLE of ϕ is taken as the adjusted estimator for the log-odds ratio β_2 in (2.3) with $k = 2$. In (2.7), ϕ is the log-odds ratio conditional on X, assuming this quantity is constant for all X. This assumption may or may not be correct; even if it were, ϕ is generally different from β_2 in (2.3). Tsiatis et al. (2007, Section 2) discuss this point in more detail.

To derive alternative methods, we begin by describing our assumed semiparametric statistical model for the full data (Y, X, Z), which is a characterization of the class of all joint densities for (Y, X, Z) that could have generated the data. We seek methods that perform well over as large a class as possible; thus, we assume that densities in this class involve no restrictions beyond the facts that $Z \perp\!\!\!\perp X$, guaranteed by randomization; that $\pi_g = P(Z = g)$, $g = 1, \ldots, k$, are known; and that β is defined through a specification on the conditional distribution of Y given Z as in (2.1)–(2.5). We thus first describe the conditional density of Y given Z. Under (2.3) and (2.4), this density is completely specified in terms of θ, while (2.5) describes only one aspect of the conditional distribution, the mean, in terms of θ, and (2.1) and (2.2) make no restrictions on the conditional distribution of Y given Z. To represent all such situations, we assume that this conditional density may be written as $p_{Y|Z}(y|z; \theta, \eta)$, where η is an additional nuisance parameter possibly needed to describe the density fully. For (2.3) and (2.4), η is null, as the density is already entirely characterized. For (2.1), (2.2), and (2.5), η is infinite-dimensional, as these specifications do not impose any additional constraints on what the density might be, so any density consistent with these models is possible.

Under the above conditions, we assume that all joint densities for (Y, X, Z) may be written, in obvious notation, as $p_{Y,X,Z}(y, x, z; \theta, \eta, \psi, \pi) = p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi) p_Z(z; \pi)$, where $p_Z(z; \pi)$ is completely specified, as $\pi = (\pi_1, \ldots, \pi_k)^T$ is known, and satisfy the constraints

$$\text{(i)} \qquad \int p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi)\, dx = p_{Y|Z}(y|z; \theta, \eta), \tag{2.8}$$

$$\text{(ii)} \qquad \int p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi)\, dy = p_X(x). \tag{2.9}$$

The joint density involves an additional, possibly infinite-dimensional nuisance parameter ψ, needed to include in the class all joint densities satisfying (i) and (ii). Here, $p_X(x)$ is any arbitrary marginal density for the covariates, and (ii) follows because $Z \perp\!\!\!\perp X$. In Section 2.5.1, we demonstrate that a rich class of joint distributions for (Y, X, Z) may be identified such that X is correlated with Y and $Z \perp\!\!\!\perp X$ [condition (ii)] that also satisfy condition (i). Because the joint density involves both finite (paramet-

ric) and infinite-dimensional components, it represents a semiparametric statistical model (see Tsiatis, 2006, Section 1.2).

2.2 Estimating Functions for Treatment Parameters Using Auxiliary Covariates

We now derive consistent, asymptotically normal estimators for θ, and hence β, in a given $p_{Y|Z}(y \mid z; \theta, \eta)$ and using the iid data (Y_i, X_i, Z_i), $i = 1 \ldots, n$, under the semiparametric framework satisfying (2.22) and (2.23). To do this, we identify the class of all estimating functions for θ based on (Y, X, Z) leading to all estimators for θ that are consistent and asymptotically normal under this framework. An estimating function is a function of a single observation and parameters used to construct estimating equations yielding an estimator for the parameters.

When the data on auxiliary covariates X are not taken into account, estimating functions for θ based only on (Y, Z) in models like those in (2.1)–(2.5) leading to consistent, asymptotically normal estimators are well known. For example, the OLS estimator for $\theta = \beta$ in the linear regression model (2.1) may be obtained by considering the estimating function

$$m(Y, Z; \theta) = \{1, I(Z = 2)\}^T \{Y - \beta_1 - \beta_2 I(Z = 2)\}, \quad \theta = \beta = (\beta_1, \beta_2)^T. \tag{2.10}$$

and solving the estimating equation $\sum_{i=1}^n m(Y_i, Z_i; \theta) = 0$ in θ. The OLS estimator for β_2 so obtained equals the usual difference in sample means. Likewise, with $\theta = \beta = (\beta_1, \ldots, \beta_k)^T$ and $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$, the usual logistic regression MLE for β in (2.3) is obtained by solving $\sum_{i=1}^n m(Y_i, Z_i; \theta) = 0$, where the estimating function $m(Y, Z; \theta)$ is equal to

$$\{1, I(Z = 2), \ldots, I(Z = k)\}^T [Y - \text{expit}\{\beta_1 + \beta_2 I(Z = 2) + \cdots + \beta_k I(Z = k)\}]. \tag{2.11}$$

The estimating functions (2.10) and (2.11) are unbiased; i.e., have mean zero assuming that (2.1) and (2.3), respectively, are correct. Under regularity conditions, unbiased estimating functions lead to consistent, asymptotically normal estimators (e.g., Carroll et al., 2006, Section A.6).

Our key result is that, given a semiparametric model $p_{Y,X,Z}(y, x, z\,; \theta, \eta, \psi, \pi)$ based on a specific $p_{Y|Z}(y|z; \theta, \eta)$ and satisfying (2.22) and (2.23), and given a fixed unbiased estimating function $m(Y, Z; \theta)$ $(r \times 1)$ for θ, such as (2.10) or (2.11), members of the class of all unbiased estimating functions for θ, and hence β, using all of (Y, X, Z) may be written as

$$m^*(Y, X, Z; \theta) = m(Y, Z; \theta) - \sum_{g=1}^{k} \{I(Z = g) - \pi_g\} a_g(X), \tag{2.12}$$

where $a_g(X)$, $g = 1, \ldots, k$, are arbitrary r-dimensional functions of X. Because $Z \perp\!\!\!\perp X$, the second term in (2.12) has mean zero; thus, (2.12) is an unbiased estimating function based on (Y, X, Z). When $a_g(X) \equiv 0$, $g = 1, \ldots, k$, (2.12) reduces to the original estimating function, which does not take account of auxiliary covariates, and solving $\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0$ leads to the unadjusted estimator $\widehat{\theta} = (\widehat{\beta}^T, \widehat{\gamma}^T)^T$ to which it corresponds. Otherwise, (2.12) "augments" $m(Y, Z; \theta)$ by the second term. With appropriate choice of the $a_g(X)$, the augmentation term exploits correlations between Y and elements of X to yield an estimator for θ solving $\sum_{i=1}^{n} m^*(Y_i, X_i, Z_i; \theta) = 0$ that is relatively more efficient than $\widehat{\theta}$. The proof of (2.12) is based on applying principles of semiparametric theory and is given in Section 2.5.2.

Full advantage of this result may be taken by identifying the optimal estimating function within class (2.12), that for which the elements of the corresponding estimator for θ have smallest asymptotic variance. This estimator for β thus yields the greatest efficiency gain over $\widehat{\beta}$ among all estimators with estimating functions in class (2.12) and hence more efficient inferences on treatment comparisons. By standard arguments for M-estimators (e.g., Stefanski and Boos, 2002), an estimator for θ corresponding to an estimating function of form (2.12) is consistent and asymptotically normal with asymptotic covariance matrix

$$\Delta^{-1}\Gamma(\Delta^{-1})^T, \quad \Gamma = E\Big([\, m(Y, Z; \theta_0) - \sum_{g=1}^{k} \{I(Z = g) - \pi_g\} a_g(X) \,]^{\otimes 2} \Big), \tag{2.13}$$

where θ_0 is the true value of θ, $u^{\otimes 2} = uu^T$, and $\Delta = E\{-\partial/\partial\theta^T\, m(Y, Z; \theta)\}|_{\theta=\theta_0}$. Thus, to find the optimal estimating function, one need only consider Γ in (2.13) and

determine $a_g(X)$, $g = 1, \ldots, k$, leading to Γ_{opt}, say, such that $\Gamma - \Gamma_{opt}$ is nonnegative definite. For given $m(Y, Z; \theta)$, it is shown in Section 2.5.3 that Γ_{opt} takes $a_g(X) = E\{m(Y, Z; \theta) \,|\, X, Z = g\}$, $g = 1, \ldots, k$. Thus, in general, the optimal estimator in class (2.12) is the solution to

$$\sum_{i=1}^{n} \Big[m(Y_i, Z_i; \theta) - \sum_{g=1}^{k} \{I(Z_i = g) - \pi_g\} E\{m(Y, Z; \theta) \,|\, X_i, Z = g\} \Big] = 0. \qquad (2.14)$$

In the case of β_2 in (2.1), (2.14) yields the optimal estimator in (16) of Tsiatis et al. (2007).

2.3 Implementation of Improved Estimators

The optimal estimator in class (2.12) solving (2.14) depends on the conditional expectations $E\{m(Y, Z; \theta) \,|\, X_i, Z = g\}$, $g = 1, \ldots, k$, the forms of which are of course unknown. Thus, to obtain practical estimators, we first consider a general adaptive strategy based on postulating regression models for these conditional expectations, which involves the following steps:

(1) Solve the original estimating equation $\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0$ to obtain the unadjusted estimator $\widehat{\theta}$. For each subject i, obtain the values $m(Y_i, g; \widehat{\theta})$ for each $g = 1, \ldots, k$.

(2) Note that the $m(Y_i, g; \widehat{\theta})$ are $(r \times 1)$. For each treatment group $g = 1, \ldots, k$ separately, based on the r-variate "data" $m(Y_i, g; \widehat{\theta})$ for i in group g, develop a parametric regression model $E\{m(Y, g; \widehat{\theta}) \,|\, X, Z = g\} = q_g(X, \zeta_g) = \{q_{g1}(X, \zeta_{g1}), \ldots, q_{gr}(X, \zeta_{gr})\}^T$, where $\zeta_g = (\zeta_{g1}^T, \ldots, \zeta_{gr}^T)^T$; i.e., such that $q_{gu}(X, \zeta_{gu})$, $u = 1, \ldots, r$, are regression models for each component of $m(Y, g; \widehat{\theta})$. We recommend an approach analogous to that in Leon et al. (2003, Section 4) where the $q_{gu}(X, \zeta_{gu})$ are represented as $\{1, c_{gu}^T(X)\}^T \zeta_{gu}$, $u = 1, \ldots, r$, and $c_{gu}(X)$ are vectors of basis functions in X that may include polynomial terms in elements of X, interaction terms, splines, and so on. This offers considerable latitude for achieving representations that can approximate the true conditional expectations, and hence predictions derived from them, well. We also recommend obtaining estimates

$\widehat{\zeta}_g = (\widehat{\zeta}_{g1}^T, \ldots, \widehat{\zeta}_{gr}^T)^T$ via OLS separately for each $u = 1, \ldots, r$, as, by a generalization of the argument in Leon et al. (2003, Section 4), this will yield the most efficient estimator for θ in step (3) below when the $q_g(X, \zeta_g)$ are of this form.

For each subject $i = 1, \ldots, n$, form predicted values $q_g(X_i, \widehat{\zeta}_g)$ for each $g = 1, \ldots, k$.

(3) Using the predicted values from step (2), form the augmented estimating equation

$$\sum_{i=1}^{n} \Big[m(Y_i, Z_i; \theta) - \sum_{g=1}^{k} \{ I(Z_i = g) - \pi_g \} q_g(X_i, \widehat{\zeta}_g) \Big] = 0 \tag{2.15}$$

and solve for θ to obtain the final, adjusted estimator $\widetilde{\theta}$. We recommend substituting $\widehat{\pi}_g = n^{-1} \sum_{i=1}^{n} I(Z_i = g)$ for π_g, $g = 1, \ldots, k$, in (2.15).

The foregoing three-step algorithm applies to very general $m(Y, Z; \theta)$. Often,

$$m(Y, Z; \theta) = A(Z, \theta)\{ Y - f(Z; \theta) \} \tag{2.16}$$

for some $A(Z, \theta)$ with r rows and some $f(Z, \theta)$, as in (2.10) and (2.11). Here, a simpler, "direct" implementation strategy is possible. Note that $E\{m(Y, Z; \theta) \,|\, X, Z = g\} = A(g, \theta)\{E(Y|X, Z = g) - f(g; \theta)\}$; thus, for each $g = 1, \ldots, k$, based on the data (Y_i, X_i) for i in group g, we may postulate parametric regression models $E(Y|X, Z = g) = q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}\zeta_g$, for a vector of basis functions $c_g(X)$, and obtain OLS estimators $\widehat{\zeta}_g$, $g = 1, \ldots, k$. Then form for each $i = 1, \ldots, n$ the corresponding predicted values for $E\{m(Y, Z; \theta) \,|\, X, Z = g\}$ as $q_g(X_i, \widehat{\zeta}_g, \theta) = A(g, \theta)\{q_g^*(X_i, \widehat{\zeta}_g) - f(g, \theta)\}$, where we emphasize that, here, $q_g(X_i, \widehat{\zeta}_g, \theta)$, $g = 1, \ldots, k$, are functions of θ. Substituting the $q_g(X_i, \widehat{\zeta}_g, \theta)$ (and $\widehat{\pi}_g$, $g = 1, \ldots, k$) in (2.15), solve the resulting equation in θ directly to obtain $\widetilde{\theta}$.

Several observations follow from semiparametric theory. Although we advocate representing $E\{m(Y, Z; \theta) \,|\, X, Z = g\}$ or $E(Y|X, Z = g)$, $g = 1, \ldots, k$, by parametric models, consistency and asymptotic normality of $\widetilde{\theta}$ hold regardless of whether or not these models are correct specifications of the true $E\{m(Y, Z; \theta) \,|\, X, Z = g\}$ or $E(Y|X, Z = g)$. Thus, the proposed methods are not parametric, as their validity does not depend on parametric assumptions. The theory also shows that, in either

implementation strategy, if the q_g are specified and fitted via OLS as described above, then, by an argument similar to that in Leon et al. (2003, Section 4), $\widetilde{\theta}$ is guaranteed to be relatively more efficient than the corresponding unadjusted estimator. Moreover, under these conditions, although ζ_g and π_g, $g = 1, \ldots, k$, are estimated, $\widetilde{\theta}$ will have the same properties asymptotically as the estimator that could be obtained if the limits in probability of the $\widehat{\zeta}_g$ were known and substituted in (2.14) and if the true π_g were substituted, regardless of whether the q_g are correct or not. In the direct strategy, if Y is discrete, it is natural to instead posit the $q_g^*(X, \zeta_g)$ as generalized linear models; e.g., logistic regression for binary Y, and fit these using iteratively reweighted least squares (IRWLS). Although the previous statements do not necessarily hold exactly, in our experience, they hold approximately. Regardless of whether or not the q_g are represented by parametric linear models and fitted by OLS, if the representation chosen contains the true form of $E\{m(Y, Z; \theta)|X, Z = g)$ or $E(Y|X, Z = g)$, respectively, then $\widetilde{\theta}$ is asymptotically equivalent to the optimal estimator solving (2.14). In general, the closer the predictions from these models are to the true functions of X, the closer $\widetilde{\theta}$ will come to achieving the precision of the optimal estimator. Because β is contained in θ, all of these results apply equally to $\widetilde{\beta}$.

Because in either implementation strategy $\widetilde{\theta}$ solving (2.15) is an M-estimator, the sandwich method (e.g., Stefanski and Boos, 2002) may be used to obtain a sampling covariance matrix for $\widetilde{\theta}$, from which standard errors for functions of $\widetilde{\beta}$ may be derived. This matrix is of form (2.13), with expectations replaced by sample averages evaluated at the estimates and $a_g(X)$ replaced by the predicted values using the q_g, $g = 1, \ldots, k$.

The regression models q_g in either implementation, which are the mechanism by which covariate adjustment is incorporated, are determined separately by treatment group and are developed independently of reference to the adjusted estimator $\widetilde{\beta}$. Thus, estimation of β could be carried out by a generalization of the "principled" strategy proposed by Tsiatis et al. (2007, Section 4) in the context of a two-arm trial and inference on β_2 in (2.1), where development of the models q_g would be undertaken by analysts different from those responsible for obtaining $\widetilde{\theta}$ to lessen concerns over possible bias, as discussed in Section 1.1.

2.4 Improved Hypothesis Tests

The principles in Section 2.2 may be used to construct more powerful tests of null hypotheses of no treatment effects by exploiting auxiliary covariates. The key is that, under a general null hypothesis H_0 involving s degrees of freedom, a usual test statistic T_n, say, based on the data (Y_i, Z_i), $i = 1, \ldots, n$, only is asymptotically equivalent to a quadratic form; i.e.,

$$T_n \approx \left\{ n^{-1/2} \sum_{i=1}^{n} \ell(Y_i, Z_i) \right\}^T \Sigma^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \ell(Y_i, Z_i) \right\}, \tag{2.17}$$

where $\ell(Y, Z)$ is a $(s \times 1)$ function of (Y, Z), discussed further below, such that $E_{H_0}\{\ell(Y, Z)\} = 0$, with E_{H_0} denoting expectation under H_0; and $\Sigma = E_{H_0}\{\ell(Y, Z)^{\otimes 2}\}$.

When the notion of "treatment effects" may be formulated in terms of β in a model like (2.1)–(2.5), the null hypothesis is typically of the form $H_0 : C\beta = 0$, where C is a $(s \times p)$ contrast matrix. E.g., in (2.2), C is (2×3) with rows $(1, -1, 0)$ and $(1, 0, -1)$. When inference on H_0 is based on a Wald test of the form $T_n = (C\widehat{\beta})^T (n^{-1}\widehat{\Sigma})^{-1} C\widehat{\beta}$, where $\widehat{\beta}$ is an unadjusted estimator corresponding to an estimating function $m(Y, Z; \theta)$, and $n^{-1}\widehat{\Sigma}$ is an estimator for the covariance matrix of $C\widehat{\beta}$, $\ell(Y, Z) = CB\, m(Y, Z, \theta_0)$. Here, B is the $(p \times r)$ matrix equal to the first p rows of $[E_{H_0}\{-\partial/\partial\theta^T m(Y_i, Z_i; \theta)\}|_{\theta=\theta_0}]^{-1}$, and θ_0 is the value of θ under H_0.

In other situations, the null hypothesis may not refer to a parameter like β in a given model. For example, the null hypothesis for a k-arm trial may be $H_0 : S_1(u) = \cdots = S_k(u) = S(u)$, where $S_g(u) = 1 - P(Y \leq u | Z = g)$, and $S(u) = 1 - P(Y \leq u)$. A popular test in this setting is the Kruskal-Wallis test, which reduces to the Wilcoxon rank sum test for $k = 2$. Letting $n_g = \sum_{i=1}^{n} I(Z_i = g)$ and $\overline{R}_g$ be the average of the overall ranks for subjects in group g, the test statistic is $T_n = 12 \sum_{g=1}^{k} n_g \{\overline{R}_g - (n+1)/2\}^2 / \{n(n+1)\}$. By results in van der Vaart (1998, Section 12.2), it may be shown that T_n is asymptotically equivalent to a statistic of the form (2.17), where $\ell(Y, Z)$ is $(k-1 \times 1)$ with gth element $\{I(Z = g) - \pi_g\}\{S(Y) - 1/2\}$.

To motivate the proposed more powerful tests, we consider the behavior of T_n in (2.17) under a sequence of local alternatives H_{1n} converging to H_0 at rate $n^{-1/2}$. Typically, under suitable regularity conditions, $n^{-1/2} \sum_{i=1}^{n} \ell(Y_i, Z_i)$ in (2.17) converges

under the sequence H_{1n} to a $\mathcal{N}(\tau, \Sigma)$ random vector for some τ, so that T_n has asymptotically a noncentral χ^2_s distribution with noncentrality parameter $\tau^T \Sigma^{-1} \tau$. To obtain a more powerful test, then, we wish to construct a test statistic with noncentrality parameter as large as possible. Based on the developments in Section 2.2, we consider test statistics of the form

$$T_n^* = \left\{ n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i) \right\}^T \Sigma^{*\,-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i) \right\}, \tag{2.18}$$

$$\ell^*(Y, X, Z) = \ell(Y, Z) - \sum_{g=1}^{k} \{ I(Z = g) - \pi_g \} a_g(X), \tag{2.19}$$

where $\Sigma^* = E_{H_0}\{\ell^*(Y, X, Z)^{\otimes 2}\}$. The second term in (2.19) has mean zero by randomization under H_0 or any alternative. Accordingly, it follows under the sequence of alternatives H_{1n} that $n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i)$ converges in distribution to a $\mathcal{N}(\tau, \Sigma^*)$ random vector, so that T_n^* in (2.18) has an asymptotic χ^2_s distribution with noncentrality parameter $\tau^T \Sigma^{*\,-1} \tau$. These results suggest that, to maximize the noncentrality parameter and thus power, we wish to find the particular Σ^*, Σ^*_{opt}, say, that makes $\Sigma^{*\,-1}_{opt} - \Sigma^{*\,-1}$ non-negative definite for all Σ^*, which is equivalent to making $\Sigma^* - \Sigma^*_{opt}$ non-negative definite for all Σ^*. This corresponds to finding the optimal choice of $a_g(X)$, $g = 1, \ldots, k$, in (2.19). By an argument similar to that leading to (2.14), the optimal choice is $a_g(X) = E\{\ell(Y, Z) | X, Z = g\}$ for $g = 1, \ldots, k$.

These developments suggest an implementation strategy analogous to that in Section 2.3:

(1) For the test statistic T_n, determine $\ell(Y, Z)$ and substitute sample quantities for any unknown parameters to obtain $\widehat{\ell}(Y_i, Z_i)$, $i = 1, \ldots, n$. E.g., for $H_0 : C\beta = 0$ in model (2.2), with C (2×3) as above, $m(Y, Z, \theta) = \{I(Z = 1), I(Z = 2), I(Z = 3)\}^T \{Y - \beta_1 I(Z = 1) - \beta_2 I(Z = 2) - \beta_3 I(Z = 3)\}$, $\theta = (\beta_1, \beta_2, \beta_3)^T$. Under H_0, $\theta_0 = (\mu, \mu, \mu)^T$, say, so that $m(Y, Z, \theta_0) = \{I(Z = 1), I(Z = 2), I(Z = 3)\}^T (Y - \mu)$, and

$$\ell(Y, Z) = \begin{pmatrix} \pi_1^{-1} I(Z = 1) - \pi_2^{-1} I(Z = 2) \\ \pi_1^{-1} I(Z = 1) - \pi_3^{-1} I(Z = 3) \end{pmatrix} (Y - \mu). \tag{2.20}$$

As μ is unknown, $\widehat{\ell}(Y_i, Z_i)$ is obtained by substituting $n^{-1}\sum_{i=1}^n Y_i$ for μ. We recommend substituting $\widehat{\pi}_g$ for π_g, $g = 1, 2, 3$, in (2.20), as above. Similarly, for the Kruskal-Wallis test, $\widehat{\ell}(Y_i, Z_i) = \{I(Z = g) - \widehat{\pi}_g\}\{\widehat{S}(Y_i) - 1/2\}$, $\widehat{S}(u) = n^{-1}\sum_{i=1}^n I(Y_i \geq u)$.

(2) For each treatment group $g = 1, \ldots, k$ separately, treating the $\widehat{\ell}(Y_i, Z_i)$ for subjects i in group g as s-variate "data," develop a regression model $E\{\widehat{\ell}(Y, g)|X, Z = g) = q_g(X, \zeta_g) = \{q_{g1}(X, \zeta_{g1}) \ldots, q_{gs}(X, \zeta_{gs})\}^T$ by representing each component $q_{gu}(X, \zeta_{gu})$, $u = 1, \ldots, s$, by the parametric "basis function" approach in Section 2.3; estimate each ζ_{gu} by OLS to obtain $\widehat{\zeta}_g$; and form predicted values $q_g(X_i, \widehat{\zeta}_g)$, $i = 1, \ldots, n$.

(3) Using the predicted values from step (2), form

$$\widehat{\ell}^*(Y_i, X_i, Z_i) = \widehat{\ell}(Y_i, Z_i) - \sum_{g=1}^{k}\{I(Z_i = g) - \widehat{\pi}_g\}q_g(X_i, \widehat{\zeta}_g) \tag{2.21}$$

and substitute these values into (2.18). Estimate Σ^* in (2.18) by $\widehat{\Sigma}^* = n^{-1}\sum_{i=1}^n \widehat{\ell}^*(Y_i, X_i, Z_i)^{\otimes 2}$. Compare the resulting test statistic $\widehat{T}_n^*$ to the χ_s^2 distribution.

As in Section 2.3, there is no effect asymptotically of estimating ζ_g and π_g, $g = 1, \ldots, k$, so that $\widehat{T}_n^*$ will achieve the same power asymptotically as if the limits in probability of $\widehat{\zeta}_g$ and the true π_g were substituted. Notably, the test based on $\widehat{T}_n^*$ will be asymptotically more powerful than the corresponding unadjusted test against any sequence of alternatives.

The approach of Tangen and Koch (1999) to modifying the Wilcoxon test for two treatments is in a similar spirit to this general approach.

2.5 Details

2.5.1 Demonstration of the Existence of Joint Densities Satisfying the Semiparametric Model

At the end of Section 2.1, the semiparametric model framework within which we derive the proposed methods is stated, and is repeated here for convenience.

The data from a clinical trial are denoted by (Y_i, X_i, Z_i), $i = 1, \ldots, n$, assumed iid across i, where Y denotes the response of interest; X is a vector of baseline auxiliary covariates; and $Z = 1, \ldots, k$ depending on to which of k possible treatment groups subject i was randomized, with randomization probabilities $P(Z = g) = \pi_g$, $g = 1, \ldots, k$, $\sum_{g=1}^{k} \pi_g = 1$. Randomization guarantees that $Z \perp\!\!\!\perp X$. We assume that interest focuses on a parameter β involved in characterizing treatment comparisons, defined in the context of a model for the conditional density of Y given Z, $p_{Y|Z}(y|z; \theta, \eta)$, where $\theta = (\beta^T, \gamma^T)^T$. Here, γ represents a finite-dimensional vector of possible additional parameters, and η is a finite- or infinite-dimensional nuisance parameter required to describe fully the class of models under consideration. Examples of such models are discussed in Section 2.1.

The semiparametric model introduced in Section 2.1 consists of all joint densities $p_{Y,X,Z}(y, x, z\,;\, \theta, \eta, \psi, \pi) = p_{Y,X|Z}(y, x \,|\, z; \theta, \eta, \psi) p_Z(z; \pi)$ such that the conditional densities for (Y, X) given Z, denoted $p_{Y,X|Z}(y, x \,|\, z; \theta, \eta, \psi)$, satisfy

$$\text{(i)} \qquad \int p_{Y,X|Z}(y, x \,|\, z; \theta, \eta, \psi)\, dx = p_{Y|Z}(y|z; \theta, \eta), \tag{2.22}$$

$$\text{(ii)} \qquad \int p_{Y,X|Z}(y, x \,|\, z; \theta, \eta, \psi)\, dy = p_X(x), \tag{2.23}$$

where $p_X(x)$ refers to any arbitrary marginal density for the auxiliary covariates, and (ii) follows because $Z \perp\!\!\!\perp X$. The additional nuisance parameters ψ and η together are used to specify conditional densities satisfying (i) and (ii).

We now demonstrate that joint distributions for (Y, X, Z) satisfying conditions (i) and (ii) in (2.22) and (2.23) may be constructed. For simplicity, we consider scalar Y and a two-armed trial, $k = 2$; extensions to vector Y and arbitrary $k \geq 2$ are straightforward. Begin with a given marginal density $p_X(x)$ for X and the conditional density $p_{Y|Z}(y|z; \theta, \eta)$ of interest. A joint distribution for (Y, X, Z) may then be developed through the following steps:

1. Generate X from $p_X(x)$.

2. Generate W_0 and W_1 from any arbitrary conditional densities $p_{W_0|X}(w_0 \,|\, x)$ and

$p_{W_1|X}(w_1 \,|\, x)$, where, clearly, W_0 and W_1 have marginal densities

$$p_{W_k}(w_k) = \int p_{W_k|X}(w_k \,|\, x) p_X(x)\, dx, \quad k = 1, 2.$$

A transformation of these variables is used in step 4 below to derive the response variable Y.

3. Generate a random Bernoulli Z random variable (taking values 1 and 2) independently of X, W_0, and W_1, with "success" probability $P(Z = 2) = \pi_2$. (Here, $P(Z = 1) = \pi_1 = 1 - \pi_2$.)

4. Let $F_{W_k}(u) = P(W_k \leq u)$, $k = 1, 2$, be the cumulative distribution functions (cdfs) for W_k, $k = 1, 2$, and write $F_{Y|Z=k}(u; \theta, \eta) = P(Y \leq u \,|\, Z = k; \theta, \eta)$, $k = 1, 2$, the cdfs corresponding to $p_{Y|Z}(y|z; \theta, \eta)$. Generate Y as

$$Y = I(Z = 1) F^{-1}_{Y|Z=1}\{F_{W_1}(W_1); \theta, \eta\} + I(Z = 2) F^{-1}_{Y|Z=2}\{F_{W_2}(W_2); \theta, \eta\}.$$

This construction guarantees that $Z \perp\!\!\!\perp X$, that the conditional distribution of Y given Z has the required density $p_{Y|Z}(y|z; \theta, \eta)$, and allows for flexible relationships for (Y, X) given Z. The derivation may be generalized straightforwardly to vector Y, as in the case of longitudinal response.

2.5.2 Derivation of Estimating Functions for Treatment Effect

We consider the semiparametric framework given at end of Section 2.1, restated in Section 2.5.1, and apply the principles of semiparametric theory to derive Equation (2.12).

Before we present the detailed argument, we summarize the general approach. Under the semiparametric theory perspective, one views estimating functions as elements of the Hilbert space $\mathcal{H}$ consisting of all functions $h(Y, X, Z)$ such that $E\{h(Y, X, Z)\} = 0$ and $E\{h(Y, X, Z)^T h(Y, X, Z)\} < \infty$ (e.g., Tsiatis, 2006, Chapter 2). The advantage of considering estimating functions as elements of $\mathcal{H}$ is that geometric principles may be used to derive the form of all estimating functions and to

assess the relative efficiencies of the estimators corresponding to them. The derivation makes use of the critical result that, under suitable regularity conditions, all estimating functions for estimators of finite-dimensional parameters of interest in semiparametric models are orthogonal to the so-called nuisance tangent space, a certain linear subspace of $\mathcal{H}$ (see Tsiatis, 2006, Chapter 4, for general discussion). Thus, the argument involves characterizing the nuisance tangent space and its orthogonal complement, in which estimating functions for a particular problem lie. In our case, then, the key to deriving semiparametric estimators for the parameter θ in our framework is to describe the nuisance tangent space and find the form of elements in its orthogonal complement, which will be of the form of estimating functions used to construct estimating equations for θ based on (Y, X, Z).

We remark that the argument given in the rest of this Section is not simply a special case of a general theory. Rather, it results from applying the semiparametric theory perspective above to this problem. The argument subsumes and represents a significant advance beyond that given by Leon et al. (2003) and in Appendix A.2 of Davidian, Tsiatis, and Leon (2005) for the particular case of estimating the difference of $k = 2$ treatment means, β_2, defined in Equation (2.1).

We present the argument for the particular case scalar Y and infinite dimensional η; similar developments are possible in the cases of multivariate Y and null η. Formally, the nuisance tangent space we seek is defined as the mean-square closure of parametric submodel nuisance tangent spaces. A parametric submodel is a finite-dimensional parametric model that

(a) Is contained in the semiparametric model and

(b) Contains the truth; i.e., the distribution that generates the data.

The parametric submodel nuisance tangent space is the space spanned by the nuisance score vector of the parametric submodel. Here, we denote such a parametric submodel as

$$p_{Y,X|Z}(y, x|z; \theta, \xi_\eta, \xi_\psi),$$

satisfying conditions analogous to (2.22) and (2.23), i.e.,

$$\text{(i)} \qquad \int p_{Y,X|Z}(y,x|z;\theta,\xi_\eta,\xi_\psi)dx = p_{Y|Z}(y|z;\theta,\xi_\eta) \tag{2.24}$$

$$\text{(ii)} \qquad \int p_{Y,X|Z}(y,x|z;\theta,\xi_\eta,\xi_\psi)dy = p_X(x;\theta,\xi_\eta,\xi_\psi), \tag{2.25}$$

where $p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0}) = p_{0Y,X|Z}(y,x|z)$, the true density of Y,X given Z (i.e., that generating the data); ξ_η is a finite-dimensional parameter defined so that $p_{Y|Z}(y|z;\theta,\xi_\eta)$ is a parametric submodel for the semiparametric model $p_{Y|Z}(y|z;\theta,\eta)$; and ξ_ψ is an additional finite dimensional parameter describing the joint distribution of Y and X given Z such that (i) and (ii) in (2.24) and (2.25) are satisfied.

The parametric submodel nuisance tangent space is the space spanned by the nuisance score vector $\{S^T_{\xi_\eta}(Y,X,Z), S^T_{\xi_\psi}(Y,X,Z)\}^T$, where

$$S_{\xi_\eta}(y,x,z) = \frac{\partial}{\partial \xi_\eta}\log\{p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\},$$

and similarly for $S_{\xi_\psi}(Y,X,Z)$. Thus, the parametric submodel nuisance tangent space is made up of elements $\{B_1 S_{\xi_\eta}(Y,X,Z) + B_2 S_{\xi_\psi}(Y,X,Z)\}$, where B_1 and B_2 are conformable matrices.

Denote the nuisance tangent space for the semiparametric model $p_{Y|Z}(y|z;\theta,\eta)$ by Λ_η. By definition, any element spanned by the parametric-submodel nuisance score vector $S^*_{\xi_\eta}(Y,Z)$, where $S^*_{\xi_\eta}(y,z) = \dfrac{\partial}{\partial \xi_\eta}\log\{p_{Y|Z}(y|z;\theta_0,\xi_{\eta_0}\}$, is an element of Λ_η.

From (2.24), we get

$$\log\left\{\int p_{Y,X|Z}(y,x|z;\theta,\xi_\eta,\xi_\psi)dx\right\} = \log\left\{p_{Y|Z}(y|z;\theta,\xi_\eta)\right\} \tag{2.26}$$

so that

$$\frac{\partial}{\partial \xi_\eta}\log\left\{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx\right\} = \frac{\partial}{\partial \xi_\eta}\log\left\{p_{Y|Z}(y|z;\theta_0,\xi_{\eta_0})\right\}$$

and thus

$$\begin{aligned} B_1\frac{\partial}{\partial \xi_\eta}\log\left\{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx\right\} &= B_1\frac{\partial}{\partial \xi_\eta}\log\left\{p_{Y|Z}(y|z;\theta_0,\xi_{\eta_0})\right\} \\ &= B_1 S^*_{\xi_\eta}(y,z). \end{aligned}$$

Under regularity conditions, the left-hand side of the above equation is equal to

$$
\begin{aligned}
& B_1 \frac{\int \frac{\partial}{\partial \xi_\eta} p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx}{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx} \\
& = B_1 \frac{\int \frac{\partial}{\partial \xi_\eta} \log\{p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\} p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx}{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx} \\
& = B_1 E\{S_{\xi_\eta}(Y,X,Z)|Y=y,Z=z\}.
\end{aligned}
$$

Thus, $B_1 E\{S_{\xi_\eta}(Y,X,Z)|Y,Z\} = B_1 S^*_{\xi_\eta}(Y,Z) \in \Lambda_\eta$.

Similarly, taking the derivative of both sides of (2.26) with respect to ξ_ψ and evaluating them at the truth, we get $\frac{\partial}{\partial \xi_\psi} \log\left\{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx\right\} = 0$, which leads to $B_2 E\{S_{\xi_\psi}(Y,X,Z)|Y,Z\} = 0$.

Combining the arguments above, it follows that any element in the submodel nuisance tangent space, $h(Y,X,Z) = B_1 S_{\xi_\eta}(Y,X,Z) + B_2 S_{\xi_\psi}(Y,X,Z)$, must satisfy the condition

$$E\{h(Y,X,Z)|Y,Z\} \in \Lambda_\eta.$$

From (2.25), we get

$$\log\left\{\int p_{Y,X|Z}(y,x|z;\theta,\xi_\eta,\xi_\psi)dy\right\} = \log\{p_X(x;\theta,\xi_\eta,\xi_\psi)\} \tag{2.27}$$

$$
\begin{aligned}
& \frac{\partial}{\partial \xi_\eta} \log\left\{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dy\right\} = \frac{\partial}{\partial \xi_\eta} \log\{p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\} \\
& \frac{\int \frac{\partial}{\partial \xi_\eta} \log\{p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\} p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dy}{\int p_{Y,X|Z}(y,x|z;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dy} \\
& = \frac{\partial}{\partial \xi_\eta} \log\{p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\},
\end{aligned}
$$

and $E\{S_{\xi_\eta}(Y,X,Z)|X,Z\} = S^*_{\xi_\eta}(X)$, where $S^*_{\xi_\eta}(x) = \frac{\partial}{\partial \xi_\eta} \log\{p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\}$. $S^*_{\xi_\eta}(X)$ has expectation 0 by the following argument. We have $\int p_X(x;\theta,\xi_\eta,\xi_\psi)dx = 1$,

which implies that

$$\frac{\partial}{\partial \xi_\eta} \int p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx = \int \frac{\partial}{\partial \xi_\eta} p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx = 0,$$

so that

$$\int \frac{\partial}{\partial \xi_\eta} \log\{p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\} p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})dx = 0,$$

which implies that $E\{S^*_{\xi_\eta}(X)\} = 0$, as required.

Similarly, taking the derivative of both sides of (2.27) with respect to ξ_ψ and evaluating them at the truth, we get $E\{S_{\xi_\psi}(Y,X,Z)|X,Z\} = S^*_{\xi_\psi}(X)$, where

$$S^*_{\xi_\psi}(x) = \frac{\partial}{\partial \xi_\psi} \log\left\{p_X(x;\theta_0,\xi_{\eta_0},\xi_{\psi_0})\right\},$$

and $E\{S^*_{\xi_\psi}(X)\} = 0$.

Therefore, if we define Λ_x as the space of all mean zero functions of X, i.e., $\Lambda_x = \{h(X): \ E\{h(X)\} = 0\}$, then any element in the submodel nuisance tangent space, $h(Y,X,Z) = B_1 S_{\xi_\eta}(Y,X,Z) + B_2 S_{\xi_\psi}(Y,X,Z)$, must also satisfy the condition:

$$E\{h(Y,X,Z)|X,Z\} \in \Lambda_x.$$

To summarize, we have demonstrated that any element $h(Y,X,Z)$ that is spanned by the score vector $\{S^T_{\xi_\eta}(Y,X,Z), S^T_{\xi_\psi}(Y,X,Z)\}^T$ must satisfy

$$\text{(a).} \quad E\{h(Y,X,Z)|Y,Z\} \in \Lambda_\eta, \tag{2.28}$$

$$\text{(b).} \quad E\{h(Y,X,Z)|X,Z\} \in \Lambda_x. \tag{2.29}$$

With these relationships in mind, we conjecture that the nuisance tangent space consists of all functions $h(Y,X,Z)$ satisfying conditions (a) and (b) given in (2.28) and (2.29). We denote such as a space by $\Lambda^{(conj)}$. We can easily show that the space of functions satisfying (2.28) is given by $\Lambda_\eta + \Lambda_1$, where

$$\Lambda_1 = \{h_1(Y,X,Z): \ E\{h_1(Y,X,Z|Y,Z)\} = 0\}, \tag{2.30}$$

and the space of functions satisfying (2.29) is given by $\Lambda_x + \Lambda_2$, where

$$\Lambda_2 = \{h_2(Y,X,Z): \ E\{h_2(Y,X,Z|X,Z)\} = 0\}. \tag{2.31}$$

Consequently, the conjectured nuisance tangent space is $\Lambda^{(conj)} = (\Lambda_\eta + \Lambda_1) \cap (\Lambda_x + \Lambda_2)$. We have already proven that any element in a parametric submodel nuisance tangent space must belong to $\Lambda^{(conj)}$; in addition, it can be shown that the space $\Lambda^{(conj)}$ is closed. Therefore, the nuisance tangent space $\Lambda \subset \Lambda^{(conj)}$.

To prove that $\Lambda^{(conj)}$ is truly the nuisance tangent space, we need to show that any element in $\Lambda^{(conj)}$ can be represented as some element or a limit of elements from some parametric submodel nuisance tangent spaces. Consider some arbitrary bounded element $h(Y, X, Z) \in \Lambda^{(conj)}$; namely,

$$h(Y, X, Z) = h_\eta(Y, Z) + h_1(Y, X, Z) = h_x(X) + h_2(Y, X, Z) \tag{2.32}$$

for some $h_\eta(Y, Z) \in \Lambda_\eta$, $h_1(Y, X, Z) \in \Lambda_1$, $h_x(X) \in \Lambda_x$, and $h_2(Y, X, Z) \in \Lambda_2$. We will construct the parametric submodels in three steps.

Step 1. Because $h_\eta(Y, Z) \in \Lambda_\eta$, $h_\eta(Y, Z)$ is either the corresponding score vector of some parametric submodel $p_{Y|Z}(y, |z; \theta_0, \xi_\eta)$ or the limiting score vector of a sequence of parametric submodels $p_{Y|Z}(y, |z; \theta_0, \xi_{\eta j})$. For simplicity, we first assume the former case; that is,

$$\frac{\partial}{\partial \xi_\eta} \log\{p_{Y|Z}(y|z; \theta_0, \xi_{\eta_0})\} = h_\eta(y, z).$$

Without loss of generality, we can assume that this submodel contains the truth when $\xi_\eta = \xi_{\eta_0} = 0$. From here on θ is taken to equal θ_0 (the truth) and hence will be suppressed in the notation.

We begin by considering an approximation to the parametric-submodel given by

$$p^*_{Y,X|Z}(y, x|z, \xi_\eta) = p_{0Y,X|Z}(y, x|z)[1 + \xi_\eta^T\{h_\eta(y, z) + h_1(y, x, z)\}]. \tag{2.33}$$

This is a proper density function as long as ξ_η is chosen sufficiently close to 0, and it contains the truth if ξ_η is chosen to be 0. By construction, the score vector is given by $\{h_\eta(Y, Z) + h_1(Y, X, Z)\}$. We may show that the submodel given by (2.33) satisfies

condition (ii) as follows:

$$\begin{aligned}
&\int p^*_{Y,X|Z}(y,x|z,\xi_\eta)dy = \int p_{0Y,X|Z}(y,x|z)[1+\xi_\eta^T\{h_\eta(y,z)+h_1(y,x,z)\}]dy \\
= &\int p_{0Y,X|Z}(y,x|z)[1+\xi_\eta^T\{h_x(x)+h_2(y,x,z)\}]dy \\
= &\; p_{0X}(x)+\xi_\eta^T h_x(x)p_{0X}(x)+\xi_\eta^T p_{0X}(x)\int h_2(y,x,z)p_{0Y|X,Z}(y|x,z)dy \\
= &\; p_{0X}(x)\{1+\xi_\eta^T h_X(x)\},
\end{aligned}$$

where the last equality follows because, by definition, the conditional expectation of $h_2(Y,X,Z)$ given (X,Z) is 0. This argument shows that X and Z are independent by this submodel. Therefore, this submodel satisfies condition (ii).

However, this submodel does not satisfy condition (i), because

$$\begin{aligned}
&\int p^*_{Y,X|Z}(y,x|z,\xi_\eta)dx = \int p_{0Y,X|Z}(y,x|z)[1+\xi_\eta^T\{h_\eta(y,z)+h_1(y,x,z)\}]dx \\
= &\; p_{0Y|Z}(y|z)+p_{0Y|Z}\xi_\eta^T[h_\eta(y,z)+E\{h_1(Y,X,Z)|Y=y,Z=z\}] \\
= &\; p_{0Y|Z}(y|z)\{1+\xi_\eta^T h_\eta(y,z)\} \neq p_{Y|Z}(y|z,\xi_\eta).
\end{aligned}$$

Step 2. In order to derive a model that satisfies conditions (i) and (ii) while still leading to the same score vector, we consider the following construction. Take the random vector (V,X,Z) which has density $p^*_{Y,X|Z}(v,x|z,\xi_\eta)$ as defined by (2.33). The idea is to perturb the random variable V slightly to ensure that the transformed random variable Y has conditional density $p_{Y|Z}(y|z,\xi_\eta)$ while not affecting the independence of X and Z or the score vector. Toward that end, define

$$(Y,X,Z) = \{G(V,Z,\xi_\eta),X,Z\}, \tag{2.34}$$

where $G\{V,Z,\xi_\eta\} = F_2^{-1}\{F_1(V|Z,\xi_\eta)|Z,\xi_\eta\}$; $F_1(y|Z,\xi_\eta)$ is the cdf for $p_{0Y|Z}(y|z)\{1+\xi_\eta^T h_\eta(y,z)\}$; and $F_2(y|Z,\xi_\eta)$ is the cdf for $p_{Y|Z}(y|z,\xi_\eta)$. By construction, the conditional distribution of Y given Z is $p_{Y|Z}(y|z,\xi_\eta)$, and the conditional density of X given Z does not change, i.e., $X \perp\!\!\!\perp Z$. Therefore, by construction, this submodel satisfies conditions (i) and (ii). In addition, when $\xi_\eta = 0$, $G\{V,Z,\xi_\eta=0\} = V$, and $p_{Y,X|Z}(y,x|z,\xi_\eta=0) = p_{0Y,X|Z}(y,x|z)$; i.e., contains the truth.

Next, we derive the density of $(Y, X|Z)$, i.e., $p_{Y,X|Z}(y, x|z, \xi_\eta)$, and show that the score vector of this density is still $\{h_\eta(Y, Z) + h_1(Y, X, Z)\}$ as required. As $Y = F_2^{-1}\{F_1(V|Z, \xi_\eta)|Z, \xi_\eta\}$, we obtain $V = F_1^{-1}\{F_2(Y|Z, \xi_\eta)|Z, \xi_\eta\}$. Consequently,

$$\frac{dV}{dY} = \frac{p_{Y|Z}(y|z, \xi_\eta)}{p_{0Y|Z}(v|z)\{1 + \xi_\eta^T h_\eta(v, z)\}}.$$

Using the change of variable formula, the density of $(Y, X|Z)$ is

$$\begin{aligned}
&p_{Y,X|Z}(y, x|z, \xi_\eta) = p_{0Y,X|Z}(v, x|z)[1 + \xi_\eta^T\{h_\eta(v, z) + h_1(v, x, z)\}]\frac{dV}{dY} \\
&= p_{0Y|Z}(v|z)p_{0X|Y,Z}(x|v, z)[1 + \xi_\eta^T\{h_\eta(v, z) + h_1(v, x, z)\}\frac{p_{Y|Z}(y|z, \xi_\eta)}{p_{0Y|Z}(v|z)\{1 + \xi_\eta^T h_\eta(v, z)\}} \\
&= p_{0X|Y,Z}(x|v, z)p_{Y|Z}(y|z, \xi_\eta)\frac{1 + \xi_\eta^T\{h_\eta(v, z) + h_1(v, x, z)\}}{\{1 + \xi_\eta^T h_\eta(v, z)\}}, \qquad (2.35)
\end{aligned}$$

where $v = F_1^{-1}\{F_2(y|z, \xi_\eta)|z, \xi_\eta\}$.

Now, we will derive the score vector of $p_{Y,X|Z}(y, x|z, \xi_\eta)$.

$$\begin{aligned}
&\frac{\partial}{\partial \xi_\eta}\log\{p_{Y,X|Z}(y, x|z, \xi_{\eta_0})\} = \frac{\frac{\partial}{\partial v}\{p_{0X|Y,Z}(x|v, z)\} \cdot \frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0}}{p_{0X|Y,Z}(x|y, z)} \\
&+\frac{\partial}{\partial \xi_\eta}\log\{p_{Y|Z}(y|z, \xi_{\eta_0})\} \\
&+\frac{h_\eta(v, z) + h_1(v, x, z) + \xi_\eta^T\frac{\partial}{\partial \xi_\eta}\{h_\eta(v, z) + h_1(v, x, z)\}}{1 + \xi_\eta^T\{h_\eta(v, z) + h_1(v, x, z)\}}|_{\xi_\eta=0} \\
&-\frac{h_\eta(v, z) + \xi_\eta^T\frac{\partial h_\eta(v, z)}{\partial \xi_\eta}}{1 + \xi_\eta^T h_\eta(v, z)}|_{\xi_\eta=0}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\frac{\partial}{\partial v}\{p_{0X|Y,Z}(x|v, z)\} \cdot \frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0}}{p_{0X|Y,Z}(x|y, z)} + h_\eta(y, z) + h_\eta(y, z) + h_1(y, x, z) - h_\eta(y, z) \\
&= h_\eta(y, z) + h_1(y, x, z) + \frac{\frac{\partial}{\partial v}\{p_{0X|Y,Z}(x|v, z)\} \cdot \frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0}}{p_{0X|Y,Z}(x|y, z)}.
\end{aligned}$$

In the above argument, we have used the facts that when $\xi_\eta = 0$, $v = y$, and that $\frac{\partial}{\partial \xi_\eta} \log\{p_{Y|Z}(y|z, \xi_{\eta_0})\} = h_\eta(y, z)$. Note that if $\frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} = 0$, then the score vector is $h_\eta(y, z) + h_1(y, x, z)$ as needed. So in the following we will show that $\frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} = 0$.

First note, under suitable regularity conditions,

$$\frac{\partial F_2(y|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0} = \frac{\partial}{\partial \xi_\eta}\int_{-\infty}^{y} p_{Y|Z}(u|z,\xi_\eta)du|_{\xi_\eta=0} = \int_{-\infty}^{y} \frac{\partial}{\partial \xi_\eta} p_{Y|Z}(u|z,\xi_\eta)|_{\xi_\eta=0}\, du$$
$$= \int_{-\infty}^{y} p_{0Y|Z}(u|z)\frac{\partial}{\partial \xi_\eta}\log\{p_{Y|Z}(u|z,\xi_\eta)\}|_{\xi_\eta=0}du = \int_{-\infty}^{y} p_{0Y|Z}(u|z)h_\eta(u,z)du.$$

Similarly,

$$\begin{aligned}\frac{\partial F_1(y|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0} &= \int_{-\infty}^{y} \frac{\partial}{\partial \xi_\eta} p_{0Y|Z}(u|z)\{1 + \xi_\eta^T h_\eta(u|z)\}|_{\xi_\eta=0}\, du \\ &= \int_{-\infty}^{y} p_{0Y|Z}(u|z)h_\eta(u,z)\, du.\end{aligned}$$

Consequently, $\frac{\partial F_1(y|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0} = \frac{\partial F_2(y|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0}$.

By construction, $y = F_2^{-1}\{F_1(v|z,\xi_\eta)|z,\xi_\eta\}$. Thus $F_2(y|z,\xi_\eta) = F_1(v|z,\xi_\eta)$, and it follows that

$$\frac{\partial F_2(y|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0} = \frac{\partial F_1(v|z,\xi_\eta)}{\partial v}\cdot\frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} + \frac{\partial F_1(v|z,\xi_\eta)}{\partial \xi_\eta}|_{\xi_\eta=0}.$$

Notice that, when $\xi_\eta = 0$, we have $v = y$, and $\frac{F_2(y|z,\xi_\eta)}{\xi_\eta}|_{\xi_\eta=0} = \frac{F_1(y|z,\xi_\eta)}{\xi_\eta}|_{\xi_\eta=0}$, so that

$$\frac{\partial F_1(v|z,\xi_\eta)}{\partial v}\cdot\frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} = 0,$$

and it follows that

$$p_{0Y|Z}(y|z)\cdot\frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} = 0, \quad \text{which implies} \quad \frac{\partial v}{\partial \xi_\eta}|_{\xi_\eta=0} = 0.$$

Therefore, we have constructed a submodel given by (2.35) that satisfies conditions (i) and (ii). In addition, this submodel has a score vector equal to $h_\eta(Y, Z) + h_1(Y, X, Z) = h_x(X) + h_2(Y, X, Z)$, which is arbitrarily chosen from the conjecture space.

Step 3. Recall that in the above arguments, we have assumed that $h_\eta(Y,Z)$ is the score vector of some parametric submodel $p_{Y|Z}(y|z;\theta_0,\xi_\eta)$. Under this assumption, it has been demonstrated that a bounded element $h(Y,X,Z)$ can be represented as an element from some parametric submodel nuisance tangent space. More generally, $h_\eta(Y,Z)$ may be the limit of score vectors of a sequence of parametric submodels $p_{jY|Z}(y|z;\theta_0,\xi_{\eta j})$; i.e.,

$$\lim_{j\to\infty}\frac{\partial}{\partial\xi_\eta}\log\{p_{jY|Z}(y|z;\theta_0,\xi_{\eta_0})\}=h_\eta(y,z).$$

In this situation, almost identical arguments to those used above can be used to construct a sequence of submodels that satisfies condition (i) and (ii), and also the limit of the corresponding score vectors is $h_\eta(Y,Z)$.

Combining the above arguments, we have shown that any bounded element in $\Lambda^{(conj)}$ can be represented as some element or a limit of elements from some parametric submodel nuisance tangent spaces. As any element in $\Lambda^{(conj)}$ is either bounded or limit of bounded elements, it follows that the nuisance tangent space is

$$\Lambda=\Lambda^{(conj)}=(\Lambda_\eta+\Lambda_1)\cap(\Lambda_x+\Lambda_2). \tag{2.36}$$

As we argued earlier, estimating functions used to derive estimating equations that lead to semiparametric estimators for θ are orthogonal to the nuisance tangent space. Therefore, we now derive the orthogonal complement to the nuisance tangent space. To do so, we use result that that the orthogonal complement of the sum of two linear spaces is equal to the intersection of the orthogonal complements. That is, if H_1, H_2 are closed linear subspaces contained in the Hilbert space $\mathcal{H}$, then

$$(H_1+H_2)^\perp=H_1^\perp\cap H_2^\perp, \tag{2.37}$$

We derived the nuisance tangent space Λ as given in (2.36). Therefore, using (2.37) we thus have that the orthogonal complement of the nuisance tangent space is

$$\Lambda^\perp=(\Lambda_\eta^\perp\cap\Lambda_1^\perp)+(\Lambda_x^\perp\cap\Lambda_2^\perp). \tag{2.38}$$

We examine the components making up the sum of spaces in (2.38) separately.

Complement of Λ_1. This space is given by

$$\Lambda_1^{\perp} = \{h(Y,Z), E\{h(Y,Z)\} = 0\}. \tag{2.39}$$

Proof: Suppose $E\{h(Y,Z)\} = 0$, and $h_1(Y,X,Z) \in \Lambda_1$, i.e., $E\{h_1(Y,X,Z|Y,Z)\} = 0$. Then

$$\begin{aligned} E\{h_1^T(Y,X,Z)h(Y,Z)\} &= E[E\{h_1^T(Y,X,Z)h(Y,Z)|Y,Z\}] \\ &= E[h^T(Y,Z)E\{h_1(Y,X,Z|Y,Z)\}] = 0. \end{aligned}$$

The last equality follows as $E\{h_1(Y,X,Z|Y,Z)\} = 0$ by assumption. Therefore, any mean-zero function of (Y,Z), $h(Y,Z)$, is orthogonal to Λ_1.

To finish the proof, we also must prove that any $h \in \mathcal{H}$ can be written as $h_1 + h_2$, where $h_1 \in \Lambda_1$, and $h_2 \in \{h(Y,Z), E\{h(Y,Z)\} = 0\}$. For any $h(Y,X,Z)$, construct $h_2(Y,Z) = E\{h(Y,X,Z)|Y,Z\}$, and $h_1(Y,X,Z) = h(Y,X,Z) - h_2(Y,Z)$. It is easy to verify that the constructed $h_1 \in \Lambda_1$ and $h_2 \in \{h(Y,Z), E\{h(Y,Z)\} = 0\}$.

Complement of $(\Lambda_\eta^{\perp} \cap \Lambda_1^{\perp})$. Because $\Lambda_1^{\perp} = \{h(Y,Z), E\{h(Y,Z)\} = 0\}$, then the space $(\Lambda_\eta^{\perp} \cap \Lambda_1^{\perp})$ consists of all elements which belong to the Hilbert space $\mathcal{H}_{Y,Z} = \{h(Y,Z) : E\{h(Y,Z)\} = 0\}$, i.e., all mean zero functions of (Y,Z), that are orthogonal to the nuisance tangent space Λ_η. This is precisely the orthogonal complement of the nuisance tangent space for the parametric submodel $p_{Y|Z}(y|z;\theta,\eta)$. Consequently, the space $(\Lambda_\eta^{\perp} \cap \Lambda_1^{\perp})$ is the space from which estimating functions for θ are derived without the consideration of the auxiliary covariates. Therefore, we call this space the space of estimating functions $\mathcal{E} = (\Lambda_\eta^{\perp} \cap \Lambda_1^{\perp})$.

Complement of $(\Lambda_x^{\perp} \cap \Lambda_2^{\perp})$. The same techniques used to find the space $\Lambda_1^{\perp}$ may be used here to prove that $\Lambda_2^{\perp} = \{h(X,Z) : E\{h(X,Z)\} = 0\}]$. Consequently, $(\Lambda_x^{\perp} \cap \Lambda_2^{\perp})$ consists of all mean-zero functions of X and Z that are orthogonal to functions of X. That is,

$$(\Lambda_x^{\perp} \cap \Lambda_2^{\perp}) = \{h(X,Z) : E\{h(X,Z)|X\} = 0\}. \tag{2.40}$$

Proof : if $E\{h(X,Z)|X\} = 0$, and $h_x(X) \in \Lambda_x$, then

$$E\{h^T(X,Z)h_x(X)\} = E[E\{h^T(X,Z)|X\}h_x(X)] = 0.$$

That is, $h(X,Z)$ is orthogonal to Λ_x. Moreover, any mean-zero functions of (X,Z) can be written as $h_1 + h_2$, where $h_1 \in \Lambda_x$, and $h_2 \in \{h(X,Z) : E\{h(X,Z)|X\} = 0\}$. For any $h(X,Z)$ which has mean zero, construct $h_1(X) = E\{h(X,Z)|X\}$, and $h_2(X,Z) = h(X,Z) - h_1(X)$. Clearly $E\{h_1(X)\} = 0$, and $E\{h_2(X,Z)|X\} = 0$ as required.

We refer to this space as the Augmentation Space, denoted by $\mathcal{A}$. Therefore, we have shown that the space orthogonal to the nuisance tangent space is given by $\Lambda^{\perp} = \mathcal{E} + \mathcal{A}$. Thus, the class of estimating functions for θ (and hence β) based on all the data (Y,X,Z) lies in this space, so that an estimating function for θ may be written as the sum of an estimating function $m(Y,Z;\theta)$ based on (Y,Z) alone (an element of $\mathcal{E}$) and an element of $\mathcal{A}$.

Accordingly, we characterize elements of $\mathcal{A}$. All functions of X and Z may be written as $\sum_{g=1}^{k} I(Z=g)a_g(X)$ for arbitrary functions $a_g(X)$, $g = 1,\ldots,k$. Thus, we can write any function $h(X,Z)$ satisfying $E\{h(X,Z)|X\} = 0$ as

$$\begin{aligned} h(X,Z) &= \sum_{g=1}^{k} I(Z=g)a_g(X) - E\left\{ \sum_{g=1}^{k} I(Z=g)a_g(X) \middle| X \right\} \\ &= \sum_{g=1}^{k} \{I(Z=g) - \pi_g\} a_g(X). \end{aligned} \tag{2.41}$$

Thus the form of all estimating functions is

$$m(Y,Z;\theta) + \sum_{g=1}^{k} \{I(Z=g) - \pi_g\} a_g(X),$$

which may be written equivalently in the form given in Equation (2.12).

2.5.3 Derivation of Optimal Estimating Function

The choice of functions $a_g(X)$, $g = 1,\ldots,k$ resulting in the optimal estimator, i.e., an estimator solving Equation(2.12) such that its variance is as small as possible, may be deduced from Theorem 4.5 of Tsiatis (2006). Alternatively, we derive such $a_g(X)$ directly. By the principles in Chapter 3 of Tsiatis (2006), the element of $\mathcal{E} + \mathcal{A}$

with smallest variance for a given $m(Y,Z;\theta) \in \mathcal{E}$ is the projection of $m(Y,Z;\theta)$ onto $\mathcal{A}$. Thus, we wish to find $a_g^*(X)$, $g = 1,\ldots,k$, such that

$$E\left(\left[m(Y,Z;\theta) - \sum_{g=1}^{k}\{I(Z=g) - \pi_g\}a_g^*(X)\right]\left[\sum_{g=1}^{k}\{I(Z=g) - \pi_g\}a_g(X)\right]\right) = 0$$

for all $a_g(X)$, $g = 1,\ldots,k$. Taking $a_g(X) = 0$ for $g \neq j$, we thus wish to find $a_g^*(X)$, $g = 1,\ldots,k$, such that

$$E\left(\left[E\{m(Y,Z;\theta)|X,Z\} - \sum_{g=1}^{k}\{I(Z=g) - \pi_g\}a_g^*(X)\right]\{I(Z=j) - \pi_j\}\middle| X\right) = 0 \tag{2.42}$$

for each $j = 1,\ldots,k$. It is straightforward to show that, writing $E\{m(Y,Z;\theta)|X,Z\} = \sum_{g=1}^{k} I(Z=g)E\{m(Y,g;\theta)|X,Z=g\}$, (2.42) implies that we must have

$$E\{m(Y,j;\theta)|X,Z=j\} - a_j^*(X) - \sum_{g=1}^{k}\left[E\{m(Y,g)|X,Z=g\} - a_g^*(X)\right]\pi_g = 0 \tag{2.43}$$

for all $j = 1,\ldots,k$. Expression (2.43) is satisfied when

$$a_g^*(X) = E\{m(Y,g;\theta)|X,Z=g\}, \quad g = 1,\ldots,k,$$

yielding the estimating function in Equation (2.14).

Chapter 3

Empirical Studies and Applications (I)

3.1 Simulation Studies

3.1.1 Estimation

We report results of several simulations, each based on 5000 Monte Carlo data sets. Tsiatis et al. (2007, Section 6) carried out extensive simulations in the particular case of (2.1); thus, we focus here on estimation of quantities other than differences of treatment means.

In the first set of simulations, we considered $k = 2$, a binary response Y, and

$$\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 I(Z = 2), \tag{3.1}$$

so that β_2 is the log-odds ratio for treatment 2 relative to treatment 1, the parameter of interest; and $\theta = \beta = (\beta_1, \beta_2)^T$. For each scenario, we generated Z as Bernoulli with $P(Z = 1) = P(Z = 2) = 0.5$ and covariates $X = (X_1, \ldots, X_8)^T$ such that X_1, X_3, $X_8 \sim \mathcal{N}(0, 1)$; X_4 and X_6 were Bernoulli with $P(X_4 = 1) = 0.3$ and $P(X_6 = 1) = 0.5$; and $X_2 = 0.2X_1 + 0.98U_1$, $X_5 = 0.1X_1 + 0.2X_3 + 0.97U_2$, and $X_7 = 0.1X_3 + 0.99U_3$, where $U_\ell \sim \mathcal{N}(0, 1)$, $\ell = 1, 2, 3$. We then generated Y as Bernoulli according to $\text{logit}\{P(Y = 1|Z = g, X)\} = \alpha_{0g} + \alpha_g^T X$, $g = 1, 2$, with α_{0g} and α_g chosen to yield

mild, moderate, and strong association between Y and X within each treatment, as follows. Using the coefficient of determination R^2 to measure the strength of association, $R^2 = (0.18, 0.16)$ for treatments (1,2) in the "mild" scenario, with $(\alpha_{01}, \alpha_{02}) = (0.25, -0.8)$, $\alpha_1 = (0.8, 0.5, 0, 0, 0, 0, 0, 0)^T$, and $\alpha_2 = (0.3, 0.7, 0.3, 0.8, 0, 0, 0, 0)^T$; $R^2 = (0.32, 0.33)$ in the "moderate" scenario, with $(\alpha_{01}, \alpha_{02}) = (0.38, -0.8)$, $\alpha_1 = (1.2, 1.0, 0, 0, 0, 0, 0, 0)^T$, and $\alpha_2 = (0.5, 1.3, 0.5, 1.5, 0, 0, 0, 0)^T$; and $R^2 = (0.43, 0.41)$ in the "strong" scenario, with $(\alpha_{01}, \alpha_{02}) = (0.8, -0.8)$, $\alpha_1 = (1.5, 1.8, 0, 0, 0, 0, 0, 0)^T$ and $\alpha_2 = (1.0, 1.3, 0.8, 2.5, 0, 0, 0, 0)^T$. Thus, in all cases, $X_1, \ldots, X_4$ are covariates "important" for adjustment while $X_5, \ldots, X_8$ are "unimportant." For each data set, $n = 600$, and, we fitted (3.1) by IRWLS to (Y_i, Z_i), $i = 1, \ldots, n$, to obtain the unadjusted estimate of β. We also estimated β by the proposed methods using the direct implementation strategy, where the models $q_g^*(X, \zeta_g)$ for each $g = 1, 2$ in the augmentation term were developed six ways:

Aug. 1 $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) =$ "true," fit by OLS
Aug. 2 $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) = X$, fit by OLS
Aug. 3 $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) =$ "true," fit by IRWLS
Aug. 4 $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X) = X$, fit by IRWLS
Aug. 5 $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X)$ by OLS with forward selection
Aug. 6 $\text{logit}\{q_g^*(X, \zeta_g)\} = \{1, c_g^T(X)\}^T \zeta_g$, $c_g(X)$ by IRWLS with forward selection

where "true" means that $c_g(X)$ contained only X_ℓ, $\ell = 1, \ldots, 4$, for which the corresponding element of α_g was not zero (i.e., using the "true important covariates" for each g); and in Aug. 5 and 6 forward selection from linear terms in $X_1, \ldots, X_8$ for linear or logistic regression was used to determine each $q_g^*(X, \zeta_g)$, with entry criterion 0.05. Aug. 3, 4, and 6 demonstrate performance when nonlinear models and methods other than OLS are used. We also estimated β_2 by estimating ϕ in (2.7) via IRWLS two ways: Usual 1, where only the "important" covariates $X_1, \ldots, X_4$ were included in the model; and Usual 2, where the subset of $X_1, \ldots, X_8$ to include was identified via forward selection with entry criterion 0.05.

Table 3.1 shows modest to considerable gains in efficiency for the proposed estimators, depending on the strength of the association. The estimators are unbiased, and

associated confidence intervals achieve the nominal level. In contrast, the usual adjustment based on (2.7) leads to biased estimation of β_2, considerable efficiency loss, and unreliable intervals. This is a consequence of the fact that β_2 is an unconditional measure of treatment effect while ϕ is defined conditional on X; this distinction does not matter when the model for Y is linear but is important when it is nonlinear, as is (2.7) (see, e.g., Robinson et al., 1998).

In the second set of simulations, we again took $k = 2$ and focused on β_2, the difference in treatment slopes in the linear mixed model (2.4). In each scenario, we generated for each $i = 1, \ldots, n = 200$ Z_i as Bernoulli with $P(Z = 1) = P(Z = 2) = 0.5$; X_{1i}, X_{2i}, X_{3i} as above; and subject-specific intercept $\beta_{0i} = 0.5 + 0.2X_{1i} + 0.5X_{2i} + b_{0i}$ and slope $\beta_{1i} = \alpha_{0g} + \alpha_{1g}X_{1i}^2 + \alpha_{2g}X_{2i} + \alpha_{13}X_{3i} + b_{1i}$, where $(\alpha_{01}, \alpha_{02}) = (1.0, 1.3)$, $(b_{0i}, b_{1i})^T \sim \mathcal{N}(0, D)$, with $D_{11} = 1$, $D_{12} = 0.2$, and $D_{22} = 0.4$, so that $\text{corr}(b_{0i}, b_{1i}) = 0.5$. We generated $m_i = 9, 10, 11$ with equal probabilities; took $t_{ij} = 2(j - 1)$ for $j = 1, \ldots, m_i$; and generated $Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}$, $j = 1, \ldots, m_i$, where $e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2 = 16)$. Writing $\alpha_g = (\alpha_{1g}, \alpha_{2g}, \alpha_{3g})^T$, we took $\alpha_1 = (0.2, 0.2, 0)^T$ and $\alpha_2 = (0.2, 0, 0.2)^T$, yielding R^2 values between subject-specific slopes and covariates of $(0.11, 0.14)$ in the two groups, for "mild" association; $\alpha_1 = (0.13, 0.1, 0)^T$ and $\alpha_2 = (0.13, 0, 0.15)^T$, $R^2 = (0.24, 0.24)$, for "moderate" association; and $\alpha_1 = (0.28, 0.25, 0)^T$ and $\alpha_2 = (0.28, 0, 0.25)^T$, $R^2 = (0.36, 0.36)$, for "strong" association. For each data set, we obtained the unadjusted estimate for θ by fitting (2.4) using SAS `proc mixed` (SAS Institute, 2006). For (2.4), $m(Y, Z; \theta)$ has components of form (2.16) for α and β and more complicated components quadratic in Y for D and σ_e^2. For simplicity, because the estimators for (α, β) and (D, σ_e^2) are uncorrelated, we fixed D and σ_e^2 at the unadjusted analysis estimates in the components of $m(Y, Z; \theta)$ for (α, β), as asymptotically this will not impact precision of the estimators for (α, β), and used the direct implementation strategy based on the components for (α, β) only. We considered three variants on the proposed methods, all with each element of $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}\zeta_g$ fitted by OLS: Aug 1., taking $c_g(X) = (1, X_1^2, X_2, X_3)^T$, corresponding to the form of the true relationship; Aug 2., with $c_g(X) = (1, X_1, X_2, X_3)^T$, so not exploiting the quadratic relationship in X_1; and Aug 3., with $c_g(X) = (1, X_1, X_1^2, X_2, X_3)^T$, including an

unneeded linear effect of X_1. Writing now $X_i = (X_{1i}, X_{2i}, X_{3i})^T$, we also estimated β_2 by the estimate of ϕ from fitting via `proc mixed` the linear mixed model $Y_{ij} = \alpha_{00} + \alpha_{01}^T X_i + (\alpha_{10} + \alpha_{11}^T X_i + \phi Z_i) t_{ij} + b_{0i} + b_{1i} t_{ij} + e_{ij}$, denoted as Usual; such a model, with linear covariate effects only, might be prespecified in a trial protocol (e.g., Grouin et al., 2004). Table 3.2 shows that the proposed methods lead to relatively more efficient estimators when quadratic terms in X_1 are included in the $q_g^*(X, \zeta_g)$.

3.1.2 Testing

We carried out simulations based on 10,000 Monte Carlo data sets involving $k = 3$ and the Kruskal-Wallis test. For each data set, we generated for each of $n = 200$ or 400 subjects Z with $P(Z = g) = 1/3$, $g = 1, 2, 3$, and (Y, X) with joint distribution of (Y, X) given Z bivariate normal with mean $\{\beta_1 I(Z = 1) + \beta_2 I(Z = 2), 0\}^T$ and covariance matrix $\text{vech}(1, \rho, 1)$, where $\rho = 0.25, 0.50, 0.75$ corresponds to mild, moderate, and strong association between covariate and response. Under the null hypothesis, we set $\beta_1 = \beta_2 = 0$; simulations under the alternative involved $\beta_1 = 0.25$, $\beta_2 = 0.4$. For each data set, we calculated the unadjusted Kruskal-Wallis test statistic T_n and the proposed statistic $\widehat{T}_n^*$ using the strategy in Section 2.4, with each component of the $s = 2$-dimensional models $q_g(X, \zeta_g)$ in (2.21) represented as $q_{gu}(X, \zeta_{gu}) = \{1, c_{gu}^T(X)\}^T \zeta_{ug}$, $u = 1, 2$, $c_{gu}(X) = (X, X^2)^T$. Each statistic was compared to the 0.95 quantile of the χ_2^2 distribution. Table 3.3 shows that the proposed procedure yields greater power than the unadjusted test while achieving the nominal level, where the extent of improvement depends on the strength of the association between Y and X, as expected.

Table 3.1: Simulation results for estimation of the log-odds ratio β_2 for treatment $Z = 2$ relative to $Z = 1$ in (3.1) based on 5,000 Monte Carlo data sets. "Unadjusted" refers to the unadjusted estimator based on the data on (Y, Z) only, "Aug. a" for $a = 1, \ldots, 6$ refers to estimators based on the data on (Y, X, Z) using the strategy in Section 2.3, and "Usual b" for $b = 1, 2$ refers to direct logistic regression adjustment, as described in the text. MC bias is Monte Carlo bias, MC SD is Monte Carlo standard deviation, Ave. SE is the average of estimated standard errors obtained using the sandwich formula (2.13), Cov. Prob. is the MC coverage probability of 95% Wald confidence intervals, and Rel. Eff. is the Monte Carlo mean squared error for the unadjusted estimator divided by that for the indicated estimator.

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Association			
Unadjusted	-0.494	0.002	0.168	0.166	0.948	1.00
Aug. 1	-0.494	-0.001	0.156	0.153	0.948	1.16
Aug. 2	-0.494	0.000	0.156	0.153	0.944	1.15
Aug. 3	-0.494	0.000	0.156	0.153	0.946	1.16
Aug. 4	-0.494	0.000	0.156	0.152	0.943	1.15
Aug. 5	-0.494	-0.001	0.156	0.153	0.945	1.16
Aug. 6	-0.494	0.000	0.156	0.153	0.946	1.16
Usual 1	-0.494	-0.091	0.185	0.182	0.922	0.66
Usual 2	-0.494	-0.090	0.185	0.182	0.922	0.66
			Moderate Association			
Unadjusted	-0.490	0.001	0.165	0.165	0.948	1.00
Aug. 1	-0.490	-0.002	0.140	0.139	0.950	1.39
Aug. 2	-0.490	-0.002	0.141	0.139	0.949	1.38
Aug. 3	-0.490	-0.001	0.139	0.138	0.948	1.41
Aug. 4	-0.490	-0.001	0.140	0.137	0.945	1.40
Aug. 5	-0.490	-0.002	0.140	0.139	0.949	1.39
Aug. 6	-0.490	-0.001	0.140	0.138	0.946	1.40
Usual 1	-0.490	-0.218	0.203	0.201	0.813	0.31
Usual 2	-0.490	-0.219	0.204	0.201	0.813	0.31
			Strong Association			
Unadjusted	-0.460	0.004	0.164	0.165	0.954	1.00
Aug. 1	-0.460	0.000	0.132	0.131	0.952	1.55
Aug. 2	-0.460	0.000	0.132	0.131	0.950	1.54
Aug. 3	-0.460	0.001	0.129	0.128	0.948	1.61
Aug. 4	-0.460	0.001	0.130	0.127	0.945	1.60
Aug. 5	-0.460	0.000	0.132	0.131	0.951	1.55
Aug. 6	-0.460	0.001	0.129	0.127	0.947	1.61
Usual 1	-0.460	-0.321	0.223	0.220	0.695	0.18
Usual 2	-0.460	-0.322	0.224	0.220	0.695	0.17

Table 3.2: Simulation results for estimation of β_2 in the linear mixed model (2.4) using the usual unadjusted method, the proposed augmented methods denoted by "Aug. a" for a=1,2,3, and the "Usual" method, as described in the text, based on 5,000 Monte Carlo data sets. Entries are as in Table 3.1.

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Association			
Unadjusted	0.300	0.000	0.100	0.099	0.951	1.00
Aug. 1	0.300	-0.001	0.095	0.094	0.951	1.10
Aug. 2	0.300	-0.001	0.100	0.097	0.945	1.00
Aug. 3	0.300	-0.001	0.096	0.094	0.950	1.08
Usual	0.300	-0.001	0.100	0.097	0.944	1.00
			Moderate Association			
Unadjusted	0.300	0.000	0.107	0.106	0.949	1.00
Aug. 1	0.300	-0.001	0.097	0.095	0.951	1.22
Aug. 2	0.300	0.000	0.106	0.103	0.945	1.02
Aug. 3	0.300	-0.001	0.097	0.095	0.952	1.21
Usual	0.300	-0.001	0.105	0.101	0.946	1.04
			Strong Association			
Unadjusted	0.300	0.000	0.116	0.115	0.950	1.00
Aug. 1	0.300	-0.001	0.098	0.096	0.951	1.41
Aug. 2	0.300	0.000	0.114	0.111	0.943	1.03
Aug. 3	0.300	-0.001	0.098	0.096	0.951	1.39
Usual	0.300	-0.001	0.113	0.109	0.944	1.06

Table 3.3: Empirical size and power of the usual Kruskal-Wallis test T_n (unadjusted) and the proposed test $\widehat{T}_n^*$ based on 10,000 Monte Carlo replications. Each entry in the columns labeled T_n and $\widehat{T}_n^*$ is the number of times out of 10,000 that each test rejected the null hypothesis of "no treatment effects" under the corresponding scenario.

		Null		Alternative	
ρ	n	T_n	$\widehat{T}_n^*$	T_n	$\widehat{T}_n^*$
0.25	200	0.05	0.05	0.51	0.54
	400	0.05	0.05	0.83	0.85
0.50	200	0.05	0.05	0.51	0.64
	400	0.05	0.05	0.83	0.92
0.75	200	0.05	0.05	0.51	0.85
	400	0.05	0.05	0.83	0.99

3.2 Applications

3.2.1 PURSUIT Clinical Trial

The Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy (PURSUIT) study (Harrington, 1998) was an international multi-center clinical trial involving subjects with acute coronary syndromes to compare the anti-coagulant therapy Integrilin plus heparin and aspirin to heparin and aspirin (control) alone on the basis of the binary composite endpoint death or myocardial infarction at 30 days. We consider data from 5,710 patients and focus on the log-odds ratio for Integrilin relative to control. Thirty-five baseline covariates were recorded, including age, height, weight, gender, race, geographic region, smoking status, diastolic and systolic blood pressure, creatine kinase and creatine kinase-MB ratios, disease history (e.g., angina, diabetes, congestive heart failure hypercholesterolemia, hypertension, renal insufficiency, peripheral vascular disease), and treatment history (e.g., percutaneous coronary intervention within 72 hours of randomization; use of calcium blockers, beta blockers, and digoxin; prior coronary artery bypass

graft).

The unadjusted estimate of the log-odds ratio based on (3.1), $\widehat{\beta}_2$, is -0.174 with standard error 0.073. To calculate the augmented estimator based on (3.1), we used the direct implementation strategy and took $q_g^*(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g$, $g = 1, 2$, with $c_g(X)$ including main effects of all 35 covariates, and fitted the models by OLS. The resulting estimate $\widetilde{\beta}_2 = -0.163$, with standard error 0.071. For these data, the relative efficiency of the proposed estimator to the unadjusted, computed as the square of the ratio of the estimated standard errors, is 1.06. For binary response, substantial increases in efficiency via covariate adjustment are not likely; thus, this admittedly modest improvement is encouraging.

3.2.2 AIDS Clinical Trials Group Protocol 175

AIDS Clinical Trials Group (ACTG) 175 (Hammer et al., 1996) randomized HIV-infected subjects to $k = 4$ different antiretroviral regimens with equal probabilities: zidovudine(ZDV) monotherapy ($g = 1$), ZDV+didanosine (ddI, $g = 2$), ZDV+zalcitabine ($g = 3$), and ddI monotherapy ($g = 4$). We consider data on 2139 subjects and the continuous response CD4 count (cells/mm^3, Y) at 20$\pm$5 weeks post-randomization and focus on comparisons based on differences of the four treatment means. Twelve baseline auxiliary covariates were considered: continuous variables: CD4 count (cells/mm^3), CD8 count (cells/mm^3), age (years), weight (kg), Karnofsky score (scale of 0-100), and indicator variables for hemophilia, homosexual activity, history of intravenous drug use, race (0=white, 1=non-white), gender (0=female), antiretroviral history (0=naive, 1=experienced), and symptomatic status (0=asymptomatic).

We consider the extension of model (2.2) to $k = 4$ treatments, so that $\theta = \beta = (\beta_1, \ldots, \beta_4)^T$, $\beta_g = E(Y|Z = g)$, $g = 1, \ldots, 4$. The standard unadjusted estimator for β is the vector of sample averages; these are $(336.14, 403.17, 372.04, 374.32)^T$ for $g = (1, 2, 3, 4)$, with standard errors $(5.68, 6.84, 5.90, 6.22)^T$. Using the direct implementation strategy with each element of $q_g^*(X, \zeta_g)$ represented using $c_g(X)$ containing all linear terms in the 12 covariates, the proposed methods yield $\widetilde{\beta} = (333.85, 403.83, 370.43,$

$376.45)^T$, with standard errors obtained via the sandwich method as $(4.61, 5.93, 4.89, 5.11)^T$. This is of course one realization of data; however, it is noteworthy that the standard errors for the proposed estimator correspond to relative efficiencies of 1.51, 1.33, 1.46 and 1.48, respectively.

Similar results obtain for hypothesis testing. The three usual unadjusted Wald tests for pairwise differences in means between ZDV monotherapy ($g = 1$) and each other regimen yield T_n for each comparison of 56.85, 19.22, and 20.55 for comparing groups 2, 3, and 4 against group 1, respectively; the corresponding proposed statistics $\widehat{T}_n^*$ are 98.27, 35.28, and 46.75. Of course, all test statistics reflect very strong evidence in favor of real differences in each case; however, notably, the "augmented" test statistics are much larger in each case. We also carried out the standard unadjusted three-degree-of-freedom Wald test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ and the unadjusted Kruskal-Wallis test for $H_0 : S_1(u) = \cdots = S_4(y) = S(u)$, as well as their adjusted counterparts; here, we used $c_{gu}(X)$ containing linear and quadratic terms in the continuous components of X and linear terms in the binary elements. The unadjusted and adjusted Wald statistics are 59.40 and 109.58, respectively; the unadjusted and adjusted Kruskal-Wallis statistics are 49.04 and 100.53; and all are to be compared to χ^2_3 critical values. Again, although the evidence against the null hypotheses is overwhelming even without adjustment, the proposed test statistics are considerably larger.

The sample size in this trial was very large, so that all analyses are easily able to uncover treatment differences. To demonstrate that such results are possible in cases where the evidence is less clear-cut, we repeated these analyses on the data from a random subset of $n = 124$ subjects. The three pairwise unadjusted Wald test statistics are 2.23, 1.45, and 2.18, with p-values 0.07, 0.11, 0.07; the corresponding adjusted statistics are 4.77, 4.87, and 10.12, with p-values 0.01, 0.01, and < 0.001. Likewise, the three-degree-of-freedom unadjusted Wald and Kruskal-Wallis statistics (p-values) are 3.36 (0.34) and 2.35 (0.50), while the adjusted versions are 11.87 (0.01) and 4.59 (0.20). Although it is certainly not guaranteed that smaller p-values will be obtained for any given realization of data, these results demonstrate that the proposed adjustment methods are capable of effecting such improvements.

Chapter 4

"Smooth" Semiparametric Regression Analysis for Arbitrarily Censored Time-to-Event Data

4.1 SNP Representation of a Survival Density

Gallant and Nychka (1987) gave a mathematical description of a class $\mathcal{H}$ of k-dimensional "smooth" densities that are sufficiently differentiable to rule out "unusual" features such as jumps or oscillations but that may be skewed, multi-modal, or fat- or thin-tailed. When $k = 1$, $\mathcal{H}$ includes almost any density that is a realistic model for a (possibly transformed), continuous time-to-event random variable and excludes implausible candidates. For $k = 1$, densities $h \in \mathcal{H}$ may be expressed as an infinite Hermite series $h(z) = P_\infty^2(z)\psi(z)$ plus a lower bound on the tails, where $P_\infty(z) = a_0 + a_1 z + a_2 z^2 + \cdots$ is an infinite-dimensional polynomial; $\psi(z)$ is the "standardized" form of a known density with a moment generating function, the "base density;" and $h(z)$ has the same support as $\psi(z)$. The base density is almost always taken as $\mathcal{N}(0, 1)$, but need not be (see below). For practical use, the lower bound is ignored and the polynomial is truncated, yielding the so-called SNP

representation

$$h_K(z) = P_K^2(z)\psi(z), \quad P_K(z) = a_0 + a_1 z + z_2 z^2 + \cdots + a_K z^K, \quad \boldsymbol{a} = (a_0, a_1, \ldots, a_K)^T, \tag{4.1}$$

With $\boldsymbol{a}$ such that $\int h_K(z)\,dz = 1$ and K suitably chosen, $h_K(z)$ provides a basis for estimation of $h(z)$. The SNP representation has been widely used, particularly in econometric applications. Section 4.4.1 gives more detail on the SNP and its properties.

Zhang and Davidian (2001) noted that requiring $\int h_K(z)\,dz = \int P_K^2(z)\psi(z)\,dz = 1$ is equivalent to requiring $E\{P_K^2(U)\} = \boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = 1$, where U has density ψ, and $\boldsymbol{A}$ is a known positive definite matrix easily calculated for given ψ. Thus, $\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = \boldsymbol{c}^T\boldsymbol{c} = 1$, suggesting the spherical transformation $c_1 = \sin(\phi_1)$, $c_2 = \cos(\phi_1)\sin(\phi_2)$, $\ldots$, $c_K = \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\sin(\phi_K)$, $c_{K+1} = \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\cos(\phi_K)$ for $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K$. Section 4.4.2 presents examples of this formulation. Thus, for fixed K, (4.1) is "parameterized" in terms of $\boldsymbol{\phi}$ ($K \times 1$) and we write $h_K(z;\boldsymbol{\phi}) = P_K^2(z;\boldsymbol{\phi})\psi(z)$; estimation of the finite-dimensional "parameter" $\boldsymbol{\phi}$ leads to an estimator for $h(z)$.

With $K = 0$ in (4.1), $P_K^2(z) \equiv 1$, and $h_K(z)$ reduces to the base density; i.e., $h_K(z) = \psi(z)$. Values $K > 1$ control the extent of departure from ψ and hence flexibility for approximating the true $h(z)$ (K is not the number of components in a mixture). Several authors (e.g., Fenton and Gallant, 1996; Zhang and Davidian, 2001) have shown that $h_K(z;\boldsymbol{\phi})$ with $K \leq 4$ can well-approximate a diverse range of true densities.

We now describe how we use (4.1) to approximate the assumed "smooth" density $f_0(t)$ of a continuous, positive, time-to-event random variable T_0 with survival function $S_0(t) = P(T_0 > t)$, $t > 0$. As T_0 is positive, we assume that we may write

$$\log(T_0) = \mu + \sigma Z, \quad \sigma > 0, \tag{4.2}$$

where Z takes values in $(-\infty, \infty)$. We consider two formulations that together are sufficiently rich to support an excellent approximation to virtually any $f_0(t)$. In (4.2), it is natural to assume that Z has density $h \in \mathcal{H}$ that may be approximated

by (4.1) with $\mathcal{N}(0,1)$ base density $\psi(z) = \varphi(z)$ for suitably chosen K, so that T_0 is lognormally distributed when $K = 0$. Although this can approximate very skewed or close-to-exponential f_0 with large enough K, an alternative formulation better suited to this case is to assume that $Z^* = e^Z$ has density $h \in \mathcal{H}$ that may be approximated by (4.1) with standard exponential base density $\psi(z) = \mathcal{E}(z) = e^{-z}$, so that Z has an extreme value distribution (Kalbfleisch and Prentice, 2002, sec. 2.2.1) when $K = 0$. As discussed in Section 4.2, we propose choosing the representation (normal or exponential) and associated K best supported by the data.

In both cases, approximations for $f_0(t)$ and $S_0(t)$ follow straightforwardly. Under the normal base density representation, for fixed K and $\boldsymbol{\theta} = (\mu, \sigma, \boldsymbol{\phi}^T)^T$, we have for $t > 0$

$$\begin{aligned} f_{0,K}(t;\boldsymbol{\theta}) &= (t\sigma)^{-1} P_K^2\{(\log t - \mu)/\sigma;\, \boldsymbol{\phi}\}\, \varphi\{(\log t - \mu)/\sigma\}, \\ S_{0,K}(t;\boldsymbol{\theta}) &= \int_{(\log t-\mu)/\sigma}^{\infty} P_K^2(z;\boldsymbol{\phi})\, \varphi(z)\, dz. \end{aligned} \tag{4.3}$$

Because $P_K^2(z;\boldsymbol{\phi})$ may be written as $\sum_{k=0}^{2K} d_k z^k$, where the d_k are functions of the elements of $\boldsymbol{\phi}$, $S_{0,K}(t;\boldsymbol{\theta})$ in (4.19) may be written as a linear combination of integrals of the form $I(k,c) = \int_c^\infty z^k \varphi(z) dz$ that satisfy $I(k,c) = c^{k-1}\varphi(c) + (k-1)I(k-2,c)$ for $k \geq 2$, where $I(0,c) = 1 - \Phi(c)$, $I(1,c) = \varphi(c)$, and $\Phi(\cdot)$ is the $\mathcal{N}(0,1)$ cumulative distribution function (cdf). For the exponential base density representation, we have approximations

$$\begin{aligned} f_{0,K}(t;\boldsymbol{\theta}) &= (\sigma e^{\mu/\sigma})^{-1}\, t^{(1/\sigma-1)}\, P_K^2\{(t/e^\mu)^{1/\sigma};\boldsymbol{\phi}\}\, \mathcal{E}\{(t/e^\mu)^{1/\sigma}\} \\ S_{0,K}(t;\boldsymbol{\theta}) &= \int_{(t/e^\mu)^{1/\sigma}}^{\infty} P_K^2(z;\boldsymbol{\phi})\, \mathcal{E}(z)\, dz, \end{aligned} \tag{4.4}$$

where, similar to the normal base case, the integral in (4.4) may be calculated using the recursion $I(k,c) = c^k \mathcal{E}(c) + kI(k-1,c)$, $k > 0$, with $I(0,c) = e^{-c}$. Note, then, that for fixed K, except for the need for a routine to calculate the normal cdf, the approximations of $f_0(t)$ and $S_0(t)$ using either base density representation are in a closed form depending on the "parameter" $\boldsymbol{\theta}$, whose finite dimension depends on K. This offers computational advantages and makes handling of arbitrary censoring patterns straightforward, as we demonstrate next.

4.2 Censored Data Regression Analysis Based on SNP

4.2.1 Popular Regression Models

Let $\boldsymbol{X}_i$ be a vector of time-independent covariates and T_i be the event time, with $(T_i, \boldsymbol{X}_i)$ independent and identically distributed (iid) for $i = 1, \ldots, n$. The usual PH model is

$$\lambda(t|\boldsymbol{X};\boldsymbol{\beta}) = \lim_{\delta\to 0^+} \delta^{-1} P(t \le T < t+\delta \,|\, T \ge t, \boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{X}^T\boldsymbol{\beta}), \quad t > 0, \tag{4.5}$$

where $\lambda_0(t)$ is the baseline hazard function corresponding to $\boldsymbol{X} = \boldsymbol{0}$. Letting $S(t|\boldsymbol{X};\boldsymbol{\beta})$ $= P(T > t|\boldsymbol{X})$ be the conditional survival function for T given $\boldsymbol{X}$, it is straightforward (Kalbfleisch and Prentice, 2002, sec. 4.1) to show that $S(t|\boldsymbol{X};\boldsymbol{\beta}) = S_0(t)^{\exp(\mathbf{X}^T\boldsymbol{\beta})}$, where $S_0(t) = \exp\{-\int_0^t \lambda_0(u)\,du\}$ is the baseline survival function associated with $\lambda_0(t)$. Usually, $\lambda_0(t)$ is left completely unspecified, whereupon (4.5) is a semiparametric model, and $\boldsymbol{\beta}$ characterizing the hazard relationship is estimated via partial likelihood (PL; Kalbfleisch and Prentice, 2002, sec. 4.2). We instead impose the mild restriction that $S_0(t)$ is the survival function of a random variable T_0 satisfying (4.2) with density $f_0(t)$, and that $f_0(t)$ and $S_0(t)$ may be approximated by either (4.19) or (4.4). Letting the conditional density of $T|\boldsymbol{X}$ be $f(t \,|\, \boldsymbol{X};\boldsymbol{\beta})$, we obtain approximations to $S(t \,|\, \boldsymbol{X};\boldsymbol{\beta})$ and $f(t \,|\, \boldsymbol{X};\boldsymbol{\beta})$ for fixed K given by

$$S_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}) = S_{0,K}(t;\boldsymbol{\theta})^{\exp(\mathbf{X}^T\boldsymbol{\beta})}, \quad f_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}) = e^{\mathbf{X}^T\boldsymbol{\beta}}\lambda_{0,K}(t;\boldsymbol{\theta})S_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}), \tag{4.6}$$

where $\lambda_{0,K}(t;\boldsymbol{\theta}) = f_{0,K}(t;\boldsymbol{\theta})/S_{0,K}(t;\boldsymbol{\theta})$. As we demonstrate shortly, the approximations in (4.6) may be substituted into a likelihood function appropriate for the censoring pattern of interest, upon which estimation of $(\boldsymbol{\beta},\boldsymbol{\theta})$ and choice of K and the base density may be based.

We propose a similar formulation for the usual AFT model

$$\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + e_i, \quad e_i \text{ iid.} \tag{4.7}$$

Rather than taking the distribution of the "errors" e_i to be completely unspecified, we assume that $e_i = \log(T_{0i})$, where T_0 has survival function $S_0(t)$ and "smooth" density $f_0(t)$ that may be approximated by (4.19) or (4.4). For fixed K, this leads to approximations to the conditional survival and density functions of $T|\boldsymbol{X}$, $S(t \,|\, \boldsymbol{X}; \boldsymbol{\beta})$ and $f(t|\boldsymbol{X}; \boldsymbol{\beta})$, given by

$$\begin{aligned} S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) &= S_{0,K}(te^{-\mathbf{X}^T\boldsymbol{\beta}}; \boldsymbol{\theta}), \\ f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) &= e^{-\mathbf{X}^T\boldsymbol{\beta}} f_{0,K}(te^{-\mathbf{X}^T\boldsymbol{\beta}}; \boldsymbol{\theta}). \end{aligned} \tag{4.8}$$

The same principle may be applied to the PO model, which in its usual form assumes

$$\frac{S(t \,|\boldsymbol{X};\, \boldsymbol{\beta})}{1 - S(t \,|\boldsymbol{X};\, \boldsymbol{\beta})} = \left\{ \frac{S_0(t)}{1 - S_0(t)} \right\} \exp(-\boldsymbol{X}^T\boldsymbol{\beta}), \tag{4.9}$$

where $S_0(t)$ is the baseline survival function, assumed to have density $f_0(t)$, and $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ is the conditional survival function given $\boldsymbol{X}$ with density $f(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$. Model (4.9) implies $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}) = S_0(t)/\{e^{\mathbf{X}^T\boldsymbol{\beta}} + S_0(t)(1 - e^{\mathbf{X}^T\boldsymbol{\beta}})\}$; thus, assuming $S_0(t)$ and $f_0(t)$ may be approximated by (4.19) or (4.4), $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ and $f(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ may be approximated by

$$\begin{aligned} S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) &= S_{0,K}(t; \boldsymbol{\theta}) a_{0,K}^{-1}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}), \\ f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) &= f_{0,K}(t; \boldsymbol{\theta}) e^{\mathbf{X}^T\boldsymbol{\beta}} a_{0,K}^{-2}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) \end{aligned} \tag{4.10}$$

for fixed K, where $a_{0,K}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) = e^{\mathbf{X}^T\boldsymbol{\beta}} + S_{0,K}(t; \boldsymbol{\theta})(1 - e^{\mathbf{X}^T\boldsymbol{\beta}})$.

We may now exploit these developments. Assuming as usual that the censoring mechanism is independent of T given $\boldsymbol{X}$, we demonstrate when T may be (i) interval-censored, known only to lie in an interval $[L, R]$; (ii) right-censored at L (set $R = \infty$); or (iii) observed (set $T = L = R$). For (i) and (ii), $\Delta = 0$; else, $\Delta = 1$ (iii). With iid data $(L_i, R_i, \Delta_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, assuming that $f(t \,|\, \boldsymbol{X})$ and $S(t \,|\, \boldsymbol{X})$ may be represented as in (4.6), (4.8), or (4.10), for fixed K, the loglikelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$, conditional on the $\boldsymbol{X}_i$, is $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$

$$= \sum_{i=1}^{n} \Big[\Delta_i \log\{f_K(L_i \,|\, \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta})\} + (1 - \Delta_i) \log\{S_K(L_i \,|\, \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta}) - S_K(R_i \,|\, \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta})\} \Big].$$

For fixed K, base density, and model, $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ may be maximized in $(\boldsymbol{\beta}, \boldsymbol{\theta})$ using standard optimization routines; we use the SAS IML optimizer `nlpqn` (SAS Institute, 2006). Choice of starting values is critical for ensuring that the global maximum is reached. In Section 4.4.3 we recommend an approach where $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ is maximized for each of several starting values found by fixing $\boldsymbol{\phi}$ over a grid and using "automatic" rules to obtain corresponding starting values for $(\mu, \sigma, \boldsymbol{\beta})$. Although elements of $\boldsymbol{\phi}$ are restricted to certain ranges, unconstrained optimization virtually always yields a valid transformation so that $\int h_K(z; \boldsymbol{\phi}) dz = 1$. The declared estimates correspond to the solution(s) yielding the largest $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Following other authors (e.g., Gallant and Tauchen, 1990; Zhang and Davidian, 2001), for a given model (PH, AFT, PO), we propose selecting adaptively the K-base density combination by inspection of an information criterion over all combinations of base density (normal, exponential) and $K = 0, 1, \ldots, K_{\max}$. Our extensive studies show $K_{\max} = 2$ is generally sufficient to achieve an excellent fit. With $q = \dim(\boldsymbol{\beta}, \boldsymbol{\theta})$, criteria of the form $-\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta}) + qc$ have been advocated, with small values preferred. Ordinarily, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn (HQ) criteria take $c = 1$, $\log(n)/2$, and $\log\{\log(n)\}$, respectively; AIC tends to select "larger" models and BIC "smaller" models, with HQ intermediate. As noted by Kooperberg and Clarkson (1997, Section 3), with censored data, dependence of c on n may be suspect; for right-censored data, replacing n by $d =$ number of failures has been proposed (e.g., Volinsky and Raftery, 2000), although a similar adjustment under interval censoring is not obvious. It is nonetheless common practice to base c on n. We have found in the current context that replacing n by d has little effect on the K-base density choice. We use HQ with $c = \log\{\log(n)\}$ in the sequel.

The SNP approach is an alternative to traditional semiparametric methods such as PL when one is willing to adopt the assumption of a "smooth" density $f_0(t)$. The formulation also supports selection among competing models (e.g., PH, AFT, PO): that for which the chosen K-base density combination yields the most favorable value of the information criterion may be viewed as "best supported" by the data. This may be used objectively or in conjunction with other evidence, e.g., the outcome of

a formal test of the proportional hazards assumption (e.g., Gray, 2000; Lin, Zhang, and Davidian, 2006), in adopting a final model.

To obtain standard errors and confidence intervals for the estimator for $\boldsymbol{\beta}$, the parameter ordinarily of central interest, as well as for any other functional of the conditional distribution of $T|\boldsymbol{X}$ based on a final selected representation, we follow other authors and treat the chosen K, base density, and model as predetermined. That is, we approximate the sampling variance of the resulting estimator $\widehat{\boldsymbol{\beta}}$ or of any functional $d(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ via the inverse "information matrix" acting as if the chosen $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ were the loglikelihood under a predetermined parametric model. This matrix is readily obtained from optimization software. For $\widehat{\boldsymbol{\beta}}$, the square root of the relevant diagonal element of this matrix yields immediately our proposed standard error; for general functionals, we use the delta method. Assuming that these quantities have approximately normal sampling distributions, $100(1-\alpha)\%$ Wald confidence intervals may be constructed as the estimate $\pm$ the normal critical value $\times$ the estimated standard error. Although the choice of K and base density is made adaptively, which would seem to invalidate this practice, results cited in Section 4.4.1 propsupport it, and simulations in Section 5.1 demonstrate that this approach yields reliable inferences in realistic sample sizes.

Several useful byproducts follow from the SNP approach. Selection of a model with $K = 0$ suggests evidence favoring the parametric model implied by the chosen base density; e.g., the AFT model with $K = 0$ and normal base density corresponds to assuming T given $\boldsymbol{X}$ is lognormally distributed. Because "smooth" estimates of baseline densities and survival functions are immediate, predictors of survival probabilities and calculation of associated confidence intervals as in Cheng et al. (1997) are easily handled.

4.3 More Complex Models

4.3.1 Extension of the AFT Model to "Heteroscedastic Errors"

For transformed event-time models such as (4.7), a standard assumption is that the deviations e_i are iid, made in virtually all studies of these models (a recent exception is Huang, Ma, and Xie, 2005). Stare, Heinzl, and Harrell (2000) discuss the potential for biased inference on $\boldsymbol{\beta}$ if this is violated. The SNP approach readily handles so-called "heteroscedastic errors" and provides a mechanism for testing departures from the iid assumption, which may be difficult to detect graphically (Stare et al., 2000).

In (4.21), the SNP representation implies that $E\{\log(T_i)|\boldsymbol{X}_i\} = \{\mu + \sigma E(Z_i)\} + \boldsymbol{X}^T\boldsymbol{\beta}$ and $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$, where $E(Z_i)$ and $\text{var}(Z_i)$ are calculated assuming either Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, so that under a fixed K-base density combination are known functions of the corresponding $\boldsymbol{\phi}$. This suggests an equivalent formulation with "centered errors;" i.e., writing Equation (4.2) instead as $\log(T_0) = \mu + \sigma\{Z - E(Z)\}$ and again taking $e_i = \log(T_{0i})$ in (4.21) yields $\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \mu + \sigma\{Z_i - E(Z_i)\}$, so the mean is reparameterized as $E\{\log(T_i)|\boldsymbol{X}_i\} = \mu + \boldsymbol{X}_i^T\boldsymbol{\beta}$ while still $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$. Viewing $\{Z_i - E(Z_i)\}$ as a mean-zero deviation, then, permits the immediate extension of (4.21) given by

$$\log(T_i) = \boldsymbol{X}^T\boldsymbol{\beta} + e_i, \quad e_i = \mu + \sigma v(\boldsymbol{X}_i, \boldsymbol{\alpha})\{Z_i - E(Z_i)\}, \tag{4.11}$$

where $v(\boldsymbol{x}, \boldsymbol{\alpha}) > 0$ for all $\boldsymbol{x}$ is a parametric variance function such that $v(\boldsymbol{x}, \boldsymbol{\alpha}) \equiv 1$ if $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{\alpha} = \boldsymbol{0}$, so that $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)v^2(\boldsymbol{X}_i, \boldsymbol{\alpha})$. Although it may not be possible to postulate a "correct" model $v(\boldsymbol{x}, \boldsymbol{\alpha})$, a parsimonious, flexible variance function may be a useful way to capture at least the predominant features of potential heterogeneity (Carroll and Ruppert, 1988, Ch. 3). E.g., a model popular in ordinary regression for this purpose is $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$ (or similar form depending on a subset of $\boldsymbol{x}$). Again assuming Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, it is straightforward to derive SNP approximations to the conditional survival and density functions of $T|\boldsymbol{X}$ based on (4.22), as we now show.

In what follows, we present the conditional survival and density functions of T given $\boldsymbol{X}$, suppressing the subscript i. Considering the case where Z_i in (4.22) is taken to have the standard normal base density SNP representation, letting

$$r = \frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z),$$

the conditional density and survival distribution are given by

$$\begin{aligned} f_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1} P_K^2(r)\, \varphi(r), \\ S_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\, \varphi(z)\, dz. \end{aligned} \tag{4.12}$$

The integral in (4.23) may be calculated straightforwardly using the recursive formulæ given after Equation (4.19). The term $E(Z)$ may written as a function of $\boldsymbol{\phi}$ as before.

For the SNP representation using the standard exponential base density, we assume that the density of $Z^* = e^Z$ may be approximated by this representation. That is, the density of Z^* is represented as $h_K(z^*) = P_K^2(z^*)\mathcal{E}(z^*) = (a_0 + a_1 z^* + \cdots + a_K z^{*K})^2 e^{-z^*}$. Let

$$r = \exp\left\{\frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z)\right\}.$$

It may then be shown that the conditional density and survival function of $T|\boldsymbol{X}$ are

$$\begin{aligned} f_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1} r P_K^2(r)\mathcal{E}(r), \\ S_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\, \mathcal{E}(z)\, dz. \end{aligned} \tag{4.13}$$

Again, the integral in (4.24) is calculable by the recursion. As r involves $E(Z) = E\{\log(Z^*)\}$, we present explicitly this calculation for $K = 0, 1, 2$, where as before a_0, a_1, a_2 are the coefficient in the polynomial $P_K(z)$, which are in turn expressed in terms of $\boldsymbol{\phi}$. With Euler's constant $\gamma = 0.57721566490153286060$, defining $H_1 = -\gamma$, $H_2 = 1 - \gamma$,$H_3 = 3 - 2\gamma$, $H_4 = 11 - 6\gamma$, $H_5 = 50 - 24\gamma$, we have for $K = 0$, $E(Z) = -\gamma = H_1$; for $K = 1$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + a_1^2 H_3$; and for $K = 2$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + (2a_0 a_2 + a_1^2) H_3 + 2a_1 a_2 H_4 + a_2^2 H_5$.

In fitting this model, one may include $\boldsymbol{\alpha}$ as an additional parameter to be estimated; typically, $\boldsymbol{\alpha}$ will be of low dimension (1 or 2). As noted by Stare et al. (2000), graphical displays that are standard diagnostic tools for detecting heteroscedasticity

in ordinary uncensored regression (Carroll and Ruppert, 1988) can be misleading, so it is not prudent to rely on such techniques to suggest starting values. As we propose "working" variance models such as the exponential model for which $\boldsymbol{\alpha} = \mathbf{0}$ corresponds to no heterogeneity, we suggest using $\boldsymbol{\alpha} = \mathbf{0}$ as the starting value in the "wave" of fits across the grid of $\boldsymbol{\phi}$. Upon inspection of the results, a second "wave" may be undertaken using a new starting value for $\boldsymbol{\alpha}$. This process may be iterated until the analyst feels confident that the procedure has "zeroed in" on a reasonable fit.

Of course, (4.22) no longer has the usual AFT property that time is simply rescaled relative to baseline by a function of covariates. See Hsieh (1996) for an interpretation of (4.22) when $\boldsymbol{X}$ is a vector of treatment indicators and $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$, allowing different location and scale for each treatment, and the goal is to test for homogeneity of scale, corresponding here to $\boldsymbol{\alpha} = \mathbf{0}$. More generally, the SNP-based model offers a convenient framework for detecting heterogeneity, alerting the analyst that standard methods may be inappropriate.

4.3.2 Extension of the AFT Model to Time-Dependent Covariates

Time-to-event regression analyses involving time-dependent covariates are commonplace in practice; see Kalbfleisch and Prentice (2002, sec. 6.3) for a discussion of the care that must be taken in this setting. Due to ease of implementation, analysts routinely default to the Cox model (which no longer has proportional hazards); however, alternative models are available, but are rarely used. Cox and Oakes (1984, sec. 5.2) define an AFT model in this case, which we describe for scalar such covariate $X(t)$; see also Robins and Tsiatis (1992). For a subject with covariate $X(t)$ and event time T, the model assumes that time evolves relative to the time T_0 the subject would have had if $X(t) \equiv 0$ according to a monotone transformation $T_0 = \int_0^T \exp\{\beta X(u)\}du = \Psi\{\overline{X}(T), \beta\}$, where $\overline{X}(t) = \{X(s), 0 \leq s \leq t\}$ is the covariate history to t, assumed independent of T_0. If T_0 has survival function $S_0(t)$ with density $f_0(t)$ and hazard function $\lambda_0(t)$, it is conventional to express the model

in terms of the hazard for T given the covariate history, which we denote in obvious notation as

$$\lambda\{t|\overline{X}(t)\} = \lambda_0[\Psi\{\overline{X}(t),\beta\}]\dot{\Psi}\{\overline{X}(t),\beta\} = \lambda_0\left[\int_0^t \exp\{\beta X(u)\}du\right]\exp\{\beta X(t)\}, \tag{4.14}$$

where $\dot{\Psi}(u,\beta) = d\Psi(u,\beta)/du$. Ordinarily, $\lambda_0(t)$ is left completely unspecified (e.g., Robins and Tsiatis, 1992; Lin and Ying, 1995). If the analyst is willing to assume $f_0(t)$ is "smooth," so that it and $S_0(t)$ may be represented by SNP as in Equations (4.19) and (4.4), it should be clear that the conditional hazard in (4.14) may be approximated by $\lambda_{0,K}(t;\boldsymbol{\theta}) = f_{0,K}(t;\boldsymbol{\theta})/S_{0,K}(t;\boldsymbol{\theta})$. In the case of right-censored data, then, where now L is a right-censoring time if $\Delta = 0$ and an event time if $\Delta = 1$, with iid data $\{L_i, \Delta_i, \overline{X}_i(L_i)\}$, $i =, 1, \ldots, n$, the loglikelihood for fixed K and base density, $\ell_K(\beta,\boldsymbol{\theta})$, for $(\beta,\boldsymbol{\theta})$ conditional on covariate history satisfies

$$\begin{aligned}\exp\{\ell_K(\beta,\boldsymbol{\theta})\} = \textstyle\prod_{i=1}^n \left(\lambda_{0,K}[\Psi\{\overline{X}_i(V_i),\beta\};\boldsymbol{\theta}]\dot{\Psi}_i\{\overline{X}_i(V_i),\beta\}\right)^{\Delta_i} \\ \exp\left\{-\textstyle\int_0^{\Psi\{\overline{X}_i(V_i),\beta\}} \lambda_{0,K}(u;\boldsymbol{\theta})\,du\right\}.\end{aligned} \tag{4.15}$$

Extension to multivariate $\boldsymbol{X}(t)$ and time-independent covariates $\boldsymbol{Z}$; i.e., $\Psi\{\boldsymbol{X}^T(T),\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\delta}\} = \int_0^T \exp\{\boldsymbol{X}^T(u)\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{\delta}\}du$, is straightforward. A similar formulation holds for the PH model.

It is worth noting that other models, e.g. for interval censored data with time dependent covariates (Sparling, Younes, and Lachin, 2006) may also be placed in the SNP framework.

4.4 Details

4.4.1 Properties of the SNP Density Estimator

In this section, we give more detail on the SNP density estimator, review work establishing its properties, and describe what is known about its performance when it has been embedded in various complex statistical models. We refer the reader to the references cited, especially Gallant and Nychka (1987) and Fenton and Gallant (1996, 1996b), for technical details and further developments.

The SNP density estimator is a truncation (or sieve) estimator based on a Hermite series expansion and was originally introduced by Gallant and Nychka (1987) in the context of representing the nonparametric part of nonlinear structural models popular in econometric analysis. These models can be rather complicated and would ordinarily also include a finite-dimensional parametric component, as in the semiparametric time-to-event regression models we consider. Since its introduction, the SNP has been used in numerous applications with great success, where it has been embedded in various complex statistical models involving possibly numerous additional parameters of interest. These include in econometric models for stock volatility (Gallant, Hansen, and Tauchen, 1990), as a model for a bivariate distribution in binary choice models for labor-force participation (Gabler, Laisney, and Lechner, 1993), as the underpinning of methods for nonlinear time series analysis (Gallant and Tauchen, 1990; Gallant, Rossi, and Tauchen, 1993), and as a representation of the density of a vector of random effects in various mixed effects (e.g., Davidian and Gallant, 1992, 1993; Zhang and Davidian, 2001; Chen, Zhang, and Davidian, 2002) and joint longitudinal-survival data models (Song, Davidian, and Tsiatis, 2002). In all of these settings, empirical studies suggest that, via a maximum likelihood approach analogous to that proposed previously for the semiparametric time-to-event regression, fitting is computationally stable and feasible and valid inferences may be obtained, as discussed further below.

Gallant and Nychka (1987) considered the general case of a k-variate density in statistical models where both the density and a finite-dimensional vector of parameters are to be estimated. They described the class $\mathcal{H}$ in which the true density f_0 is assumed to lie in terms of a weighted Sobolev norm, depending on the number of derivatives f_0 is assumed to possess, and they provided a rigorous statement of the conditions under which the SNP estimators for f_0 and other parameters should be consistent in some sense for the true values, assuming the parametric part of the model is correctly specified. In particular, they showed that, as long as the truncation rule (choice of K) is such that $K = K_n$, say, converges to infinity with n, the SNP density estimator is consistent with respect to Sobolev norm and that this implies that functionals of the true density, such as the distribution function, as well as the finite-dimensional parameters in the model, are also estimated consistently. See Gal-

lant and Nychka (1987) for technical details and discussion and Davidian and Gallant (1993, Section 3) for a summary. From a practical point of view, a main consideration in the use of SNP as a representation for the density of a model component is the degree of smoothness the true density is thought to enjoy as reflected by the degree of differentiability it is thought to possess, as for other density estimation methods.

Estimation of f_0 in the case $k = 1$, has been studied in some detail. Fenton and Gallant (1996) specialized the consistency results of Gallant and Nychka (1987) to the univariate case when estimation of f_0 is to be based on an iid sample from f_0, and they carried out an extensive battery of empirical studies demonstrating the ability of the SNP density estimator to approximate a wide range of true densities, including some exhibiting rather extreme behavior. They and other authors mentioned below focused on the estimator based on the normal base density, as it has been used extensively in econometric applications. They noted that, for $k = 1$, the class $\mathcal{H}$ of densities defined by Gallant and Nychka is spanned by

$$\mathcal{H}_n = \left\{ f_n : \; f_n(z, \boldsymbol{a}) = \left(\sum_{j=0}^{K_n} a_j z^j \right)^2 e^{-z^2/2} + \epsilon_0 \varphi(z) \right\}, \tag{4.16}$$

where $\varphi(z)$ is the standard normal density, and $\boldsymbol{a}$ are such that $\int f_n(z, \boldsymbol{a})\, dz = 1$; choices other than $e^{-z^2/2}$ and $\varphi(z)$ are also permitted. In (4.16), ϵ_0 is a small positive number, and K_n depends on n; it is possible to rewrite (4.16) in terms of Hermite polynomials. As discussed by Gallant and Nychka (1987) and Davidian and Gallant (1993), the second term in (4.16) acts as a lower bound that governs tail behavior, ensuring that $\int \log f_n(z, \boldsymbol{a}) f_0(z)\, dz$ exists for all $f_n \in \mathcal{H}_n$, required in order to establish the results in Gallant and Nychka (1987); see this paper for further discussion. The lower bound is usually ignored in practice, and vast empirical evidence has shown that this practice leads to reasonable results..

Fenton and Gallant (1996b) established rates of convergence in L_1 where $K_n = O(n^\alpha)$ for $\alpha > 0$. Coppejans and Gallant (2002) derived the convergence rate under the Hellinger metric and investigated the use of cross-validation as an alternative to information-criterion based selection of the truncation point. As noted by Kim (2007), an SNP estimator may not achieve the optimal convergence rate established

by Stone (1990) for log-spline density estimators; however, it has several advantages, including computational ease and convenience; a straightforward means of simultaneous estimation of finite-dimensional parameters when the density is part of an overall semiparametric model; and the ability to evaluate whether or not the parametric model corresponding to the base density is sufficient to represent the data. Fenton and Gallant (1996b, Erratum) note that, while it is not possible to demonstrate that SNP density estimators have the same convergence rate as kernel density estimators, the extensive available empirical evidence suggests that they are qualitatively and asymptotically similar to kernel estimators.

Regarding asymptotic normality of estimators for finite-dimensional parameters and functionals in SNP-based semiparametric models, formal, theoretical results for general semiparametric models are not available. As noted by Kim (2007), this is probably because of the fact that the SNP density estimator is "parametric" for any fixed degree of truncation. There is extensive empirical evidence in different statistical models (e.g., Gallant and Tauchen, 1990; Zhang and Davidian, 2001; Song et al., 2002), as well as theoretical evidence in specific settings (e.g., Eastwood and Gallant, 1991; Fan, Zhang, and Zhang, 2001) that, if one treats the degree of truncation as fixed, so that the model involves a finite-dimensional "parameter," standard errors and confidence intervals may be constructed using standard parametric asymptotic theory. As shown by Eastwood and Gallant (1991) in a simpler setting, this requires that the degree of truncation be chosen adaptively; these authors show that the use of information-criterion-based (so adaptive) truncation rules will result in such inferences being asymptotically correct. As noted by Coppejans and Gallant (2002) and Kim (2007), the practice of basing inferences on standard parametric large sample theory following adaptive choice of the truncation point is widely accepted to yield reasonable inferences in general problems and is standard in applications in analyses based on SNP.

In summary, two decades of experience suggest that use of SNP to represent ordinarily unspecified or latent components of general complex statistical models, as proposed for the specific case of semiparametric time-to-event regression models, leads to reliable inferences under conditions similar to those assumed for competing

approaches.

As noted in the Discussion, a rigorous proof of the theoretical properties of the SNP approach is an open problem. We conjecture that it should be possible to prove that the SNP-based estimator for $\boldsymbol{\beta}$ is root-n consistent. For the PH and PO models, which are members of the linear transformation model class, this is true when one is completely nonparametric with respect to the unknown baseline distribution and uses nonparametric maximum likelihood to estimate it in these models. Thus, we expect that, under appropriate conditions, it is true for the SNP approach as well. Our simulation results do not contradict this supposition. We conjecture that this is also true for the AFT model, as it is possible to show such results for, e.g., rank-based methods. This model is a bit more problematic than the other two in that a fully efficient approach where one is completely nonparametric about the unknown survival distribution would require the support points of the distribution to depend on $\boldsymbol{\beta}$. We suspect that the undercoverage of Wald confidence intervals for $\boldsymbol{\beta}$ we report on in this case for smaller samples may be related to this structural phenomenon somehow.

4.4.2 Parametrization of the SNP Representation

In this section, we give a more detailed description of how the "standard" SNP density representation in Equation (4.1) may be parameterized in terms of $\boldsymbol{\phi}$. See Zhang and Davidian (2001) for the general case. For fixed K and base density $\psi(z)$, the representation is

$$h_K(z) = P_K^2(z)\,\psi(z) = (a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2\,\psi(z), \tag{4.17}$$

subject to constraint

$$\int (a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2\,\psi(z)\,dz = 1. \tag{4.18}$$

Let $\boldsymbol{a} = (a_0, a_1, \cdots, a_K)^T$ of length $K+1$, define $\boldsymbol{w} = (1, z, z^2, \ldots, z^K)$, and define the random vector $\boldsymbol{W} = (1, Z, Z^2, \ldots, Z^K)^T$, where Z is a random variable with density $\psi(z)$. Then note that we can write the polynomial squared in (4.17) as

$$(a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2 = \boldsymbol{a}^T \boldsymbol{w}\boldsymbol{w}^T \boldsymbol{a}.$$

Therefore, the constraint (4.18) is equivalent to requiring that

$$\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = 1, \quad \boldsymbol{A} = E(\boldsymbol{W}\boldsymbol{W}^T).$$

For $\psi(z)$ either the standard normal or exponential densities, the matrix $\boldsymbol{A}$ is known and positive definite, so that we can write $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ for some positive definite matrix $\boldsymbol{B}$. Thus, write $\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = \boldsymbol{a}\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{a}$, so that with $\boldsymbol{c} = \boldsymbol{B}\boldsymbol{a}$, $\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} = \boldsymbol{c}^T\boldsymbol{c} = 1$. Thus, $\boldsymbol{c}$ lies on the unit sphere, which suggests the spherical transformation

$$\begin{aligned}
c_1 &= \sin(\phi_1), \\
c_2 &= \cos(\phi_1)\sin(\phi_2), \\
&\vdots \\
c_K &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\cos(\phi_K), \\
c_{K+1} &= \cos(\phi_1)\cos(\phi_2)\cdot\cos(\phi_{K-1})\cos(\phi_K),
\end{aligned}$$

given in Section 4.1, where $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_K)^T$, $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K-1$, $0 \leq \phi_k \leq 2\pi$.

To demonstrate how this transformation works, we give two explicit examples. In the first example, suppose $K = 2$ and let $\phi(z)$ be the standard normal density. In this case, $\boldsymbol{c} = (c_1, c_2, c_3)^T$, and $c_1 = \sin(\phi_1)$, $c_2 = \cos(\phi_1)\sin(\phi_2)$, $c_3 = \cos(\phi_1)\cos(\phi_2)$, so that $\boldsymbol{\phi} = (\phi_1, \phi_2)^T$. It is straightforward to show that

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix},$$

in which case

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} \text{ and } \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix}.$$

Now

$$\boldsymbol{a} = \boldsymbol{B}^{-1}\boldsymbol{c} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sin(\phi_1) \\ \cos(\phi_1)\sin(\phi_2) \\ \cos(\phi_1)\cos(\phi_2) \end{pmatrix}. \tag{4.19}$$

Thus note that we can express the polynomial in (4.17) in terms of $\boldsymbol{\phi}$ as $a_0 + a_1 z + a_2 z^2 = \boldsymbol{a}^T(1, z, z^2)^T$, where $\boldsymbol{a}$ is given in (4.19). This may be substituted in (4.17) to give the representation $h_2(z; \boldsymbol{\phi})$ in terms of $\boldsymbol{\phi}$.

As a second example, take again $K = 2$ but with $\psi(z)$ the standard exponential density. Again we have $\boldsymbol{a} = (a_0, a_1, a_2) = \boldsymbol{B}^{-1}\boldsymbol{c}$, where $\boldsymbol{c}$ is as before. It is straightforward to show that now

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 6 \\ 2 & 6 & 24 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{pmatrix}, \quad \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1/2 \end{pmatrix}.$$

4.4.3 Achieving the Global Maximum/Starting Values

In this section, we describe the approaches we have used successfully to obtain starting values for parameters for maximizing the SNP loglikelihood for each model (AFT, PH, or PO) for fixed K and base density. For a given K and base density, the corresponding SNP likelihood $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ involves the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\mu, \sigma, \boldsymbol{\phi}^T)^T$, and maximization requires starting values for all of these parameters. The SNP loglikelihood typically is quite complex and is replete with local maxima. Thus, we require a procedure that offers assurance that the global maximum has been identified. This suggests using "waves" of starting values, as has been proposed with SNP in other contexts (e.g., Gallant and Tauchen, 1990). We thus obtain different sets of starting values that hopefully traverse a likely region of the parameter space where the global maximum lies by fixing $\boldsymbol{\phi}$ at each value over a grid of possible values and then deriving corresponding starting values for the remaining parameters $(\mu, \sigma, \boldsymbol{\beta}^T)^T$ (μ, σ, β) depending on the model (PH, AFT, PO) being fitted, as we describe shortly. For each set of starting values so obtained, $\ell_K(\beta, \boldsymbol{\theta})$ is maximized. The maximizing values of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ leading to the largest value of $\ell_K(\beta, \boldsymbol{\theta})$ are assumed to yield to the global maximum and are taken to be the final estimates. Often, many of the sets of starting values will lead to the same maximized value of $\ell_K(\beta, \boldsymbol{\theta})$ and the same estimates, engendering confidence that the global maximum has indeed been identified. We have found that, although elements of $\boldsymbol{\phi}$ are restricted to certain

ranges, as long as the grid of starting values is chosen as recommended, one may use unconstrained optimization of $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ with assurance that the resulting estimates are such that $h_K(z; \boldsymbol{\phi})$ evaluated at the estimates is a valid density.

Our recommended grid points become less dense as K increases owing to the increasing computational cost of repeated maximizations. For $K = 0$, there is no $\boldsymbol{\phi}$, and starting values for $(\mu, \sigma, \boldsymbol{\beta}^T)^T$ may be found as described below, where $E(Z)$ and var(Z) are known constants. Because for $K > 0$ each element of $\boldsymbol{\phi}$ must satisfy $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K$, for $K = 1$ we choose the grid to be the 16 values in $(-1.5, -1.3, -1.1, \cdots, 1.3, 1.5)$. For $K = 2$, we fix $\boldsymbol{\phi} = (\phi_1, \phi_2)$ over 16 values of $(-1.5, -0.5, 0.5, 1.5) \times (-1.5, -0.5, 0.5, 1.5)$. We have demonstrated in our simulations and applications that choosing the grid points in this way yields reliable results (i.e., plausible estimates that appear to represent the global maximum) with feasible computation times.

Indeed, computation times are entirely manageable. For example, the typical time to fit all three models (PH, AFT, PO) to one data set with $n = 200$ and 25% right censoring using our SAS implementation, where maximizations are carried out using the SAS IML optimizer `nlpqn`, including maximization at each set of starting values for each K-base density combination for each model followed by selection of the preferred model-K-base density combination using HQ, is 100 seconds on a 1.73 GHz PC.

AFT Model

As in Equation (4.7), the AFT model is

$$\log(T_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + e_i, \quad e_i \text{ iid.} \tag{4.20}$$

The SNP approach represents the AFT model (4.7) as

$$\log(T_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + e_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \mu + \sigma Z_i, \tag{4.21}$$

where e_i and Z_i are iid, and the density of Z_i may be well-approximated by the two proposed SNP formulations. To get a rough estimate of $(\mu, \sigma, \boldsymbol{\beta})$ for each fixed $\boldsymbol{\phi}$,

we pretend that the e_i follows a normal distribution and fit (4.21) using SAS `proc lifereg` to obtain estimates of $\boldsymbol{\beta}$ and the mean and variance of e_i, which we denote by $\boldsymbol{\beta}_e$, μ_e, and σ_e^2, respectively. We use $\boldsymbol{\beta}_e$ as the starting value for $\boldsymbol{\beta}$ and obtain starting values of μ and σ by solving the equations

$$\begin{aligned} \mu_e &= \mu + \sigma E(Z) \\ \sigma_e^2 &= \sigma^2 \mathrm{var}(Z), \end{aligned}$$

for μ and σ. Here, $E(Z)$ and $\mathrm{var}(Z)$ are functions of $\boldsymbol{\phi}$ for each K-base density combination ($K > 0$) and hence for a given $\boldsymbol{\phi}$ grid point are fixed constants. E.g., for the standard normal base density and $K = 2$, $E(Z) = 2a_0a_1 + 6a_1a_2$ and $\mathrm{var}(z) = a_0^2 + 3(2a_0a_2 + a_1^2) + 15a_2^2 - \{E(Z)\}^2$, where $\boldsymbol{a}$ is a function of $\boldsymbol{\phi}$ as in Section 4.4.2 and hence is fixed once $\boldsymbol{\phi}$ is fixed.

When $K = 0$, there is no $\boldsymbol{\phi}$. To obtain multiple starting values, we solve for μ and σ as above, where $E(Z)$ and $\mathrm{var}(Z)$ are known constants for both base densities. We use three sets of starting values: the solution (μ, σ) so determined, $(\mu - \sigma/2, \sigma)$, and $(\mu + \sigma/2, \sigma)$.

PH Model

To obtain a starting value for $\boldsymbol{\beta}$, we use Cox's partial likelihood method implemented in SAS `proc phreg`. The procedure `proc phreg` also gives an estimate of the baseline survival function $S_0(t)$. To obtain starting values for μ and σ for a fixed $\boldsymbol{\phi}$, we pretend that $\log(T_0)$ in Equation (4.2) is normally distributed, so that T_0 is lognormal. Now $E(T_0) = \int_0^\infty S_0(t)dt$ and $E(T_0^2) = \int_0^\infty 2tS_0(t)dt$, and by substituting the estimated baseline survival function into these expressions, we obtain estimates of $E(T_0)$ and $E(T_0^2)$. This calculation is simple, as the estimated baseline survival function is a step function and thus the two integrals reduce to summations. If we denote the mean and variance of $\log(T_0)$ as μ_e and σ_e^2, using the relationships $E(T_0^m) = \exp(m\mu_e + m^2\sigma_e^2/2)$, $m = 1, 2$, we may obtain rough estimates of μ_e and σ_e^2 by solving two equations. Once these are obtained, we may proceed as described before for the AFT model to find starting values for μ and σ for each $K \geq 0$.

PO Model

Similar to the procedure for the AFT model, we first assume a parametric model for the baseline event time to estimate $\boldsymbol{\beta}$. In order to use a standard SAS procedure to fit a PO model, we exploit the fact that when the "errors" in an AFT model, e_i, $i = 1, \ldots, n$, are iid with a logistic distribution, this model is also a PO model. That is, by letting e_i in (4.21) be iid with a logistic distribution, we are equivalently specifying a PO model with baseline event time T_0 from a log-logistic distribution. Thus, we may use SAS `proc lifereg` to obtain estimates we denote as $\boldsymbol{\beta}_{aft}$, μ_l and σ_l, where the subscript "aft" indicates that the fitted coefficient is with respect to the AFT model, and subscript l indicates μ_l and σ_l are parameters characterizing a logistic distribution. Obtaining estimates of the mean and variance of e_i, denoted by μ_e and σ_e as before, is straightforward by using the relationships $\mu_e = \mu_l$ and $\sigma_e^2 = \pi^2\sigma_l^2/3$. Starting values for for μ and σ may be obtained in the same way as described previously for $K \geq 0$. As for the starting value for $\boldsymbol{\beta}$, the coefficient the coefficient corresponding to the PO model, one can easily derive that $\boldsymbol{\beta}$ is equal to $-\boldsymbol{\beta}_{aft}/\sigma_l$, and thus the obvious approach is to substitute the fitted values from `proc lifereg` into this expression.

4.4.4 Extension of the AFT Model to "Heteroscedastic Errors"

For transformed event-time models such as (4.7), a standard assumption is that the deviations e_i are iid, made in virtually all studies of these models (a recent exception is Huang, Ma, and Xie, 2005). Stare, Heinzl, and Harrell (2000) discuss the potential for biased inference on $\boldsymbol{\beta}$ if this is violated. The SNP approach readily handles so-called "heteroscedastic errors" and provides a mechanism for testing departures from the iid assumption, which may be difficult to detect graphically (Stare et al., 2000).

In (4.21), the SNP representation implies that $E\{\log(T_i)|\boldsymbol{X}_i\} = \{\mu + \sigma E(Z_i)\} + \boldsymbol{X}^T\boldsymbol{\beta}$ and $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$, where $E(Z_i)$ and $\text{var}(Z_i)$ are calculated assuming either Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, so that under a fixed K-base

density combination are known functions of the corresponding $\boldsymbol{\phi}$. This suggests an equivalent formulation with "centered errors;" i.e., writing (2) in the main paper instead as $\log(T_0) = \mu + \sigma\{Z - E(Z)\}$ and again taking $e_i = \log(T_{0i})$ in (4.21) yields $\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \mu + \sigma\{Z_i - E(Z_i)\}$, so the mean is reparameterized as $E\{\log(T_i)|\boldsymbol{X}_i\} = \mu + \boldsymbol{X}_i^T\boldsymbol{\beta}$ while still $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$. Viewing $\{Z_i - E(Z_i)\}$ as a mean-zero deviation, then, permits the immediate extension of (4.21) given by

$$\log(T_i) = \boldsymbol{X}^T\boldsymbol{\beta} + e_i, \quad e_i = \mu + \sigma v(\boldsymbol{X}_i, \boldsymbol{\alpha})\{Z_i - E(Z_i)\}, \tag{4.22}$$

where $v(\boldsymbol{x}, \boldsymbol{\alpha}) > 0$ for all $\boldsymbol{x}$ is a parametric variance function such that $v(\boldsymbol{x}, \boldsymbol{\alpha}) \equiv 1$ if $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{\alpha} = \boldsymbol{0}$, so that $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)v^2(\boldsymbol{X}_i, \boldsymbol{\alpha})$. Although it may not be possible to postulate a "correct" model $v(\boldsymbol{x}, \boldsymbol{\alpha})$, a parsimonious, flexible variance function may be a useful way to capture at least the predominant features of potential heterogeneity (Carroll and Ruppert, 1988, Ch. 3). E.g., a model popular in ordinary regression for this purpose is $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$ (or similar form depending on a subset of $\boldsymbol{x}$). Again assuming Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, it is straightforward to derive SNP approximations to the conditional survival and density functions of $T|\boldsymbol{X}$ based on (4.22), as we now show.

In what follows, we present the conditional survival and density functions of T given $\boldsymbol{X}$, suppressing the subscript i. Considering the case where Z_i in (4.22) is taken to have the standard normal base density SNP representation, letting

$$r = \frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z),$$

the conditional density and survival distribution are given by

$$\begin{aligned} f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1} P_K^2(r)\, \varphi(r), \\ S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\, \varphi(z)\, dz. \end{aligned} \tag{4.23}$$

The integral in (4.23) may be calculated straightforwardly using the recursive formulæ given after Equation (3) in the main paper. The term $E(Z)$ may written as a function of $\boldsymbol{\phi}$ as before.

For the SNP representation using the standard exponential base density, we assume that the density of $Z^* = e^Z$ may be approximated by this representation. That is, the density of Z^* is represented as $h_K(z^*) = P_K^2(z^*)\mathcal{E}(z^*) = (a_0 + a_1 z^* + \cdots + a_K z^{*K})^2 e^{-z^*}$. Let

$$r = \exp\left\{\frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z)\right\}.$$

It may then be shown that the conditional density and survival function of $T|\boldsymbol{X}$ are

$$\begin{aligned} f_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1} r P_K^2(r)\mathcal{E}(r), \\ S_K(t \mid \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\, \mathcal{E}(z)\, dz. \end{aligned} \tag{4.24}$$

Again, the integral in (4.24) is calculable by the recursion described in the main paper. As r involves $E(Z) = E\{\log(Z^*)\}$, we present explicitly this calculation for $K = 0, 1, 2$, where as before a_0, a_1, a_2 are the coefficient in the polynomial $P_K(z)$, which are in turn expressed in terms of $\boldsymbol{\phi}$. With Euler's constant $\gamma = 0.57721566490153286060$, defining $H_1 = -\gamma$, $H_2 = 1 - \gamma$,$H_3 = 3 - 2\gamma$, $H_4 = 11 - 6\gamma$, $H_5 = 50 - 24\gamma$, we have for $K = 0$, $E(Z) = -\gamma = H_1$; for $K = 1$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + a_1^2 H_3$; and for $K = 2$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + (2a_0 a_2 + a_1^2) H_3 + 2a_1 a_2 H_4 + a_2^2 H_5$.

In fitting this model, one may include $\boldsymbol{\alpha}$ as an additional parameter to be estimated; typically, $\boldsymbol{\alpha}$ will be of low dimension (1 or 2). As noted by Stare et al. (2000), graphical displays that are standard diagnostic tools for detecting heteroscedasticity in ordinary uncensored regression (Carroll and Ruppert, 1988) can be misleading, so it is not prudent to rely on such techniques to suggest starting values. As we propose "working" variance models such as the exponential model for which $\boldsymbol{\alpha} = \mathbf{0}$ corresponds to no heterogeneity, we suggest using $\boldsymbol{\alpha} = \mathbf{0}$ as the starting value in the "wave" of fits across the grid of $\boldsymbol{\phi}$. Upon inspection of the results, a second "wave" may be undertaken using a new starting value for $\boldsymbol{\alpha}$. This process may be iterated until the analyst feels confident that the procedure has "zeroed in" on a reasonable fit.

Of course, (4.22) no longer has the usual AFT property that time is simply rescaled relative to baseline by a function of covariates. See Hsieh (1996) for an interpretation of (4.22) when $\boldsymbol{X}$ is a vector of treatment indicators and $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$,

allowing different location and scale for each treatment, and the goal is to test for homogeneity of scale, corresponding here to $\boldsymbol{\alpha} = \mathbf{0}$. More generally, the SNP-based model offers a convenient framework for detecting heterogeneity, alerting the analyst that standard methods may be inappropriate.

Chapter 5

Empirical Studies and Applications (II)

5.1 Simulation Studies

We report on simulations to evaluate performance of the SNP approach. For all SNP fits, we considered $K_{\max} = 2$ and both the normal and exponential base densities.

In the first set of scenarios, data were generated under the PH model (4.5) with continuous covariate X uniformly distributed on $(0, 1)$ and 25% independent uniform right censoring, with true baseline hazard $\lambda_0(t)$ corresponding to a lognormal with mean 2.9 and scale 0.66; a Weibull with shape 0.9 and scale 25.0; a gamma with shape and scale 2.0; and a log-mixture of normals found by exponentiating draws from the bimodal normal mixture $0.3\mathcal{N}(0.2, , 0.36)+0.7\mathcal{N}(1.8, 0.36)$. In all cases, the true value of $\beta = 2.0$, $n = 200$, and 1000 Monte Carlo (MC) data sets were generated. For each, the PH model was fitted by PL via SAS `proc phreg` and by the SNP approach, with comparable results, as shown in Table 5.1(a). The SNP-based AFT and PO models (4.7) and (4.9) were also fitted to each data set, and Table 5.1(a) summarizes how often HQ selected each model. Percentages do not necessarily add to 100% across the three models; because fits with $K = 0$ and exponential base density lead to the same value of HQ for the AFT and PH models, HQ supports more than one model when this configuration is selected, so the percentages reflect the proportions of times

this occurred. Under the Weibull, selection of the AFT or PH model corresponds to choosing a PH model (Kalbfleisch and Prentice, 2002, p. 44). "Correct" indicates the percentage of data sets for which HQ supported selection of the (true) PH model under these conditions and indicates that inspection of HQ across SNP fits of competing models can be useful for deducing the appropriate model, a capability that grows with sample size. The true PH model was identified over 90% of the time for all distributions when $n = 500$.

Informally, the information on $\lambda_0(t)$ and β is roughly "orthogonal," so it is not unexpected that imposing smoothness assumptions on $\lambda_0(t)$ and fitting the SNP-based PH model does not yield increased precision for estimating β relative to PL. For the PH model, the real advantage of the SNP approach is the ease with which it handles interval- and other arbitrarily-censored data. Under the gamma and log-mixture-normal scenarios, we generated interval-censored data for each subject by drawing five random examination times, where the times between each were independently lognormally distributed, and then generated independently an event time from the PH model. This led to the percentages of right- and interval-censored data in Table 5.1(a). Results of fitting the PH model by SNP for 1000 MC data sets with $n = 200$ show that the approach leads to reliable inferences.

The second set of scenarios involved a true PO model (4.9) with X and either independent 25% uniform right censoring or interval censoring as above, $\beta = 2.0$ or -2.0, $n = 200$, and 1000 data sets generated with $f_0(t)$ lognormal with mean and scale 13.8 and 0.53, log-mixture-normal from $0.3\mathcal{N}(1.2, 0.36) + 0.7\mathcal{N}(-1.8, 0.36)$, and Weibull with shape and scale 1.0 and 5.0. From Table 5.1(b), when the true PO model is fitted via SNP, reliable inferences on β obtain, and HQ is able to identify the true PO model well except for the Weibull; for this case, performance improves with increasing sample size.

In the third set of scenarios, data were generated from the AFT model (4.7) with X as above; $\beta = 2.0$; and $f_0(t)$ lognormal with mean 0.5 and scale 1.31, Weibull with shape 2.0 and scale 16.0, gamma with shape and scale 2.0, and the log-mixture of normals $0.3\mathcal{N}(1.2, , 0.36) + 0.7\mathcal{N}(-1.8, 0.36)$ (bimodal). For each of 1000 data sets with independent uniform right censoring, the AFT model was fitted via SNP;

the Buckley-James method; and the rank-based method of Jin et al. (2003), with both logrank and Gehan-type weight functions, using the `R` function `aft.fun` (Zhou, 2006). Table 5.2 shows that the SNP method yields reliable inferences and compares favorably to competing semiparametric methods, achieving marked relative gains in efficiency in some cases. With $n = 200$ and 50% censoring, the SNP procedure continues to perform well. Undercoverage of SNP Wald confidence intervals in the log-mixture-normal case is resolved for $n = 500$. The PH and PO models were also fitted. In all but the gamma scenario, HQ strongly supports the AFT model; increasing to $n = 500$ in the gamma case vastly improves identification of the correct model. The similarity of the gamma distribution to a Weibull may be responsible for the difficulty the criterion has distinguishing the AFT and PH models for the smaller sample size.

Byproducts of the SNP approach for any model are estimates of the corresponding density $f_0(t)$ and survival function $S_0(t)$. Figure 5.1 shows the 1000 estimates of $S_0(t)$ under two of the AFT scenarios, demonstrating that its true form can be recovered with impressive accuracy.

We also carried out simulations for the true AFT model for gamma and log-mixture-normal scenarios under interval censoring, each with 1000 data sets generated as for the PH model to yield the censoring patterns in Table 5.2. The AFT model was fitted to each using the SNP approach; as for PH, the results demonstrate the reliable performance of the method, with undercoverage of confidence intervals for $n = 200$ under the log-mixture-normal.

For all three model scenarios, Table 5.3 presents for selected configurations the number of times each K-base density combination was chosen by HQ when fitting the true model. Not surprisingly, the normal base density is chosen most often when $f_0(t)$ is lognormal and log-mixture normal, and the exponential base density is preferred for the Weibull and gamma.

Undercoverage of Wald intervals in some instances with $n = 200$ in Table 5.2 suggests that the delta method approximation may be less reliable for the AFT model than for PH and PO. We thus investigated replacing delta method standard errors by those from a nonparametric bootstrap, where the K-base density for the AFT model

is chosen by HQ for each bootstrap data set. E.g., for the interval censored log-mixture scenario, for the first 300 of the 1000 MC data sets, the MC mean, standard deviation, average of delta method standard errors, and associated coverage are 2.05, 0.26, 0.23, and 90.3, respectively. The MC average of bootstrap standard errors using 50 bootstrap samples and coverage of the associated interval are 0.28 and 94.3, suggesting that this approach can correct underestimation of sampling variation.

The foregoing results involved simple models with a single covariate in order to allow reporting for a number of scenarios with straightforward interpretation. Given failure to achieve nominal coverage for some settings for the AFT model, further evaluation of the SNP-based approach in this case is warranted. Moreover, demonstration of computational stability and feasibility of the proposed methods for all three models under more complex conditions is required. Accordingly, we carried out additional simulations. We report on representative scenarios, each involving 1000 MC data sets with $n = 200$ or 500 and 25% independent uniform right censoring and generated from the AFT model (4.7), where $\boldsymbol{X} = (X_1, X_2, X_3)^T$ with X_1 distributed as uniform on (0,2), X_2 Bernoulli with $P(X_2 = 1) = 0.5$, and $X_3 \sim \mathcal{N}(0.5, 1)$ and the true value of $\boldsymbol{\beta} = (2.5, 0.5, -0.8)^T$. Table 5.4 shows results for fitting the AFT model when the true $f_0(t)$ was lognormal with mean 54.6 and scale 7.3, gamma with with shape 2.0 and scale 6.0, and log-mixture of normals $0.3\mathcal{N}(1.2, , 0.36) + 0.7\mathcal{N}(-1.8, 0.36)$ (bimodal). In all scenarios, no computational issues were encountered for any data sets, and performance is similar to that for the simpler models above, with analogous undercoverage of Wald intervals for components of $\boldsymbol{\beta}$ in some cases. HQ chose K-base density combinations in proportions similar to those in Table 5.3 in all cases. For the gamma and log-mixture scenarios with $n = 200$, we used a nonparametric bootstrap with 50 bootstrap replicates as described above to obtain alternative standard errors for the first 300 MC data sets; results are indicated by an asterisk in Table 5.4 and suggest that, as above, use of bootstrap standard errors to form Wald intervals yields reasonable performance.

We also carried out analogous simulations under the PH and PO models. Here, we show results of two representative, analogous simulations when the true model is the PH or PO model. Each is based on 1000 MC data sets with $n = 200$ and

25% independent uniform right censoring. In each case, $\boldsymbol{X} = (X_1, X_2, X_3)^T$ were generated as before.

In the first scenario, data were generated from the PH model in (4.5) with $\boldsymbol{\beta} = (1.2, 1.0, 0.2)^T$ and with true $\lambda_0(t)$ corresponding to a gamma with with shape 2.0 and scale 6.0. In the second scenario, data were generated from the PO model in (4.9) with $\boldsymbol{\beta} = (1.2, -1.0, 0.2)^T$ and with true $f_0(t)$ a log-mixture of normals $0.3\mathcal{N}(10, , 0.6) + 0.7\mathcal{N}(8, 0.6)$. Table 5.6 shows the results, where the PH model was also fitted using PL, which are qualitatively similar to those with a single covariate for these models. Again, computation was stable and straightforward in every situation we tried, and, as in the single covariate case reported above, coverage of delta method intervals achieved the nominal level for both models.

We carried out a small simulation (100 data sets for each scenario) to demonstrate its value for accommodating and detecting heterogeneity of the "errors" in the AFT model using (4.22). For each data set with $n = 200$, iid Z_i were generated from the (bimodal) normal mixture $0.3\mathcal{N}(0.21, 0.36) + 0.7\mathcal{N}(-0.9, 0.36)$; X_i were generated as uniform on $(0, 1)$ as before; and T_i were generated from either (4.7) or (4.22) with $\mu = -0.9$ and $\beta = 2.0$, subject to independent uniform 30% right censoring. In scenario I, T_i were generated from the usual AFT model (4.7) with $\sigma = 1$, and (4.7) was fitted via SNP. Scenario II was the same as I, except we fitted (4.22) with $v(x, \alpha) = \exp(x\alpha)$. In scenarios III and IV, data were generated from (4.22) with $\sigma = 0.4$ and $v(x, \alpha) = \exp(x)$ ($\alpha = 1.0$); (4.7) was fitted in III and (4.22) with $v(x, \alpha) = \exp(x\alpha)$ was fitted in IV. In scenario V, T_i were from (4.22) with $\sigma = 1$ and $v(x, \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 x$, $\boldsymbol{\alpha} = (0.4, 0.7)^T$, but (4.22) was fitted with $v(x, \alpha) = \exp(x\alpha)$ as in IV, so misspecifying v. In II, IV, and V, α was estimated along with β and $\boldsymbol{\theta}$. Table 5.5 shows the results. I and IV show that the SNP method yields reliable performance when the correct model is fitted, while II shows that departures from homogeneity may be detected. III shows that failure to take account of heterogeneity has dire consequences, and V demonstrates that the exponential model can detect heterogeneity even if the working variance model is not of the correct functional form.

To illustrate the feasibility of implementing of the AFT model with time-dependent covariates using the SNP approach, we conducted a simulation with 1000 MC data

sets and $n = 200$ generated to mimic a heart transplant scenario (e.g., Lin and Ying, 1995). For each i, a $U(0, 600)$ waiting time W_i was generated, and T_{0i} was generated independently from a gamma distribution with shape 10 and scale 40. With $X_i(t) = 0$ for $t < W_i$ and $X_i(t) = 1$ for $t \geq W_i$, the event time T_i was computed according to the transformation $T_{0i} = \int_0^{T_i} \exp\{\beta X_i(u)\}du$ with $\beta = -1.0$ and was possibly right censored by an independently generated $U(0, 600)$ censoring time, yielding about 30% censoring. Maximizing the SNP-based loglikelihood (4.15) yielded MC mean estimated β of -1.00, with MC standard deviation and average of estimated delta method standard errors both equal to 0.08, and MC coverage of the nominal 95% Wald interval for β of 93.0%.

Overall, the simulations demonstrate that the SNP approach is computationally straightforward and yields reliable performance under the "smoothness" assumption and provides a tool for practical model selection.

Table 5.1: Simulation results based on 1000 Monte Carlo data sets when the true model is the PH or PO model with baseline density $f_0(t)$. Mean is Monte Carlo mean of the 1000 estimates of β when the true model is fitted, SD is their Monte Carlo standard deviation, SE is the average of the 1000 estimated delta method standard errors, and CP is Monte Carlo coverage probability, expressed as a percent, of 95% Wald confidence intervals. For right-censored data, SNP and PL indicate fitting using the SNP approach with K and the base density chosen via HQ and partial likelihood, respectively. All of the AFT, PH, and PO models were fitted to each data set under right-censoring; the columns AFT, PH, and PO indicate the percentage of 1000 data sets for which that model was chosen based on HQ, and Correct indicates the percentage of data sets supporting the PH model; see the text. For interval-censored data, only SNP was used. (a) PH model: true value of $\beta = 2.0$ in all cases. (b) PO model: true value of $\beta = 2.0$ (lognormal, Weibull) or $\beta = -2.0$ (log-mixture).

$f_0(t)$	n	Cens.	Method	Mean	SD	SE	CP	AFT	PH	PO	Cor.
				(a) True PH model							
				Right-censored data							
lognormal	200	25%	SNP	2.02	0.32	0.31	95.4	9.4	86.5	5.2	86.5
			PL	2.00	0.32	0.32	96.3				
Weibull	200	25%	SNP	2.02	0.31	0.31	95.5	86.7	88.0	2.8	97.2
			PL	2.01	0.32	0.32	96.3				
gamma	200	25%	SNP	2.06	0.32	0.31	94.3	66.8	81.6	6.3	81.6
			PL	2.02	0.32	0.32	95.2				
log-mixture	200	25%	SNP	2.04	0.34	0.33	94.9	2.4	73.6	24.0	73.6
			PL	2.02	0.33	0.33	95.5				
				Interval-censored data							
gamma	200	18% right, 82% interval	SNP	2.04	0.30	0.29	93.9				
log-mixture	200	20% right, 80% interval	SNP	2.04	0.30	0.30	94.8				
				(b) True PO model							
				Right-censored data							
lognormal	200	25%	SNP	2.01	0.18	0.18	94.6	0.7	2.2	97.1	97.1
Weibull	200	25%	SNP	2.00	0.46	0.45	94.6	29.4	19.7	65.1	65.1
log-mixture	200	25%	SNP	−1.99	0.46	0.45	95.3	1.4	15.4	83.2	83.2
				Interval-censored data							
log-mixture	200	20% right, 80% interval	SNP	−2.01	0.48	0.47	95.7				

Table 5.2: Simulation results based on 1000 Monte Carlo data sets when the true model is the AFT model with baseline density $f_0(t)$. For right-censored data, SNP, BJ, Gehan, and LR indicate fitting using the SNP approach with K and the base density chosen via HQ, the Buckley-James method, and the rank-based method of Jin et al. (2003) using Gehan-type and log-rank weight functions, respectively. All other entries are as in Table 5.1. For interval-censored data, only SNP was used. True value of $\beta = 2.0$ in all cases.

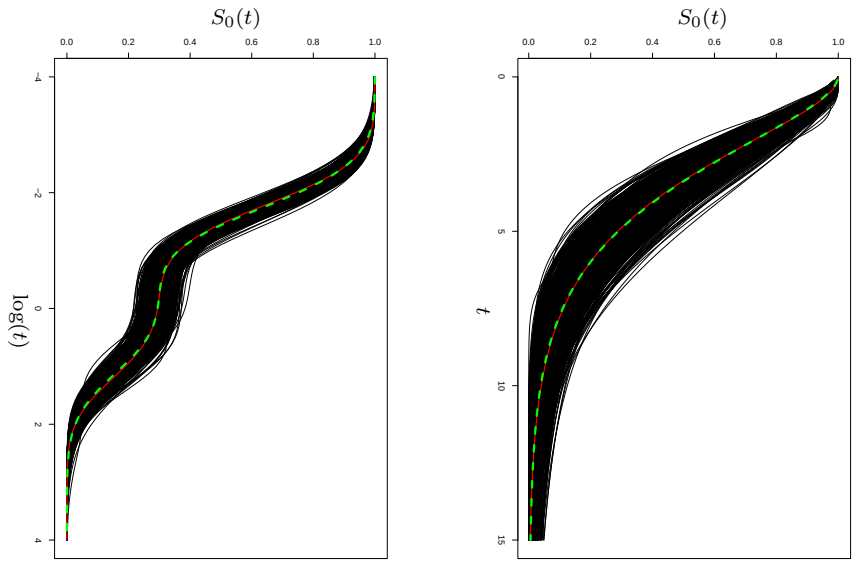
$f_0(t)$	n	Cens.	Method	Mean	SD	SE	CP	AFT	PH	PO	Cor.
				Right-censored data							
lognormal	200	25%	SNP	2.02	0.27	0.26	94.4	80.2	8.0	12.3	80.2
			BJ	2.02	0.27	0.26	94.2				
			Gehan	2.02	0.27	0.28	95.3				
			logrank	2.01	0.29	0.29	94.7				
	200	50%	SNP	2.02	0.30	0.30	93.3				
Weibull	200	25%	SNP	2.00	0.14	0.15	95.9	71.0	88.7	1.1	98.9
			BJ	2.00	0.18	0.19	96.3				
			Gehan	2.00	0.17	0.17	95.7				
			logrank	2.00	0.14	0.15	96.1				
	200	50%	SNP	2.01	0.20	0.20	94.8				
gamma	200	25%	SNP	2.00	0.20	0.19	94.0	65.0	66.4	11.4	65.0
			BJ	2.00	0.22	0.23	96.0				
			Gehan	2.00	0.21	0.22	95.4				
			logrank	2.00	0.20	0.21	95.4				
	200	50%	SNP	2.00	0.26	0.24	92.9				
	500	25%	SNP	2.00	0.06	0.06	94.0	98.8	1.2	0.0	98.8
log-mixture	200	25%	SNP	1.99	0.19	0.18	91.9	100.0	0.0	0.0	100.0
			BJ	1.99	0.42	0.28	80.5				
			Gehan	1.99	0.29	0.29	95.6				
			logrank	2.00	0.41	0.43	96.5				
	200	50%	SNP	1.98	0.23	0.22	91.5				
	500	25%	SNP	2.00	0.05	0.05	94.8				
	500	50%	SNP	2.00	0.06	0.06	93.5				
				Interval-censored data							
gamma	200	20% right, 80% interval	SNP	2.01	0.22	0.21	92.2				
gamma	500	16% right, 84% interval	SNP	2.00	0.06	0.06	94.7				
log-mixture	200	17% right, 83% interval	SNP	2.05	0.27	0.23	90.3				
log-mixture	500	17% right, 83% interval	SNP	2.00	0.07	0.06	94.0				

Table 5.3: Numbers of times each K-base density combination was chosen by the HQ criterion when fitting the true model (PH, AFT, PO) for selected configurations in Tables 5.1 and 5.2.

Base Density			Standard Normal			Standard Exponential		
			K			K		
$f_0(t)$	n	Cens.	0	1	2	0	1	2
			True PH Model					
lognormal	200	25%	873	43	19	11	33	21
Weibull	200	25%	9	0	35	854	74	28
gamma	200	25%	140	12	33	644	143	28
log-mixture	200	25%	239	26	721	0	0	14
gamma	200	18% right, 82% interval	302	28	8	645	12	5
log-mixture	200	20% right, 80% interval	505	62	241	9	6	177
			True AFT Model					
lognormal	200	25%	873	49	18	10	30	20
Weibull	200	25%	3	0	24	890	54	29
gamma	200	25%	65	16	49	624	212	34
log-mixture	200	25%	0	153	847	0	0	0
gamma	200	20% right, 80% interval	223	32	17	640	68	20
log-mixture	200	18% right, 82% interval	0	567	432	0	0	1
			True PO Model					
lognormal	200	25%	878	81	19	0	6	16
Weibull	200	25%	31	3	40	830	82	14
log-mixture	200	25%	0	225	774	0	0	1
log-mixture	200	18% right, 82% interval	0	620	350	0	6	24

Table 5.4: Simulation results for the SNP approach based on 1000 Monte Carlo data sets when the true model is the AFT model with baseline density $f_0(t)$ and multiple covariates under right censoring. Entries are as in Table 5.1. The True $\boldsymbol{\beta}$ column gives the true values of the elements of $\boldsymbol{\beta}$. Entries with an asterisk (*) at the sample size indicate results for the first 300 Monte Carlo data sets, for which both delta method and nonparametric bootstrap standard errors were used, where SE_{boot} and CP_{boot} denote the average of bootstrap standard error and Monte Carlo coverage probability expressed as a percent, of 95% Wald confidence intervals using the bootstrap standard errors, respectively. For each scenario, N_K and E_K, $K = 0, 1, 2$, indicate the number of times the configuration of normal (N) or exponential (E) base density with the indicated K was chosen by HQ.

$f_0(t)$	n	Cens.	True $\boldsymbol{\beta}$	Mean	SD	SE	CP	SE_{boot}	CP_{boot}
lognormal	200	25%	2.5	2.51	0.27	0.26	94.8		
			0.5	0.49	0.15	0.15	93.5		
			-0.8	-0.81	0.30	0.29	94.6		
	($N_0 = 853$, $N_1 = 64$, $N_2 = 19$, $E_0 = 8$, $E_1 = 38$, $E_2 = 18$)								
gamma	200	25%	2.5	2.50	0.16	0.11	93.3		
			0.5	0.50	0.06	0.06	92.3		
			-0.8	-0.80	0.12	0.11	92.1		
	($N_0 = 50$, $N_1 = 16$, $N_2 = 60$, $E_0 = 594$, $E_1 = 237$, $E_2 = 43$)								
	200*	25%	2.5	2.50	0.11	0.11	93.3	0.13	96.0
			0.5	0.51	0.07	0.06	92.7	0.07	95.0
			-0.8	-0.79	0.12	0.11	91.3	0.13	96.3
	500	25%	2.5	2.50	0.07	0.07	93.7		
			0.5	0.50	0.04	0.04	93.1		
			-0.8	-0.80	0.07	0.07	93.8		
	($N_0 = 0$, $N_1 = 3$, $N_2 = 68$, $E_0 = 418$, $E_1 = 464$, $E_2 = 47$)								
log-mixture	200	25%	2.5	2.49	0.10	0.09	94.1		
			0.5	0.50	0.06	0.05	92.8		
			-0.8	-0.80	0.11	0.10	91.8		
	($N_0 = 0$, $N_1 = 197$, $N_2 = 803$, $E_0 = 0$, $E_1 = 0$, $E_2 = 0$)								
	200*	25%	2.5	2.49	0.09	0.09	95.3	0.10	96.7
			0.5	0.50	0.06	0.05	93.0	0.06	94.0
			-0.8	-0.80	0.11	0.10	91.7	0.11	93.0
	500	25%	2.5	2.49	0.06	0.06	94.7		
			0.5	0.50	0.03	0.03	94.7		
			-0.8	-0.80	0.07	0.06	94.5		
	($N_0 = 0$, $N_1 = 16$, $N_2 = 984$, $E_0 = 0$, $E_1 = 0$, $E_2 = 0$)								



(a) (b)

Figure 5.1: SNP estimates of $S_0(t)$ for the AFT model based on 1000 Monte Carlo data sets, with the true $S_0(t)$ (white solid line) and average of 1000 estimates (dashed line) superimposed. (a) log-normal mixture scenario with $n = 500$, 50% right censoring. (b) gamma scenario with $n = 200$, 25% right censoring.

Table 5.5: Simulation results based on 100 Monte Carlo data sets under different scenarios involving possible "heteroscedastic" errors in the AFT model (4.22). Scenarios I–V are described in the text; table entries are as described in the tables for estimation of each of the parameters μ, β, α.

	μ (true=-0.9)				β (true=2.0)				α			
	Mean	SD	SE	CP	Mean	SD	SE	CP	Mean	SD	SE	CP
I	-0.92	0.15	0.14	0.95	2.03	0.18	0.18	0.94				
II	-0.91	0.14	0.15	0.94	2.01	0.22	0.23	0.96	-0.03	0.18	0.17	0.95
III	-0.75	0.11	0.09	0.66	1.51	0.22	0.12	0.08				
IV	-0.90	0.06	0.07	0.96	2.01	0.15	0.16	0.93	0.97	0.15	0.16	0.94
V	-0.91	0.06	0.07	0.97	2.03	0.16	0.18	0.95	0.99	0.16	0.17	

5.2 Applications

5.2.1 Cancer and Leukemia Group B Protocol 8541

Lin et al. (2006) discuss Cancer and Leukemia Group B (CALGB) protocol 8541, a randomized clinical trial comparing survival for high, moderate, and low dose regimens of cyclophosphamide, adriamycin, and 5-flourouacil (CAF) in women with early stage, node-positive breast cancer. Following the primary analysis, interest focused on the prognostic value of baseline characteristics. We consider estrogen receptor (ER) status; ER-positive tumors are more likely to respond to anti-estrogen therapies than those that are ER-negative. ER status is available for 1437 of the 1479 subjects, of whom 64% were ER-positive, with 64% right-censored survival times. Figure 1 of Lin et al. (2006) suggests that the relationship of survival to ER status does not exhibit proportional hazards, a finding corroborated by their spline-based test for departures from proportional hazards (p-value< 0.001).

We fit the AFT, PH, and PO models with binary covariate X_i (=1 if ER-positive) using SNP; here and in Section 5.2.2 we considered the normal and exponential base densities and $K_{\max} = 2$. The HQ criterion was 10177, 10197, and 10192 for the preferred K-base density combinations for AFT, PH, and PO, respectively. Both the PL and SNP fits of the PH model yielded an estimated hazard ratio of 0.77. The AFT

Table 5.6: Simulation results based on 1000 Monte Carlo data sets when the true model is PH or PO. Table entries are as described in the tables for estimation of $\boldsymbol{\beta}$ (3×1).

$f_0(t)$	n	Cens.	Method	True $\boldsymbol{\beta}$	Mean	SD	SE	CP
				PH model				
gamma	200	25%	SNP	1.2	1.24	0.31	0.30	94.7
				1.0	1.05	0.18	0.18	94.2
				0.2	0.21	0.09	0.09	94.9
				$(N_0 = 137,\ N_1 = 8,\ N_2 = 31,\ E_0 = 681,\ E_1 = 115,\ E_2 = 28)$				
			PL	1.2	1.22	0.31	0.30	94.8
				1.0	1.03	0.18	0.18	95.0
				0.2	0.21	0.09	0.09	95.4
				PO model				
log-mixture	200	25%	SNP	1.2	1.24	0.24	0.23	95.4
				-1.0	-1.02	0.27	0.27	95.3
				0.2	0.21	0.13	0.13	95.4
				$(N_0 = 5,\ N_1 = 125,\ N_2 = 850,\ E_0 = 0,\ E_1 = 0,\ E_2 = 20)$				

model is best supported by the data, consistent with the evidence discrediting the PH model. The preferred AFT fit takes $K = 1$ with normal base density, with an estimate of β of 0.45 (SE 0.08). Here, the effect of a covariate is multiplicative on time itself rather than on the hazard, leading to the interpretation that failure times for ER-positive women are "decelerated" relative to those for ER-negative women: the probability that an ER-positive woman survives to time t is the same as the probability that an ER-negative woman survives to time $0.64t$, so that, roughly, being ER-negative curtails survival times by 64% relative to being ER-positive. A possible explanation is that ER-positive women may have received anti-estrogen therapy during follow-up, enhancing their survival.

For demonstration of analysis under a more complex model in practice, we fit PH, PO, and AFT models to the CALGB 8541 data involving a linear predictor in several covariates $\boldsymbol{X}$. As the primary analysis found no difference between the high and moderate doses of CAF, with both superior to the low dose, we considered the treatment indicator $X_1 = 1$ if high-moderate dose and 0 if low dose. We also included $X_2 = 1$ if the woman was ER-positive, $= 0$ otherwise; $X_3 = 1$ if the woman was post-menopausal, $= 0$ otherwise; $X_4 =$ tumor size (cm); and $X_5 =$ number of histologically positive lymph nodes found. Letting $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^T$, we fit the SNP-based PH, PO, and AFT models to the data from the 1429 subjects for whom all five covariates are available; for comparison, we also fit the PH model via PL using SAS `proc phreg`, and the AFT model assuming $f_0(t)$ is lognormal using SAS `proc lifereg`. The results are shown in Table 5.7. Note that for the AFT model, HQ chooses the normal base density but with $K = 1$, suggesting that, if one assumes this model, the parametric lognormal model is not appropriate. Estimates and standard errors for the SNP-based (via the delta method) and traditional fits of the PH and AFT models are comparable. Looking across models, the HQ criterion indicates support for the PO model, with normal baseline density $f_0(t)$, over the other two models.

Table 5.7: Fits to the CALGB data. Base-K shows the base density-K combination chosen by the HQ criterion for the indicated model, and HQ gives the value of the criterion for the preferred choice. Est denotes the estimate of the corresponding component of $\boldsymbol{\beta}$, and SE denotes either delta method (SNP) or usual (PL, likelihood) standard errors.

Model	Method	Base-K	HQ		X_1	X_2	X_3	X_4	X_5
PH	SNP	normal-1	10033	Est	−0.234	−0.269	−0.104	0.181	0.058
				SE	0.091	0.090	0.089	0.036	0.006
	PL			Est	−0.239	−0.271	−0.111	0.182	0.057
				SE	0.091	0.090	0.089	0.036	0.006
PO	SNP	normal-0	10016	Est	−0.303	−0.402	−0.177	0.231	0.090
				SE	0.114	0.113	0.111	0.046	0.010
AFT	SNP	normal-1	10019	Est	0.185	0.339	0.146	−0.140	−0.058
				SE	0.074	0.072	0.071	0.030	0.007
	lognormal ML			Est	0.206	0.292	0.116	−0.148	−0.057
				SE	0.076	0.074	0.073	0.031	0.007

5.2.2 Breast Cosmesis Study

The famous breast cosmesis data (Finkelstein and Wolfe, 1985) involve time to cosmetic deterioration of the breast in early breast cancer patients who received radiation alone ($X = 0$, 46 patients) or radiation+adjuvant chemotherapy ($X = 1$, 48 patients). Deterioration times were right-censored for 38 women. Times for the 56 women experiencing deterioration were interval-censored due to its evaluation only at intermittent clinic visits. Numerous authors have used these data to demonstrate methods for interval censored data.

We fitted the AFT, PH, and PO models using the SNP approach, obtaining HQ values of 309, 309, 317, respectively, for the chosen K-base density combination, supporting AFT and PH. The preferred fit for each uses $K = 0$ and the exponential base density; this configuration is equivalent to a Weibull regression model, for which the PH and AT models are the same. This is consistent with the adoption of the PH (e.g., Goetghebeur and Ryan, 2000; Betensky et al., 2002) or AFT (e.g., Tian

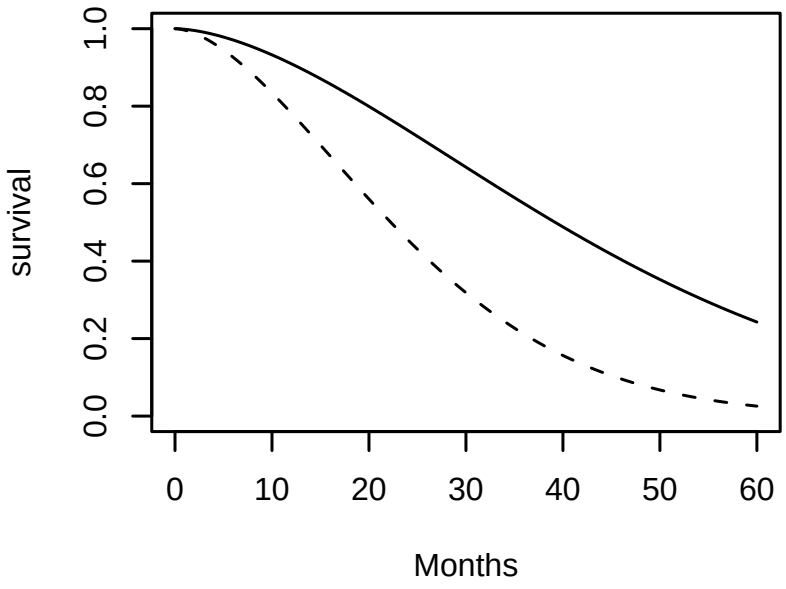


Figure 5.2: Estimated survival functions for time to cosmetic deterioration for the radiation only group (solid line) and radiation+chemotherapy group (dashed line) based on the SNP fit of the PH (AFT) model to the breast cosmesis study data.

and Cai, 2006) models by many authors. The SNP estimate of $\beta = 0.95$ (SE 0.280) in (4.5) is consistent with the results from several methods for fitting the PH model with interval censored data reported by Goetghebeur and Ryan (2000) and Betensky et al. (2002). The corresponding Wald statistic for testing $\beta = 0$ is 3.35, in line with the score statistic of Finkelstein (1986) of 2.86 and Wald statistics implied in Table 2 of Goetghebeur and Ryan (2000). Figure 5.2 shows the SNP estimates of $S(t \mid X = 0)$ and $S(t \mid X = 1)$ based on the PH fit; compare to Figure 1 of Goetghebeur and Ryan (2000).

Chapter 6

Discussion

6.1 Discussion (I)

In Chapters 2 and 3, we have proposed a general approach to using auxiliary baseline covariates to improve the precision of estimators and tests for general measures of treatment effect and general null hypotheses in the analysis of randomized clinical trials by using semiparametric theory.

We identify the optimal estimating function involving covariates within the class of such estimating functions based on a *given* $m(Y, Z; \theta)$. For differences of treatment means or measures of treatment effect for binary outcomes, this estimating function in fact leads to the efficient estimator for the treatment effect. In more complicated models, e.g., repeated measures models, we do not identify the optimal estimating function among *all* possible. Our experience in other problems suggests that gains over the methods here would be modest.

The use of model selection techniques, such as forward selection in our simulations, to determine covariates to include in the augmentation term models should have no effect asymptotically on the properties of the estimators for θ. However, such effects may be evident in smaller samples, requiring a "correction" to account for failure of the asymptotic theory to represent faithfully the uncertainty due to model selection. Investigation of how approaches to inference after model selection (e.g., Hjort and Claeskens, 2003; Shen, Huang and Ye, 2004) may be adapted to this setting would

be a fruitful area for future research.

6.2 Discussion (II)

In Chapters 4 and 5, we have proposed a general framework for regression analysis of arbitrarily censored time-to-event data under the mild assumption of a "smooth" density for model components ordinarily left unspecified under a semiparametric perspective. The methods are straightforward to implement using standard optimization software, and computation is stable across a range of conditions. A SAS macro is available from the first author. Although we focused on the PH, AFT, and PO models, the approach allows any competing models, such as generalizations of (4.7), models with nonlinear covariate effects, and linear transformation models to be placed in a common framework, providing a basis for model selection. Standard errors and Wald confidence intervals may be obtained using standard parametric asymptotic theory in most cases; however, this approximation is less reliable for the AFT model, so we recommend using a nonparametric bootstrap with small samples/numbers of failures in this case. A rigorous proof of consistency and asymptotic normality of the estimators for $\boldsymbol{\beta}$ and functionals of $f_0(t)$ in the general censored-data regression formulation here is an open problem.

It should be possible to adapt the approach to problems involving both censoring and truncation (Joly et al., 1998; Pan and Chappell, 2002). Because with the SNP representation $f(t|\boldsymbol{X};\boldsymbol{\beta})$ and $S(t|\boldsymbol{X};\boldsymbol{\beta})$ are in "parametric" form, the likelihood function is straightforward under the usual assumption that censoring and truncation are independent of event time.

A further advantage, not illustrated here, is that an efficient rejection sampling algorithm for simulation from a fitted SNP density is available (Gallant and Tauchen, 1990). This may be used to simulate draws from the fit of $f_0(t)$ under the preferred model and hence draws of T_i from $f(t \mid \boldsymbol{X})$ for any $\boldsymbol{X}$, allowing any functional of this distribution to be approximated.

Bibliography

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21,** 263–275.

Betensky, R.A., Rabinowitz, D., and Tsiatis, A.A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88,** 703–711.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59,** 570–579.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* London: Chapman and Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Boca Raton: Chapman and Hall/CRC.

Chen, J., Zhang, D., and Davidian, M. (2002) Generalized linear mixed models with flexible distributions of random effects for longitudinal data. *Biostatistics* **3,** 347–360.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation

models with censored data. *Biometrika* **89,** 659–668.

Chen, Y. Q. and Jewell, N. P. (2001). On a general class of semiparametric hazards regression model. *Biometrika* **88,** 677–702.

Chen, Y. Q. and Wang, M. C. (2000). Analysis of accelerated hazards model. *Journal of American Statistical Association* **95,** 608–618.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformed models with censored data. *Biometrika* **82,** 835–842.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of American Statistical Association* **92,** 227–235.

Coppejans, M. and Gallant, A. R./ (2002). Cross-validated SNP density estimates. *Journal of Econometrics* **110,** 27–65.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Davidian, M. and Gallant, A.R. (1992) Smooth nonparametric maximum likelihood for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics* **20,** 529–556.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with Discussion). *Statistical Science* **20,** 261–301.

Eastwood, B. J. and Gallant, A. R. (1991). Adaptive rules for seminonparametric estimators that achieve asymptotic normality. *Econometric Theory* **7,** 307–340.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and

Wilks phenomenon. *Annals of Statistics* **29,** 153–193.

Fenton, V. M. and Gallant, A. R. (1996b). Convergence rates of SNP density estimators. *Econometrica* **64,** 719–727.

Fenton, V. M. and Gallant, A. R. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics* **7**4, 77–118.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Finkelstein, D. and Wolfe, R. M. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41,** 933–945.

Gabler, S., Laisney, F., and Lechner, M. (1993) Seminonparametric estimation of binary choice models with an application to labor-force participation. *Journal of Business and Economic Statistics* **11,** 61–80.

Gallant, A.R., Hansen, L.P., and Tauchen, G.E. (1990) Using conditional moments of asset payoffs to infer volatility of intertemporal marginal rates of substitution. *Journal of Econometrics* **45,** 141–180.

Gallant, A. R. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55,** 363–390.

Gallant, A.R., Rossi, P.E., and Tauchen, G.E. (1993) Nonlinear dynamic structures. *Econometrica* **61,** 871–907.

Gallant, A. R. and Tauchen, G. E. (1990). A nonparametric approach to nonlinear time series analysis: Estimation and simulation. In *New Directions in Time Series Analysis, Part II,* D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M.S. Taqqu (eds.) New York: Springer, pp. 71–92.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* **56,** 1139–1144.

Gray, R. J. (2000). Estimation of regression parameters and the hazard function in transformed linear survival models. *Biometrics* **56,** 571–576.

Grouin, J. M., Day, S., and Lewis, J. (2004). Adjustment for baseline covariates: An introductory note. *Statistics in Medicine* **23,** 697–699.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C., for the AIDS Clinical Trials Group Study 175 Study Team. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335,** 1081–1089.

Harrington, R. A., for the PURSUIT Investigators. (1998). Inhibition of platelet glycoprotein IIb/IIIa with eptifibatide in patients with acute coronary syndromes without persistent ST-segment elevation. *New England Journal of Medicine* **339,** 436–443

Hauck, W. W., Anderson, S., and Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* **19,** 249–256.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98,** 879–899.

Hsieh, F. (1996). Empirical process approach in a two-sample location-scale model with censored data. *Annals of Statistics* **24,** 2705–2719.

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates.*Biometrics* **62,** 813–820.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341–353.

Jin, Z., Lin, D. Y., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika* **93,** 147–162.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition.* New York: John Wiley and Sons.

Kim, K. I. (2007). Uniform convergence rate of the seminonparametric density estimator and testing for similarity of two unknown densities. *Econometrics Journal* **10,** 1–34.

Koch, G. G., Tangen, C. M., Jung, J. W., and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17,** 1863–1892.

Komárek, A., Lesaffre, E., and Hilton, J.F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* **14,** 726–745.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* **53,** 1485–1494.

Leon, S. Tsiatis, A. A., and Davidian, M. (2003). Semiparametric efficient estimation of treatment effect in a pretest-posttest study. *Biometrics* **59,** 1046–1055.

Lesaffre, E. and Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* **22,** 3586–3596.

Lin, D.Y. and Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics* **1,** 77–90.

Lin, D. Y. and Ying, Z. (1995). Semiparametric inference for the accelerated failure life model with time-dependent covariates. *Journal of Statistical Planning and Inference* **44,** 47–63.

Lin, J., Zhang, D., and Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics* **62,** 803–812.

Lin, J. S. and Wei, L. J. (1992). Linear regression analysis based on Buckley-James estimating equation. *Biometrics* **48,** 679–681.

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

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56,** 199–203.

Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics* **58,** 64–70.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine* **21,** 2917–2930.

Robins, J. and Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79,** 311–319.

Robinson, L. D., Dorroh, J. R., Lein, D., and Tiku, M. L. (1998). The effects of covariate adjustment in generalized linear models. *Communications in Statistics, Theory and Methods* **27**, 1653–1675.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics* **18,** 303–328.

SAS Institute, Inc. (2006). *SAS Online Doc 9.1.3.* Cary, NC: SAS Institute, Inc.

Satten, G. A., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93,** 318–327.

Scharfstein, D. O., Tsiatis, A. A., Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis* **4,** 355–391.

Senn, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8,** 467–475.

Shen, X., Huang, H. C., and Ye, J. (2004). Inference after model selection. *Journal of the American Statistical Association* **99,** 751–762.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach for joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Sparling, Y. H., Younes, N., and Lachin, J. M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* **7,** 599–614.

Stare, J., Heinzl, H., and Harrell, F. (2000). On the use of Buckley and James least squares regression for survival data. In *New Approaches in Applied Statistics*, A. Ferilgoj and A. Mrvar (eds). *Metodološki zveszki* **16,** Ljubljana: Faculty of Social Sciences, pp. 125–134.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56**, 29–38.

Stone, C. J. (1990). Large sample inference for log-spline models. *Annals of Statistics* **18,** 717-741.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* New York: Springer.

Tangen, C. M. and Koch, G. G. (1999). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics* **9,** 307–338.

Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93,** 329–342.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics* **18,** 354–372.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled

yet flexible approach. *Statistics in Medicine*, in press.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Volinsky, C.T. and Raftery, A.E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56,** 256–262.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77,** 845–851.

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association* **94,** 125–136.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study for a treatment effect in a pretest-posttest trial. *The American Statistician* **56,** 29–38.

Zhang, D. and Davidian, M. (2001) Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57,** 795–802.

Zhou, M. (2006). The `rankreg` package.
`http://cran.r-project.org/src/contrib/Descriptions/rankreg.html`