Robust and Semiparametric Statistical Modeling for Cancer Research

by

John D. Rice

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2015

Doctoral Committee:

Professor Jeremy M. G. Taylor, Co-Chair
Professor Alexander Tsodikov, Co-Chair
Assistant Professor Rafael Meza
Professor Bin Nan

Acknowledgements

I would first of all like to thank my doctoral committee, Professors Jeremy Taylor, Alex Tsodikov, and Bin Nan, of the Department of Biostatistics; and Assistant Professor Rafael Meza, of the Department of Epidemiology. Their comments and questions have proven very helpful in the course of my research.

Special thanks in particular to my co-chairs Jeremy and Alex: Jeremy introduced me to the field of cancer in biostatistics and guided me through my early years at Michigan, while Alex has taught me nearly everything I know about survival analysis. I am incredibly grateful to both Jeremy and Alex for their patience and advice over the course of my studies and research.

Thanks as well to Professor Bhramar Mukherjee, who, in addition to getting me involved in gene-environment interaction research early in my time at Michigan, has been extremely helpful with career guidance and many other issues arising toward the end of my term as a PhD student.

Although I would never have made it to this point without several of my fellow students, I would like to thank one in particular. Jared Foster, who was initially assigned to me as a mentor when I first arrived at Michigan, became a close friend during my time here. He has provided both invaluable moral support as well as academic assistance with classwork and my dissertation research.

I am grateful to Michael Sabel for providing the melanoma sentinel lymph node biopsy data used in Chapter 1. I would also like to thank Douglas Miller for the use of the rat PCH data in Chapter 3.

This research was supported in part by National Institutes of Health grant T32 CA-83654. Support was also provided by grant 5U01CA157224 (CISNET) from the National Cancer Institute.

Contents

Acknowledgements ii

List of Tables vi

List of Figures vii

List of Appendices viii

Introduction 1

Chapter 1: Locally Weighted Score Estimation for Quantile Classification in Binary Regression Models 3

1.1 Introduction . . . 3
1.1.1 Background . . . 4
1.1.2 Locally weighted score equation approach . . . 5
1.1.3 Metrics of local accuracy . . . 6
1.2 Weighted score estimation . . . 7
1.2.1 Estimating equations . . . 7
1.2.2 Robustness and rationale of the method . . . 8
1.3 Selection of bandwidth . . . 12
1.4 Simulation study . . . 14
1.4.1 Design . . . 14
1.4.2 Results . . . 15
1.5 Melanoma data analysis . . . 20
1.6 Discussion . . . 22
Appendices . . . 24
A Asymptotic bias and variance under misspecification . . . 24
B Weight function from weighted likelihood . . . 27

Chapter 2: Semiparametric Time-to-Event Modeling in the Presence of a Latent Progression Event 29

2.1 Introduction . . . 29
2.2 Model and likelihood . . . 31
2.2.1 Data structure and notation . . . 31
2.2.2 Model . . . 32
2.2.3 Likelihood . . . 33

2.3 Nonparametric maximum likelihood estimation . . . 35
2.3.1 Functional derivative and score equations . . . 35
2.3.2 EM algorithm . . . 36
2.4 Simulation study . . . 37
2.5 SEER prostate cancer data analysis . . . 41
2.6 Discussion . . . 44
Appendices . . . 47
C Derivation of marginal survival and density functions . . . 47
D EM algorithm for estimation of baseline hazard . . . 48
E Prediction of survival function for the latent event . . . 54
F Asymptotic properties . . . 56

Chapter 3: Partial Likelihood Estimation for Continuous Outcomes with Excess Zeros in a Random-threshold Damage-resistance Model 62
3.1 Introduction . . . 62
3.1.1 Modeling data with excess zeros in the outcome . . . 62
3.1.2 Left censoring and the retro-hazard function . . . 64
3.1.3 Damage manifestation and resistance processes . . . 65
3.2 The competitive damage/resistance model . . . 68
3.2.1 Specification of the model . . . 68
3.2.2 Rationale for use of the retro-hazard function . . . 69
3.2.3 Parameterization . . . 70
3.3 Semiparametric estimation based on partial likelihood . . . 71
3.3.1 Counting process formulation . . . 71
3.3.2 Derivation of the partial likelihood . . . 71
3.4 Simulation study . . . 74
3.5 Rat PCH data analysis . . . 79
3.5.1 Data description and background . . . 79
3.5.2 Results . . . 81
3.6 Discussion . . . 84
Appendices . . . 86
G Derivation of NPMLE of the retro-hazard . . . 86
H Marginal model for observed damage . . . 88
I Score components and observed information matrix . . . 90

Conclusion 93

References 96

List of Tables

1.1 Simulation results: local error rate 17
1.2 Simulation results: optimal bandwidths 18
1.3 Summary of WSE analysis of melanoma data set 22
2.1 Simulation results, unidentifiable scenario 40
2.2 Simulation results, identifiable scenario 41
2.3 Summary of analysis of SEER prostate cancer data 42
3.1 Simulation results: logistic part of model 75
3.2 Simulation results: positive part of model 77
3.3 Relative mean-square errors (MSE) for simulated data 78
3.4 Parameter estimates for rat PCH data 81

List of Figures

1.1 Implicit objective functions used in weighted score estimation 9
1.2 SBRQ objective functions 11
1.3 Contours of constant probability for simulations 15
2.1 Function estimates: identifiable scenario 38
2.2 Function estimates: unidentifiable scenario 39
2.3 SEER data analysis: conditional survival functions for onset of metastasis . . 44
3.1 Alternative threshold models 65
3.2 Histogram and boxplot of the outcome for rat PCH data 80
3.3 Observed and fitted values for rat PCH data 83

List of Appendices

A Asymptotic bias and variance under misspecification 24
B Weight function from weighted likelihood 27
C Derivation of marginal survival and density functions 47
D EM algorithm for estimation of baseline hazard 48
E Prediction of survival function for the latent event 54
F Asymptotic properties 56
G Derivation of NPMLE of the retro-hazard 86
H Marginal model for observed damage 88
I Score components and observed information matrix 90

Introduction

In the application of biostatistical methodology to cancer studies, there is a desire to use methods with fewer or less restrictive assumptions, which often lead to more easily generalizable conclusions. In this dissertation, we approach this problem from two basic angles: in the first chapter, we make use of robust estimation procedures to reduce contamination of a model fit due to model misspecification; in the second and third chapters, we apply semiparametric methods to allow for increased flexibility of model fitting with nonparametrically estimated functions.

The first chapter deals with robust modeling of binary responses with the goal of improving classification at an arbitrary probability threshold dictated by the particular application. Specifically, for the linear logistic model, we solve a set of locally weighted score equations, using a kernel-like weight function centered at the threshold. The bandwidth for the weight function is selected by cross validation of a novel hybrid loss function that combines classification error and a continuous measure of divergence between observed and fitted values; other possible cross-validation functions based on more common binary classification metrics are also examined. This work has much in common with robust estimation, but differs from previous approaches in this area in its focus on prediction, specifically classification into high- and low-risk groups. Simulation results are given showing the reduction in error rates that can be obtained with this method when compared with maximum likelihood estimation, especially under certain forms of model misspecification. Analysis of a melanoma data set is presented to illustrate the use of the method in practice.

The second chapter addresses the difficulties inherent in investigating time to cancer

onset when only time to diagnosis can be observed in population studies of cancer incidence. In cancer research, interest frequently centers on factors influencing a latent event that must precede a terminal event. In practice it is often impossible to observe the latent event precisely, making inference about this process difficult. We propose a joint model for the unobserved time to the latent and terminal events, with the two events linked by the baseline hazard. Covariates enter the model parametrically as linear combinations that multiply, respectively, the hazard for the latent event and the hazard for the terminal event conditional on the latent one. We derive an EM algorithm for estimation of the baseline hazard, which allows for closed-form Breslow-type estimators at each iteration, drastically reducing computational time compared with maximizing the marginal likelihood directly. The parametric part of the model is estimated by maximizing the profile likelihood. We present simulation studies to illustrate the finite-sample properties of the method; its use in practice is demonstrated in the analysis of a prostate cancer data set.

In the third chapter, we apply methodology originally used in survival analysis to model semicontinuous data. Continuous outcome data with a proportion of observations equal to zero arises frequently in biomedical studies. Typical approaches involve two-part models, with one part a logistic model for the probability of observing a zero and some parametric continuous distribution for modeling the positive part of the data. We propose a semi-parametric model based on a biological system with competing damage manifestation and resistance processes. This allows us to derive a partial likelihood based on the retro-hazard function, leading to a flexible procedure for modeling continuous data with a point mass at zero. A simulation study is presented to examine the properties of the method in finite samples. We apply the method to a data set consisting of pulmonary capillary hemorrhage area in laboratory rats subjected to diagnostic ultrasound.

Chapter 1: Locally Weighted Score Estimation for Quantile Classification in Binary Regression Models

1.1 Introduction

This chapter develops a class of robust estimators in binary regression models with the goal of increasing predictive accuracy at a given classification threshold. The motivation for this work is a situation arising in the practice of oncology in which it must be decided on the basis of some clinical variables whether or not to perform a sentinel lymph node biopsy in melanoma patients (Mocellin et al., 2009). Given a classification threshold $p^* \in (0, 1)$, the decision is made based on whether the predicted probability of metastasis in the lymph node is greater than p^* (perform biopsy) or less than p^* (do not perform biopsy). There are two types of errors possible in this setting: missing a metastatic cancer by not performing the biopsy (false negative) or performing an unnecessary biopsy (false positive). We assume throughout this chapter that the ultimate objective of the analysis is classification of future subjects into high- and low-risk groups based on the threshold p^*, and that p^* is dictated by considerations specific to the application.

1.1.1 Background

Although a great deal of research has been done on the problem of predicting future binary outcomes in a population based on a sample from that population, most authors dealing primarily with classification (as opposed to regression) have focused on median classification, in which a positive response is predicted if the estimated response probability exceeds $p^* = 0.5$. In one of the few articles dealing with general quantile classification, Mease et al. (2007) propose an over- and undersampling method for boosted classification trees, where the relative amount of over- or undersampling is determined by the classification threshold. (It is from Mease et al., 2007, that we borrow the term "quantile" as used in this chapter.) In other research, classification thresholds play a secondary role: Kordas (2006) discusses the use of the smoothed maximum score estimator of Horowitz (1992) for the binary choice model at quantiles other than the median; both authors assume a latent continuous response variable that is dichotomized as the observed binary outcome. Wang et al. (2008) advocate estimation using an asymmetric version of the hinge loss function (Hastie et al., 2009), for which losses incurred by a false positive and false negative differ, in order to provide interval estimates of response probabilities. Support vector machines (SVM) are also a popular method for classification of binary responses; Dmochowski et al. (2010) point out that the case of unequal costs for false negatives and false positives may be dealt with by assigning class weights in the SVM procedure in accordance with these costs.

To introduce notation, suppose we observe a sample $(y_i, \mathbf{x}_i'), y_i \in \{0, 1\}, i = 1, \ldots, n$, and assume a model in which each response is related to a vector of covariates $\mathbf{x}_i$ by $P(Y = 1|\mathbf{x}_i) = G(\mathbf{x}_i'\boldsymbol{\beta})$, $P(Y = 0|\mathbf{x}_i) = 1 - G(\mathbf{x}_i'\boldsymbol{\beta}) \equiv \overline{G}(\mathbf{x}_i'\boldsymbol{\beta})$, where $G(\cdot)$ is a known link function. We want to find an estimate $\hat{\boldsymbol{\beta}}$ such that $G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$ is most accurate for $P(Y = 1|\mathbf{x}_i)$ near p^*, but is not necessarily as accurate when $P(Y = 1|\mathbf{x}_i)$ is not near p^*. In other words, we sacrifice global goodness of fit for improved local goodness of fit around p^*, where "local goodness of fit" is defined in equation (1.3) below.

1.1.2 Locally weighted score equation approach

Since we are interested in robustness to model misspecification away from p^*, we propose obtaining regression estimates by solving a set of locally weighted score equations, using a kernel-like weight function centered about p^*. There has been some research on robust estimation in binary models, but it has not generally focused on prediction. Pregibon (1982) recommends maximizing a modified log-likelihood function that "tapers" contributions with large deviance residuals (see McCullagh and Nelder, 1989, for a detailed discussion of residuals in generalized linear models, including the logistic model). A thorough development of M-estimation in the context of logistic regression is given by Carroll and Pederson (1993). Ruckstuhl and Welsh (2001) analyze a form of misspecification familiar to the literature on robust estimation for continuous responses, in which the assumed probability mass function is "contaminated" by an unknown pmf with some probability.

We consider the following form of misspecification: we fit an assumed model

$$P(Y = 1|\mathbf{x}_i) = G(\mathbf{x}_i'\boldsymbol{\beta}), \tag{1.1}$$

while the true model is instead

$$p(\mathbf{x}_i) \equiv P(Y = 1|\mathbf{x}_i) = G\left[\mathbf{x}_i'\boldsymbol{\beta}_0 + Q(\mathbf{x}_i, \boldsymbol{\beta}_0)\right] \tag{1.2}$$

where Q is an unknown function of the covariates and $\boldsymbol{\beta}_0$. We make two assumptions regarding Q:

(i) $Q(\mathbf{x}_i, \boldsymbol{\beta}_0) \equiv Q_i = 0$ when $\mathbf{x}_i'\boldsymbol{\beta}_0 = \eta^*$, where $\eta^* = G^{-1}(p^*)$

(ii) $|Q_i|$ is increasing in $|\mathbf{x}_i'\boldsymbol{\beta}_0 - \eta^*|$.

In other words, we are supposing that at least around the threshold of interest, the assumed model is correctly specified, while it may differ from the true model when the linear predictor is not close to η^*.

1.1.3 Metrics of local accuracy

We define a function that measures *local* accuracy,

$$\epsilon(\mathbf{x}_i) = \mathbb{1}\left[p(\mathbf{x}_i) \geq G(\eta^*)\right] \mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} < \eta^*\right) + \mathbb{1}\left[p(\mathbf{x}_i) < G(\eta^*)\right] \mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} \geq \eta^*\right), \tag{1.3}$$

and refer to $\mathbb{E}\left[\epsilon(\mathbf{X})\right]$ as the *local error rate* (LER), where an error occurs if the underlying true response probability $p(\mathbf{x}_i)$ is not on the same side of p^* that $G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$ is. This does not correspond to the expectation of any loss function, implying that it cannot be estimated in practice, at least in our scenario with ungrouped binary data. It may be possible, however, to estimate (1.3) in the setting of Ruckstuhl and Welsh (2001) with $m_i \geq 2$. In this case, where $Y|\mathbf{x}_i$ is distributed binomially with probability $p(\mathbf{x}_i)$ and number of trials m_i, we would have two estimates of $p(\mathbf{x}_i)$: y_i/m_i and $G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$. An estimate of $\mathbb{E}\left[\epsilon(\mathbf{X})\right]$ could then be constructed as

$$\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{1}\left[y_i/m_i \geq G(\eta^*)\right] \mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} < \eta^*\right) + \mathbb{1}\left[y_i/m_i < G(\eta^*)\right] \mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} \geq \eta^*\right)\right\}. \tag{1.4}$$

For this estimator to be sensible, we would need both $n \to \infty$ and $m_i \to \infty$ for all $i = 1, \ldots, n$.

The structure of this chapter is as follows. In Section 1.2, we describe the weighted score estimator and some of its properties. In Section 1.3, we suggest a method for choosing the degree of locality for the weight function. Section 1.4 presents a simulation study. Section 1.5 applies the method to a melanoma data set.

1.2 Weighted score estimation

1.2.1 Estimating equations

As stated previously, we wish to find locally accurate estimates of $\boldsymbol{\beta}_0$. As in Carroll and Pederson (1993), we obtain estimates of $\boldsymbol{\beta}$ by solving the weighted score equations

$$\mathbf{0} = \sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i \left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})\right]. \tag{1.5}$$

Here, $w_h(\cdot)$ is some unimodal function, symmetric about η^*, that attains a maximum of 1 at $\eta = \eta^*$, as in Copas (1995). The parameter h determines the degree of locality of the model fit, and may be thought of as a bandwidth parameter (see Kordas, 2006, for a closely related use of the term "bandwidth"). We use the Gaussian kernel weight function

$$w_h(\eta) = \exp\left[-\frac{(\eta - \eta^*)^2}{2h^2}\right]. \tag{1.6}$$

Define $\hat{\boldsymbol{\beta}}_h$ as the solution to equation (1.5), which may be obtained using the Newton-Raphson algorithm. We refer to this as the weighted score estimator (WSE).

Intuitively, for an appropriate choice of h, the solution to equation (1.5) should be a more accurate estimate of the true parameter vector under the misspecified model (1.2), since for small values of h we will be downweighting the regions of the data that are not reflective of the linear part of (1.2), $\mathbf{x}_i'\boldsymbol{\beta}_0$. When $h \to \infty$, the solution to equation (1.5) corresponds to the maximum likelihood estimator (MLE) for the linear logistic model, since in this case $w_h(\eta) \to 1$ for all η. The idea of downweighting observations inconsistent with an assumed model is not new, and has been considered in the context of binary response models by Pregibon (1982), Carroll and Pederson (1993), Copas (1988), and Ruckstuhl and Welsh (2001). However, none of these articles has addressed the robustness issue from the perspective of what we might call *locally correct* model specification, as measured by equation (1.3).

Supposing the model is correctly specified (i.e., $Q_i = 0$ for all i), consistency follows immediately for fixed h, as this estimator falls in the Mallows class (Carroll and Pederson, 1993) in which weights do not depend on the outcome. Moreover, we know from an application of standard robust estimation theory given by Carroll and Pederson (1993) that the asymptotic variance of the weighted score estimators is approximately

$$\boldsymbol{\mathcal{I}}_n^{-1}(\boldsymbol{\beta}_0)\mathbf{V}_n(\boldsymbol{\beta}_0)\boldsymbol{\mathcal{I}}_n^{-1}(\boldsymbol{\beta}_0), \tag{1.7}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ and $\mathbf{V}(\boldsymbol{\beta})$ are, respectively,

$$\boldsymbol{\mathcal{I}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta})$$
$$\mathbf{V}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_h^2(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta})$$

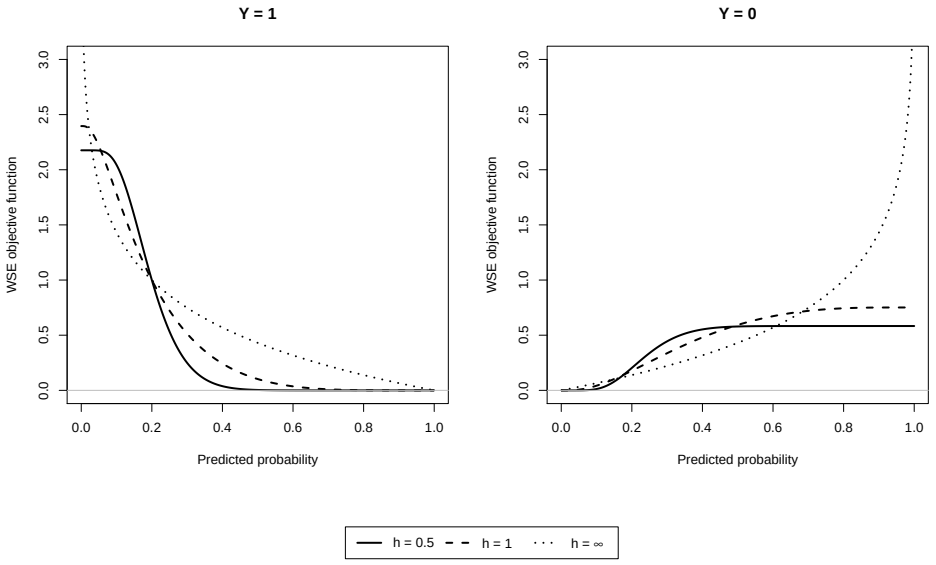
where a superscript (j) denotes the jth derivative of a function with respect to its argument. (See Appendix A for a development of the asymptotic distributions of the proposed estimators when $Q_i \neq 0$.)

1.2.2 Robustness and rationale of the method

Our method may be thought of as a generalization of the robust estimators of Carroll and Pederson (1993) in that our weight function $w_h(\cdot)$ depends on the classification threshold, allowing for improved accuracy at any p^*, while the weighting schemes they consider implicitly assume an interest in conventional median classification since they downweight observations with very high or low fitted probabilities (i.e., probabilities far from $p^* = 0.5$). We make the weight a function of the linear predictor rather than of the predicted probability so as to avoid issues regarding the restricted range of the mean of $Y|\mathbf{x}_i$, which for binary response models is $(0, 1)$.

The motivation for our specific approach comes from Copas (1995), who maximizes a

Figure 1.1. This figure depicts the contribution of a single observation to the implicit objective functions used in weighted score estimation, scaled such that the loss is 1 at $p^* = 0.2, y = 1$; specifically, the vertical axis is $-\tilde{l}_h\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}\right) / -\tilde{l}_h\left(1, \eta^*\right)$. These are obtained by numerical evaluation of equation (1.8). The x-axis of each panel corresponds to a predicted probability for one observation; the curves shown represent the loss or cost associated with that predicted probability for either $y = 1$ or $y = 0$ when using equation (1.5) to obtain estimates of $\boldsymbol{\beta}$.



form of local likelihood for continuous models where locality is determined by proximity to some value t in the outcome space. However, in our binary setting, this is not appropriate, as it would result in only two weights, one for $y_i = 0$ and another for $y_i = 1$. Directly adapting this method to weight individual contributions to the log-likelihood instead by some function of $\mathbf{x}_i'\boldsymbol{\beta} \in \mathbb{R}$ is also problematic, as (combined with the requirement for consistency of the resultant estimator) it leads to estimating equations formally identical to (1.5), but with a weight function in (1.5) that may be negative and is not maximized at η^* (see Appendix B, where we show that the weighted likelihood with correction for bias in the estimating equations turns out to be just a weighted score equation with an undesirable weight function).

Alternatively, we may think of the contribution of a single observation to the objective

function implied by the weighted score (1.5), given by

$$\tilde{l}_h\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}\right) = \int_{y_i}^{G\left(\mathbf{x}_i'\boldsymbol{\beta}\right)} w_h\left(G^{-1}(\pi)\right) \frac{y_i - \pi}{G^{(1)}\left(G^{-1}(\pi)\right)}\, d\pi, \tag{1.8}$$

where G is assumed to be monotone, so that $\frac{d}{d\pi}G^{-1}(\pi) = 1/G^{(1)}\left(G^{-1}(\pi)\right)$. Minimization of $-\sum_{i=1}^{n} \tilde{l}_h\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}\right)$ with respect to $\boldsymbol{\beta}$ results in the same estimators obtained by solving (1.5). The gain in robustness due to the WSE approach is apparent from Figure 1.1, which is a plot of $-\tilde{l}_h\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}\right) / -\tilde{l}_h\left(1, \eta^*\right)$ versus $G\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$: in the usual maximum likelihood case ($h = \infty$) the loss is unbounded. By contrast, we see from this figure that for $h < \infty$, the maximum possible loss is finite. Furthermore,

$$\lim_{h\to 0} \frac{-\tilde{l}_h\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}\right)}{h} \propto y_i\, \mathbb{1}(\mathbf{x}_i'\boldsymbol{\beta} < \eta^*) + \frac{p^*}{1-p^*}(1 - y_i)\, \mathbb{1}(\mathbf{x}_i'\boldsymbol{\beta} \geq \eta^*),$$

which is a weighted misclassification error. Thus the WSE is a compromise between the efficiency of the MLE and the robustness of a classification approach.

The objective function corresponding to the smoothed maximum score estimator of Horowitz (1992) and Kordas (2006) is shown in Figure 1.2 for purposes of comparison; these authors use $\tau = 1 - p^*$ to define the quantile of interest. Their objective function is

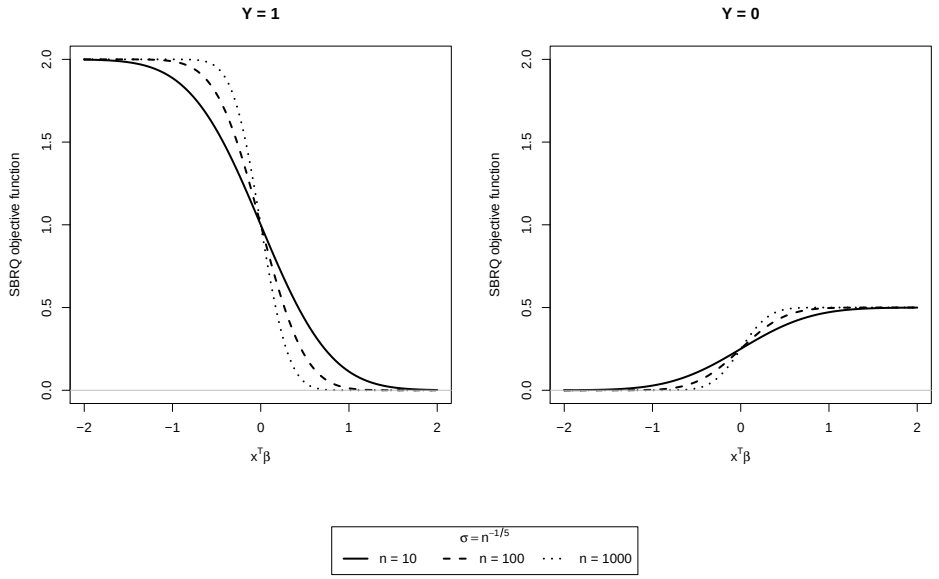
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - p^*)K\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_n}\right), \tag{1.9}$$

where $K(\cdot)$ is an integrated kernel function (i.e., a cdf), $\sigma_n \sim n^{-1/5}$ is a bandwidth, and the maximizer of (1.9) gives the smoothed binary regression quantile estimator. As Kordas (2006) notes, maximizing (1.9) is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{p^*}\left(y_i - K\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_n}\right)\right), \tag{1.10}$$

which is the loss depicted in Figure 1.2. Here, $\rho_{p^*}(u) = [\mathbb{1}(u \geq 0) - p^*]u$. To place this in our

Figure 1.2. This figure depicts the objective functions of Kordas (2006) for $\tau = 1 - p^* = 0.8$. These curves correspond to equation (1.10) and are scaled such that the loss is 1 at $p^* = 0.2, y = 1$.



context, this means that a false positive ($\mathbf{x}_i'\boldsymbol{\beta} \geq 0, y_i = 0$) implies a loss of approximately p^*, while a false negative ($\mathbf{x}_i'\boldsymbol{\beta} < 0, y_i = 1$) implies a loss of approximately $1 - p^*$; the approximation becomes exact as $\sigma_n \to 0$. This can be seen in the relative heights of the curves in Figure 1.2, which echo the patterns depicted for the objective functions associated with the weighted score estimators in Figure 1.1.

We point out here that there is an interesting connection between our method and the work of Kordas (2006), in which the objective function is smoothed by replacing an indicator function with a cdf, given here in equations (1.9) and (1.10). Clearly, the corresponding gradient will involve a pdf, which is identical in form to our weight functions apart from a scaling factor. Specifically, to maximize equation (1.9) (using the normal cdf as the kernel function K), we must solve

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - p^*) \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_n}\right)$$

$$= \frac{1}{n\sigma_n} \sum_{i=1}^{n} \phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_n}\right) \mathbf{x}_i(y_i - p^*). \tag{1.11}$$

Compare (1.11) with the weighted score equations (1.5), which we may rewrite as

$$\mathbf{0} = \sqrt{2\pi} \sum_{i=1}^{n} \phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - \eta^*}{h}\right) \mathbf{x}_i \left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})\right], \tag{1.12}$$

since we are using the Gaussian kernel given by (1.6).

1.3 Selection of bandwidth

A key element of the proposed method is the value of the parameter h, which is similar to the bandwidth of the smoothed indicator function in Kordas (2006) and Horowitz (1992). We may think of the selection of the bandwidth as a way of dealing with model misspecification locally: if, for a given threshold, the specified model is consistent with the truth globally, then h should be large, as we will be able to gain information from all observations regardless of distance from p^*. On the other hand, if the specified model is *not* consistent with the truth everywhere, then h should be small, in order to minimize the contamination of the model fit resulting from misspecification.

Likelihood cross validation is recommended by Eguchi and Copas (1998) as a way to choose h in local likelihood estimation. However, in their setting, the outcome is continuous; furthermore, their weights are a function of the outcome itself rather than a monotone transformation of its estimated conditional mean (as in our case). Irizarry (2001) deals with a similar issue, and bases his selection on modified information criteria, i.e., functions that are based on the negative maximized log-likelihood plus a penalty term. However, his weights are solely functions of the covariates, with the unknown parameter playing no role.

These methods are primarily aimed at assessing overall goodness of fit, as opposed to the local goodness of fit in which we are interested. Although ideally we would like to select h

to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon(\mathbf{x}_i), \tag{1.13}$$

where $\epsilon(\mathbf{x}_i)$ is defined as in equation (1.3), we have seen that this quantity is not estimable from ungrouped binary data. Therefore, we shall attempt to minimize (1.13) indirectly. To do this, we recommend minimizing a cross-validated version of a hybrid error rate (see Wahba, 2002, for a discussion of a similar cost function designed to compromise between SVM and logistic regression).

Defining $\hat{\boldsymbol{\beta}}_h^{(-i)}$ as the estimate of $\boldsymbol{\beta}$ obtained by leaving out the ith observation and solving (1.5), we choose h as

$$\begin{aligned}\hat{h} = \arg\min_{h>0} -2\sum_{i=1}^{n}\Big[&y_i \log G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)}\right)\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^*\right)\\ &+ (1-y_i)\log \overline{G}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)}\right)\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} \geq \eta^*\right)\Big].\end{aligned} \tag{1.14}$$

This error is a combination of classification error, from the terms $\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^*\right)$ and $\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} \geq \eta^*\right)$; and deviance loss, from the terms $\log G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)}\right)$ and $\log \overline{G}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)}\right)$. Note that this is equal to the deviance (or minus twice the log likelihood) for all subjects with $\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} \geq \eta^*\right) \neq y_i$. Due to the presence of indicators this will be a piecewise constant function with jump discontinuities; however, as we will be using a grid search to find the optimal value of h, these discontinuities will present no numerical problems.

The flexibility of the method is such that we might also choose h to optimize any desired predictive metric, such as the negative predictive value (NPV):

$$\hat{h} = \arg\max_{h>0} \frac{\sum_{i=1}^{n}(1-y_i)\,\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^*\right)}{\sum_{i=1}^{n}\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^*\right)} \tag{1.15}$$

or specificity

$$\hat{h} = \arg\max_{h>0} \frac{\sum_{i=1}^{n}(1-y_i)\,\mathbb{1}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^*\right)}{\sum_{i=1}^{n}(1-y_i)}. \tag{1.16}$$

In particular applications, these may be of greater interest than the form of local accuracy defined by (1.3).

1.4 Simulation study

1.4.1 Design

We examine the performance of the proposed method by generating data based on model (1.2). First, X_{i1}, X_{i2} are generated iid standard normal. Then we obtain Z_{i1}, Z_{i2} as follows:

$$Z_{i1} = \frac{X_{i1}}{\sqrt{2}} - \frac{X_{i2}}{\sqrt{2}}, \quad Z_{i2} = \frac{X_{i1}}{\sqrt{2}} + \frac{X_{i2}}{\sqrt{2}}.$$

The linear predictor is

$$\eta_i = \eta^* + \omega_2 Z_{i2} \left[1 + \gamma \left(e^{-\omega_1 \operatorname{sgn}(Z_{i2}) Z_{i1}} - 1\right)\right], \tag{1.17}$$

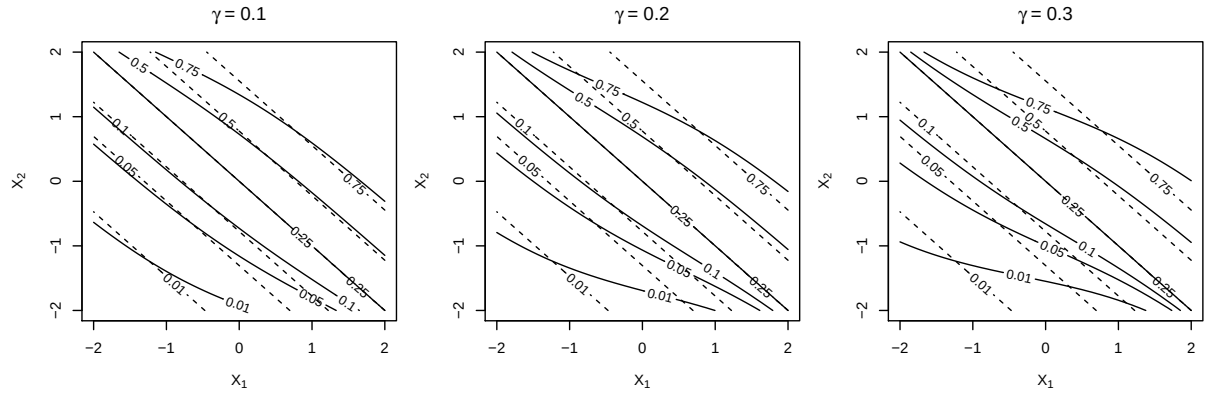
the response is then generated as Bernoulli with mean $e^{\eta_i}/(1+e^{\eta_i})$. The parameter γ is used to control the degree of departure from the linear logistic model outside of the region around η^*, with larger values leading to more nonlinearity in the true model (see Figure 1.3). For these simulations, $\gamma \in \{0, 0.1, 0.2, 0.3\}, \omega_1 = 1, \omega_2 = 2$.

This simulation design was chosen in part so that for a given p^*, changes in γ have minimal effect on the marginal mean of the outcome. At the same time, we believe that the resulting contours of constant probability shown in Figure 1.3 are realistic in the sense that we might encounter shapes of this sort in real data.

Two sample sizes were examined, $n \in \{200, 500\}$. In each case, one sample was used for fitting the weighted score models and selecting h, while a separate sample of the same size was held out for calculation of error rates.

To find the optimal value of h for each data set, we use a grid search of (1.14), (1.15), and (1.16) at 50 equally spaced points of $h^{-1} \in [0, 2]$; we expect these metrics to give

Figure 1.3. Contours of constant probability for simulations, $p^* = 0.25$. Solid lines are the true model; dashed lines represent the linear logistic model for reference, although it is only correct at the thick solid line, $\mathbb{E}(Y|\mathbf{x}_i) = p^*$.



better performance than the cross-validated deviance (which we also include for purposes of comparison) because they explicitly depend on the threshold. Because of the computational expense of true leave-one-out cross validation, we instead use 5-fold cross validation. To further increase computational efficiency as well as numerical stability, when performing the cross validation, we do not fit a full model for each fold, but rather use a one-step Newton approximation.

We compare the proposed method with two existing methods which produce as fitted values binary estimates of $\mathbb{1}\left[\mathbb{E}(Y|\mathbf{x}_i) \geq p^*\right]$. The weighted SVM approach of Dmochowski et al. (2010) assigns class weights as $1-p^*$ for the $y_i = 1$ class and p^* for the $y_i = 0$ class. We refer to this method as the weighted SVM (wSVM) method. The smoothed binary regression quantiles (SBRQ) method of Kordas (2006) with $\tau = 1 - p^*$ is a semiparametric regression procedure that makes minimal assumptions regarding the latent error distribution of the responses. Both methods produce fitted values of either 1 or 0, corresponding to whether or not the predicted probability exceeds p^*.

1.4.2 Results

Performance was evaluated by computing the error rate (1.13) on the validation samples for $h = \hat{h}$ and $h = \infty$, corresponding to the particular WSE model minimizing (1.14) and the

MLE, respectively. Their averages across 1000 simulated data sets for each γ are displayed in Table 1.1.

Table 1.1. Simulation results: LER (1.13) averaged over 1000 simulations; for a single observation this is given by equation (1.3). With the exception of the column labeled Dev., which refers to likelihood cross validation, the cross-validation functions to choose h are based on equations (1.14), or the hybrid error rate (HER); (1.15), the negative predictive value (NPV); and (1.16), the specificity (Spec.).

				CV functions (WSE)					
n	p^*	γ	MLE	Dev.	HER	NPV	Spec.	SBRQ	wSVM
200	0.10	0.0	0.058	0.060	0.062	0.062	0.061	0.103	0.101
		0.1	0.062	0.062	0.062	0.062	0.062	0.088	0.101
		0.2	0.077	0.074	0.073	0.070	0.070	0.082	0.103
		0.3	0.087	0.083	0.080	0.078	0.078	0.082	0.104
	0.25	0.0	0.046	0.046	0.048	0.048	0.048	0.072	0.074
		0.1	0.047	0.046	0.048	0.048	0.048	0.064	0.069
		0.2	0.056	0.054	0.054	0.053	0.053	0.063	0.073
		0.3	0.065	0.061	0.060	0.057	0.057	0.060	0.075
	0.50	0.0	0.039	0.040	0.042	0.042	0.043	0.066	0.065
		0.1	0.042	0.042	0.043	0.043	0.043	0.057	0.062
		0.2	0.048	0.047	0.046	0.046	0.046	0.052	0.064
		0.3	0.053	0.051	0.050	0.049	0.048	0.051	0.065
500	0.10	0.0	0.039	0.040	0.042	0.042	0.041	0.065	0.072
		0.1	0.047	0.045	0.046	0.045	0.045	0.056	0.070
		0.2	0.064	0.059	0.055	0.054	0.055	0.054	0.072
		0.3	0.080	0.075	0.066	0.063	0.066	0.053	0.074
	0.25	0.0	0.029	0.030	0.032	0.032	0.031	0.047	0.052
		0.1	0.034	0.033	0.033	0.033	0.032	0.039	0.051
		0.2	0.046	0.043	0.041	0.040	0.039	0.038	0.052
		0.3	0.058	0.055	0.049	0.047	0.048	0.038	0.055
	0.50	0.0	0.024	0.025	0.026	0.027	0.026	0.039	0.043
		0.1	0.029	0.028	0.029	0.028	0.028	0.034	0.042
		0.2	0.040	0.038	0.036	0.035	0.034	0.031	0.043
		0.3	0.049	0.046	0.042	0.041	0.040	0.030	0.045

Table 1.2. Simulation results: optimal bandwidths, the harmonic mean of the chosen h across 1000 simulations. For the WSE method, each column heading indicates the cross-validation function optimized to choose h: with the exception of the column labeled Dev., which refers to likelihood cross validation, the cross-validation functions to choose h are based on equations (1.14), or the hybrid error rate (HER); (1.15), the negative predictive value (NPV); and (1.16), the specificity (Spec.). For the SBRQ method, this is calculated based on Theorem 2 in Horowitz (1992).

			CV functions (WSE)				
n	p^*	γ	Dev.	HER	NPV	Spec.	SBRQ
200	0.10	0.0	2.587	2.102	1.855	1.891	0.561
		0.1	2.699	1.986	1.938	1.911	0.553
		0.2	3.019	1.926	1.833	1.999	0.554
		0.3	3.343	1.963	1.900	2.073	0.547
	0.25	0.0	2.441	1.632	1.457	1.482	0.562
		0.1	2.580	1.633	1.450	1.456	0.566
		0.2	2.740	1.654	1.519	1.570	0.572
		0.3	2.965	1.650	1.533	1.530	0.584
	0.50	0.0	2.462	1.626	1.443	1.244	0.562
		0.1	2.690	1.532	1.401	1.229	0.582
		0.2	2.936	1.509	1.455	1.243	0.588
		0.3	2.890	1.589	1.468	1.288	0.595
500	0.10	0.0	2.299	1.504	1.541	1.709	0.472
		0.1	2.664	1.550	1.617	1.808	0.467
		0.2	3.280	1.476	1.614	1.808	0.461
		0.3	4.003	1.473	1.598	1.861	0.462
	0.25	0.0	2.347	1.335	1.419	1.361	0.456
		0.1	2.684	1.447	1.423	1.411	0.461
		0.2	3.065	1.393	1.405	1.423	0.469
		0.3	3.849	1.403	1.394	1.462	0.475
	0.50	0.0	2.526	1.354	1.368	1.184	0.454
		0.1	2.802	1.399	1.366	1.262	0.467
		0.2	3.030	1.339	1.362	1.244	0.480
		0.3	3.745	1.407	1.438	1.266	0.493

We see from this table that using the hybrid error rate to select h leads to substantial improvement over the MLE in the LER, assuming that there is some degree of misspecification ($\gamma > 0$). There is an even greater improvement in performance of the WSE method relative to the wSVM, likely due to the fact that the true model is close to the specified parametric model in this case. This implies that the hybrid error rate is a fair surrogate for the LER.

There is relatively little difference, indeed, between the LER for models obtained using each of the CV functions we examined: use of the hybrid error rate, the NPV, or the specificity seems to result in similar performance gains over the MLE, to the extent that no single CV function can be recommended over the others. This points to another strength of our method: the investigators may choose a cross-validation function that best reflects their interests in a particular classification metric and expect to see substantial improvements in the local error rate relative to competing methods. This is in contrast to using likelihood cross validation to select the bandwidth, which clearly does not provide any advantage over the MLE.

Also apparent from Table 1.1 is that for large degrees of misspecification, the SBRQ method slightly outperforms the WSE method; this difference becomes more pronounced with larger sample sizes. One possible explanation for its superior performance with larger γ is that SBRQ makes fewer parametric assumptions about the data. The wSVM method, by contrast, is always the worst performing of the methods we have examined, probably because it makes even weaker assumptions regarding the data-generating mechanism.

Recall, however, that SBRQ and wSVM do not produce estimated probabilities, while the WSE method does. This means that it is possible with our proposed method to examine how close to the classification threshold any individual subject might be, ultimately providing more information to the investigator than do the competing methods. Additionally, when there is no misspecification, the SBRQ and wSVM methods perform very poorly relative to the MLE, while the WSE method loses little in comparison. This is a further advantage of our approach if we assume that in practice we are unlikely to encounter extremely misspecified

models.

Table 1.2 gives a summary of the distribution of selected h values across the simulated data sets. The harmonic mean is shown because $h = \infty$ has a nonzero probability of being chosen. Using the cross-validated deviance to choose h leads to increasing choices of h with increasing γ, regardless of the classification threshold, which is likely the explanation for its failure to show improvement over the MLE in Table 1.1. This is in contrast to the case with each of the metrics we recommend for the selection of h, which do not exhibit this pattern. Similarly, the selected values of the bandwidth parameter for the method of Kordas (2006) show little variability with respect to changes in γ.

Additionally, it is clear from Table 1.2 that for the CV functions given by equations (1.14), (1.15), and (1.16), the chosen bandwidth is decreasing with sample size. This is a desirable feature of the methodology, as with larger samples we are in effect able to focus on a smaller region about the threshold, and thereby obtain larger gains in performance relative to the MLE.

1.5 Melanoma data analysis

In order to illustrate the use of the proposed method in practice, we apply it to a melanoma data set originally analyzed by Sabel et al. (2012). The data consist of the binary outcome and covariate values on $n = 2244$ melanoma patients who had undergone a sentinel lymph node biopsy (SLNB). The outcome is equal to 1 if metastasis was found on SLNB, 0 otherwise. The covariates are age, mitotic rate, Breslow depth, ulceration, regression, and high Clark level (defined as IV or V); the last three are binary, the first three continuous. Mitotic rate and Breslow depth were log transformed prior to model fitting.

One of the principal goals of the original analysis was to evaluate the predictive models of Mocellin et al. (2009). Two cutoff values were considered in Sabel et al. (2012) in the application of the logistic regression model estimated by Mocellin et al. (2009) to classification

on a new data set, $p^* = \{0.1, 0.2\}$. In the current analysis, we apply the WSE methodology to this data at these same two thresholds. We chose $h^{-1} \in [0, 2]$ from a grid of 1000 equally spaced points according to either (1.14), (1.15), or (1.16) with 10-fold cross validation. As a computational note, using the one-step Newton approximation during cross validation results in rapid execution of the entire WSE methodology: the total time to select the optimal bandwidth for two thresholds and three CV functions was approximately five minutes.

Table 1.3 gives a summary of the chosen h values and the effect of the WSE method on classification for this data set. We see improvements with the WSE fits versus the MLE in each of the metrics for which we have conducted cross validation, especially in the case of $p^* = 0.1$, where we are able to improve the negative predictive value from 0.935 to 0.947 and the specificity from 0.093 to 0.140. For $p^* = 0.2$ we still see improvements over the MLE, although the gains in this case are less pronounced.

Another aspect of the analysis in which we are able to improve on the simple logistic fit is in the number of negative predictions (NNP), which Mocellin et al. (2009) cited as an important part of the intended use of their predictive models; again, this seems to be much more dramatic for $p^* = 0.1$ than $p^* = 0.2$. Of course, by itself, NNP is fairly useless as an indicator of a model's usefulness, but given that we are also improving on a certain predictive metric simultaneously, an increase in NNP means fewer patients subjected to unnecessary biopsies, which was one goal of the original analysis.

The NNP also suggests a reason for the difference in the analysis between the two thresholds considered. Specifically, nearly half of the data is classified into the low-risk group at this threshold regardless of fit (MLE or WSE). This indicates that the "center" of the data in some sense lies in this region. Indeed, the marginal probability of a positive biopsy result for this data set is approximately 25%, so that the WSE methodology, when used with thresholds in the neighborhood of 0.25, will not have much room for improvement over the MLE.

Table 1.3. This table summarizes the results of the WSE analysis of the melanoma data set. The two values of p^* given here correspond to the values considered by Sabel et al. (2012). Three cross-validation functions were examined: the negative predictive value (NPV), specificity, and the hybrid error rate (although these results are not given in this table as h was chosen to be ∞; that is for that metric, the MLE was deemed to be the best fitting model). The table gives the value of each of these metrics for both the linear logistic model ($h = \infty$) as well as the optimal model (where h is given by the "Chosen bandwidth" column); the final two columns give the number of negative predictions (NNP) associated with each model fit.

		Chosen	Metric		NNP	
CV function	p^*	bandwidth	$h = \infty$	$h = \hat{h}$	$h = \infty$	$h = \hat{h}$
NPV	0.1	1.44	0.935	0.947	168	208
	0.2	0.76	0.850	0.853	999	995
Specificity	0.1	1.07	0.093	0.140	168	252
	0.2	1.13	0.505	0.512	999	1013

1.6 Discussion

In this chapter, we have developed the weighted score estimation framework, which constitutes a robust estimation method drawing inspiration from the local likelihood approach of Copas (1995), in order to improve quantile classification for parametric binary response models. We have dealt theoretically with model misspecifcation of a particular type, namely data generated according to (1.2) while fitting a model of form (1.1). This has allowed us to demonstrate the ability of the method to reduce contamination of the model fit due to deviations from the assumed model away from the threshold of interest p^*. We have also proposed a novel loss function in equation (1.14) as a way to assess predictive accuracy when interest is in classification at an arbitrary threshold p^*, and have shown through simulation studies the increased accuracy with respect to the LER metric that is possible when an *a priori* classification threshold is incorporated into the parameter estimation procedure.

It might not be immediately apparent why we are not bypassing the bandwidth selection problem altogether and simply minimizing the hybrid error rate to estimate $\boldsymbol{\beta}$. We have instead suggested a two-stage procedure beginning with a localized extension of maximum likelihood estimation because many loss functions commonly used in the estimation of binary

response models result in predicted probabilities being forced to 0 or 1 when optimized directly (Mease et al., 2007). The degree of locality for the WSE model is then chosen based on a measure that more directly addresses our interest in classification. One advantage of this approach is that virtually any measure of predictive accuracy can be used to select h, so that, for example, a researcher interested in maximizing the NPV of a model (e.g., Mocellin et al., 2009) may use that as the cross-validation function instead of the hybrid error rate that we have introduced in this chapter.

Numerical instability occurring when attempting to fit WSE models with small h is one possible avenue for future research. In many of the simulated data sets, the cross-validation function was still decreasing at the smallest h for which the fitting algorithm converged. This indicates that the chosen values of h might be biased upwards in some cases, and this in turn could explain the lack of a decreasing trend in the chosen values of h with increasing degrees of model misspecification. We have conjectured that some values of h for which the model fitting procedure fails might correspond essentially to eliminating too much of the data, so a weight function with heavier tails (such as the Cauchy kernel) could be helpful in this regard.

Appendices

A Asymptotic bias and variance under misspecification

In this section, we present a sketch of the derivation of the asymptotic distribution of the WSE under this form of model misspecification. We assume the usual regularity conditions involving existence of derivatives, expectations, and limits in probability (for details, see, e.g., Lehmann, 2004).

Assume now that $Q_i \neq 0$. An asymptotic theory for local likelihood estimators under a similar form of model misspecification has been developed by Eguchi and Copas (1998), but they are dealing with a continuous outcome and limit attention to misspecification that is of order $n^{-1/2}$. Furthermore, because in their setting the model is asymptotically correctly specified, they assume that $h \to \infty$ at some specified rate, whereas we hold h constant in our analysis.

In order to render (1.2) easier to work with on the probability scale, we approximate $P(Y = 1|\mathbf{x}_i)$ by a Taylor expansion about the model we are fitting:

$$\begin{aligned} P(Y = 1|\mathbf{x}_i) &= G\left(\mathbf{x}_i'\boldsymbol{\beta}_0 + Q_i\right) \\ &= G(\mathbf{x}_i'\boldsymbol{\beta}_0) + Q_i G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}_0) + o\left(Q_i^2\right) \\ &\approx G(\mathbf{x}_i'\boldsymbol{\beta}_0) + Q_i G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}_0). \end{aligned} \tag{A1}$$

The expectation of the score will not be $\mathbf{0}$ when $Q_i \neq 0$ for all i, implying inconsistency of the estimators.

After using a Taylor expansion about $\boldsymbol{\beta}_0$ to approximate the estimating equations, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \approx \left[-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{U}_i(\boldsymbol{\beta}_0)\right]^{-1}\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{U}_i(\boldsymbol{\beta}_0)\right]. \tag{A2}$$

We note that

$$\left[-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{U}_i(\boldsymbol{\beta}_0)\right]^{-1} \xrightarrow{p} \boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\beta}_0) \tag{A3}$$

by the weak law of large numbers and continuous mapping;

$$\boldsymbol{\mathcal{J}}_n(\boldsymbol{\beta}) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta})\left[w_h(\mathbf{x}_i'\boldsymbol{\beta}) - Q_i w_h^{(1)}(\mathbf{x}_i'\boldsymbol{\beta})\right]$$

is the finite-sample analogue of $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$. Likewise, by the central limit theorem,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{U}_i(\boldsymbol{\beta}_0) - \mathbf{S}_n(\boldsymbol{\beta}_0)\right] \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \mathbf{T}(\boldsymbol{\beta}_0)\right), \tag{A4}$$

where

$$\begin{aligned}
\mathbf{S}_n(\boldsymbol{\beta}) &\approx \frac{1}{n}\sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i Q_i G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}) \xrightarrow{p} \mathbf{S}(\boldsymbol{\beta}),\\
\mathbf{T}_n(\boldsymbol{\beta}) &\approx \frac{1}{n}\sum_{i=1}^{n} w_h^2(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i'\left[G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}) + Q_i G^{(2)}(\mathbf{x}_i'\boldsymbol{\beta})\right] \xrightarrow{p} \mathbf{T}(\boldsymbol{\beta}).
\end{aligned}$$

Here we have made use of a further Taylor expansion to approximate the variance of $Y|\mathbf{x}_i$ under misspecification. Therefore, by Slutsky's theorem, we have from (A3) and (A4) that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \xrightarrow{\mathcal{L}} N\left(\boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\beta}_0)\mathbf{S}(\boldsymbol{\beta}_0), \boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\beta}_0)\mathbf{T}(\boldsymbol{\beta}_0)\boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{\beta}_0)\right). \tag{A5}$$

Note that due to the approximation in (A1) as well as the further approximation for $\mathrm{Var}(Y|\mathbf{x}_i)$, (A5) is also approximate, even asymptotically. The accuracy of these approximations of course depends on the magnitude of Q_i; that is, on the degree to which the true model deviates from the assumed model. The advantage to the use of these approximations is that they allow us to give very simple expressions for the quantities involved in the asymptotic distribution of the WSEs under misspecification. Specifically, for example, the approximations for $\boldsymbol{\mathcal{J}}_n(\boldsymbol{\beta})$ and $\mathbf{S}_n(\boldsymbol{\beta})$ show that the magnitude of the bias will depend on

the proportion of observations with $p(\mathbf{x}_i)$ near p^*: since we have assumed that $|Q_i|$ increases with $|\mathbf{x}_i'\boldsymbol{\beta}_0 - \eta^*|$, the bias will be greater when the distribution of the covariates is such that many observations have a true response probability far from p^*. This makes intuitive sense, but these asymptotic formulae give analytical support to such intuition and provide explicit quantitative descriptions of the local accuracy of the weighted score estimators.

B Weight function from weighted likelihood

We show in this section that adapting the method of Copas (1995) to weight individual contributions to the log-likelihood by some function of $\mathbf{x}_i'\boldsymbol{\beta} \in \mathbb{R}$ leads to estimating equations formally identical to (1.5), but with the addition of a multiplicative factor that yields an undesirable weight function on the level of the score. This is due to the consistency requirement that the estimating equations be unbiased at the correctly specified model.

Suppose, then, that we wish to maximize a weighted log-likelihood of the form

$$\ell_h(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[w_h(\mathbf{x}_i'\boldsymbol{\beta}) l_i(\boldsymbol{\beta}) - M_i(\boldsymbol{\beta}) \right], \tag{B1}$$

where $l_i(\boldsymbol{\beta}) = \log[G(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \overline{G}(\mathbf{x}_i'\boldsymbol{\beta})^{1-y_i}]$ is the binary likelihood. Following Copas (1995), the correction term in (B1) is $M_i(\boldsymbol{\beta})$, a function with gradient

$$\begin{aligned} \frac{\partial M_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial w_h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathbb{E}\left[l_i(\boldsymbol{\beta}) \right] \\ &= \frac{\partial w_h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} [G(\mathbf{x}_i'\boldsymbol{\beta}) \log G(\mathbf{x}_i'\boldsymbol{\beta}) + \overline{G}(\mathbf{x}_i'\boldsymbol{\beta}) \log \overline{G}(\mathbf{x}_i'\boldsymbol{\beta})]. \end{aligned} \tag{B2}$$

This guarantees consistency of the resulting estimators, as the score will have expectation $\mathbf{0}$. To obtain the weighted maximum likelihood estimators, we then need to solve

$$\begin{aligned} \frac{\partial \ell_h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[\frac{\partial w_h(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} l_i(\boldsymbol{\beta}) + w_h(\mathbf{x}_i'\boldsymbol{\beta}) \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial M_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right] \\ &= \sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}) \left\{ -\mathbf{x}_i \left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - \eta^*}{h^2} \right) l_i(\boldsymbol{\beta}) + \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right. \\ &\quad \left. + \mathbf{x}_i \left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - \eta^*}{h^2} \right) \mathbb{E}\left[l_i(\boldsymbol{\beta}) \right] \right\} \\ &= \sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}) \left\{ -\mathbf{x}_i \left(\frac{\mathbf{x}_i'\boldsymbol{\beta} - \eta^*}{h^2} \right) \left(l_i(\boldsymbol{\beta}) - \mathbb{E}\left[l_i(\boldsymbol{\beta}) \right] \right) + \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \\ &= \sum_{i=1}^{n} w_h(\mathbf{x}_i'\boldsymbol{\beta}) \left(-\frac{1}{h^2} (\mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\eta^*}{h^2} \mathbf{x}_i'\boldsymbol{\beta} + 1 \right) \mathbf{U}_i(\boldsymbol{\beta}) \end{aligned}$$

$$= \mathbf{0}.$$

We see from this that the estimating equations take the form of a weighted score function, where the weight is no longer $w_h(\mathbf{x}_i'\boldsymbol{\beta})$, but rather $w_h(\mathbf{x}_i'\boldsymbol{\beta})\,[-h^{-2}(\mathbf{x}_i'\boldsymbol{\beta})^2 + h^{-2}\eta^*(\mathbf{x}_i'\boldsymbol{\beta}) + 1]$. The requirement for consistency of the local likelihood estimators that we have unbiased estimating equations has led us to a weight function on the level of the score that is not only not maximized at η^*, but may in fact be negative as well.

Chapter 2: Semiparametric Time-to-Event Modeling in the Presence of a Latent Progression Event

2.1 Introduction

The analysis of time-to-event data in areas of biomedical research dealing with progressive diseases is particularly challenging because a large part of the disease process is not observed. For example, a disease is typically diagnosed only when symptoms reach the point where a patient seeks medical attention, or the disease is detected through some screening program, with the point of onset of detectable disease being unobserved. Another example, studied in detail in this chapter, is metastatic progression in prostate cancer, where cancer-specific death occurs due to metastasis, whose onset is unobserved. It is usually the case that we are interested in the effect of covariates in time to onset of disease rather than time to diagnosis, so we are confronted with a combination of the usual right censoring characteristic of survival analysis generally (Kalbfleisch and Prentice, 2002) as well as left- or interval censoring due to our particular application (Dejardin et al., 2010; Tsodikov et al., 1995).

To make these concepts concrete, assume a well defined starting point and let T_1 denote

the time to the terminal event, and let T_0 denote the time to the latent event, which must occur before T_1. We assume that T_1 is observed, but is subject to right censoring, which is indicated by $\Delta = 0$; $\Delta = 1$ means that the terminal event occurs at time T_1. The time T_0, by contrast, is never observed; therefore generally we must rely on T_1 and the structure of the model to inform us about the distribution of T_0.

Our work draws on the literature of frailty models, which involve an unobserved random variable that modifies the hazard of an event. Prior work on this subject dates back to Vaupel et al. (1979), who introduced the concept of frailty variables in life-table analysis and assumed a gamma distribution. A more general distribution for the frailty variable is investigated in Hougaard (1986). Oakes (1989) gives a number of examples of different possible frailty distributions while focusing on the dependence structure induced by such models. Zeng and Lin (2007) provide a unifying theory for inference in semiparametric survival models, but are interested primarily in shifting from the commonly used gamma frailty to normally distributed random effects. Horowitz (1999) proposes a method for fully nonparametric estimation of the baseline hazard and frailty distribution, but confines himself mostly to uncensored data.

An important difference in our model, however, is that the frailty is no longer a random variable but rather a stochastic process $N_0(t)$ that jumps from 0 to 1 at the time of the latent event. While there has been some work done on such models (e.g., Gjessing et al., 2003), the overwhelming body of research considers only frailties that are properly random variables, that is, are fixed at time 0. Hu and Tsodikov (2014b) develop a similar model for cancer progression that also makes use of a jump process as the frailty. However, they did not devise an efficient EM approach, the key contribution of this chapter. Also, theirs is a marked survival response model where the latent event does not necessarily precede the terminal event.

The multi-state model proposed by Dejardin et al. (2010) with two events, progression (of cancer) and death, is similar to ours in that it also assumes a (recurrent) ordering to

the two events, with progression necessarily preceding death. However, they specify a parametric form for the baseline hazard, while we propose a method to estimate this function nonparametrically. Frydman and Szarek (2009) also propose a multi-state Markov model and derive nonparametric maximum likelihood estimators, but in their scenario there is no natural ordering to the two events: one is assumed to be nonfatal but related to the disease process, while the other is death. Lin et al. (1999) also deal with events that have a recurrent ordering in time with the goal of jointly modeling the gap time distribution between serial events, the primary statistical problem being dependent censoring induced by the time ordering of the events.

In each of these articles, all events are at least partially observable. In our model, by contrast, the latent event is by definition never observed. Another unique feature of our model is that the frailty term is linked to the observed event process through the infinite-dimensional common parameter, the baseline hazard. It is possible to maximize the marginal likelihood directly, but computational issues resulting from the large number of parameters in semiparametric models make this option unattractive (Tsodikov, 2003). Instead we propose and derive an EM algorithm for estimating the baseline hazard for this model. Inference for the parametric part of the model is based on standard profile likelihood theory for semiparametric models (Murphy and van der Vaart, 2000; Tsodikov, 2003).

2.2 Model and likelihood

2.2.1 Data structure and notation

There are two events associated with our model, latent and terminal. The time to the latent event is denoted as T_0 and is never observed; time to the terminal event is T_1. By definition, the latent event must precede the terminal one: $T_0 \leq T_1$. There is a censoring time C that is (conditional on covariates) independent of T_0 and T_1. We observe $(X, \Delta, \mathbf{z}')$, where $X = \min\{T_1, C\}$ and $\Delta = \mathbb{1}(X = T_1)$ and $\mathbf{z}$ is a vector of covariates; $\mathbb{1}(\cdot)$ is the indicator

function, taking the value 1 if $\cdot$ is true and 0 otherwise. The maximum follow-up time is τ.

Note that if $\Delta = 1$, we must have $T_0 \leq T_1 \leq C$. However, if $\Delta = 0$, then either

(i) $C \leq T_0$ or

(ii) $T_0 \leq C \leq T_1$.

Thus, we are unable to tell from observed data whether or not the latent event has occurred in the case of a censored observation.

2.2.2 Model

As in Dejardin et al. (2010), we formulate our model in two parts. The first is the marginal hazard of the latent event $d\Lambda_0$, and the second is the conditional hazard of the terminal event given time to the latent event $d\Lambda_1$.

$$\begin{aligned} d\Lambda_0(t|\mathbf{z}) &= \lim_{h\to 0} \frac{P\left(T_0 \in [t, t+h) | T_0 \geq t, \mathbf{z}\right)}{h} \\ &= \mu \, dH(t) \qquad (2.1) \\ d\Lambda_1(t|T_0 = t_0, \mathbf{z}) &= \lim_{h\to 0} \frac{P\left(T_1 \in [t, t+h) | T_1 \geq t, T_0 = t_0, \mathbf{z}\right)}{h} \\ &= \mathbb{1}(t > t_0)\eta \, dH(t). \qquad (2.2) \end{aligned}$$

The latent and terminal events may be thought of as two events in a reccurent-events model: when $\eta = \mu$ for a subject, these will be two events in a Poisson process. For $\eta > \mu$, the terminal event will be accelerated following the latent event (relative to such a Poisson process), while for $\eta < \mu$ the reverse is true. The baseline hazard $H(\cdot)$ models the temporal pattern of the disease progression (see Hu and Tsodikov, 2014b, for a mechanistic justification and detailed discussion). Covariates $\mathbf{z}$ will enter the model through μ and η: specifically, for $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_\eta', \boldsymbol{\beta}_\mu'\right)'$, we have $\eta = \eta(\boldsymbol{\beta}) = e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}_\eta}$ and $\mu = \mu(\boldsymbol{\beta}) = e^{\mathbf{z}'\boldsymbol{\beta}_\mu}$. For notational simplicity we refer to a single covariate vector $\mathbf{z}$ and assume that it contains all covariates

relevant to the model, while noting that it would be possible to restrict some components of either $\boldsymbol{\beta}_\eta$ or $\boldsymbol{\beta}_\mu$ to be zero.

An interesting feature of this model is that when the parameters μ and η are specified as exponential functions of linear combinations of the covariates, an identifiability problem emerges. Specifically, when the model contains the same set of covariates in both μ and η, for any given set of parameters $(\beta_0, \boldsymbol{\beta}'_\eta, \boldsymbol{\beta}'_\mu)'$, we have the same marginal distribution of time to the observed event (integrating over the unobserved time to the latent event) with $(-\beta_0, \boldsymbol{\beta}'_\mu, \boldsymbol{\beta}'_\eta)'$. The source of the issue is the fact that summands are exchangeable while the sum is fixed. External considerations are needed to fully identify the model.

The main methodological contribution of this chapter is the derivation of an EM algorithm for estimation of the baseline hazard in this class of joint models, presented in Appendix D, for which the non-terminal event is never observed and must precede the terminal event. The model belongs to a class of dynamic stochastic process frailty models where the distribution of the frailty process and the conditional model have a common infinite-dimensional parameter. No EM solutions are available for this class of problems to the best of our knowledge. As this is based on a full-likelihood approach, it is asymptotically fully efficient, in contrast to Breslow estimators based on martingale estimating equations.

2.2.3 Likelihood

The likelihood for a single subject with observed data (X, Δ) conditional on time to the latent event $T_0 = t_0$ is

$$L_0 = [\mathbb{1}(t_0 < X)\eta \, dH(X)]^{\Delta} \, e^{-\mathbb{1}(t_0 < X)\eta[H(X) - H(t_0)]}. \tag{2.3}$$

The marginal survival function for the terminal event is the expectation of (2.3) over the distribution of T_0 for a censored observation:

$$G_*(X) = \frac{1}{\eta - \mu}\left[\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}\right]. \tag{2.4}$$

The marginal density function is given by the expectation of (2.3) for a failed observation:

$$g_*(X) = \eta\mu \frac{e^{-\mu H(X)} - e^{-\eta H(X)}}{\eta - \mu}\, dH(X). \tag{2.5}$$

The marginal hazard function is $\lambda_* = g_*/G_*$. (See Appendix C for details.)

Using (2.4) and (2.5), we can write the marginal log-likelihood associated with this model in counting process form:

$$\begin{aligned}\ell\left(H(t);\boldsymbol{\beta}\right) &= \sum_{i=1}^{n} \ell_i\left(H(t);\boldsymbol{\beta}\right) \\ &= \sum_{i=1}^{n} \int_0^\tau \left\{\left[\log \gamma_i\left(H(t);\boldsymbol{\beta}\right) + \log dH(t)\right]\, dN_i(t) - Y_i(t)\gamma_i\left(H(t);\boldsymbol{\beta}\right)\, dH(t)\right\},\end{aligned} \tag{2.6}$$

where

$$\gamma_i\left(H(t);\boldsymbol{\beta}\right) = \eta_i\mu_i \frac{e^{-\mu_i H(t)} - e^{-\eta_i H(t)}}{\eta_i e^{-\mu_i H(t)} - \mu_i e^{-\eta_i H(t)}}$$

and $\mu_i = e^{\mathbf{z}_i'\boldsymbol{\beta}_\mu}, \eta_i = e^{\beta_0 + \mathbf{z}_i'\boldsymbol{\beta}_\eta}$. Note that the marginal hazard using this notation is

$$\gamma_i\left(H(t);\boldsymbol{\beta}\right) dH(t).$$

We define the martingale $dM_i(t)$ based on observed counting processes $N_i(t)$ with respect to filtration $\mathcal{F}(t-) = \sigma\{N_i(s), Y_i(s), \mathbf{z}_i : s \in [0,t), i = 1,\ldots,n\}$ as

$$dM_i(t) = dN_i(t) - Y_i(t)\gamma_i\left(H(t);\boldsymbol{\beta}\right) dH(t),$$

since

$$\begin{aligned}\mathbb{E}[dN_i(t)|\mathcal{F}_i(t-)] &= Y_i(t)P[dN_i(t)=1|Y_i(t)=1] \\ &= Y_i(t)\gamma_i\left(H(t);\boldsymbol{\beta}\right)dH(t)\end{aligned}$$

by definition of the hazard rate under the true model. Note that $Y_i(t) = \mathbb{1}(X_i \geq t)$ is the at-risk function for subject i.

2.3 Nonparametric maximum likelihood estimation

2.3.1 Functional derivative and score equations

Define derivatives of γ_i with respect to H and $\boldsymbol{\beta}$ as

$$\dot{\gamma}_{i,H}\left(H(t);\boldsymbol{\beta}\right) = \frac{\partial\gamma_i\left(H(t);\boldsymbol{\beta}\right)}{\partial dH(t)} \tag{2.7}$$

$$\dot{\gamma}_{i,\boldsymbol{\beta}}\left(H(t);\boldsymbol{\beta}\right) = \frac{\partial\gamma_i\left(H(t);\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}, \tag{2.8}$$

respectively. The functional derivative (2.7) is as described in Hu and Tsodikov (2014a, Section 3.2) and corresponds to taking the derivative with respect to a jump in H at time t when H is a step function. Generally, for a linear functional of the form $J(f) = \int_0^t \varphi(x)\,df(x)$, the functional derivative is defined as

$$\begin{aligned}\delta_s J &\equiv \frac{\partial J}{\partial df(s)} \\ &= \int_0^t \varphi(x)\,d\frac{\partial(f+\epsilon g)}{\partial\epsilon}\bigg|_{\epsilon=0,g=\mathbb{1}(x>s)} \\ &= \int_0^t \varphi(x)\,d\,\mathbb{1}(x>s) \\ &= \varphi(s)\,\mathbb{1}(t\geq s).\end{aligned}$$

Using this definition, we have

(i) $\delta_s H(t) = \frac{\partial H(t)}{\partial dH(s)} = \mathbb{1}(t \geq s)$

(ii) $\delta_s \log dH(t) = \frac{\partial \log dH(t)}{\partial dH(s)} = \frac{1}{dH(t)} \frac{\partial dH(t)}{\partial dH(s)} = \frac{1}{dH(t)} d\,\mathbb{1}(t \geq s)$.

To obtain the score equations, first we differentiate the log-likelihood with respect to the infinite-dimensional parameter $H(\cdot)$, making use of the identity $\delta_s f(t) = f(t)\delta_s \log f(t)$ to write in terms of the martingale $dM_i(t)$:

$$\mathcal{U}_{dH(s)} = \sum_{i=1}^{n} \left[\int_s^\tau \frac{\dot{\gamma}_{i,H}\left(H(t);\boldsymbol{\beta}\right)}{\gamma_i\left(H(t);\boldsymbol{\beta}\right)} dM_i(t) + \frac{dM_i(s)}{dH(s)} \right]. \tag{2.9}$$

In order to facilitate the asymptotic analysis, we replace s by a dummy variable x and integrate this expression to obtain the alternative form of the score:

$$\mathcal{U}_{H(s)} = \sum_{i=1}^{n} \int_0^\tau \left[\frac{\dot{\gamma}_{i,H}\left(H(t);\boldsymbol{\beta}\right)}{\gamma_i\left(H(t);\boldsymbol{\beta}\right)} H(s \wedge t) + \mathbb{1}(t < s) \right] dM_i(t). \tag{2.10}$$

Define $\varepsilon_i(t, s; H, \boldsymbol{\beta}) = \frac{\dot{\gamma}_{i,H}(H(t);\boldsymbol{\beta})}{\gamma_i(H(t);\boldsymbol{\beta})} H(s \wedge t) + \mathbb{1}(t < s)$. As shown in Hu and Tsodikov (2014a, Supplementary Materials B), the linear transform $\int_0^\tau \varepsilon_i(t, s; H, \boldsymbol{\beta})\, dM_i(t)$ is a martingale as a process in s under the true model when $\varepsilon_i(t, s; H, \boldsymbol{\beta})$ does not depend on s for $t < s$, as is the case here.

The score function for the regression parameters $\boldsymbol{\beta}$ is

$$\mathcal{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \int_0^\tau \frac{\dot{\gamma}_{i,\boldsymbol{\beta}}\left(H(t);\boldsymbol{\beta}\right)}{\gamma_i\left(H(t);\boldsymbol{\beta}\right)} dM_i(t). \tag{2.11}$$

2.3.2 EM algorithm

Using the relative expectation approach of Tsodikov (2003), we have derived the EM algorithm for this problem; details are given in Appendix D. This results in the following

equation that defines iterations over $k = 0, 1, 2, \ldots$, that converges $dH^{(k)} \to \widehat{dH}$ as $k \to \infty$.

$$0 = \sum_{i=1}^{n} \left\{ \frac{dN_i(s)}{dH^{(k+1)}(s)} - \Psi_i^{(k)}(s) + \left[\frac{dH^{(k)}(s)}{dH^{(k+1)}(s)} - 1 \right] \theta_i^{(k)}(s) \right\}, \tag{2.12}$$

where

$$\Psi_i^{(k)}(s) = Y_i(s) \frac{\eta^{1-\Delta_i} \mu e^{-\mu H^{(k)}(X_i)} - \eta \mu^{1-\Delta_i} e^{-\eta H^{(k)}(X_i)}}{\eta^{1-\Delta_i} e^{-\mu H^{(k)}(X_i)} - \mu^{1-\Delta_i} e^{-\eta H^{(k)}(X_i)}}$$

and

$$\theta_i^{(k)}(s) = (\eta - \mu) \mu^{1-\Delta_i} \frac{Y_i(s) e^{-\eta H^{(k)}(X_i) + (\eta - \mu) H^{(k)}(s)} + (1 - \Delta_i) \left[1 - Y_i(s) \right] e^{-\mu H^{(k)}(s)}}{\eta^{1-\Delta_i} e^{-\mu H^{(k)}(X_i)} - \mu^{1-\Delta_i} e^{-\eta H^{(k)}(X_i)}}.$$

Equation (2.12) constitutes a self-consistency equation (Tsodikov, 2003). Note that at the solution, i.e., when $dH^{(k)} = dH^{(k+1)} = \widehat{dH}$, the second term in (2.12) disappears, leaving the score equation corresponding to the marginal likelihood for the time to the terminal event.

Solving (2.12) for the next-iteration hazard, we obtain a Breslow-type expression

$$dH^{(k+1)}(s) = \frac{\sum_{i=1}^{n} dN_i(s) + \left[\sum_{i=1}^{n} \theta_i^{(k)}(s) \right] dH^{(k)}(s)}{\sum_{i=1}^{n} \left[\Psi_i^{(k)}(s) + \theta_i^{(k)}(s) \right]}. \tag{2.13}$$

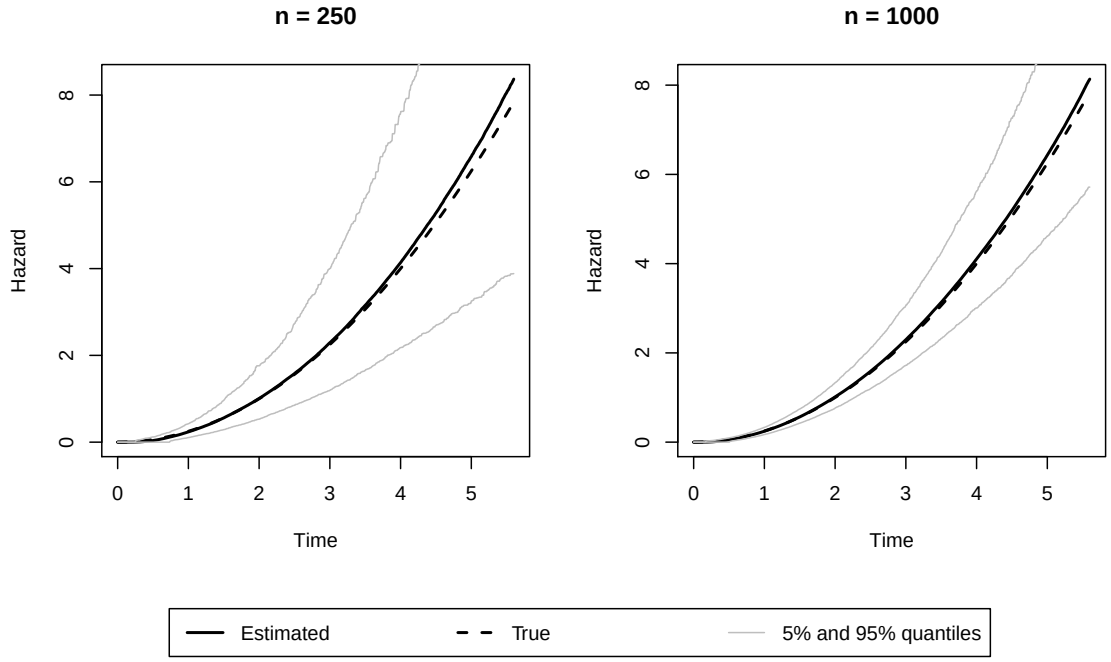
The second term in the numerator may be thought of as the imputed $\sum_{i=1}^{n} dN_{0i}(s)$, while the denominator could be called an effective imputed at-risk process for the combined latent and terminal failures.

2.4 Simulation study

This section presents a simulation study to illustrate our method. The simulation settings are as follows. The baseline hazard was $H(x) = \frac{1}{4}x^2$. The true parameter vectors were $\boldsymbol{\beta}_{\eta 0} = (-1, 1, -2)'$ and $\boldsymbol{\beta}_{\mu 0} = (2, -1)'$. The censoring distribution was $U(0, \tau)$, $\tau = 7$.

Because of the potential identifiability issue with this model, we chose two scenarios from

Figure 2.1. Function estimates over 1000 simulated data sets for the identifiable scenario. The solid black line is the mean of the estimated hazard functions, while the truth is given as the dashed black line.



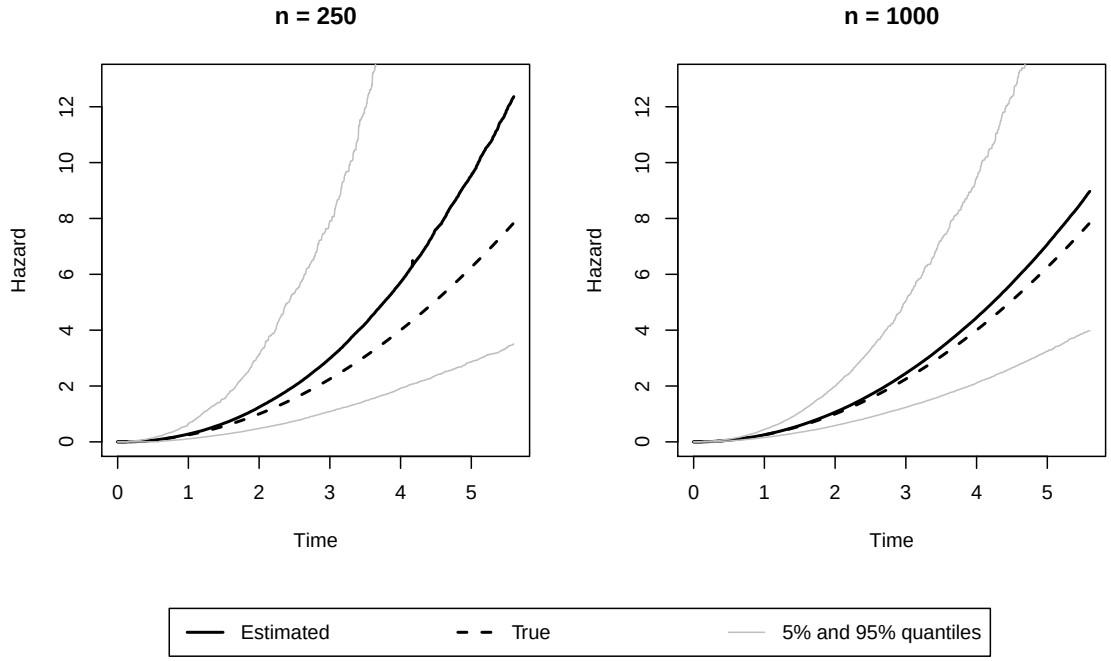
which to simulate data:

Unidentifiable case For these simulations, the covariates were $Z_1 \sim N(0,1), Z_2 \sim B(0.25)$, with $\log \eta_i = (1, z_{1i}, z_{2i})'\boldsymbol{\beta}_{\eta 0}$ and $\log \mu_i = (z_{1i}, z_{2i})'\boldsymbol{\beta}_{\mu 0}$. This gives rise to unidentifiability in the model, as changing the sign of the intercept and exchanging the roles of μ and η will result in an identical marginal distribution of time to the terminal event.

Identifiable case For these simulations, an additional covariate, $Z_3 \sim B(0.5)$, was included. The model for η is the same as in the unidentifiable case, but now $\log \mu_i = (z_{3i}, z_{2i})'\boldsymbol{\beta}_{\mu 0}$. The model is now identifiable, and only one solution to the likelihood equations exists.

We chose two sample sizes, $n = 250$ and $n = 1000$, to see the effects on estimation of a relatively small and large sample, respectively. For each sample size in each of the two scenarios, 1000 data sets were generated. Initial values were chosen by fitting either one (unidentifiable

Figure 2.2. Function estimates over 1000 simulated data sets for the unidentifiable scenario. The solid black line is the mean of the estimated hazard functions, while the truth is given as the dashed black line. For this scenario, a 0.5% trim was used with the mean due to the presence of a few outlying function estimates.



scenario) or two (identifiable scenario) Cox models. We used the R function `optim()` to maximize the profile likelihood function; the L-BFGS-B method (which allows box constraints for restricting the intercept estimate to be negative) was used for the unidentifiable case, while the BFGS method was used for the identifiable case. Standard errors were obtained from the numerically evaluated Hessian matrix at the solution.

The results of the simulation study are summarized in Tables 2.1 and 2.2. Beginning with Table 2.1, we see that bias for most parameters in the models decreases with increasing sample size, although for the parameters associated with the continuous covariate Z_1 the bias seems to be negligible even for the smaller sample size. The bias for the intercept parameter, however, seems to be markedly more persistent with increasing sample size, indicating the difficulty in estimating this parameter when the model is unidentifiable.

As we noted a small number of outliers in the parameter estimates for some of the simulated data sets, we chose to compare the estimated standard errors not only with the

Table 2.1. Simulation results for the unidentifiable scenario. The empirical standard deviation (SD), median absolute deviation (MAD), and average standard error of the estimates are normalized by the true values of the parameters so as to provide a more equitable basis for comparisons between them.

n	Model	Covariates	Truth	Avg. est.	Emp. SD	Emp. MAD	Avg. SE
250	η	(Intercept)	-1	-0.905	1.014	1.009	1.204
		$Z_1 \sim N(0,1)$	1	1.030	0.302	0.265	0.340
		$Z_2 \sim B(0.25)$	-2	-1.825	0.470	0.299	0.339
	μ	$Z_1 \sim N(0,1)$	2	2.046	0.486	0.281	0.298
		$Z_2 \sim B(0.25)$	-1	-1.209	1.719	1.543	1.389
1000	η	(Intercept)	-1	-0.935	0.703	0.655	0.631
		$Z_1 \sim N(0,1)$	1	0.979	0.172	0.175	0.172
		$Z_2 \sim B(0.25)$	-2	-1.965	0.192	0.122	0.133
	μ	$Z_1 \sim N(0,1)$	2	2.059	0.194	0.142	0.149
		$Z_2 \sim B(0.25)$	-1	-0.987	0.995	0.704	0.688

standard deviation (SD) but also the median absolute deviation (MAD), for which the scale factor, $1/\Phi^{-1}(3/4) \approx 1.483$, was chosen to ensure consistency of the MAD for the standard deviation of a normally distributed random variable (Rousseeuw and Croux, 1993). This is appropriate since our estimators are asymptotically normal (by the results of Appendix F), so that any outliers are simply due to numerical issues arising during the estimation process.

From this perspective, then, we may evaluate the agreement between the estimated standard errors and the true variability of the estimators. In general, the mean of the estimated standard errors falls between the SD and the MAD of the estimated parameters. With the larger sample size, we see much better agreement between the mean estimated SE and the MAD in particular, suggesting that the asymptotic approximation of the covariance matrix for the profile likelihood is good for samples of this size.

Table 2.2 exhibits much the same patterns, although in this case, we do not find difficulty in estimating the intercept as we did in the unidentifiable scenario. However, we do see a clearer trend in the variability of estimates in the η model as compared with the μ model: the SD and MAD of the estimates (which generally agree closely with the mean SE estimates) are substantially greater for the parameters in the μ part of the model. This is sensible,

Table 2.2. Simulation results for the identifiable scenario. The empirical standard deviation (SD), median absolute deviation (MAD), and average standard error of the estimates are normalized by the true values of the parameters so as to provide a more equitable basis for comparisons between them. Note that the only difference between the simulations for this scenario and the unidentifiable scenario is one of the covariates in the μ part of the model.

n	Model	Covariates	Truth	Avg. est.	Emp. SD	Emp. MAD	Avg. SE
250	η	(Intercept)	-1	-0.900	0.513	0.460	0.487
		$Z_1 \sim N(0,1)$	1	1.066	0.179	0.155	0.160
		$Z_2 \sim B(0.25)$	-2	-1.849	0.394	0.276	0.289
	μ	$Z_3 \sim B(0.5)$	2	2.156	0.653	0.449	0.684
		$Z_2 \sim B(0.25)$	-1	-1.163	1.544	1.178	1.538
1000	η	(Intercept)	-1	-0.998	0.247	0.239	0.239
		$Z_1 \sim N(0,1)$	1	1.015	0.073	0.070	0.071
		$Z_2 \sim B(0.25)$	-2	-1.994	0.135	0.131	0.125
	μ	$Z_3 \sim B(0.5)$	2	2.138	0.348	0.257	0.316
		$Z_2 \sim B(0.25)$	-1	-0.981	0.660	0.566	0.582

as the μ part of the model corresponds to the latent event, which we are unable to observe directly.

2.5 SEER prostate cancer data analysis

To illustrate the use of the proposed method on a real data set, we apply it to SEER registry data on prostate cancer. We examined the survival data from the Detroit SEER registry with year of diagnosis with prostate cancer between 1983 and 2003, which consisted of data on 47,187 men. We included two binary covariates: race (0 if white or 1 if black) and a dichotomized time of diagnosis (0 if pre- or 1 if post-1988, the year PSA screening was introduced). Of these subjects, 26.2% were black and 89.4% were diagnosed in the PSA era (1988 or later); 7.8% died of cancer during the follow-up period.

In prostate cancer, metastasis (the latent event) must occur prior to death due to cancer (the terminal event); the time origin is the point of diagnosis with cancer. In SEER, no detailed post-treatment followup is available, so the time at which the disease becomes metastatic, even in the symptomatic sense, is unknown.

Table 2.3. Parameter estimates (standard errors) from analysis of SEER prostate cancer data. The Cox model estimates are shown for purposes of comparison with the estimates for the η part of the joint model, which pertains to time to death due to cancer given metastasis.

	Parameter	Cox	Joint
Death	Intercept	—	2.136 (0.233)
	Black	0.338 (0.036)	0.301 (0.114)
	Dx post-1988	-0.845 (0.038)	-1.813 (0.381)
Onset	Black	—	0.137 (0.097)
	Dx post-1988	—	0.401 (0.357)

The results of the conventional analysis (involving a simple Cox model fit) as well as the proposed method are shown in Table 2.3. We display only the positive-intercept model, but recall that due to the lack of identifiability of the sign of the intercept, we would obtain the same fit to the data with a negative intercept of the same magnitude and an exchange of the roles of η and μ. We emphasize that this choice of model is possible only through the use of external information; no statistical justification can be made, as both models fit the observed data equally well.

The rationale for our choice of the positive intercept model is that it results in the correct sign for the effect of PSA screening on time to death. In the positive intercept model, this coefficient estimate is negative, which agrees with the Cox model's estimate in sign if not magnitude. Moreover, it is an accepted scientific view that PSA screening prolongs time to death, if only due to the artifact of lead-time bias.

The model allows us to make predictions of the distribution of the time to the latent event, given observed data on a subject. Specifically, for the survival functions, we have (see Appendix E for details)

$$G(t_0|X,\Delta=0)=\begin{cases}\dfrac{\eta e^{-\mu H(X)}-\mu e^{-\eta H(X)+(\eta-\mu)H(t_0)}}{\eta e^{-\mu H(X)}-\mu e^{-\eta H(X)}}, & t_0<X\\ \dfrac{(\eta-\mu)e^{-\mu H(t_0)}}{\eta e^{-\mu H(X)}-\mu e^{-\eta H(X)}}, & t_0\geq X,\end{cases} \tag{2.14}$$

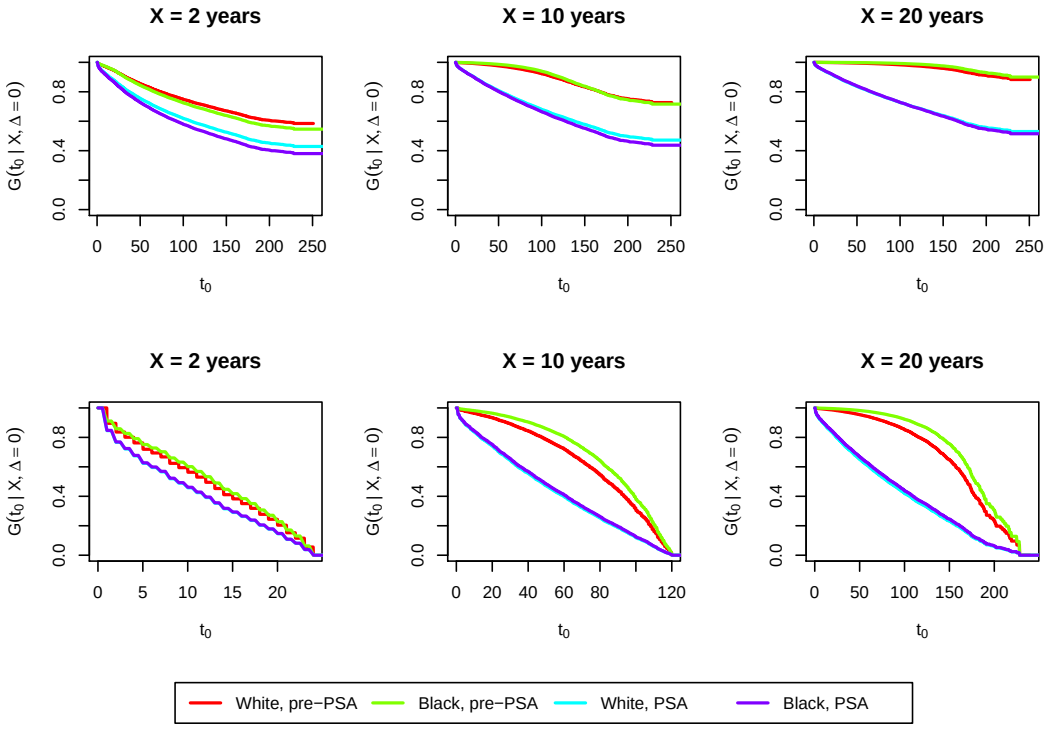
for a censored subject, and

$$G(t_0|X, \Delta = 1) = \begin{cases} \dfrac{e^{(\eta-\mu)H(X)} - e^{(\eta-\mu)H(t_0)}}{e^{(\eta-\mu)H(X)} - 1}, & t_0 < X \\ 0, & t_0 \geq X. \end{cases} \tag{2.15}$$

for a failed subject (that is, a subject who has experienced the terminal event).

This model leads to the plots shown in Figure 2.3. These are the conditional survival functions for onset of metastasis given time to death or censoring, shown in equations (2.14) and (2.15) above. An immediately evident feature of these curves is their ordering, with the PSA-screened population below the non-screened population, indicating an earlier onset of metastasis for these subjects. This may be explained with reference to the selection effect caused by screening (Zelen and Feinleib, 1969): under PSA screening, tumors are detected earlier and treated, in some sense removing these cases from the population. Due to the length bias, these tumors will be generally less aggressive, so the remainder (i.e., the cases which would be depicted in these plots, conditioning on death) will be relatively more agressive, with earlier metastasis.

Figure 2.3. Conditional survival functions for onset of metastasis given observed data (time to death and censoring indicator)—positive-intercept model. Top row is for a hypothetical subject censored at the times indicated, while the bottom row is for a subject who dies at these times. The x-axis of each panel is scaled in months since diagnosis.



2.6 Discussion

In this chapter, we have presented a method for jointly modeling time to a latent event and time to the terminal event, in a semiparametric framework, when the latent event is unobservable and must precede the terminal event. Our approach involves an EM algorithm for estimating the baseline hazard, the derivation of which constitutes the chief methodological contribution of our work. The method is generalizable, in the sense that it is not specific to a certain data structure, but can instead be used with any survival data for which there is interest in factors affecting time to an unobserved event which is known to precede an observed event.

An alternative to our method is a weighted Breslow-type estimator (Chen, 2009) that is also asymptotically fully efficient, although this estimator needs an extension to our class of

problems. However, the weighted Breslow method does not enjoy the property of monotonic convergence intrinsic to our EM approach. This property makes the EM approach a stable, computationally efficient solution that handles the curse of dimensionality in a closed form. In fact, EM will converge even in the unidentifiable case without restricting the sign of the intercept term.

The latent frailty process considered in this chapter is rather primitive (one jump). A generalization to other types of disease progression processes such as continous growth or compartmental progression through stages is an interesting topic for future research. A number of other extensions are possible as well. For example, in certain data sets, partial information on the time to the latent event may be available. In such cases, it would be possible to modify the likelihood function to account for this additional data; in so doing, we would likely be able to substantially increase the precision of our estimates. As was seen in our real data analysis, the parameters pertaining to the unobserved event (metastasis of prostate cancer) were not found to be statistically significant, but if we were able to augment our data set with patients for whom time to metastasis was observed, we would perhaps reduce the estimated standard errors associated with these parameters.

Another important aspect of modeling not addressed in detail in this chapter is the issue of model checking. One simple graphical check of the model fit would involve comparing the marginal survival curves (obtained after getting estimates of the parameters in the model as well as the baseline hazard) with Kaplan-Meier curves of the time to the terminal event. These should be very similar to each other, so any differences could be evidence of problems with the assumed model.

Relaxation of the common baseline hazard assumption would be useful as well. One simple approach would be to apply a monotone parametric transformation to the baseline cumulative hazard in either the marginal hazard of the latent event or the conditional hazard of the terminal event given the latent event. For example, instead of $\mathbb{1}(t > t_0)\eta\,[H(t) - H(t_0)]$, we might replace H by H^α, for some $\alpha > 0$. This would allow for a wide range of alternative

shapes in the hazard for the terminal event relative to the hazard for the latent event.

Appendices

C Derivation of marginal survival and density functions

The marginal survival function for the terminal event is the expectation of (2.3) over the distribution of T_0 for a censored observation:

$$\begin{aligned}
G_*(X) &= \mathbb{E}\left[e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}\right] \\
&= \int e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= \int_0^{H(X)} \mu e^{-\eta H(X)+(\eta-\mu)H(t_0)}\, dH(t_0) + \int_{H(X)}^{\infty} \mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= e^{-\eta H(X)}\frac{\mu}{\eta-\mu}\left[e^{(\eta-\mu)H(X)}-1\right] + e^{-\mu H(X)} \\
&= \frac{1}{\eta-\mu}\left[\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}\right].
\end{aligned}$$

The marginal density function is given by the expectation of (2.3) for a failed observation:

$$\begin{aligned}
g_*(X) &= \mathbb{E}\left[\mathbb{1}(T_0<X)\eta\, dH(t) e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}\right] \\
&= \int_0^{H(X)} \eta\, dH(X) e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= \eta\mu\, dH(X) e^{-\eta H(X)} \int_0^{H(X)} e^{(\eta-\mu)H(t_0)}\, dH(t_0) \\
&= \eta\mu e^{-\eta H(X)}\frac{e^{(\eta-\mu)H(X)}-1}{\eta-\mu}\, dH(X) \\
&= \eta\mu\frac{e^{-\mu H(X)}-e^{-\eta H(X)}}{\eta-\mu}\, dH(X).
\end{aligned}$$

Note that the presence of the term $\mathbb{1}(T_0 < X)$ outside of the exponential renders the integrand zero for $t_0 \geq X$.

D EM algorithm for estimation of baseline hazard

This section presents our derivation of the EM algorithm for our model, which is based on the methods of Tsodikov (2003). Note that the functional derivative

$$\begin{aligned}
\mathcal{U}_0(s) &= \delta_s \log L_0 + \delta_s \log P_0 \\
&= \delta_s \log \left\{ [\mathbb{1}(T_0 < X)\eta dH(X)]^{\Delta} e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]} \right\} \\
&\quad + \delta_s \log \left\{ \mu dH(T_0) e^{-\mu H(T_0)} \right\} \\
&= \frac{\mathbb{1}(X = s)\Delta}{dH(s)} - \mathbb{1}(T_0 < X)\eta \left[\mathbb{1}(X \geq s) - \mathbb{1}(T_0 \geq s) \right] \\
&\quad + \frac{\mathbb{1}(T_0 = s)}{dH(s)} - \mu\, \mathbb{1}(T_0 \geq s) \\
&= \frac{dN(s)}{dH(s)} - \eta \tilde{Y}(s) + \frac{dN_0(s)}{dH(s)} - \mu Y_0(s) \qquad \text{(D1)}
\end{aligned}$$

is the conditional (on T_0) score for a single observation. Here we define

$$\begin{aligned}
\tilde{Y}(s) &= \mathbb{1}(T_0 < X) \left[\mathbb{1}(X \geq s) - \mathbb{1}(T_0 \geq s) \right] \\
&= \mathbb{1}(T_0 \leq s < X),
\end{aligned}$$

while $Y_0(s) = \mathbb{1}(T_0 \geq s)$ and $dN_0(s) = \mathbb{1}(T_0 = s)$.

The rest of this section is organized as follows. First we derive the E step, for the censored and failed cases respectively, then we derive the M step, which has a simple closed-form expression reminiscent of the weighted Breslow-type estimators of Chen (2009).

E step

Consider first the case where $\Delta = 0$, that is, a censored observation at time X. The unconditional score is

$$\mathcal{U}(s) = \mathbb{E}\left[-\eta \tilde{Y}(s) - \mu Y_0(s) + \frac{dN_0(s)}{dH^{(k+1)}(s)} \middle\| L_0^{(k)} \right]$$

$$= -\frac{\mathbb{E}\left[\eta\tilde{Y}(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} - \frac{\mathbb{E}\left[\mu Y_0(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} + \frac{1}{dH^{(k+1)}(s)}\frac{\mathbb{E}\left[dN_0(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]}. \tag{D2}$$

Since the index $(k+1)$ does not appear under the $\mathbb{E}$ operator in (D2), we drop iteration indices for the calculation of the following expectations. Note that $\mathbb{E}[L_0]$, appearing in all denominators in (D2), is just the marginal survival function (2.4); denote this as $S(t) = \mathbb{E}\left[e^{-\mathbb{1}(T_0<t)\eta[H(t)-H(T_0)]}\right]$.

First we calculate $\mathbb{E}\left[\tilde{Y}(s)L_0\right]$:

$$\begin{aligned}
\mathbb{E}\left[\tilde{Y}(s)L_0\right] &= \int \mathbb{1}(t_0 < s < X)e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= \mathbb{1}(X > s)\mu \int_0^{H(s)} e^{-\eta H(X)+\eta H(t_0)-\mu H(t_0)}\, dH(t_0) \\
&= Y(s)\frac{\mu}{\eta-\mu}\left[e^{-\eta H(X)+(\eta-\mu)H(s)} - e^{-\eta H(X)}\right].
\end{aligned} \tag{D3}$$

Next we calculate $\mathbb{E}[Y_0(s)L_0]$. To do this we must consider two cases: first, the case where $s \le X$, and second, the case where $s > X$.

- $s \le X$

$$\begin{aligned}
\mathbb{E}[Y_0(s)L_0] &= \mathbb{E}\left[e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}Y_0(s)\right] \\
&= \int e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\,\mathbb{1}(t_0 > s)\mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= \mu e^{-\eta H(X)} \int_{H(s)}^{H(X)} e^{(\eta-\mu)H(t_0)}\, dH(t_0) + \int_{H(X)}^{\infty} \mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= \frac{\mu}{\eta-\mu}e^{-\eta H(X)}\left[e^{(\eta-\mu)H(X)} - e^{(\eta-\mu)H(s)}\right] + e^{-\mu H(X)} \\
&= \frac{\mu}{\eta-\mu}\left[e^{-\mu H(X)} - e^{-\eta H(X)+(\eta-\mu)H(s)}\right] + e^{-\mu H(X)}.
\end{aligned}$$

- $s > X$

$$\mathbb{E}[Y_0(s)L_0] = \mathbb{E}\left[e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}Y_0(s)\right]$$

$$
\begin{aligned}
&= \int e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\,\mathbb{1}(t_0>s)\mu e^{-\mu H(t_0)}\,dH(t_0)\\
&= \int_{H(s)}^{\infty} \mu e^{-\mu H(t_0)}\,dH(t_0)\\
&= e^{-\mu H(s)}.
\end{aligned}
$$

In a single expression, this is

$$
\mathbb{E}\left[Y_0(s)L_0\right] = Y(s)\left\{\frac{\mu}{\eta-\mu}\left[e^{-\mu H(X)} - e^{-\eta H(X)+(\eta-\mu)H(s)}\right] + e^{-\mu H(X)}\right\} + [1-Y(s)]e^{-\mu H(s)}. \tag{D4}
$$

Finally, we calculate $\mathbb{E}\left[dN_0(s)L_0\right]$. This is simple since the integral is only positive at a single point, $t_0 = s$:

$$
\begin{aligned}
\mathbb{E}\left[dN_0(s)L_0\right] &= \int dN_0(s)e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\mu e^{-\mu H(t_0)}\,dH(t_0)\\
&= e^{-\mathbb{1}(s<X)\eta[H(X)-H(s)]}\mu e^{-\mu H(s)}\,dH(s)\\
&= \mu\left\{Y(s)e^{-\eta H(X)+(\eta-\mu)H(s)} + [1-Y(s)]e^{-\mu H(s)}\right\}\,dH(s). \qquad \text{(D5)}
\end{aligned}
$$

Combining these results, when $\Delta = 0$ it may be shown that the unconditional score can be written as

$$
\begin{aligned}
\mathcal{U}(s) = &-Y(s)\frac{\eta\mu e^{-\mu H(X)} - \eta\mu e^{-\eta H(X)}}{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}}\\
&+ \left[\frac{dH^{(k)}(s)}{dH^{(k+1)}(s)} - 1\right](\eta-\mu)\frac{Y(s)\mu e^{-\eta H(X)+(\eta-\mu)H(s)} + [1-Y(s)]\mu e^{-\mu H(s)}}{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}}.
\end{aligned}
$$

It is apparent from this that when $dH^{(k)} = dH^{(k+1)}$, i.e., at convergence, the second term disappears and we are left with the marginal score,

$$
\mathcal{U}(s) = -Y(s)\frac{\eta\mu e^{-\mu H(X)} - \eta\mu e^{-\eta H(X)}}{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}}.
$$

Now we turn to the case of $\Delta = 1$, an observation failed at time X. We begin with

$$\begin{aligned}\mathcal{U}(s) &= \mathbb{E}\left[\frac{dN(s)}{dH^{(k+1)}(s)} - \eta\tilde{Y}(s) - \mu Y_0(s) + \frac{dN_0(s)}{dH^{(k+1)}(s)}\middle\| L_0^{(k)}\right] \\ &= \frac{dN(s)}{dH^{(k+1)}(s)} - \frac{\mathbb{E}\left[\eta\tilde{Y}(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} - \frac{\mathbb{E}\left[\mu Y_0(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} + \frac{1}{dH^{(k+1)}(s)}\frac{\mathbb{E}\left[dN_0(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]}. \quad \text{(D6)}\end{aligned}$$

Note that now

$$L_0 = \mathbb{1}(T_0 < X)\eta dH(X)e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}.$$

Its expectation over the distribution of T_0 is given by (2.5).

Now we proceed through the same steps as for the $\Delta = 0$ case. Some of the steps are simplified due to the $\mathbb{1}(T_0 < X)$ term, since

$$\begin{aligned}\mathbb{1}(T_0 < X)Y_0(s) &= \mathbb{1}(T_0 < X)\,\mathbb{1}(T_0 \geq s) \\ &= \mathbb{1}(s \leq T_0 < X).\end{aligned}$$

First we calculate $\mathbb{E}\left[\eta\tilde{Y}(s)L_0\right]$. We may ignore the term $\mathbb{1}(T_0 < X)$ appearing in L_0 since

$$\begin{aligned}\mathbb{1}(T_0 < X)\tilde{Y}(s) &= \mathbb{1}(T_0 < X)^2\left[\mathbb{1}(X \geq s) - \mathbb{1}(T_0 \geq s)\right] \\ &= \mathbb{1}(T_0 < X)\left[\mathbb{1}(X \geq s) - \mathbb{1}(T_0 \geq s)\right] \\ &= \mathbb{1}(T_0 \leq s < X).\end{aligned}$$

Now we have

$$\begin{aligned}\mathbb{E}\left[\eta\tilde{Y}(s)L_0\right] &= \eta\int \mathbb{1}(t_0 < s < X)\eta dH(X)e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]}\mu e^{-\mu H(t_0)}\,dH(t_0) \\ &= \eta^2\mu e^{-\eta H(X)}dH(X)Y(s)\int_0^{H(s)} e^{(\eta-\mu)H(t_0)}\,dH(t_0) \\ &= Y(s)\eta^2\mu\frac{e^{-\eta H(X)+(\eta-\mu)H(s)} - e^{-\eta H(X)}}{\eta-\mu}dH(X). \quad \text{(D7)}\end{aligned}$$

Now we calculate $\mathbb{E}\left[\mu Y_0(s) L_0\right]$:

$$\begin{aligned}
\mathbb{E}\left[\mu Y_0(s) L_0\right] &= \mathbb{E}\left[\mu Y_0(s)\, \mathbb{1}(T_0 < X) \eta dH(X) e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}\right] \\
&= \eta\mu^2 e^{-\eta H(X)} dH(X) Y(s) \int_{H(s)}^{H(X)} e^{(\eta-\mu)H(t_0)}\, dH(t_0) \\
&= Y(s)\eta\mu^2 \frac{e^{-\mu H(X)} - e^{-\eta H(X)+(\eta-\mu)H(s)}}{\eta-\mu} dH(X). \qquad \text{(D8)}
\end{aligned}$$

The calculation of $\mathbb{E}\left[dN_0(s) L_0\right]$ is very similar in the failed case to what it is in the censored case. Note first that $dN_0(s)\, \mathbb{1}(T_0 < X) = \mathbb{1}(T_0 = s)\, \mathbb{1}(T_0 < X) = \mathbb{1}(T_0 = s)\, \mathbb{1}(s < X) = Y(s)\, \mathbb{1}(T_0 = s)$. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[dN_0(s) L_0\right] &= Y(s)\eta dH(X) \int dN_0(s) e^{-\mathbb{1}(t_0<X)\eta[H(X)-H(t_0)]} \mu e^{-\mu H(t_0)}\, dH(t_0) \\
&= Y(s)\eta\mu dH(X) e^{-\mathbb{1}(s<X)\eta[H(X)-H(s)]} e^{-\mu H(s)}\, dH(s) \\
&= \eta\mu dH(X) Y(s) \left\{Y(s) e^{-\eta H(X)+(\eta-\mu)H(s)} + [1-Y(s)] e^{-\mu H(s)}\right\} dH(s) \\
&= \eta\mu dH(X) Y(s) e^{-\eta H(X)+(\eta-\mu)H(s)}\, dH(s). \qquad \text{(D9)}
\end{aligned}$$

Combining equations (D7), (D8), and (D9), we have for the unconditional score

$$\begin{aligned}
\mathcal{U}(s) = \frac{dN(s)}{dH^{(k+1)}(s)} &- Y(s) \frac{\mu e^{-\mu H(X)} - \eta e^{-\eta H(X)}}{e^{-\mu H(X)} - e^{-\eta H(X)}} \\
&+ \left[\frac{dH^{(k)}(s)}{dH^{(k+1)}(s)} - 1\right] \frac{(\eta-\mu) Y(s) e^{-\eta H(X)+(\eta-\mu)H(s)}}{e^{-\mu H(X)} - e^{-\eta H(X)}}.
\end{aligned}$$

We see from this that if we are at a fixed point of the algorithm, so that $dH^{(k)} = dH^{(k+1)}$, the last term above disappears and we again obtain the marginal score,

$$\mathcal{U}(s) = \frac{dN(s)}{dH^{(k+1)}(s)} - Y(s) \frac{\mu e^{-\mu H(X)} - \eta e^{-\eta H(X)}}{e^{-\mu H(X)} - e^{-\eta H(X)}}.$$

M step

Denote the marginal score, for $\Delta_i \in \{0, 1\}$, as

$$\mathcal{U}_i(s) = \frac{dN_i(s)}{dH^{(k+1)}(s)} - \Psi_i^{(k)}(s), \tag{D10}$$

where

$$\Psi_i^{(k)}(s) = Y_i(s)\frac{\eta^{1-\Delta_i}\mu e^{-\mu H^{(k)}(X_i)} - \eta\mu^{1-\Delta_i}e^{-\eta H^{(k)}(X_i)}}{\eta^{1-\Delta_i}e^{-\mu H^{(k)}(X_i)} - \mu^{1-\Delta_i}e^{-\eta H^{(k)}(X_i)}}.$$

Note that the scores derived above consist of the marginal score plus a "correction term,"

$$\left[\frac{dH^{(k)}(s)}{dH^{(k+1)}(s)} - 1\right]\theta_i^{(k)}(s),$$

where $\theta_i^{(k)}(s)$ is given by

$$\theta_i^{(k)}(s) = (\eta - \mu)\mu^{1-\Delta_i}\frac{Y_i(s)e^{-\eta H^{(k)}(X_i)+(\eta-\mu)H^{(k)}(s)} + (1-\Delta_i)\left[1 - Y_i(s)\right]e^{-\mu H^{(k)}(s)}}{\eta^{1-\Delta_i}e^{-\mu H^{(k)}(X_i)} - \mu^{1-\Delta_i}e^{-\eta H^{(k)}(X_i)}}. \tag{D11}$$

Now, we want to solve

$$0 = \sum_{i=1}^{n}\left\{\frac{dN_i(s)}{dH^{(k+1)}(s)} - \Psi_i^{(k)}(s) + \left[\frac{dH^{(k)}(s)}{dH^{(k+1)}(s)} - 1\right]\theta_i^{(k)}(s)\right\},$$

which implies that the Breslow-type estimator is

$$dH^{(k+1)}(s) = \frac{\sum_{i=1}^{n} dN_i(s) + \left[\sum_{i=1}^{n}\theta_i^{(k)}(s)\right]dH^{(k)}(s)}{\sum_{i=1}^{n}\left[\Psi_i^{(k)}(s) + \theta_i^{(k)}(s)\right]}. \tag{D12}$$

This constitutes a self-consistency equation that is solved iteratively (Tsodikov, 2003).

E Prediction of survival function for the latent event

Prediction of time to the latent event is an important goal of analysis using models of this kind, as noted in Zeng and Lin (2006). This section presents the derivation of equations (2.14) and (2.15). We are interested in the survival function for the latent event, given observed data and estimates of η, μ, H:

$$G(t_0|X,\Delta) = \frac{\int_{H(t_0)}^{\infty} L_0(u)\mu e^{-\mu H(u)}\, dH(u)}{\int_0^{\infty} L_0(u)\mu e^{-\mu H(u)}\, dH(u)}. \tag{E1}$$

Consider first the case where $\Delta = 0$, that is, a censored observation at time X. Note that the denominator of (E1) is the expectation of L_0 with respect to the distribution of T_0. Now we calculate the numerator of (E1). Assuming first that $t_0 < X$, we have

$$\begin{aligned}
\int_{H(t_0)}^{\infty} L_0(u)\mu e^{-\mu H(u)}\, dH(u) &= \int_{H(u)}^{\infty} e^{-\mathbb{1}(u<X)\eta[H(X)-H(u)]}\mu e^{-\mu H(u)}\, dH(u) \\
&= \int_{H(t_0)}^{H(X)} \mu e^{-\eta H(X)+(\eta-\mu)H(u)}\, dH(u) + \int_{H(X)}^{\infty} \mu e^{-\mu H(u)}\, dH(u) \\
&= e^{-\eta H(X)}\frac{\mu}{\eta-\mu}\left[e^{(\eta-\mu)H(X)} - e^{(\eta-\mu)H(t_0)}\right] + e^{-\mu H(X)}.
\end{aligned} \tag{E2}$$

If, on the other hand, $t_0 \geq X$,

$$\int_{H(t_0)}^{\infty} L_0(u)\mu e^{-\mu H(u)}\, dH(u) = e^{-\mu H(t_0)}. \tag{E3}$$

Therefore, after some algebra we see that the conditional survival function for T_0 given observed X, Δ is

$$G(t_0|X,\Delta=0) = \begin{cases} \dfrac{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)+(\eta-\mu)H(t_0)}}{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}}, & t_0 < X \\ \dfrac{(\eta-\mu)e^{-\mu H(t_0)}}{\eta e^{-\mu H(X)} - \mu e^{-\eta H(X)}}, & t_0 \geq X. \end{cases}$$

Now we turn to the case of $\Delta = 1$, an observation failed at time X. Note that now

$$L_0(T_0) = \mathbb{1}(T_0 < X)\eta dH(X)e^{-\mathbb{1}(T_0<X)\eta[H(X)-H(T_0)]}.$$

Its expectation over the distribution of T_0 is given by (2.5). Now we have for the numerator of (E1)

$$\begin{aligned}\int_{H(t_0)}^{\infty} L_0(u)\mu e^{-\mu H(u)}\, dH(u) &= \eta\mu dH(X)e^{-\eta H(X)} \int_{H(t_0)}^{H(X)} e^{(\eta-\mu)H(u)}\, dH(u) \\ &= \eta\mu e^{-\eta H(X)} \frac{e^{(\eta-\mu)H(X)} - e^{(\eta-\mu)H(t_0)}}{\eta-\mu} dH(X). \qquad \text{(E4)}\end{aligned}$$

Because a failed subject at time X (i.e., $\Delta = 1$) implies that $T_0 < X$, the integrand is only nonzero on $[0, H(X)]$. The conditional survival function for the latent event in the case of a failed observation at time X is thus

$$G(t_0|X, \Delta = 1) = \begin{cases} \dfrac{e^{(\eta-\mu)H(X)} - e^{(\eta-\mu)H(t_0)}}{e^{(\eta-\mu)H(X)} - 1}, & t_0 < X \\ 0, & t_0 \geq X. \end{cases}$$

F Asymptotic properties

This section is adapted from Hu and Tsodikov (2014a, Supplementary Materials C). Let $\Omega = (\boldsymbol{\beta}', \{dH\})$; denote the dimension of $\boldsymbol{\beta}$ as p. Let $\|\cdot\|_\infty$ denote the supremum norm on the interval $[0, \tau]$; let $\|w\|_{TV}$ denote the total variation of $w(t)$ on the interval $[0, \tau]$.

Define $\mathcal{Q} = \{w(t) : \|w\|_{TV} \leq 1\}$ such that $\hat{H}(t)$ may be regarded as a bounded linear functional in $\mathcal{L}^\infty(\mathcal{Q})$, and $\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{H}(t) - H^0(t)\}$ as a random element in the metric space $\mathbb{R}^p \times \mathcal{L}^\infty(\mathcal{Q})$. We denote $\mathcal{H}$ as the compact convex set in the metric space $\mathbb{R}^p \times \mathcal{L}^\infty(\mathcal{Q})$ in which Ω^0 is contained.

Conditions:

1. The true hazard H^0 is strictly increasing and differentiable. Ω^0 is in the interior of the compact convex set $\mathcal{H}$.

2. With probability 1, the covariate process $\mathbf{z}(t)$ is left continuous with bounded total variation on $[0, \tau]$. Also, $\mathbf{z}(t)$ is linearly independent in the sense that if there exist $a(t), \mathbf{c}$ such that $a(t) + \mathbf{c}'\mathbf{z}(t) = 0$ with probability 1, then $a(t) = 0$ and $\mathbf{c} = \mathbf{0}$.

3. With probability 1, $\mathbb{E}[Y(\tau)|\mathbf{z}] > 0$ and $P(\Delta = 0, T_1 = \tau|\mathbf{z}) > 0$. In other words, the risk set will not shrink to zero at time τ.

4. The Hessian matrix $\boldsymbol{\mathcal{I}}_n$ evaluated at $\boldsymbol{\beta}^0, H^0$ is positive definite and converges in probability to $\mathcal{I}^0$, a deterministic and invertible operator.

5. *Identifiability condition.* The model is identifiable such that $H = H^0$ uniformly over Ω implies $\Omega = \Omega^0$. This will ensure that for any sequence $\Omega_n \in \mathcal{H}$,

$$\liminf_{n\to\infty} \ell(\Omega_n) \geq \ell(\Omega^0) \Rightarrow \|\Omega_n - \Omega^0\| \xrightarrow{p} 0.$$

6. *Uniform convergence condition.* For any sequence $\Omega \in \mathcal{H}$, we have uniform conver-

gence, i.e.,

$$\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.$$

Theorem 1. *Assuming regularity conditions hold, with probability 1:* $\hat{\boldsymbol{\beta}}$ *converges to* $\boldsymbol{\beta}^0$*; and* $\hat{H}$ *converges to* H^0 *uniformly on the interval* $[0, \tau]$*.*

Proof. We need to prove consistency: $\|\hat{H}(t) - H^0(t)\|_{\mathcal{L}^\infty(\mathcal{Q})} \xrightarrow{p} 0$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0| \xrightarrow{p} 0$. Since $\ell_n(\hat{\Omega}) = \sup_{\Omega \in \mathcal{H}} \ell_n(\Omega) + o_p(1)$, by Theorem 2.12 of Kosorok (2008), we have $\|\hat{\Omega} - \Omega^0\| \xrightarrow{p} 0$. As in Hu and Tsodikov (2014a, Supplementary Materials C.1), we verify conditions 5 and 6 in the following steps.

1. *Convexity and unique maximum of the likelihood function.* Recall that the marginal hazard for the ith subject may be written as $d\Lambda_*(t) = \gamma(\Omega)\, dH(t)$. Note that this is a functional that depends on the processes $H(\cdot), \mathbf{z}(\cdot)$ on the interval $[0, t]$. Let $F(t)$ be the cumulative incidence function for observed diagnosis events, and let $R(t)$ be the survival function for diagnosis subject to censoring. Note that $dF(t) = R(t)\, d\Lambda_*(t)$. Now we can write the "true" log-likelihood as

$$\ell(\Omega, \Omega^0) = \mathbb{E} \int_0^\tau \left[\log d\Lambda_*(t)\, dF^0(t) - R^0(t)\, d\Lambda_*(t) \right], \tag{F1}$$

   where F^0, R^0 denote the corresponding "true" quantities, respectively, and expectation is taken with respect to the distribution of the covariate process $\mathbf{z}(t)$.

   Now consider the negative "true" Kullback-Leibler distance:

$$\begin{aligned}
D &= \ell(\Omega, \Omega^0) - \ell(\Omega^0, \Omega^0) \\
&= \mathbb{E} \int_0^\tau \left[\log d\Lambda_*(t)\, dF^0(t) - R^0(t)\, d\Lambda_*(t) - \log dH^0_*(t)\, dF^0(t) + R^0(t)\, dH^0_*(t) \right] \\
&= \mathbb{E} \int_0^\tau \left\{ \log \frac{d\Lambda_*(t)}{dH^0_*(t)}\, dF^0(t) + R^0(t) \left[d\Lambda^0_*(t) - d\Lambda_*(t) \right] \right\} \\
&= \mathbb{E} \int_0^\tau \left\{ \log \frac{d\Lambda_*(t)}{dH^0_*(t)}\, dF^0(t) + \frac{dF^0(t)}{d\Lambda^0_*(t)} \left[d\Lambda^0_*(t) - d\Lambda_*(t) \right] \right\}
\end{aligned}$$

$$= \mathbb{E} \int_0^\tau \left\{ \log \frac{d\Lambda_*(t)}{dH_*^0(t)} \, dF^0(t) + \left[1 - \frac{d\Lambda_*(t)}{d\Lambda_*^0(t)} \right] \, dF^0(t) \right\}$$
$$= \mathbb{E} \int_0^\tau v \left(\frac{d\Lambda_*(t)}{d\Lambda_*^0(t)} \right) \, dF^0(t),$$

where $v(x) = \log x + 1 - x$ is a convex non-positive function with a unique maximum of 0 at $x = 1$. Therefore, D has a unique maximum when $d\Lambda_*(t) = dH_*^0(t)$ uniformly. Under an identifiable model, this implies that the unique maximum of D occurs at Ω^0.

2. *Identifiability condition.* Since Λ_* is assumed to be a continuous and differentiable functional of H, then so is the likelihood function $\ell(\Omega)$. Step 1 implies that $\Omega^0 = \arg\max_{\Omega \in \mathcal{H}} \ell(\Omega)$ is unique. We assume our model is identifiable in the sense that $\Lambda_* = \Lambda_*^0$ uniformly over Ω implies $\Omega = \Omega^0$ uniformly. Therefore, by Lemma 14.3 of Kosorok (2008), $\liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^0)$, i.e., the identifiability condition is satisfied.

3. *Uniform convergence condition.* Condition 1 implies that Ω is in the class of functions of bounded variation with integrable envelope, which in turn implies that $H(t)$ is bounded. Therefore, $\mathcal{H}$ is a Glivenko-Cantelli class, whose ϵ-entropy with bracketing number is bounded by A/ϵ, where A is some constant. Then by the assumption of continuity of the functionals Λ_* and ℓ, and the integrability of the envelope of Ω, the integrand in $\ell(\Omega)$ is also Glivenko-Cantelli by the preservation theorems. Therefore we may apply the uniform law of large numbers to the empirical process counterparts of D and ℓ, i.e.,

$$D_n = \ell_n(\Omega, \Omega^0) - \ell_n(\Omega^0, \Omega^0)$$

and

$$\ell_n(\Omega, \Omega^0) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ \left[\log \gamma_i \left(H(t); \boldsymbol{\beta} \right) + \log dH(t) \right] \, dN_i(t) - Y_i(t) \gamma_i \left(H(t); \boldsymbol{\beta} \right) \, dH(t) \right\}$$

such that

$$\sup_{\Omega \in \mathcal{H}} |D_n(\Omega) - D(\Omega)| \xrightarrow{p} 0, \quad \sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.$$

□

Consider a linear functional

$$n^{1/2}\left[\mathbf{a}'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)+\int_0^\tau b(t)\,d\left(\hat{H}(t)-H^0(t)\right)\right], \tag{F2}$$

where $\mathbf{a}$ is a real vector and $b(t)$ is a function with bounded total variation. Let $\mathbf{B}$ denote the vector consisting of the values of $b(t)$ evaluated at the observed failure times corresponding to the set $\{dH\}$; let $\boldsymbol{\mathcal{E}}' = (\mathbf{a}', \mathbf{B}')$.

Theorem 2. *Assuming regularity conditions hold,* $n^{1/2}\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)', \hat{H}(t)-H^0(t)\right]'$ *converges weakly to a zero-mean Gaussian process. In addition,* $n\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{I}}_n^{-1}\boldsymbol{\mathcal{E}}$ *converges in probability to the asymptotic covariance function of the linear functional* (F2)*, where* $\boldsymbol{\mathcal{I}}_n$ *is the negative Hessian matrix of the observed log-likelihood function* (2.6) *with respect to* Ω*.*

Proof. Our proof closely follows that of Hu and Tsodikov (2014a, Supplementary Materials C.2). Let $\mathcal{U}(\Omega) = \left(\mathcal{U}'_{\boldsymbol{\beta}}, \mathcal{U}_{H(s)}\right)'$ be the score, and the proposed NPMLE $\hat{\Omega}$ be the solution to the equation $\mathcal{U}(\Omega) = \mathbf{0}$. Note that in our case this solution involves the profile likelihood for $\boldsymbol{\beta}$:

$$\ell_{\mathrm{pr}}(\boldsymbol{\beta}) = \sup_H \ell(H(t); \boldsymbol{\beta}),$$

where ℓ is defined in equation (2.6) and the estimate of H is obtained using the EM algorithm we have derived. Asymptotically, this is equivalent to simply solving the marginal score, which is what we work with here.

Now let Ω^0 be the set of true parameters. Based on the martingale representation of $\mathcal{U}(\Omega^0)$ and the fact that $N_i(t), i = 1, \ldots, n$ are orthogonal, it follows by the martingale central limit theorem that $n^{-1/2}\mathcal{U}(\Omega^0)$ converges weakly to $U(t) = \left(\mathbf{U}'_{\boldsymbol{\beta}}, U_{H(t)}\right)'$, where $\mathbf{U}_{\boldsymbol{\beta}}$ is a mean-zero p-variate normal random variable and $U_{H(t)}$ is a mean-zero Gaussian process. The variance-covariance function of $U(t)$ is characterized by $\sigma^2_H(s,t;\boldsymbol{\beta}^0, H^0)$, $\sigma^2_{\boldsymbol{\beta}}(\boldsymbol{\beta}^0)$, and $\sigma^2_{H,\boldsymbol{\beta}}(t;\boldsymbol{\beta}^0, H^0)$ as derived below.

The predictable variation process for the score process $\mathcal{U}_{H(s)}$ in (2.10) (scaled by $n^{-1/2}$) is

$$\begin{aligned}\operatorname{Var}\left(n^{-1/2}\,\mathcal{U}_{H(s)}\middle|\,\mathcal{F}_{t^-}\right) &= \frac{1}{n}\operatorname{Var}\left[\sum_{i=1}^{n}\int_0^\tau \varepsilon_i(t,s;\boldsymbol{\beta},H)\,dM_i(t)\middle|\,\mathcal{F}_{t^-}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varepsilon_i^2(t,s;\boldsymbol{\beta},H)\operatorname{Var}\left[dM_i(t)|\mathcal{F}_{t^-}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varepsilon_i^2(t,s;\boldsymbol{\beta},H)Y_i(t)\gamma_i\left(H(t);\boldsymbol{\beta}\right)\,dH(t),\end{aligned}$$

which converges weakly as $n\to\infty$ to a mean-zero Gaussian process with covariance function

$$\sigma_H^2(s,t;\boldsymbol{\beta}^0,H^0) = \int_0^\tau \varepsilon(u,s;\boldsymbol{\beta},H)\varepsilon(u,t;\boldsymbol{\beta},H)P(X>u)\gamma\left(H(u);\boldsymbol{\beta}\right)\,dH(u)$$

for $s,t\in[0,\tau]$. Similarly, $n^{-1/2}\mathcal{U}_{\boldsymbol{\beta}}$ is a martingale and converges to a mean-zero Gaussian process with covariance function

$$\sigma_{\boldsymbol{\beta}}^2(\boldsymbol{\beta}^0) = \int_0^\tau \frac{\dot{\gamma}_{\boldsymbol{\beta}}^2\left(H(u);\boldsymbol{\beta}\right)}{\gamma\left(H(u);\boldsymbol{\beta}\right)}P(X>u)\,dH(u),$$

and $n^{-1/2}\mathcal{U}_{H(s),\boldsymbol{\beta}}$ is a martingale that converges to a mean-zero Gaussian process with covariance function

$$\sigma_{H,\boldsymbol{\beta}}^2(s;\boldsymbol{\beta}^0,H^0) = \int_0^\tau \varepsilon(u,s;\boldsymbol{\beta},H)\dot{\gamma}_{\boldsymbol{\beta}}(H(u);\boldsymbol{\beta})P(X>u)\,dH(u).$$

Now, let the limit in probability of the likelihood function (2.6), normalized as ℓ/n, be ℓ_0. Define a linear information operator as

$$\mathcal{I}_0(t,s) = \frac{\partial\mathcal{U}^0}{\partial\Omega} = -\left[\begin{array}{cc}\frac{\partial^2\ell_0}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2\ell_0}{\partial\boldsymbol{\beta}\partial dH(s)} \\ \frac{\partial^2\ell_0}{\partial dH(t)\partial\boldsymbol{\beta}'} & \frac{\partial^2\ell_0}{\partial dH(t)\partial dH(s)}\end{array}\right]_{\Omega=\Omega^0},$$

where $\mathcal{U}^0 = \left(\frac{\partial\ell_0}{\partial\boldsymbol{\beta}'},\frac{\partial\ell_0}{\partial dH(t)}\right)'$. The operator $\mathcal{I}_0$ acts on an arbitrary vector function element

$\Omega_s = (\boldsymbol{\beta}', dH(s))'$ as

$$\mathcal{I}_0(t,s)\Omega_s = -\begin{bmatrix} \frac{\partial^2 \ell_0}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\boldsymbol{\beta} + \int_0^\tau \frac{\partial^2 \ell_0}{\partial \boldsymbol{\beta} \partial dH(s)}\, dH(s) \\ \frac{\partial^2 \ell_0}{\partial \boldsymbol{\beta}' \partial dH(t)}\boldsymbol{\beta} + \int_0^\tau \frac{\partial^2 \ell_0}{\partial dH(t) \partial dH(s)}\, dH(s) \end{bmatrix}. \tag{F3}$$

With this notation, expanding the score $\mathcal{U}(\hat{\Omega})$ about the true parameter Ω^0, we have

$$\mathcal{I}_0(t,s) n^{1/2} \left(\hat{\Omega}_s - \Omega_s^0\right) = U(t) + o_p(1). \tag{F4}$$

Assuming that the Fredholm operator expressed by the kernel $\mathcal{I}_0$ of the Fredholm integral equation (F4) of the first kind is square integrable, and that the equation $\mathcal{I}_0\Omega = 0$ has only the trivial solution $\Omega = 0$, then equation (F4) has a unique solution. By Theorem 3.3.1 of van der Vaart and Wellner (1996), there exists an inverse information operator $\mathcal{I}_0^{-1}(s,t)$ such that

$$n^{1/2} \left(\hat{\Omega}_s - \Omega_s^0\right) = \mathcal{I}_0^{-1}(s,t)U(t) + o_p(1).$$

Upon differentiation of the equation $\mathbb{E}\left[\mathcal{U}(\Omega^0)\right] = 0$ with respect to Ω at the truth Ω^0, we obtain the usual equivalence between $\mathcal{I}_0$ represented by second derivatives and

$$\mathcal{I}_0(t,s) = \begin{bmatrix} \frac{\partial \ell_0}{\partial \boldsymbol{\beta}} \frac{\partial \ell_0}{\partial \boldsymbol{\beta}'} & \frac{\partial \ell_0}{\partial \boldsymbol{\beta}} \frac{\partial \ell_0}{\partial dH(s)} \\ \frac{\partial \ell_0}{\partial \boldsymbol{\beta}'} \frac{\partial \ell_0}{\partial dH(t)} & \frac{\partial \ell_0}{\partial dH(t)} \frac{\partial \ell_0}{\partial dH(s)} \end{bmatrix}_{\Omega=\Omega^0},$$

which represents the variance of the normalized score Gaussian process $U(t)$. Also, by the functional delta method (Kosorok, 2008, Section 2.2.4), for a differentiable functional $F(\Omega)$, $n^{1/2}\left[F(\hat{\Omega}) - F(\Omega^0)\right]$ converges weakly to a mean-zero Gaussian process with variance-covariance function $\dot{F}(\Omega^0)'\mathcal{I}_0^{-1}\dot{F}(\Omega^0)$, where $\dot{F}(\Omega) = \frac{\partial F}{\partial \Omega}$ and the operator products are defined similarly to (F3). Applying this to (F2) and replacing operator products by matrix algebra, and $\mathcal{I}_0$ by its consistent (matrix) estimator $n^{-1}\hat{\boldsymbol{\mathcal{I}}}_n$, we obtain the desired result. □

Chapter 3: Partial Likelihood Estimation for Continuous Outcomes with Excess Zeros in a Random-threshold Damage-resistance Model

3.1 Introduction

3.1.1 Modeling data with excess zeros in the outcome

The analysis of data for which the outcome exhibits a large number of zero values presents problems for conventional statistical methods. For continuous outcomes, logic suggests that no ties should occur, certainly not when they constitute a substantial proportion of the data at a boundary of the outcome space. Even when the outcome is discrete (e.g., count data), a Poisson model often is insufficient to account for the observed number of zeros.

One approach to this kind of data involves a two-part mixture model, in which one part of the model deals with the probability of the outcome taking the value zero, while the other is a conditional, generally parametrically specified, model for the strictly positive outcome

values. Much of the zero-inflated models literature is focused on count data (e.g., Lambert, 1992). However, in one of the earliest articles on the topic, Aitchison (1955) defines a model for the outcome X where

$$\begin{aligned} P(X = 0) &= \theta \\ P(X > 0) &= 1 - \theta \\ P(x < X < x + dx | X > 0) &= g(x)\, dx, \end{aligned}$$

which leads to the cdf

$$F(x) = \theta + (1 - \theta) \int_0^x g(u)\, du.$$

Aitchison (1955) examines several examples, for both continuous and discrete models for the positive part of the outcome distribution. This type of mixture model has seen a great deal of use in environmental and bioassay applications: Moulton and Halsey (1995) and Taylor et al. (2001) both use mixture models with a lognormal model for the positive values of the outcome.

Work with this kind of model has not been limited to fully parametric specifications, however. Polansky (2005) provides a nonparametric method for estimation of the distribution function associated with a "nonstandard mixture" model (meaning a model with probability mass at known discrete points) using a combination of an empirical distribution function and a kernel estimate of a distribution function, but does not address regression modeling. Zhou and Liang (2006) present a method for the analysis of skewed data with excess zeros based on a two-part model, with the probability of a zero outcome being observed following a logistic model and the continuous positive outcome's conditional mean being modeled using a nonparametrically estimated smooth link function.

Alternatively, some authors have devoted attention to specific parametric models (i.e., not based on a mixture). Siegel (1985) uses what amounts to a profile likelihood method to obtain maximum likelihood estimates for the parameters of a noncentral chi-squared

distribution with zero degrees of freedom (a distribution which contains a point mass at zero). Foster and Bravington (2013) propose a model based on an extension of the Tweedie generalized linear model.

3.1.2 Left censoring and the retro-hazard function

There is a degree of overlap between data with excess zeros and left-censored data: Moulton and Halsey (1995) and Taylor et al. (2001), for example, extend the mixture model methods described above to deal with left censoring due to detection limits. The overwhelming mass of the literature on censored data, however, is in the context of survival analysis. Most of this work is also very general, for the most part dealing with doubly- and/or interval-censored data (Turnbull, 1974; Cai and Cheng, 2004; Goetghebeur and Ryan, 2000; Finkelstein, 1986), and so cannot take advantage of the symmetry between purely left-censored data and purely right-censored survival data.

Specifically, consider a random variable T taking values on the interval $(0, \infty)$. Traditionally, in survival analysis, the baseline cumulative hazard, $H(t)$, is defined as $H(t) = -\log S(t)$, where $S(t) = P(T > t)$ is the survival function (Kalbfleisch and Prentice, 2002). This works well for right-censored data, but this is an inconvenient way to formulate the model for left-censored data.

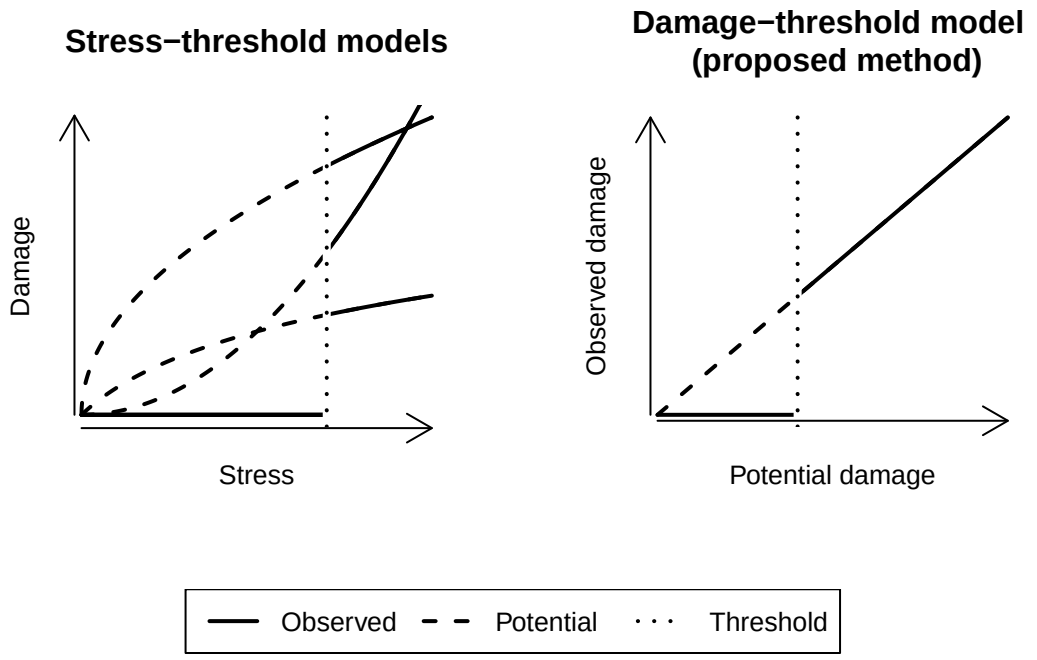
Instead, we define

$$H^*(t) \equiv -\log F(t), \tag{3.1}$$

where $F(t) = P(T \leq t)$ is the cdf. Lagakos et al. (1988) introduced a similar function for the analysis of right-truncated survival data, which they refer to as a "reverse-time hazard function." Gross and Huber-Carol (1992) further develop the ideas of the "retro-hazard," but are also primarily interested in dealing with right-truncated data. Throughout this chapter, following Gross and Huber-Carol (1992), we refer to the function H^* as the retro-hazard.

We also may write $F(t) = e^{-H^*(t)}$, implying that the pdf of T under this formulation

Figure 3.1. This figure depicts schematically the relationship between applied stress and observed damage (left panel) and potential damage and observed damage (right panel) in two alternative models. The left panel represents a model for which the threshold is on the scale of some variable associated with the applied stress. The right panel shows the model corresponding to the proposed method, for which observed damage is equal to zero up to the threshold, from which point observed damage equals potential damage; in this case, the threshold is measured on the scale of damage itself.



is $f(t) = -\frac{dH^*(t)}{dt}\, e^{-H^*(t)}$. This corresponds to the Lehmann (1953) alternative, proposed in the context of nonparametric testing of the equality of distribution functions: the cdf in a regression model based on H^* will be the baseline cdf $e^{-H^*(t)}$ raised to the power $e^{\mathbf{z}_i'\boldsymbol{\beta}}$. This directly parallels the situation with the Cox proportional hazards model (Cox, 1972), with the cdf replacing the survival function.

3.1.3 Damage manifestation and resistance processes

Our goal in this chapter is to semiparametrically model data where the outcome represents some measure of damage to a biological system, in which two competing processes are at work. On the one hand, we have the damage manifestation process, which leads to expression

of the damage in some observable form; on the other, we have the damage resistance process, which, up to a random, subject-specific threshold, may prevent the expression of the damage entirely, leading to an observed outcome of zero (see the right panel of Figure 3.1). This kind of data occurs in the context of experimental setups in which test animals are subjected to external stress and a measure of the damage caused by such pressures is obtained as the outcome (e.g., Miller, 2012).

We propose a two-part model based on the function H^*: if D_i is the random variable representing the damage expression and R_i the damage resistance capacity, then our observed data is

$$X_i = D_i \, \mathbb{1}(D_i > R_i), \tag{3.2}$$

i.e., we observe the damage D_i if and only if it exceeds the resistance capacity of the organism R_i; otherwise we observe 0 for the outcome. To be clear, R_i is never observed: the only information we have on the damage resistance capacity is whether or not R_i is exceeded by D_i. The probability model for D_i and R_i is

$$P(R_i \leq r) = e^{-\mu_i H^*(r)} \tag{3.3}$$

$$P(D_i \leq d) = e^{-\eta_i H^*(d)}. \tag{3.4}$$

We refer to this as the competitive damage-resistance (CDR) model; dependence between the observed damage X and the resistance capacity R is induced by equation (3.2).

The biological motivation for this model derives from the concept in cancer etiology of growth-promoting and growth-inhibitory signals (Weinberg, 1991). On the one hand, proto-oncogenes encourage cell proliferation, while on the other, tumor suppressor genes actively inhibit such proliferation. The failure of these tumor suppressor genes can lead to uncontrolled growth and ultimately to the development of a cancerous tumor, but in the normal course of cell functioning, these genes prevent any cancer from manifesting. In the context of our model, we may view the unobserved R_i as representative of the action of

growth-inhibitory signals; D_i, by contrast, corresponds to the action of growth-promoting signals. The event $D_i > R_i$ would then correspond to the point at which the tumor suppressor genes have failed and allowed a tumor to develop due to runaway cell proliferation.

This type of model is reminiscent of the competing risks approach in the survival literature (Prentice et al., 1978), but also bears similarities to cure models (e.g., Farewell, 1982, which is also a mixture model approach). Mechanistically, cumulative damage/shock models (Ebrahimi, 1999; Esary and Marshall, 1973) are similar; however, these authors are interested in modeling the time to failure of some system rather than a direct measure of the damage process itself.

Although not explicitly a dose-response model, our approach is similar to that of, e.g., Cox (1987) or Crump (1979). These authors, however, typically are focused on estimation of the threshold, in contrast to our situation, where the threshold is random and dependent on the subject; we are interested in estimation of the effect of covariates on the probability of exceeding this threshold. One reference in which the threshold is random is Brockhoff and Muller (1997), in which the authors make use of quasi-likelihood estimation in the analysis of repeated measures data.

The remainder of this chapter is structured as follows: in Section 3.2, we lay out the details of our model for the competing damage and resistance processes; in Section 3.3, we propose an estimator for the parametric part of the model based on a partial likelihood defined using the function H^*; Section 3.4 presents simulation results; and Section 3.5 describes the results of the application of the proposed method to a study of pulmonary capillary hemorrhage in rats exposed to diagnostic ultrasound.

3.2 The competitive damage/resistance model

3.2.1 Specification of the model

We want the marginal distribution of $X = D\,\mathbb{1}(D > R)$. Consider the transformation $(D, R) \mapsto (X, R)$. This will only be one-to-one when $D > R$; for this set, the determinant of the Jacobian of the inverse transformation is -1, implying a joint pdf of

$$\eta\mu e^{-\mu H^*(r)-\eta H^*(x)}\, dH^*(r)\, dH^*(x)\, \mathbb{1}(x > r).$$

The marginal pdf of X will then be

$$\begin{aligned} f(x) &= \int \eta\mu e^{-\mu H^*(r)-\eta H^*(x)}\, dH^*(r)\, dH^*(x)\, \mathbb{1}(x > r)\, dr \\ &= \int_{\infty}^{H^*(x)} \eta\mu e^{-\mu H^*(r)-\eta H^*(x)}\, dH^*(r)\, dH^*(x) \\ &= \eta dH^*(x) e^{-\eta H^*(x)} \int_{H^*(x)}^{\infty} -\mu e^{-\mu H^*(r)}\, dH^*(r) \\ &= \eta dH^*(x) e^{-\eta H^*(x)} \left[e^{-\infty} - e^{-\mu H^*(x)}\right] \\ &= -\eta dH^*(x) e^{-(\eta+\mu)H^*(x)}. \end{aligned}$$

Note that $\{x > r\} \Leftrightarrow \{H^*(x) < H^*(r)\}$, which we have used in the above derivation; we present this in full in order to illustrate integration with the function H^*. The marginal cdf of X is

$$P(X \leq x) = \frac{\eta e^{-(\eta+\mu)H^*(x)} + \mu}{\eta + \mu}; \tag{3.5}$$

see Appendix H for details of an alternative derivation of this marginal model. Note that for $x = 0$, the marginal cdf is equal to $\mu/(\eta + \mu)$. This corresponds to a point mass at 0 in the marginal distribution of damage. The intuition behind this is in the relative magnitudes of η and μ: the larger μ is relative to η, the greater the probability that no damage will be observed because of an increased resistance to damage.

When H^* is not specified parametrically, as in our work, this model is similar to that of Zhou and Liang (2006). Their model allows for easier interpretation of model parameters (since it is a conditional mean model) but at the cost of a more involved estimation procedure, including a bandwidth selection problem.

3.2.2 Rationale for use of the retro-hazard function

We begin this subsection with an outline of some of the properties of the retro-hazard function. From the definition of the function H^* in (3.1), it is apparent that $dH^*(t) \leq 0, t \in (0, \infty)$, so H^* must be nonincreasing. Furthermore, we may deduce that (for a proper distribution of T) since $F(0) = 0$ and $\lim_{t\to\infty} F(t) = 1$, $H^*(0) = \infty$ and $\lim_{t\to\infty} H^*(t) = 0$. The foregoing also implies that

$$H^*(t) = \int_t^\infty -dH^*(x). \tag{3.6}$$

Apart from a sign change, dH^* is equivalent to the function ρ introduced by Lagakos et al. (1988). By analogy with the limit definition of the hazard rate for survival data, we may also write

$$-dH^*(t) = \lim_{dt\to 0^+} \frac{P(t - dt < T \leq t | \, T \leq t)}{dt}.$$

The cumulative hazard function may be recovered using the equation $H(t) = -\log\left[1 - e^{-H^*(t)}\right]$.

Some discussion of the interpretation of the function H^* may be in order here. In contrast to the hazard function, for which larger values of H are associated with smaller values of the outcome variable, larger values of H^* are associated with larger values of the outcome, making this a more attractive function to work with when the outcome of interest is not time-to-event. One of the problems presented by analyzing such data with conventional survival methods and our competitive damage-resistance model is that under our model, the damage outcome is subject to left censoring. Specifically, when resistance capacity exceeds damage, the complete data will exhibit left censoring of the damage outcome, since we will

know only that it was less than the resistance capacity.

Furthermore, it is difficult to formulate the model itself when using the traditional hazard function, as the hazard would have to be infinite to correspond to an outcome "time" of zero. Using H^* instead allows for a more straightforward approach, as well as contributing to a more easily interpretable set of parameter estimates when the model is used in a regression framework (as described later in this chapter).

3.2.3 Parameterization

The parameters η and μ will incorporate covariates $\mathbf{z}_i$ as follows:

$$\eta_i = e^{\mathbf{z}_i'\boldsymbol{\beta}_\eta}, \quad \mu_i = \frac{\theta_i}{1-\theta_i}\eta_i, \quad \theta_i = \frac{e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}{1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}, \tag{3.7}$$

where $\mathbf{z}_i$ is a $p \times 1$ vector. The parameter vectors $\boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\theta$ are also each $p \times 1$ vectors, but may have elements constrained to be 0 if the corresponding covariate is not wanted in that part of the model.

This parameterization follows by defining

$$\theta_i = \frac{\mu_i}{\eta_i + \mu_i},$$

and then using a logistic link function to model θ_i. This allows for the interpretation of the intercept parameter β_0 as $\log P(D \leq R)/P(D > R)$ for a subject with covariate vector of $\mathbf{0}$.

The derivation of the partial likelihood that follows in Section 3.3 retains the original parameterization using only η and μ. This allows for simpler expressions throughout, but for implementation of the method, we will use the parameterization with η and θ.

3.3 Semiparametric estimation based on partial likelihood

3.3.1 Counting process formulation

In the setting of left-censored data (see Appendix G), recall the counting process notation of survival analysis, where $N(t)$ denotes the counting process that takes value 0 until the event occurs, then jumps to 1 (right continuous); and $Y(t)$, which takes value 1 while the subject is at risk of the event, and 0 otherwise (left continuous by convention; see Kalbfleisch and Prentice, 2002, p. 25).

For our purposes, we will imagine a reversal of the time scale (similar to the approach of Lagakos et al., 1988), and define new processes

$$N^*(t) = 1 - N(t^-) \tag{3.8}$$

$$Y^*(t) = 1 - Y(t^+). \tag{3.9}$$

The process defined by (3.8) will be left continuous, while the process defined by (3.9) will be right continuous (somewhat different from the definitions given by Gross and Huber-Carol, 1992, Sections 4.1–4.2).

3.3.2 Derivation of the partial likelihood

Based on the marginal cdf (3.5), we may now write the marginal likelihood for this data (see Appendix H for details):

$$\begin{aligned} L(\boldsymbol{\beta}; H^*) &= e^{\ell_1(\boldsymbol{\beta}) + \ell_2(\boldsymbol{\beta}; H^*)} \\ &= \prod_{i:X_i=0} \frac{\mu_i}{\eta_i + \mu_i} \prod_{i:X_i>0} -\eta_i e^{-(\eta_i+\mu_i)H^*(X_i)}\, dH^*(X_i), \end{aligned} \tag{3.10}$$

where

$$\ell_1(\boldsymbol{\beta}) = \sum_{i:X_i=0} \log \frac{\mu_i}{\eta_i + \mu_i}, \quad \ell_2(\boldsymbol{\beta}; H^*) = \sum_{i:X_i>0} \log \left[-\eta_i e^{-(\eta_i+\mu_i)H^*(X_i)}\, dH^*(X_i) \right].$$

We now consider the problem of estimating H^*, for which only the observations with $X_i > 0$ (that is, observations for which damage is observed) are relevant. The log-likelihood for these observations may be written as

$$\ell_2(\boldsymbol{\beta}; H^*) = \sum_{i:X_i>0} \left\{ \int_0^\infty \log \left[-\eta_i\, dH^*(t) \right]\, dN_i^*(t) - \int_0^\infty (\eta_i + \mu_i) Y_i^*(t)\, dH^*(t) \right\} \tag{3.11}$$

using the counting processes defined by (3.8) and (3.9). By functional differentiation of (3.10) with respect to H^*, we find that the score function is

$$\begin{aligned}
\mathcal{U}(s) &= \delta_s \log \left\{ \prod_{i:X_i>0} \left[-\eta_i e^{-(\eta_i+\mu_i)H^*(X_i)}\, dH^*(X_i) \right] \right\} \\
&= \delta_s \sum_{i:X_i>0} \left\{ \log \eta_i + \log \left[-dH^*(X_i) \right] - (\eta_i + \mu_i) H^*(X_i) \right\} \\
&= \sum_{i:X_i>0} \left[\frac{dN_i^*(s)}{dH^*(s)} - (\eta_i + \mu_i) \cdot -Y_i^*(s) \right] \\
&= \sum_{i:X_i>0} \frac{dN_i^*(s)}{dH^*(s)} + \sum_{i:X_i>0} (\eta_i + \mu_i) Y_i^*(s).
\end{aligned}$$

Note that we have used the identities (G2) and the fact that $Y_i^*(s) = \mathbb{1}(X_i \leq s)$. Furthermore, since for this model all observations greater than 0 are uncensored, $dN_i^*(s) = \mathbb{1}(X_i = s)$ when $X_i > 0$. Setting $\mathcal{U}(s) = 0$ implies a Breslow estimator of

$$\widehat{dH^*}(s) = -\frac{\sum_{i:X_i>0} dN_i^*(s)}{\sum_{i:X_i>0} (\eta_i + \mu_i) Y_i^*(s)}. \tag{3.12}$$

Substitution of (3.12) into the log-likelihood (3.11) yields

$$\begin{aligned}
\ell_2(\boldsymbol{\beta}; \widehat{H^*}) &= \int_0^\infty \sum_{i:X_i>0} \log\left[\eta_i \frac{\sum_{j:X_j>0} dN_j^*(t)}{\sum_{j:X_j>0}(\eta_j+\mu_j)Y_j^*(t)}\right] dN_i^*(t) \\
&\quad + \int_0^\infty \sum_{i:X_i>0} (\eta_i+\mu_i)Y_i^*(t)\, \frac{\sum_{j:X_j>0} dN_j^*(t)}{\sum_{j:X_j>0}(\eta_j+\mu_j)Y_j^*(t)} \\
&= \int_0^\infty \sum_{i:X_i>0} \log\left[\eta_i \frac{\sum_{j:X_j>0} dN_j^*(t)}{\sum_{j:X_j>0}(\eta_j+\mu_j)Y_j^*(t)}\right] dN_i^*(t) + \int_0^\infty \sum_{j:X_j>0} dN_j^*(t) \\
&= \text{const.} + \sum_{i:X_i>0} \int_0^\infty \left[\log\eta_i - \log \sum_{j:X_j>0} (\eta_j+\mu_j)Y_j^*(t)\right] dN_i^*(t),
\end{aligned}$$

where in the last line we have absorbed into the constant all terms not involving η or μ. Returning to (3.10), we see that

$$\begin{aligned}
L(\boldsymbol{\beta}; \widehat{H^*}) &= e^{\ell_1(\boldsymbol{\beta})+\ell_2(\boldsymbol{\beta};\widehat{H^*})} \\
&\propto \prod_{i:X_i=0} \frac{\mu_i}{\eta_i+\mu_i} \prod_{i:X_i>0} \frac{\eta_i}{\sum_{j:X_j>0}(\eta_j+\mu_j)Y_j^*(X_i)} \\
&= \prod_{i:X_i=0} \frac{\mu_i}{\eta_i+\mu_i} \prod_{i:X_i>0} \frac{\eta_i}{\sum_{j:0<X_j\le X_i}(\eta_j+\mu_j)}. \qquad (3.13)
\end{aligned}$$

This constitutes a "partial likelihood" for $\boldsymbol{\beta}$, by which we mean simply that (3.13) is proportional to the profile likelihood over H^* (see Breslow's contribution to the discussion of Cox, 1972, pp. 216–217). This implies that we may base our inferences about these parameters on

$$\ell_{\text{pr}}(\boldsymbol{\beta}) = \sum_{i:X_i=0} [\log\mu_i - \log(\eta_i+\mu_i)] + \sum_{i:X_i>0} \left[\log\eta_i - \log \sum_{j:0<X_j\le X_i} (\eta_j+\mu_j)\right]. \qquad (3.14)$$

Using the parameterization given by (3.7), we may rewrite (3.14) in the form we use for the actual estimation procedure:

$$\ell_{\text{pr}}(\boldsymbol{\beta}) = \sum_{i:X_i=0} \left[\beta_0 + \mathbf{z}_i'\boldsymbol{\beta}_\theta - \log\left(1 + e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}\right)\right] + \sum_{i:X_i>0} \left[\mathbf{z}_i'\boldsymbol{\beta}_\eta - \log \sum_{j:0<X_j\leq X_i} e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1 + e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]. \tag{3.15}$$

The variance-covariance matrix of the parameter estimates may be estimated consistently by $\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\beta}})$; see Appendix I for the derivation of $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$. A proof of the asymptotic normality of a similar estimator is given by Gross and Huber-Carol (1992); only slight modifications of their proof are necessary for our estimator.

Cook and Farewell (1999) give a similar example of the use of a partial likelihood in the analysis of left-censored data, but it is based on the conventional hazard rather than the retro-hazard. Our method can in fact be viewed as a generalization of theirs for a random left-censoring point that varies by subject and is related in a specific way to the outcome (in our case, by the proportionality of the retro-hazard functions). The authors do not, however, provide much guidance as to interpretation of the model parameters. This is further elaborated on in Farewell (1989), although the author simply changes the sign of the original outcomes in order to make use of Kaplan–Meier methodology for estimation of the cdf; the Lehmann family of alternatives is also mentioned (Farewell, 1989, pp. 288–289).

3.4 Simulation study

This section presents a simulation study to examine the finite-sample properties of the proposed method. We simulated 1000 data sets for each of four sample sizes and three intercept values; the intercept was varied in order to produce different proportions of observed zeros in the response. A baseline retro hazard of $H^*(t) = -\log\left(1 - e^{-t/10}\right)$ was used, corresponding to an exponential model. We simulated two covariates, both of which were included in each

Table 3.1. Simulation results: logistic part of model. This table depicts the bias, empirical standard deviation (ESD), and average standard error (ASE) of the parameter estimates across all simulated data sets for the part of the model pertaining to the probability of positive damage being observed.

		β_0			$\beta_{\theta 1} = 2$			$\beta_{\theta 2} = -1$		
Prop. 0	n	Bias	ESD	ASE	Bias	ESD	ASE	Bias	ESD	ASE
18%	50	-0.465	1.357	0.959	0.603	1.457	0.949	-0.180	1.389	1.122
	100	-0.177	0.620	0.556	0.248	0.598	0.519	-0.136	0.784	0.693
	250	-0.066	0.332	0.325	0.074	0.306	0.293	-0.020	0.412	0.405
	500	-0.027	0.224	0.224	0.029	0.210	0.200	-0.004	0.279	0.281
43%	50	0.015	0.591	0.545	0.329	0.909	0.657	-0.144	0.868	0.796
	100	-0.007	0.380	0.364	0.154	0.428	0.406	-0.070	0.528	0.522
	250	-0.003	0.238	0.223	0.078	0.250	0.242	-0.014	0.333	0.318
	500	-0.006	0.159	0.156	0.022	0.173	0.165	-0.002	0.220	0.222
71%	50	0.288	1.036	0.812	0.390	0.967	0.733	-0.154	1.040	0.904
	100	0.167	0.576	0.512	0.184	0.501	0.448	-0.091	0.646	0.586
	250	0.072	0.310	0.304	0.077	0.268	0.260	-0.033	0.361	0.352
	500	0.027	0.212	0.210	0.036	0.178	0.179	-0.011	0.243	0.244

part of the model: $Z_1 \sim N(0,1)$ and $Z_2 \sim B(1/2)$; the covariate vector is $\mathbf{z} = (Z_1, Z_2)'$. Then for each subject θ and η are as defined by equations (3.7).

For all simulations, $\boldsymbol{\beta}_\theta = (2, -1)'$ and $\boldsymbol{\beta}_\eta = (-1, 2)'$; the intercept β_0 was allowed to take values -2, 0, and 2, corresponding to, respectively, approximately 18%, 43%, and 71% of observations equal to 0.

The fact that the coefficients in the two parts of the model have opposite signs is intentional. This is indicative of a particular mechanistic hypothesis about the data-generating mechanism; specifically, that the covariates have the same direction of effect on both the probability of exceeding the threshold as well as the amount of damage manifested given threshold exceedance. For example, increasing values of Z_1 will lead to decreasing amounts of observed damage given that $D > R$, but also to a decreased probability of observing any damage.

Due to numerical issues resulting from complete separation of data points for the logistic part of the model with small sample sizes, a procedure based on Lesaffre and Albert

(1989) was used to detect data sets for which this was a problem; these were then excluded. Specifically, the function `separation.detection()` from the R package `brglm` was used, with option `nsteps = 10`. This results in a matrix of ratios of standard error estimates for each parameter by maximum iterations of the iteratively reweighted least squares fits for the model: if any of these diverge to infinity, separation has occurred. For our purposes, successive differences in the columns of this matrix were examined and the data set was excluded if the final difference was greater than 2 for any parameter.

This seems to have been primarily a problem for the simulation scenarios where there was either a large or a small proportion of observed zeros; the problem is most pronounced when $\beta_0 = -2$ (so that on average approximately 18% of observations were equal to 0).

Additional numerical problems (in the estimation of the positive or η part of the model) were found to be due to an insufficiently large proportion of nonzero observations, particularly in the case of the coefficient for the binary covariate. This is analogous to the situation in classical Cox regression when one of two groups has no events: the MLE of the associated coefficient is then $\pm\infty$ (Kalbfleisch and Prentice, 2002, p. 103). To address this, we excluded data sets for which the information matrix at convergence had any diagonal elements less than 10^{-3}.

In our simulations, these two problems together resulted in the exclusion of 4.5–12.5% of the data sets for $n = 50$; for $n = 100$, this range was reduced to 0.1–0.6%, while no data sets were excluded for $n \in \{250, 500\}$. Although this is clearly only an issue with smaller sample sizes, it could be dealt with using the general method proposed by Firth (1993) for penalization of the likelihood function. This approach has been used for both separation in logistic regression as well as monotone likelihood in Cox models.

The results of the simulation study for the logistic (θ) part of the model are displayed in Table 3.1, which shows that bias and variance decrease with increasing sample size, as we would expect. Bias of all parameter estimates also seems to be adversely affected by intercept values differing from 0, however. We also see good agreement between the ESD and ASE

Table 3.2. Simulation results: positive part of model. This table depicts the bias, empirical standard deviation (ESD), and average standard error (ASE) of the parameter estimates across all simulated data sets for the part of the model pertaining to the observed damage, given that damage is greater than zero.

		$\beta_{\eta 1} = -1$			$\beta_{\eta 2} = 2$		
Prop. 0	n	Bias	ESD	ASE	Bias	ESD	ASE
18%	50	-0.035	0.258	0.245	0.093	0.507	0.464
	100	-0.020	0.162	0.164	0.049	0.317	0.310
	250	-0.006	0.099	0.099	0.025	0.193	0.190
	500	-0.002	0.071	0.069	0.012	0.132	0.133
43%	50	-0.028	0.342	0.322	0.143	0.630	0.593
	100	-0.016	0.225	0.210	0.073	0.394	0.389
	250	-0.006	0.132	0.127	0.010	0.240	0.235
	500	-0.001	0.089	0.088	0.013	0.166	0.164
71%	50	-0.026	0.697	0.557	0.110	1.010	0.964
	100	-0.013	0.366	0.339	0.133	0.658	0.623
	250	-0.012	0.195	0.193	0.059	0.372	0.360
	500	-0.010	0.136	0.133	0.026	0.251	0.246

for moderate to large samples, although there does seem to be a slight underestimation of the variance of the parameter estimates for smaller samples.

Table 3.2 shows the same summary of results as Table 3.1, but for the positive (η) part of the model. In contrast to Table 3.1, it is clear that bias and variance of the parameter estimates for the η part of the model monotonically increase with increasing proportions of observed zeros, which is precisely what we would expect to occur, since this is effectively decreasing the sample size available for estimation of this part of the model. We also observe good agreement between the ESD and ASE for moderate to large samples, as was the case in Table 3.1, which indicates the adequacy of the asymptotic approximations for the covariance matrix of the parameter estimates.

We also compared our proposed method with standard methods (i.e., for which the model for probability of observing damage is not linked with the model for positive damage itself) of addressing the problem. Specifically, for each data set, we fit a standard logistic model, with the outcome being $\mathbb{1}(X_i = 0)$; this fit corresponds to the θ part of our model. For

Table 3.3. Relative mean-square errors (MSE) for simulated data. This is computed as the ratio of the MSE for the proposed method to the MSE for standard logistic (corresponding to the θ part of the model) and to a separate Cox-type regression for only the outcomes greater than 0 included (corresponding to the η part of the model).

	n			
Parameter	50	100	250	500
$\beta_0 = -2$	1.066	0.970	0.930	0.917
$\beta_{\theta 1}$	0.991	0.933	0.853	0.803
$\beta_{\theta 2}$	0.884	0.877	0.854	0.798
$\beta_{\eta 1}$	0.752	0.474	0.221	0.122
$\beta_{\eta 2}$	1.079	1.013	0.898	0.676
$\beta_0 = 0$	0.920	0.936	0.925	0.895
$\beta_{\theta 1}$	0.911	0.858	0.860	0.815
$\beta_{\theta 2}$	0.870	0.838	0.896	0.838
$\beta_{\eta 1}$	0.390	0.203	0.077	0.036
$\beta_{\eta 2}$	1.055	0.911	0.478	0.295
$\beta_0 = 2$	1.091	0.950	0.895	0.893
$\beta_{\theta 1}$	1.047	0.839	0.863	0.844
$\beta_{\theta 2}$	1.009	0.937	0.892	0.867
$\beta_{\eta 1}$	0.426	0.154	0.050	0.025
$\beta_{\eta 2}$	0.994	0.875	0.438	0.225

the subset of observations greater than zero, we fit a simple retro-hazard Cox model, which corresponds to the η part of our model. Specifically, we obtained these naive estimates as

$$\hat{\boldsymbol{\beta}}_\eta = \underset{\boldsymbol{\beta}_\eta \in \mathbb{R}^2}{\arg\max} \sum_{i:X_i>0} \left(\mathbf{z}_i' \boldsymbol{\beta}_\eta - \log \sum_{j:0<X_j \leq X_i} e^{\mathbf{z}_j' \boldsymbol{\beta}_\eta} \right).$$

This objective function may be derived using arguments similar to those given in Section 3.3.

Note that standard logistic regression with $\mathbb{1}(X_i = 0)$ as the outcome will be consistent for the true $(\beta_0, \boldsymbol{\beta}_\theta')'$, as this is fully observed data. However, since this method ignores the information available from the positive observations (quantified by the information matrix component corresponding to the covariance between $(\beta_0, \boldsymbol{\beta}_\theta')'$ and $\boldsymbol{\beta}_\eta$; see Appendix I), we will expect to see a loss in efficiency relative to our method. In the case of the positive part

of the model, by contrast, we will not expect even consistency, as we are ignoring entirely the observations with $X_i = 0$ with the simple retro-hazard Cox model.

Table 3.3 gives the ratio of the mean-square error of the parameter estimates for our proposed method to these two comparison methods. We see that our intuition, as outlined in the preceding paragraph, is borne out by the results: for $\beta_0, \boldsymbol{\beta}_\theta$, the ratio is almost always less than 1. There is also a clear trend for all parameters and all scenarios of increasing MSE of the standard methods relative to the proposed method with increasing sample size.

Focusing now on the largest sample size considered ($n = 500$), we find that for the intercept β_0, the loss in efficiency from use of the naive method seems to be about 10% regardless of the proportion of observations for which no damage was observed. For $\boldsymbol{\beta}_\theta$, there is a larger loss of efficiency, ranging from approximately 15% when $\beta_0 = 2$ to 20% when $\beta_0 = -2$. (We have used the term efficiency here because the bias component of the MSE for the naive method here will be zero asymptotically.)

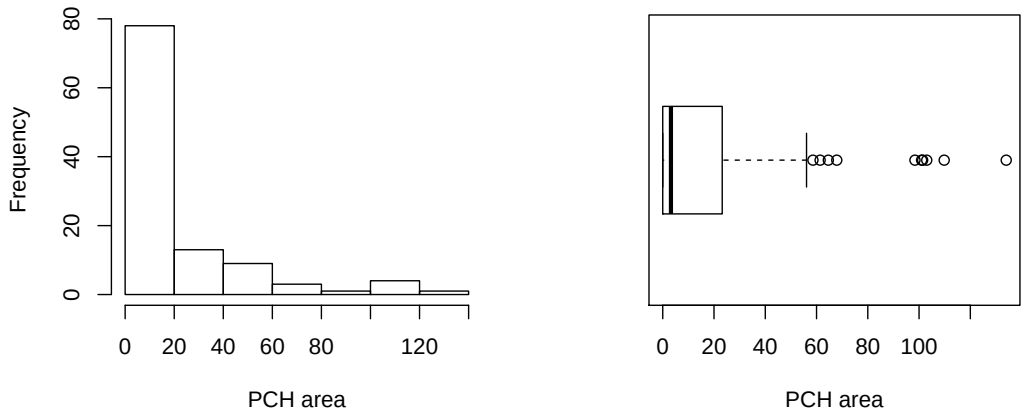
The increase in MSE for $\boldsymbol{\beta}_\eta$ estimated by the naive method is much more pronounced, reflecting the large bias resulting from the exclusion in the naive analysis of the information contained in the observations with $X_i = 0$. The trend here is to increasing MSE of the naive method relative to the proposed model with increasing proportion of observations equal to zero. This is, again, what intuition leads us to expect, as the naive method is losing more information relative to the proposed method when there are fewer positive observations.

3.5 Rat PCH data analysis

3.5.1 Data description and background

To evaluate the CDR model in practice, we applied it to a data set of 109 rats subjected to diagnostic ultrasound. From previous studies, it is known that diagnostic ultrasound (DUS) can induce pulmonary capillary hemorrhage (PCH) in rats (Miller, 2012). This is of clinical relevance for human patients because it demonstrates the potential for pulmonary injury

Figure 3.2. Histogram and boxplot of the outcome for the rat PCH data, pulmonary capillary hemorrhage area in mm^2. Both are intended to give an idea of the "clumping" of the data at 0, as well as the distribution of the positive values.



following ultrasound examinations (for example, examinations to diagnose conditions such as pulmonary edema, effusion, and embolism).

The rats in this study were evaluated at various combinations of ultrasonic frequencies (1.5, 4.5, 7.6, and 12 MHz) and peak rarefactional pressure amplitude (PRPA, referred to hereafter simply as amplitude). There was especial interest in thresholds for PCH expressed in terms of the amplitude, which makes this data particularly suitable for our method, as we explicitly model the probability of exceeding subject-specific damage thresholds as a function of covariates.

The outcome was measured area of PCH for each rat, in mm^2, obtained using photographs from a stereomicroscope with digital camera. The marginal mean of the outcome for all rats (including those with no damage) was 17.63 mm^2; when restricted to those rats with positive damage, the mean area was 26.69 mm^2; 66.1% of rats were observed to have damage (that is, PCH area > 0). It is clear from Figure 3.2, which gives some visualizations of the outcome, that these data are heavily right skewed. Additionally, as 33.9% of rats exhibited no hemorrhagic damage, there is a definite point mass at 0.

Table 3.4. Parameter estimates for the rat PCH data. The final model was chosen on the basis of visual fit to the observed data (see Figure 3.3). The column labeled "Model" denotes the part of the model to which the covariates refer: θ is the logistic model for the probability of not exceeding the resistance threshold, while η is the model for the positive responses (i.e., observed damage > 0).

Model	Covariate	Est.	SE	p-value
θ	(Intercept)	6.064	1.621	0.0002
	Amplitude	-7.696	1.658	0.0000
	Frequency	0.356	0.101	0.0004
η	Amplitude	8.290	1.009	0.0000
	Frequency (1.5 MHz: ref.)	1.000	—	—
	Frequency (4.5 MHz)	2.632	1.271	0.0384
	Frequency (7.6 MHz)	3.143	1.461	0.0314
	Frequency (12 MHz)	2.230	3.209	0.4871
	Amplitude × Frequency (4.5 MHz)	-3.907	0.836	0.0000
	Amplitude × Frequency (7.6 MHz)	-4.420	1.020	0.0000
	Amplitude × Frequency (12 MHz)	-4.257	2.096	0.0423

3.5.2 Results

The results of applying our method to this data are displayed in Table 3.4 and Figure 3.3. Two covariates (along with possible interactions) were considered in this analysis: frequency, which takes only four possible values in this data set; and amplitude. It was found that treating frequency as a categorical rather than a continuous variable in the η part of the model provided a substantial improvement in fit to the data without sacrificing too much in terms of efficiency (as measured by AIC; model comparisons not shown).

In Table 3.4, we see coefficient estimates for amplitude are large in magnitude but opposite in sign in the two parts of the model: this is sensible, recalling that we are modeling the probability of damage not being manifested with the θ part of the model; and that the η part of the model essentially scales the cdf of observed positive damage, so that more positive coefficient estimates indicate increased damage. The interpretation is that larger amplitudes lead both to increased probability of exceeding the resistance threshold as well as to increased damage once the threshold has been exceeded.

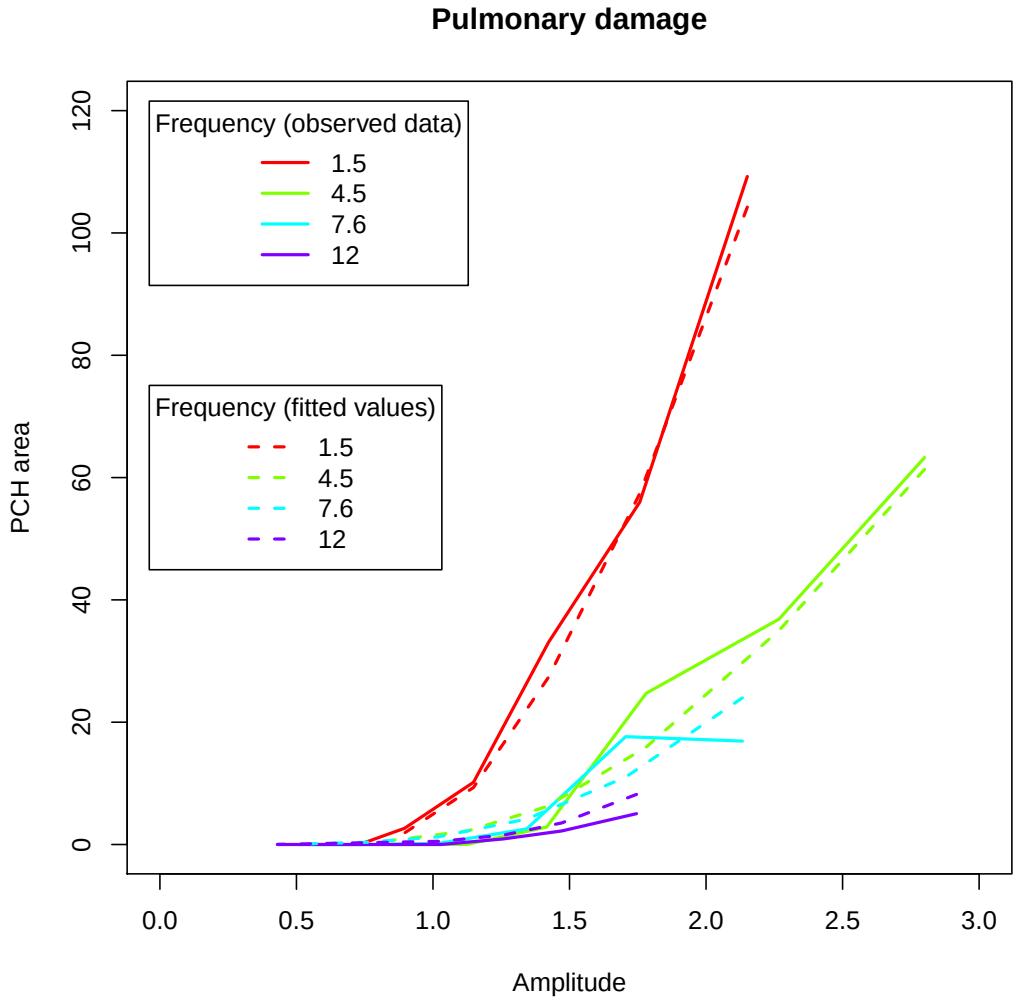
The interpretation of the effect of frequency on PCH area is somewhat more complicated,

both because it is treated as continuous in the logistic (θ) part of the model and categorical in the positive (η) part, as well as because of the inclusion of an interaction term in the positive part. However, we can say that increasing frequency leads to decreasing probability of exceeding the resistance threshold, since the coefficient estimate for this covariate in the θ part of the model is positive. Although the coefficient estimates for the frequency terms alone are all positive in the η part of the model, which would indicate an association of increasing frequency with increasing damage (given exceedance of the threshold), note that the interaction terms all have greater magnitude and negative sign. Therefore, as long as amplitude is greater than zero, the net effect of frequency will be negative, which coincides with what intuition suggests given the positive sign of this coefficient in the logistic part of the model.

It is important to note here that the model is flexible enough to allow for a covariate to be associated with an decreased chance of damage being observed, but also to contribute to an increase in damage when it is observed, and vice versa. This situation would correspond to the coefficient estimates for that covariate having the same sign in both parts of the model.

Turning now to Figure 3.3, we may observe the visual fit of the model to the data. This figure was obtained by using the parameter estimates from Table 3.4 and the Breslow-type estimator of the retro-hazard in equation (3.12). Then (3.5) gives the predicted cdf, which will be a step function; the jump sizes in this estimated cdf will correspond to an estimate of the density. If we denote this estimate as $\hat{f}_i$, then the fitted value (i.e., conditional expected damage) for subject i will be $\sum_{j:X_j>0} X_j \hat{f}_i(X_j)$. It is clear from this figure that the model provides a good fit to the data for each frequency and across amplitudes. There may be slight overestimation in the fitted values for the highest frequency, but overall we see precisely the patterns in the observed data, with smooth curves rising from 0 (no damage observed) at the lowest amplitudes.

Figure 3.3. Observed and fitted values for the rat PCH data. Curves labeled "observed data" are conditional means for the amplitude and frequency values depicted. Curves labeled "fitted values" were obtained by fitting the CDR model using the partial likelihood technique outlined in Section 3.3; the retro-hazard was then obtained using the estimation procedure given in Appendix G; finally, these elements were combined to give an estimate of the conditional density function, which was then used along with the observed damage values to obtain expectations numerically.



3.6 Discussion

In this chapter, we have proposed a model for competitive damage and damage-resistance processes in a biological system, motivated by a data set consisting of test animals subjected to an external stress expected to lead to injury. Our model, using the retro-hazard function first proposed by Lagakos et al. (1988) and later elaborated upon by Gross and Huber-Carol (1992), leads to an estimation procedure based on a partial likelihood. This procedure is fast, efficient, and does not require any distributional assumptions on the observed damage outcome.

There is, however, the issue of interpretation of the results. For the logistic part of the model, this is a simple matter, as this is the probability of observing a zero for the outcome. For the continuous part, on the other hand, interpretation is difficult: Gross and Huber-Carol (1992) do offer some suggestions for intuition in a model using the retro-hazard, but they are still in the setting of survival data, and their explanation involves a discussion of "reverse time." For this reason, we have proposed a simple procedure for obtaining fitted values (see Section 3.5), which are directly interpretable as conditional means given covariates. We may also think of the exponentiated coefficient estimates as "retro-hazard ratios."

Additionally, the assumption of a common baseline retro-hazard for both the damage and resistance systems could be questioned in a particular application. However, the inclusion of covariates in each part of the model, which may of course take the same or opposite signs, seems to allow sufficient flexibility in terms of the effect of a particular factor on the observed outcome. Indeed, it is necessary to make this assumption in order for the model to be identifiable. One justification we might suggest is that the damage and resistance processes are reacting in parallel toward the same externally applied, damaging force, and therefore should share the same retro-hazard.

Future research may examine the possibility of relaxing this assumption via inclusion of shared variables, similar to frailties in survival analysis, between the two parts of the model. Another possible direction for further study is explicit incorporation of a dose-

response relationship in the model. Currently, our approach implicitly assumes that the outcome is the response to some applied dose; however, a dynamic model for variable dose over time could be quite interesting.

Appendices

G Derivation of NPMLE of the retro-hazard

In this section, we derive the nonparametric maximum likelihood estimator (NPMLE) of H^* under the general condition of left-censored data, of which the CDR model's data structure constitutes a special case. Suppose we have $X_i = \max\{T_i, C_i\}, \Delta_i = \mathbb{1}(X_i = T_i), i = 1, \dots n$, where $T_i \sim e^{-H^*(t)}$. The likelihood for this data is

$$L(H^*) = \prod_{i=1}^{n} \left[-dH^*(X_i)\right]^{\Delta_i} e^{-H^*(X_i)}. \tag{G1}$$

Define differentiation of a linear functional J with respect to H^* as (see Hu and Tsodikov, 2014a, Section 3.2)

$$\delta_s J = \frac{\partial J}{\partial dH^*(s)}.$$

Now, differentiation of the log-likelihood proceeds using the chain rule and definition (3.6):

$$\begin{aligned}
\delta_s \log L(H^*) &= \sum_{i=1}^{n} \left\{\Delta_i \delta_s \log\left[-dH^*(X_i)\right] - \delta_s H^*(X_i)\right\} \\
&= \sum_{i=1}^{n} \Delta_i \frac{\partial \log\left[-dH^*(X_i)\right]}{\partial dH^*(s)} - \sum_{i=1}^{n} \frac{\partial H^*(X_i)}{\partial dH^*(s)} \\
&= \sum_{i=1}^{n} \frac{\Delta_i}{-dH^*(s)} \cdot \frac{\partial}{\partial dH^*(s)} \left[-dH^*(X_i)\right] - \sum_{i=1}^{n} \frac{\partial}{\partial dH^*(s)} \int_{X_i}^{\infty} -dH^*(t) \\
&= \sum_{i=1}^{n} \frac{\Delta_i}{-dH^*(s)} \cdot - \mathbb{1}(X_i = s) - \sum_{i=1}^{n} \int_0^{\infty} - \mathbb{1}(X_i \le t) \frac{\partial}{\partial dH^*(s)} dH^*(t) \\
&= \sum_{i=1}^{n} \frac{\Delta_i \, \mathbb{1}(X_i = s)}{dH^*(s)} - \sum_{i=1}^{n} \int_0^{\infty} - \mathbb{1}(X_i \le t) \, \mathbb{1}(t = s) \\
&= \sum_{i=1}^{n} \frac{\Delta_i \, \mathbb{1}(X_i = s)}{dH^*(s)} - \sum_{i=1}^{n} - \mathbb{1}(X_i \le s).
\end{aligned}$$

The important identities established here are

$$\delta_s \log\left[-dH^*(t)\right] = \frac{\mathbb{1}(t=s)}{dH^*(s)}, \quad \delta_s H^*(t) = -\,\mathbb{1}(X_i \le s). \tag{G2}$$

Setting $\delta_s \log L(H^*) = 0$ implies a Nelson–Aalen estimator

$$\widehat{dH^*}(s) = -\frac{\sum_{i=1}^n \Delta_i\, \mathbb{1}(X_i = s)}{\sum_{i=1}^n \mathbb{1}(X_i \le s)}.$$

The negative sign of the estimator indicates that these will be decrements instead of the usual increments in the classical Nelson–Aalen estimator. Otherwise, the form of the estimator is identical, with the only difference being that the "risk set" at point s is composed of observations with $X_i \le s$. Recalling the identity in equation (3.6), the estimate of H^* is

$$\widehat{H^*}(t) = -\int_t^\infty \widehat{dH^*}(s).$$

H Marginal model for observed damage

Alternatively to equations (3.3) and (3.4), the model may be defined by a conditional cdf for X given R (in addition to the marginal definition of the resistance capacity itself):

$$P(R_i \leq r) = e^{-\mu_i H^*(r)} \tag{H1}$$

$$P(X_i \leq x | R_i = r) = e^{-\eta_i H^*(r \vee x)}. \tag{H2}$$

Equation (H2) requires some justification. Specifically, we begin with a conditional hazard–like entity which will only be positive when $D > R$, that is, when damage exceeds repair capacity. We then make use of the identity (3.6) and the definition of H^* (3.1):

$$\begin{aligned}
P(X \leq d | R = r) &= \exp\left\{ -\int_d^\infty -\eta \, dH^*(u) \, \mathbb{1}(u > r) \right\} \\
&= \exp\left\{ \int_{r \vee x}^\infty \eta \, dH^*(u) \right\} \\
&= \exp\left\{ \eta \left[0 - H^*(r \vee x) \right] \right\} \\
&= e^{-\eta H^*(r \vee x)}.
\end{aligned}$$

(See Tsodikov et al., 2013, for details on stochastic process frailty models.) This allows us to obtain the marginal cdf of X as the expectation of the conditional cdf given resistance capacity R:

$$\begin{aligned}
\mathbb{E}\left[e^{-\eta H^*(R \vee x)} \right] &= \int e^{-\eta H^*(r \vee x)} \cdot \left[-\mu \, dH^*(r) e^{-\mu H^*(r)} \right] \\
&= \int_\infty^{H^*(x)} e^{-\eta H^*(x) - \mu H^*(r)} \cdot \left[-\mu \, dH^*(r) \right] + \int_{H^*(x)}^0 e^{-(\eta+\mu) H^*(r)} \cdot \left[-\mu \, dH^*(r) \right] \\
&= e^{-(\eta+\mu) H^*(x)} + \left[\frac{\mu}{\eta + \mu} e^{-(\eta+\mu) H^*(r)} \right]_{H^*(x)}^0 \\
&= e^{-(\eta+\mu) H^*(x)} + \frac{\mu}{\eta + \mu} \left[1 - e^{-(\eta+\mu) H^*(x)} \right] \\
&= \frac{\eta e^{-(\eta+\mu) H^*(x)} + \mu}{\eta + \mu}.
\end{aligned}$$

Considering the complete-data problem, when $D \leq R$ (so that $X = 0$), we have left censoring: the likelihood in this case will be $e^{-\eta H^*(R)}$. We may calculate its expected value as

$$\begin{aligned}
\mathbb{E}\left[e^{-\eta H^*(R)}\right] &= \int e^{-\eta H^*(r)} \cdot \left[-\mu\, dH^*(r) e^{-\mu H^*(r)}\right] \\
&= \int_{\infty}^{0} -\mu e^{-(\eta+\mu)H^*(r)}\, dH^*(r) \\
&= \left[\frac{\mu}{\eta+\mu} e^{-(\eta+\mu)H^*(r)}\right]_{\infty}^{0} \\
&= \frac{\mu}{\eta+\mu}, \qquad \text{(H3)}
\end{aligned}$$

which is equal to the marginal cdf at $x = 0$, i.e., the probability that $D \leq R$. If, on the other hand, we observe damage to the organism, then $X > 0$ and the likelihood is $-\eta\, dH^*(X) e^{-\eta H^*(X)}\, \mathbb{1}(X > R)$. Since we know in this case that $D = X > R$, its expectation is

$$\begin{aligned}
\mathbb{E}\left[-\eta\, dH^*(X) e^{-\eta H^*(X)}\, \mathbb{1}(X > R)\right] &= \int_{\infty}^{H^*(X)} -\eta\, dH^*(X) e^{-\eta H^*(X)} \cdot \left[-\mu\, dH^*(r) e^{-\mu H^*(r)}\right] \\
&= -\eta\, dH^*(X) e^{-\eta H^*(X)} \int_{\infty}^{H^*(X)} -\mu\, dH^*(r) e^{-\mu H^*(r)} \\
&= -\eta e^{-\eta H^*(X)} \left[e^{-\mu H^*(r)}\right]_{\infty}^{H^*(X)}\, dH^*(X) \\
&= -\eta e^{-(\eta+\mu)H^*(X)}\, dH^*(X). \qquad \text{(H4)}
\end{aligned}$$

By differentiation of (3.5), we note that (H4) is also the marginal density.

I Score components and observed information matrix

The partial likelihood is given by equation (3.15).

Score components

In the interests of more compact notation, we hereafter adopt the convention that summations over j refer to the set $\{j : 0 < X_j \leq X_i\}$. The score components are

$$
\begin{aligned}
U_0 &\equiv \frac{\partial \ell_{\mathrm{pr}}(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i:X_i=0} \frac{1}{1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}} - \sum_{i:X_i>0} \frac{\sum_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}}{\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)} \\
\mathbf{U}_\theta &\equiv \frac{\partial \ell_{\mathrm{pr}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_\theta} = \sum_{i:X_i=0} \frac{\mathbf{z}_i}{1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}} - \sum_{i:X_i>0} \frac{\sum_j \mathbf{z}_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}}{\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)} \\
\mathbf{U}_\eta &\equiv \frac{\partial \ell_{\mathrm{pr}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_\eta} = \sum_{i:X_i>0} \left[\mathbf{z}_i - \frac{\sum_j \mathbf{z}_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)}{\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)}\right].
\end{aligned}
$$

The score vector is $\mathbf{U}(\boldsymbol{\beta}) = (U_0, \mathbf{U}_\theta', \mathbf{U}_\eta')'$.

Observed information

The observed information matrix will be

$$
\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \begin{bmatrix} \mathcal{I}_{00} & \boldsymbol{\mathcal{I}}_{\theta 0}' & \boldsymbol{\mathcal{I}}_{\eta 0}' \\ \boldsymbol{\mathcal{I}}_{\theta 0} & \boldsymbol{\mathcal{I}}_{\theta\theta} & \boldsymbol{\mathcal{I}}_{\eta\theta}' \\ \boldsymbol{\mathcal{I}}_{\eta 0} & \boldsymbol{\mathcal{I}}_{\eta\theta} & \boldsymbol{\mathcal{I}}_{\eta\eta} \end{bmatrix},
$$

with component matrices derived below. $\mathcal{I}_{00}$ is a scalar; $\boldsymbol{\mathcal{I}}_{\theta 0}$ and $\boldsymbol{\mathcal{I}}_{\eta 0}$ are $p \times 1$ vectors; and $\boldsymbol{\mathcal{I}}_{\theta\theta}$, $\boldsymbol{\mathcal{I}}_{\eta\theta}$, and $\boldsymbol{\mathcal{I}}_{\eta\eta}$ are $p \times p$ matrices. Clearly, then, $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ will be a $(2p+1) \times (2p+1)$ matrix. Below, we calculate the elements of this matrix.

- Derivatives of the score with respect to β_0:

$$\mathcal{I}_{00} \equiv -\frac{\partial U_0}{\partial \beta_0} = \sum_{i:X_i=0} \frac{e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}{\left(1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}\right)^2} + \sum_{i:X_i>0} \frac{\left[\sum_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2}$$

$$\boldsymbol{\mathcal{I}}_{\theta 0} \equiv -\frac{\partial \mathbf{U}_\theta}{\partial \beta_0} = \sum_{i:X_i=0} \frac{\mathbf{z}_i e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}{\left(1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}\right)^2} + \sum_{i:X_i>0} \frac{\left[\sum_j \mathbf{z}_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2}$$

$$\boldsymbol{\mathcal{I}}_{\eta 0} \equiv -\frac{\partial \mathbf{U}_\eta}{\partial \beta_0} = \sum_{i:X_i>0} \left\{ \frac{\left[\sum_j \mathbf{z}_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right.$$
$$\left. - \frac{\left[\sum_j \mathbf{z}_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]\left[\sum_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right\}$$

- Derivatives of the score with respect to $\boldsymbol{\beta}_\theta$:

$$\boldsymbol{\mathcal{I}}_{\theta\theta} \equiv -\frac{\partial \mathbf{U}_\theta}{\partial \boldsymbol{\beta}_\theta} = \sum_{i:X_i=0} \frac{\mathbf{z}_i\mathbf{z}_i' e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}{\left(1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}\right)^2}$$
$$+ \sum_{i:X_i>0} \left\{ \frac{\left[\sum_j \mathbf{z}_j\mathbf{z}_j' e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right.$$
$$\left. - \frac{\left[\sum_j \mathbf{z}_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]^{\otimes 2}}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right\}$$

$$\boldsymbol{\mathcal{I}}_{\eta\theta} \equiv -\frac{\partial \mathbf{U}_\eta}{\partial \boldsymbol{\beta}_\theta} = \sum_{i:X_i>0} \left\{ \frac{\left[\sum_j \mathbf{z}_j\mathbf{z}_j' e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right.$$
$$\left. - \frac{\left[\sum_j \mathbf{z}_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]\left[\sum_j \mathbf{z}_j e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta+\mathbf{z}_j'\boldsymbol{\beta}_\eta}\right]'}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right\}$$

- Derivatives of the score with respect to $\boldsymbol{\beta}_\eta$:

$$\boldsymbol{\mathcal{I}}_{\eta\eta} \equiv -\frac{\partial \mathbf{U}_\eta}{\partial \boldsymbol{\beta}_\eta} = \sum_{i:X_i>0} \left\{ \frac{\left[\sum_j \mathbf{z}_j \mathbf{z}_j' e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right.$$
$$\left. - \frac{\left[\sum_j \mathbf{z}_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^{\otimes 2}}{\left[\sum_j e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1+e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]^2} \right\}$$

Conclusion

This dissertation has covered a great deal of ground in terms of subfields of statistics: the first chapter dealt with binary classification, the second with missing data in the context of survival analysis, and the third with a partial likelihood for semicontinuous data. However, each of these chapters represents a contribution to statistical methodology useful for cancer research, specifically the relaxation of parametric assumptions in statistical models.

The first chapter, addressing a robust classification method for binary data, was motivated by a melanoma study. The issue of whether or not to perform an invasive procedure, such as the sentinel lymph node biopsies dealt with in that chapter, is certainly not limited to melanoma. In the practice of oncology, the necessity to balance possible harms and benefits is of particular importance because of the high stakes involved: not only are most cancers deadly in their own right, but many treatments carry great risks as well. It is important for clinicians to be able to make the best decisions for their patients based on what researchers in their particular field deem to be the optimal p^*, that is, the best balance of risk and benefit of a treatment or procedure. Chapter 1 provides a method by which they might incorporate this balance into the estimation of logistic regression models, which are arguably among the most familiar and easy-to-understand statistical tools for a majority of clinicians.

There are a number of possibilities in terms of future research in this direction. One of these, as mentioned in Section 1.6, is the numerical stability of the method with small values of h. Another area of interest is in potential ways to directly estimate the LER with ungrouped binary data. A recent test for misspecification in binary response models

is proposed by Esarey and Pierce (2012), which involves a nonparametrically smoothed variation on the Hosmer–Lemeshow test statistic (Collett, 2003, p. 88). This approach is of particular interest to us because it is explicitly a local test of misspecification, in that it allows an investigator to see where precisely in the interval $(0, 1)$ the model predictions $G\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$ deviate from the truth as measured by the empirical mean of the outcome conditional on the predictions. It seems as if this could be used to develop an alternative method for choosing the bandwidth that would more directly capture the essence of the LER.

The progression of cancer in a mechanistic sense is at the core of the second chapter. The model we have developed there is general enough to be applied to any disease that exhibits a cancer-like growth pattern. However, it is especially useful for cancer data, as the baseline hazard shared between the latent and terminal event hazards can be viewed as a surrogate for the tumor-growth process. The original formulation of this problem had onset of cancer and diagnosis as the latent and terminal events, respectively; the assumption of common hazards is particularly delicate in this scenario, because it could be argued that different processes drive tumor initiation and tumor progression. To circumvent this issue, we have defined "onset" as onset of detectable disease, which is to say that the tumor has already begun growing by time T_0.

Specific to our application for this chapter, which was prostate cancer (see Section 2.5), the issue of PSA screening is of great interest for future research. We used a binary covariate to incorporate screening into the model, but this is not the ideal solution: for cancers diagnosed after 1988, our data set does not distinguish between cases that were clinically detected and cases that were detected via screening. This means that subjects diagnosed in the screening era are in fact a mixture of two populations, one with tumors more aggressive (on average) than pre-screening cases and one with tumors less aggressive than pre-screening cases (length bias: see Zelen and Feinleib, 1969).

One simple way to address this issue would be to find auxiliary data containing mode of diagnosis, which could then be included as another covariate in the model. Indeed, another

area of future research would be formally incorporating partial information on time to the latent event into the model. Extensions of the model to deal with the screening problem as a missing data issue would also be interesting and useful.

While motivated by the rat PCH data set studied in Section 3.5, the methodology introduced in Chapter 3 was developed originally in the context of the stochastic process frailty approach adopted in Chapter 2 (see Appendix H for this version of the development of the model). It was found that in this case, however, EM was unnecessary, as there was a closed form for the estimate of the retro-hazard, which led to the partial likelihood of Section 3.3.

The possible applications of this model to data on experiments with laboratory animals in cancer should be readily apparent: tumorigenesis subsequent to application of radiation or other external stressors, for example. Some animals would not exhibit tumors on sacrifice and would therefore result in a clump of zero values for the outcome, while for those that did, the size of the tumor would constitute positive values of the outcome variable.

However, use of this model need not be limited to experimental animal data. The method could be applied, for example, to observational studies on environmental causes of cancer in human subjects. This would be particularly interesting because cancers, especially of a specific type that would be under investigation in a study such as this, are rare in the general population, so a large majority of subjects would have an outcome value of zero. This should not present a major problem with a large enough total sample size, as the simulations in Section 3.4 demonstrate.

One extension of this work would be the incorporation of an explicit dose-response relationship. Typical dose-response studies involve a binary response (see, e.g., Cox, 1987), but our model would allow for the inclusion of magnitude of response, which could lead to much more detailed inferences regarding the effect of dose on the response variable.

References

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* **50,** 901–908.

Brockhoff, P. M. and Muller, H.-G. (1997). Random effect threshold models for dose-response relationships with repeated measurements. *Journal of the Royal Statistical Society, Series B (Methodological)* **59,** 431–446.

Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* **91,** 277–290.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B (Methodological)* **55,** 693–706.

Chen, Y.-H. (2009). Weighted Breslow-type and maximum likelihood estimation in semiparametric transformation models. *Biometrika* **96,** 591–600.

Collett, D. (2003). *Modelling Binary Data*. Chapman & Hall, second edition.

Cook, R. J. and Farewell, V. T. (1999). The utility of mixed-form likelihoods. *Biometrics* **55,** 284–288.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B (Methodological)* **50,** 225–265.

Copas, J. B. (1995). Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 221–235.

Cox, C. (1987). Threshold dose-response models in toxicology. *Biometrics* **43,** 511–523.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)* **34,** 187–220.

Crump, K. S. (1979). Dose response problems in carcinogenesis. *Biometrics* **35,** 157–167.

Dejardin, D., Lesaffre, E., and Verbeke, G. (2010). Joint modeling of progression-free survival and death in advanced cancer clinical trials. *Statistics in Medicine* **29,** 1724–1734.

Dmochowski, J. P., Sajda, P., and Parra, L. C. (2010). Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research* **11,** 3313–3332.

Ebrahimi, N. (1999). Stochastic properties of a cumulative damage threshold crossing model. *Journal of Applied Probability* **36,** 720–732.

Eguchi, S. and Copas, J. B. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society: Series B (Methodological)* **60,** 709–724.

Esarey, J. and Pierce, A. (2012). Assessing fit quality and testing for misspecification in binary-dependent variable models. *Political Analysis* **20,** 480–500.

Esary, J. D. and Marshall, A. W. (1973). Shock models and wear processes. *The Annals of Probability* **1,** 627–649.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38,** 1041–1046.

Farewell, V. T. (1989). Some comments on analysis techniques for censored water quality data. *Environmental Monitoring and Assessment* **12,** 285–294.

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

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80,** 27–38.

Foster, S. D. and Bravington, M. V. (2013). A Poisson-gamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics* **20,** 533–552.

Frydman, H. and Szarek, M. (2009). Nonparametric estimation in a Markov "illness-death" process from interval censored observations with missing intermediate transition status. *Biometrics* **65,** 143–151.

Gjessing, H. K., Aalen, O. O., and Hjort, N. L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability* **35,** 532–550.

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

Gross, S. T. and Huber-Carol, C. (1992). Regression models for truncated survival data. *Scandinavian Journal of Statistics* **19,** 193–213.

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

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60,** 505–531.

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* **67,** 1001–1028.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73,** 387–396.

Hu, C. and Tsodikov, A. (2014a). Joint modeling approach for semicompeting risks data with missing nonterminal event status. *Lifetime Data Analysis* (in press).

Hu, C. and Tsodikov, A. (2014b). Semiparametric regression analysis for time-to-event marked endpoints in cancer studies. *Biostatistics* **15,** 513–525.

Irizarry, R. A. (2001). Information and posterior probability criteria for model selection in local likelihood. *Journal of the American Statistical Association* **96,** 303–315.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, second edition.

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics* **21,** 387–407.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lagakos, S. W., Barraj, L. M., and Gruttola, V. D. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75,** 515–523.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14.

Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics* **24,** 23–43.

Lehmann, E. L. (2004). *Elements of Large Sample Theory*. Springer, corrected edition.

Lesaffre, E. and Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B (Methodological)* **51,** 109–116.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* **86,** 59–70.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, second edition.

Mease, D., Wyner, A. J., and Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research* **8,** 409–439.

Miller, D. L. (2012). Induction of pulmonary hemorrhage in rats during diagnostic ultrasound. *Ultrasound in Medicine and Biology* **38,** 1476–1482.

Mocellin, S., Thompson, J. F., Pasquali, S., Montesco, M. C., Pilati, P., Nitti, D., Saw, R. P., Scolyer, R. A., Stretch, J. R., and Rossi, C. R. (2009). Sentinel node status prediction by four statistical models: results from a large bi-institutional series ($n = 1132$). *Annals of Surgery* **250,** 964–969.

Moulton, L. H. and Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **51,** 1570–1578.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–465.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

Polansky, A. M. (2005). Nonparametric estimation of distribution functions of nonstandard mixtures. *Communications in Statistics—Theory and Methods* **34,** 1711–1724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38,** 485–498.

Prentice, R. L., Kalbfleisch, J. D., Peterson, Jr., A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88,** 1273–1283.

Ruckstuhl, A. F. and Welsh, A. H. (2001). Robust fitting of the binomial model. *Annals of Statistics* **29,** 1117–1136.

Sabel, M. S., Rice, J. D., Griffith, K. A., Lowe, L., Wong, S. L., Chang, A. E., Johnson, T. M., and Taylor, J. M. G. (2012). Validation of statistical predictive models meant to select melanoma patients for sentinel lymph node biopsy. *Annals of Surgical Oncology* **19,** 287–293.

Siegel, A. F. (1985). Modelling data containing exact zeroes using zero degrees of freedom. *Journal of the Royal Statistical Society, Series B (Methodological)* **47,** 267–271.

Taylor, D. J., Kupper, L. L., Rappaport, S. M., and Lyles, R. H. (2001). A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics* **57,** 681–688.

Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society, Series B (Methodological)* **65,** 759–774.

Tsodikov, A., Liu, L. X., Murray, S., and Park, Y.-S. (2013). Stochastic process hazard models. *Journal of the Royal Statistical Society, Series B (Methodological)* (submitted).

Tsodikov, A. D., Asselain, B., Fourque, A., Hoang, T., and Yakovlev, A. Y. (1995). Discrete strategies of cancer post-treatment surveillance: Estimation and optimization problems. *Biometrics* **51,** 437–447.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* **69,** 169–173.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16,** 439–454.

Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences* **99,** 16524–16530.

Wang, J., Shen, X., and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95,** 149–167.

Weinberg, R. A. (1991). Tumor suppressor genes. *Science* **254,** 1138–1146.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56,** 601–614.

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

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

Zhou, X.-H. and Liang, H. (2006). Semi-parametric single-index two-part regression models. *Computational Statistics and Data Analysis* **50,** 1378–1390.