AN ABSTRACT OF THE THESIS OF

DeWayne R. Derryberry for the degree of Doctor of Philosophy in Statistics presented on September 22, 1998. Title: Extensions of the Proportional Hazards Loglikelihood for Censored Survival Data.

Abstract approved: Redacted for Privacy

Paul A. Murtaugh

The semi-parametric approach to the analysis of proportional hazards survival data is relatively new, having been initiated in 1972 by Sir David Cox, who restricted its use to hypothesis tests and confidence intervals for fixed effects in a regression setting.

Practitioners have begun to diversify applications of this model, constructing residuals, modeling the baseline hazard, estimating median failure time, and analyzing experiments with random effects and repeated measures. The main purpose of this thesis is to show that working with an incompletely specified loglikelihood is more fruitful than working with Cox's original partial loglikelihood, in these applications.

In Chapter 2, we show that the deviance residuals arising naturally from the partial loglikelihood have difficulties detecting outliers. We demonstrate that a smoothed, non-parametric baseline hazard partially solves this problem. In Chapter 3, we derive new deviance residuals that are useful for identifying the shape of the baseline hazard. When these new residuals are plotted in temporal order, patterns in the residuals mirror patterns in the baseline hazard. In Chapter 4, we demonstrate how to analyze survival data having a split-plot design structure. Using a BLUP estimation algorithm, we produce hypothesis tests for fixed effects, and estimation procedures for the fixed effects and random effects.

Extensions of the Proportional Hazards Loglikelihood for Censored Survival Data

by

DeWayne R. Derryberry

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 22, 1998
Commencement June 1999

Doctor of Philosophy thesis of DeWayne R. Derryberry presented on September 22, 1998

APPROVED:

Redacted for Privacy

Major professor, representing Statistics

Redacted for Privacy

Chair of the Department of Statistics

Redacted for Privacy

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for Privacy

DeWayne R. Derryberry, Author

ACKNOWLEDGMENT

I am grateful I had Paul Murtaugh as my mentor. He gave me a relatively free hand in determining my own thesis topics, which, I suspect, required more effort on his part than if he had been more direct in handing me topics. This also required that he learn about some topics that either were not of intrinsic interest to him, or led to dead ends, or both.

When I first began working with deviance residuals Dr. Peters and Dr. Schafer were very helpful. Dr. Pierce was also very helpful, volunteering time out of his busy schedule to discuss Chapter 2 and his likelihood test-based definition of deviance residuals with me. I should also thank many faculty members who were confronted by me one week early in my research, when a went through the basement halls asking "What is a residual?" and "what is an outlier?".

When I began working with random effects and BLUP estimators Dr. Pereira and Dr. Birkes were very helpful in pointing out the Bayesian flavor of random effects models.

Finally, a significant portion of my thesis required an understanding of constrained non-linear optimization. Dr. Arthur was always available to discuss whatever problem was bothering me. In fact, looking back, I think I wasted a lot of peoples time for two and a half years, but nobody complained.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

Extensions of the Proportional Hazards Loglikelihood for Censored Survival Data

Chapter 1

1. Introduction

1.1 Survival Data

In the health sciences, subjects are often observed from the time they enter a study until some event of interest occurs, often called the failure time (unfortunately this is often a catastrophic event such as death). Subjects may enter the study after it has begun and may either drop out of the study or fail to manifest the event of interest before the study ends. For this reason, only a portion of the subjects have a reported failure time. The rest of the subjects have a censoring time, which is effectively a lower bound on their failure time. Data of this type are usually called censored survival data.

An important element in survival data is the hazard rate. Intuitively, the hazard rate is the likelihood a subject fails in the next small time increment, given that the subject has not yet failed at time t. For continuous distributions this can be expressed as:

$$\lambda(t) = f(t)/S(t), \text{ where } S(t) = 1 - F(t),$$

where f(t) is the density of the failure times, F(t) is the cumulative density function, and S(t) is the survivor function.

The Cox proportional hazards (PH) model supposes:

$$\lambda(t,z) = \lambda(t)\exp(z\beta),$$

where z is a vector of covariates and β are fixed unknown parameters. All subjects share a common underlying baseline hazard, $\lambda(t)$, and have varying hazard rates that depend multiplicatively only on covariates. This relationship is called proportional

hazards, because any one subject's hazard rate is a fixed multiple of any other subject's and this multiple is not time dependent.

1.2 The Proportional Hazards Loglikelihood

Some notation is commonly used: $\Lambda(t) = \int_0^t \lambda(s)ds$ is the cumulative hazard function; δ_i is an indicator variable: 1 for a failure or 0 for a censored time. We will assume for simplicity that the events are ordered ($i > j \Rightarrow t_i > t_j$) and there are no ties. The loglikelihood for PH survival data with n independent subjects and random censoring becomes

$$-\sum_{i=1}^{n} \Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n} \delta_i \log\{\lambda(t_i)\exp(z_i\beta)\}. \tag{1}$$

When a parametric model is specified $\Lambda(\cdot)$ and $\lambda(\cdot)$ are replaced by specific expressions. If a specific non-parametric estimate, the Breslow cumulative hazard and its associated baseline hazard, are used for $\Lambda(\cdot)$ and $\lambda(\cdot)$; then (1) becomes the partial loglikelihood of Cox (Breslow, 1974):

$$\sum_{i=1}^{n} \delta_i [z_i\beta - \log\{\sum_{j=i}^{n} \exp(z_j\beta)\}], \tag{2}$$

where the events are ordered in time.

When the cumulative hazard is specified as a step function, as with the Breslow cumulative hazard, then

$$\Lambda(t_i) = \sum_{j=1}^{i} \lambda_j \Delta_j, \tag{3}$$

where $\Delta_j = t_j - t_{j-1}$, $t_0 = 0$ and λ_j is a constant failure rate on the interval (t_{j-1}, t_j). Then

$$\begin{aligned} &-\sum_{i=1}^{n} \exp(z_i\beta) \sum_{k=1}^{i} \lambda_k \Delta_k + \sum_{i=1}^{n} \delta_i \log\{\lambda_i \exp(z_i\beta)\} \\ &= -\sum_{i=1}^{n} \lambda_i \Delta_i \sum_{k=i}^{n} \exp(z_k\beta) + \sum_{i=1}^{n} \delta_i \log\{\lambda_i \exp(z_i\beta)\}. \end{aligned} \tag{4}$$

1.3 Deviance Residuals

To compute the deviance residual for observation j, we first need to form a special loglikelihood ratio test: H_o: the current model fits at observation j; H_a: a different model is appropriate for observation j. For any model, the null loglikelihood and alternative loglikelihood differ only in that, in the alternative loglikelihood observation j has a single indicator variable instead of a covariate vector. This has the effect of artificially allowing the observed and predicted (maximum likelihood) values to be exactly equal for observation j.

If we denote this loglikelihood ratio as $\mathcal{L}_j$, then the deviance residual is

$$d_j = \text{sign}(\text{observed}_j - \text{predicted}_j) \cdot \mathcal{L}_j^{\frac{1}{2}}$$

In many cases, $\mathcal{L}_j$ is approximately χ^2_1, so d_j is approximately standard normal.

Deviance residuals measure the discrepancy between the observed and predicted values for an observation, placing the degree of discrepancy on a "nice" scale -- standard normal.

1.4 Common Themes

The next three chapters, which comprise the main work of the dissertation, share two common themes: first use of (1) instead of (2) gives opportunities to think more carefully about the choice of $\Lambda(\cdot)$ normally treated as a nuisance parameter, allows more flexibility in the selection of $\Lambda(\cdot)$, and allows derivations that are simple and intuitively clearer. Secondly, all three chapters use deviance residuals to detect outliers.

In Chapter 2 we use (1) and (3) and specify a cumulative hazard that is different from the Breslow cumulative hazard, but still non-parametric. By clustering neighboring observations to form a smoother baseline hazard, we find a deviance residual that detects outliers more effectively than do the current deviance residuals.

The main result of Chapter 3 involves using relations (3) and (4) to derive a new deviance residual. Using (3) to approximate a parametric cumulative hazard, i.e., $\Lambda(t_i) \approx \sum_{j=1}^{n} \lambda(t_j)\Delta_j$, we derive these deviance residuals for parametric models as well. This new deviance residual assumes the baseline hazard is locally constant over the time between observations. Outliers detected by these residuals indicate time segments where the observed and predicted hazard rates are very different. Using (1) and (3), we also derive a non-parametric estimate of baseline hazard that is monotone increasing.

In Chapter 4 we focus on the split-plot design in survival data. Using a hierarchical Bayesian framework and the PH model, we develop an estimation and testing procedure for the fixed effects. Our algorithm is easily implemented in current software; neither numeric integration nor Newton's method is required of the user. The simplifying ideas depend heavily on using derivatives from the Bayesian posterior joint density formed with (1) instead of (2).

Split-plots are unusual in that there are experimental units at the whole-plot level and the sub-plot level. Our method also estimates random effects, which for normal models are the deviance residuals at the whole-plot level. Using the estimated random effects, it is possible to look for outliers at the whole-plot level. In medical studies, these may be subjects that react in an unusual manner to treatment, whom it would be important to identify.

Chapter 2

Improved Outlier Detection in Survival Models

DeWayne R. Derryberry and Paul A. Murtaugh

2.1 Abstract

Deviance residuals have been recommended for outlier detection in Cox (1972) regression. The purpose of this paper is two-fold: to show that deviance residuals often do not detect outliers, and to present modified deviance residuals that better detect outliers. Simulation results compare the new residuals to those currently in use.

2.2 Introduction

2.2.1 *Survival Data*

In many medical studies, subjects with a serious medical condition (such as cancer) are observed until a catastrophic event (such as death) occurs. The response variable is the time to the event, or failure time. Because subjects may begin the study at any time and leave the study before the event of interest occurs, many responses are right censored. In such studies, the identification of outliers is important as they may represent subjects responding in an unusual manner. Lawless (1982) and Kalbfleisch and Prentice (1980) discuss such data.

Consider n independent subjects, each having a censoring time c_i and a failure time f_i, although only the earlier event is observed. Sample information for subject i consists of a triplet (t_i, δ_i, z_i), where $t_i = \text{MIN}(f_i,c_i)$, $\delta_i = I(f_i < c_i)$, and z_i is a vector of covariates. For simplicity, the paper considers one covariate, an indicator variable distinguishing treatment and control subjects. Without loss of generality, the events are sorted chronologically ($i > j$ implies $t_i > t_j$), and there are no ties.

2.2.2 *Regression Models*

Assume a hazard function of the form $\lambda(t,z) = \lambda(t)\exp(z\beta)$, where $\lambda(t)$ is the baseline hazard. Cox's (1972) partial likelihood uses this relationship, which we will call the PH (proportional hazards) model, with baseline hazard unspecified. An important parametric model, Weibull regression, uses the PH model with $\lambda(t) = \rho\lambda^{\rho}t^{\rho-1}$. For PH models with random censoring, the loglikelihood is

$$-\sum_{i=1}^{n}\Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n}\delta_i\log\{\lambda(t_i)\exp(z_i\beta)\}, \qquad (1)$$

where $\Lambda(t_i) = \int_0^{t_i}\lambda(s)ds$ is the cumulative hazard function.

The baseline hazard can be modeled as a step function, which is equivalent to modeling the cumulative hazard function as a linear spline. The interval (0,T], where T is the last event time, can be partitioned into sub-intervals of the form $(t_{m-1}, t_m]$, each with a constant hazard rate λ_m. Suppose $t_k \in (t_{m-1}, t_m]$, and let $t_0 = 0$ and $\Delta_m = t_m - t_{m-1}$. Then $\Lambda(t_k) = (t_k - t_{m-1})\lambda_m + \sum_{j=1}^{m-1}\lambda_j\Delta_j$.

Modeling the baseline hazard as a step function is not new, having been suggested by Oakes (1972), Kalbfleisch and Prentice (1973), and Breslow (1974) as a method for estimating the cumulative hazard and survivor functions. Kalbfleisch and Prentice chose to partition (0,T] independently of the data, while Oakes and Breslow based the partition on the failure times. We will select a different partition, which will generally have little effect on estimation, but which may substantially alter the estimated hazard function and deviance residuals in a few cases.

2.2.3 *Outlier Detection*

Data analysis, especially model selection, requires the identification and investigation of poorly fitted observations. Deviance residuals are recommended for outlier detection in generalized linear models (Pierce and Schafer, 1986; McCullagh and

Nelder, 1991) and survival analysis (Fleming and Harrington, 1984; Therneau and Grambsch, 1990).

If we treat failure time as the response, the i^{th} deviance residual in a PH model is

$$d_i = -\text{sign}(M_i)[2\{-M_i - \delta_i \log(\delta_i - M_i)\}]^{\frac{1}{2}}, \quad (2)$$

where

$$M_i = \delta_i - \Lambda(t_i)\exp(z_i\hat{\beta}). \quad (3)$$

Even with the baseline hazard restricted to a step function, the d_i's differ for different partitions. Some partitions have special significance. The null partition, one constant hazard rate, is exponential regression. When the interval (0,T] is partitioned at the failure times, $\Lambda(t_i)$ is estimated by the Breslow cumulative hazard function and (1) is equivalent to the partial likelihood of Cox (Breslow,1974). In this latter case, (2) yields the deviance residual discussed in the counting process literature (Fleming and Harrington, 1984), which we will call the FH deviance residual. (The deviance residuals above have a sign opposite that of similar residuals defined in the counting process literature. This is because the sign is determined by failure times, not counts.)

2.3 The Problem

A desired property of semi-parametric deviance residuals is that they mimic parametric deviance residuals when the associated parametric model is correct. FH deviance residuals do not always display this property.

Cox regression implicitly assumes a partition consisting of all failure times (Breslow, 1974) -- a model that may have almost as many nuisance parameters as observations (Kalbfleisch and Prentice, 1980, p. 79). We believe this causes over-fitting of the baseline hazard. When the first event occurs very early, the baseline hazard

becomes locally large; and when the last event is very late, the hazard becomes locally small.

Consider an uncensored first event. Maximization of (1), when the failure times form a partition, yields: $t_1\hat{\lambda}_1 = \{\exp(z_1\hat{\beta}) + ... + \exp(z_n\hat{\beta})\}^{-1}$. The event time t_1 should be identified as an outlier when $t_1 \rightarrow 0$. Because the method of estimating β is rank-based, the right hand side of the equation is unaffected by this limiting process. So $\hat{\lambda}_1 \rightarrow \infty$, i.e., the model "explains" the early failure by assigning a large local estimate of hazard. Similarly $(t_n - t_{n-1})\hat{\lambda}_n = \delta_n\exp(-z_n\hat{\beta})$, so $\hat{\lambda}_n \rightarrow 0$ as $t_n \rightarrow \infty$. The anomalous nature of the event time is absorbed in the baseline hazard, not reflected in the deviance residual. This analysis is only tractable for the first and last events and the paper will focus on these cases. It is unclear to what extent FH deviance residuals for other events are similarly affected.

2.4 A Proposed Solution

If the problem arises due to over-fitting of the baseline hazard, any thoughtful reduction in the number of nuisance parameters should be beneficial. Neighboring events can be clustered to give a pooled local estimate of the baseline hazard. Clustering is a simple non-parametric smoothing technique that, for large data sets, can substantially reduce the number of nuisance parameters.

When the first two events are clustered (and both are uncensored),

$$\hat{\lambda}_1 = 2\{\Delta_1\sum_{i=1}^{n}\exp(z_i\hat{\beta}) + \Delta_2\sum_{i=2}^{n}\exp(z_i\hat{\beta})\}^{-1}.$$

In this case, $t_1 \rightarrow 0$ implies $\hat{\lambda}_1 \rightarrow 2[\, t_2\, \{\exp(z_2\hat{\beta}) + ... + \exp(z_n\hat{\beta})\}]^{-1}$. When the local hazard rate cannot become arbitrarily large, the first event is revealed to be an outlier.

A similar argument applies to the last event. In fact, any scheme that clusters the first event with one or more subsequent events and clusters the last event with one or more preceding events has the desired property that $t_1 \rightarrow 0$ implies $d_1 \rightarrow -\infty$ (when the first event is a failure), and $t_n \rightarrow \infty$ implies $d_n \rightarrow \infty$. Clustering, while only slightly reducing the flexibility of the baseline hazard, reduces the problem of over-fitting.

We assume the true baseline hazard is a continuous function, so cluster sizes are chosen to reduce both bias and sampling error as sample size increases. Intuition suggests smaller clusters reduce bias, while larger clusters reduce sampling error. The chosen cluster size, $n^{\frac{1}{2}}$, is one of many possible clustering schemes consistent with these goals.

For any cluster of adjacent observations, if an estimate of β is available, a common constant baseline hazard λ_c can be estimated using maximum likelihood:

$$\hat{\lambda}_c = \sum_{j \text{ in cluster}} \delta_j \Big/ \sum_{j \text{ in cluster}} \Delta_j \sum_{k=j}^{n} \exp(z_k \hat{\beta}).$$

Deviance residuals constructed with the Cox regression estimate of β and this clustering scheme are MB (modified baseline) deviance residuals, and the linear spline formed by integrating this baseline hazard is the MB cumulative hazard.

2.5 Simulations

Weibull data sets were generated using the relationship $\lambda(t,z) = \rho\lambda^{\rho}t^{\rho-1}\exp(z\beta)$. The censorship rate, baseline hazard rate, and sample size were all varied. The parameters λ and β were chosen so that the control group had a mean of 1 and the treatment group had a mean of 2. Censoring times were exponential.

Each data set had one outlier, either the first or last event. If the outlier was the first event, it was constructed to be an uncensored control, having an exact deviance residual

of -5. When the outlier was the last event, it was constructed to be an uncensored treated case, having an exact deviance residual of 5. These combinations of covariates and responses probably give the FH deviance residuals the most difficulty.

For each data set, the deviance residuals were estimated using (2), but with differing methods of estimating $\Lambda(t)$ and β:

- Parametric - With ρ known and $\Lambda(t) = \lambda^{\rho}t^{\rho}$, λ and β were estimated by maximizing (1).
- FH - $\Lambda(t)$ was estimated with Breslow's cumulative hazard function, and β was estimated by Cox regression.
- MB - β was estimated by Cox regression, but $\Lambda(t)$ was estimated with the MB cumulative hazard.

Each method was evaluated by comparing its residuals to exact deviance residuals. Exact deviance residuals were computed using (2) with the parameter values for $\Lambda(t)$ and β instead of estimates. Ideally, estimated deviance residuals should mimic the exact deviance residuals for both the outlier and the n-1 non-outliers. A more modest hope is that estimated deviance residuals behave as standard normal variates for non-outliers and display a significant large deviation from normality for outliers.

2.6 Results

Figure 2.1 shows histograms for the event times for a typical data set, separated into the treatment and control group. This data set was generated with a Weibull shape parameter (ρ) of 2 and consisted of 100 observations, 7 of which were censored. Deviance residuals for this data set are presented in Figures 2.2-2.4. FH deviance residuals are plotted against exact deviance residuals in Figure 2.2. The constructed outlier, which has an exact deviance residual of 5, has an FH deviance residual of only

1.98 (2.24 standardized), giving little indication that the observation is an outlier. Figure 2.3 shows that the MB deviance residual, with a value of 2.85 (3.12 standardized), gives substantial evidence of an outlier, although the magnitude is much smaller than the true value of 5. Finally, Figure 2.4 shows that the parametric deviance residual, with a value of 4.45 (4.21 standardized), both identifies the observation as an outlier and assigns it a realistic magnitude.

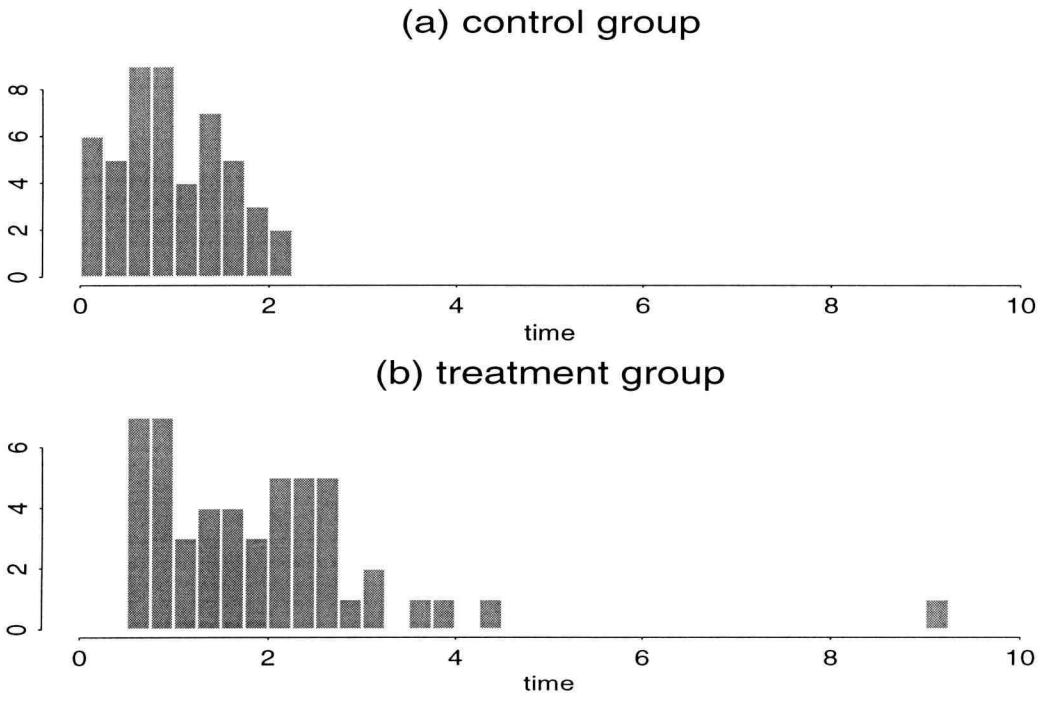

Figure 2.1 Histograms of the event times for the example data set. The extreme observation (failure time = 9.1) in the treatment group is the outlier.

For the 99 non-outliers in the data set, the estimated residuals for all three methods fall close to the 45° line, indicating a close match of the estimated residual to the exact residual (Figures 2.2-2.4). The mean absolute deviations (MAD) between estimated and

exact residual for the non-outliers were 0.108, 0.100, and 0.078 for the FH, MB and parametric deviance residuals, respectively. These results are typical of errors in estimation of the exact deviance residuals for non-outliers.

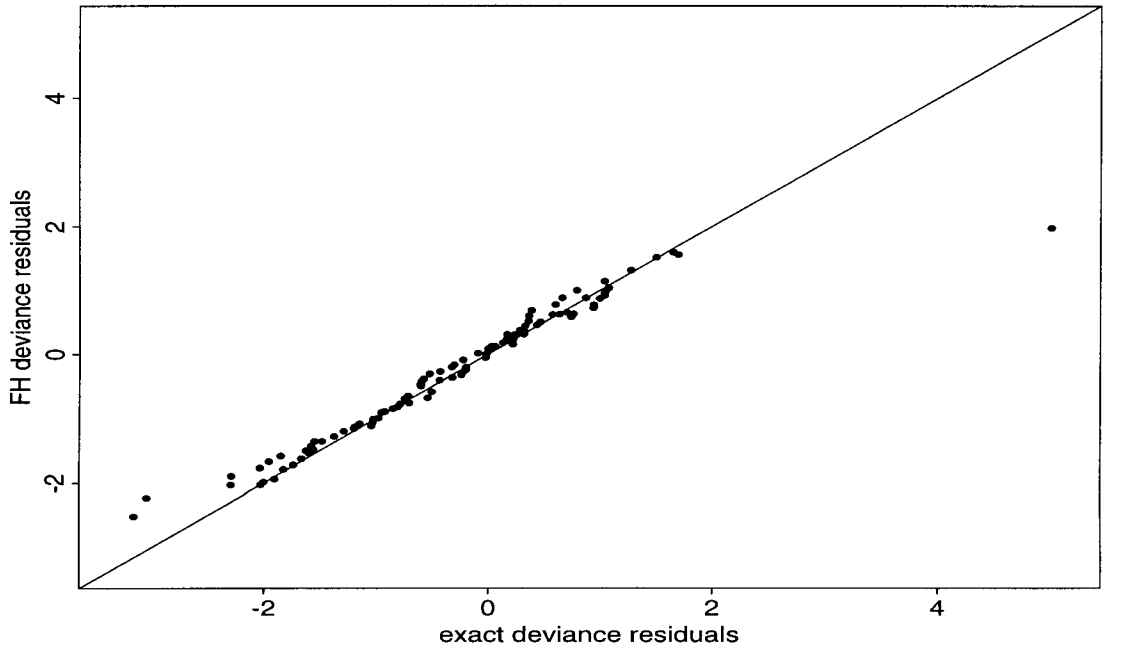

Figure 2.2 FH deviance residuals plotted against exact deviance residuals for the example data set. A 45° line has been added.

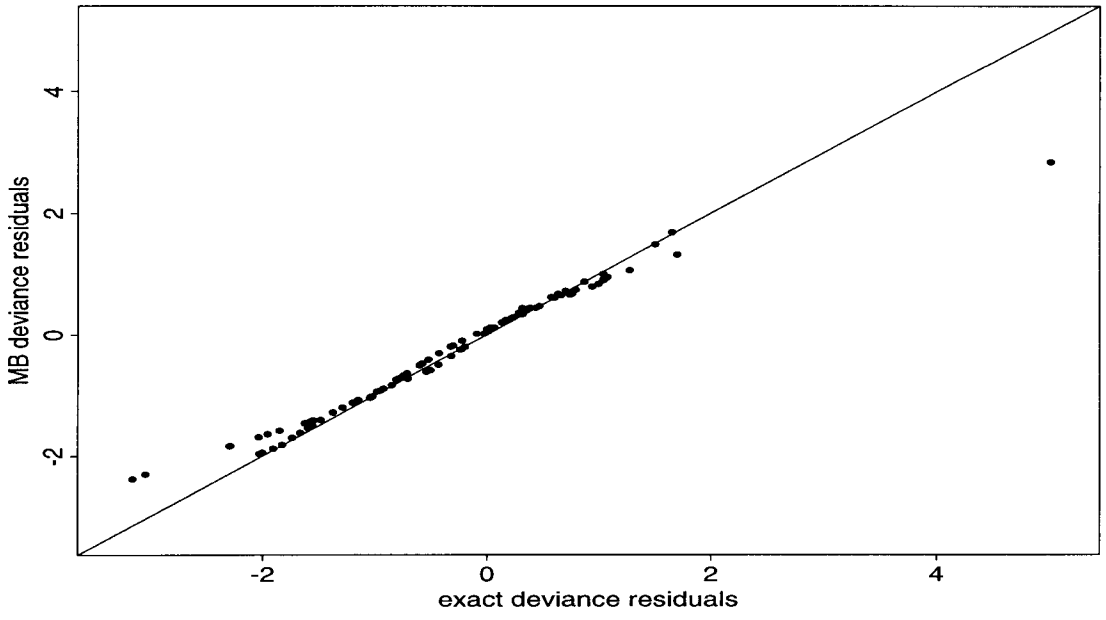

Figure 2.3 MB deviance residuals plotted against exact deviance residuals for the example data set. A 45° line has been added.

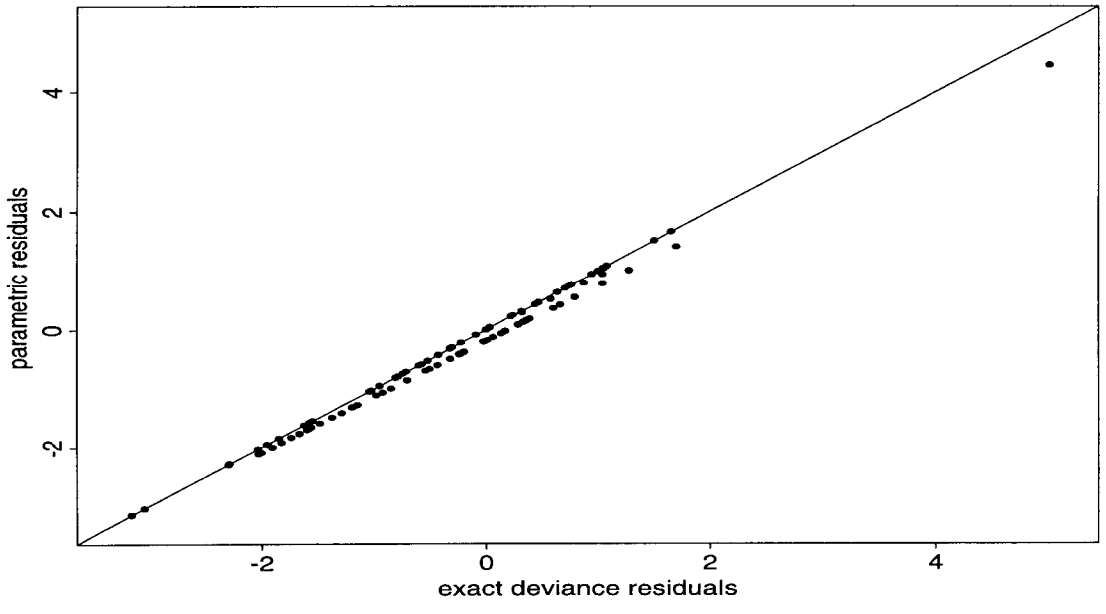

Figure 2.4 Parametric deviance residuals plotted against exact deviance residuals for the example data set. A 45° line has been added.

Table 2.1 Confidence intervals for estimated deviance residuals, exact deviance residual = -5. Sample size, n. Censorship rate, θ. Weibull parameter, ρ. Each 95% confidence interval is based on 1000 data sets.

			Method of computing estimated deviance residual		
ρ	n	θ	FH	MB	parametric
$\frac{1}{2}$	100	50%	-2.63, -2.62	-6.85, -6.84	-4.99, -4.99
		10%	-2.63, -2.63	-6.87, -6.86	-5.00, -4.99
	196	50%	-2.87, -2.87	-6.82, -6.81	-5.00, -5.00
		10%	-2.87, -2.87	-6.83, -6.82	-5.00, -4.99
1	100	50%	-2.58, -2.58	-4.97, -4.96	-5.00, -4.99
		10%	-2.58, -2.58	-4.97, -4.96	-5.00, -4.99
	196	50%	-2.83, -2.83	-4.97, -4.96	-5.00, -5.00
		10%	-2.86, -2.83	-4.98, -4.97	-5.00, -5.00
2	100	50%	-2.51, -2.51	-3.77, -3.76	-4.99, -4.98
		10%	-2.51, -2.51	-3.66, -3.66	-4.99, -4.99
	196	50%	-2.76, -2.76	-3.88, -3.86	-5.00, -4.99
		10%	-2.76, -2.76	-3.73, -3.72	-5.00, -5.00

Table 2.1 consists of confidence intervals for the deviance residual when the first event time is an outlier. The parametric residual accurately estimates the exact deviance residual. The FH deviance residual was relatively insensitive to the presence of an outlier and actually assigned a value of about -2.6 (-2.85) to the first observation, when it was an uncensored control and the sample size was 100 (196), irrespective of the failure time. The MB deviance residual performs well for constant hazards ($\rho = 1$), and gives results that depend little on sample size or censoring rate at all hazard rates.

MB deviance residuals always associate a large deviation with the outlier, but give estimates that are too small for decreasing failure rates ($\rho = \frac{1}{2}$) and too large for the increasing failure rates ($\rho = 2$). This is due to the bias introduced when estimating a monotone function with a step function. A step function will generate a measure of

central tendency for the baseline hazard on each interval. When the actual function being estimated is monotone decreasing, the step function must generally give estimates too low on the left end of each interval and too high on the right end of each interval. For a monotone increasing function the bias will be the opposite. The step function estimate of the baseline hazard for the earliest event will always be too low for decreasing failure rates and too high for increasing failure rates.

Table 2.2 Confidence interval for estimated deviance residual, exact deviance residual = +5. Sample size, n. Censorship rate, θ. Weibull parameter, ρ. Each 95% confidence interval is based on 1000 data sets.

			Method of computing estimated deviance residual		
ρ	n	θ	FH	MB	parametric
$\frac{1}{2}$	100	50%	0.89, 0.91	1.65, 1.73	3.59, 3.64
		10%	1.73, 1.75	3.16, 3.23	4.13, 4.18
	196	50%	0.96, 0.98	1.99, 2.06	4.12, 4.17
		10%	1.89, 1.91	3.73, 3.77	4.52, 4.55
1	100	50%	1.11, 1.14	2.10, 2.16	3.53, 3.58
		10%	1.84, 1.86	2.94, 2.97	4.15, 4.19
	196	50%	1.26, 1.29	2.53, 2.59	4.09, 4.14
		10%	2.06, 2.08	3.34, 3.38	4.50, 4.53
2	100	50%	1.17, 1.19	1.83, 1.87	3.24, 3.30
		10%	1.88, 1.90	2.57, 2.60	4.15, 4.19
	196	50%	1.37, 1.40	2.14, 2.18	3.90, 3.95
		10%	2.14, 2.14	2.91, 2.95	4.51, 4.55

Table 2.2 consists of confidence intervals for the deviance residual when the final event time is a constructed outlier. The FH deviance residuals are especially ineffective, rarely associating a large deviation with the outlier. Even the parametric deviance

residuals suffer somewhat, since the last observation, when an outlier, is highly influential for the estimation of β. The MB deviance residuals perform reasonably well for low levels of censorship and always outperform FH deviance residuals.

Figure 2.5 compares the correct cumulative hazard for the example data set to that used with our method (MB cumulative hazard) and that used with FH residuals (Breslow cumulative hazard). The Breslow cumulative hazard is too large at intermediate values and too flat at extreme values. Our method reduces this problem but does not fully correct it.

2.7 Discussion

For these particular parametric models and outliers, the MB deviance residuals behave more like parametric deviance residuals than do the FH deviance residuals. When there are no outliers in the data, standard asymptotic results can be used to show that parametric deviance residuals behave much like independent standard normal variates for large samples (Pierce and Schafer, 1986). When there are no outliers, FH deviance residuals have been observed to follow a standard normal distribution as well, especially when the censoring rate is not too high (Fleming and Harrington, 1984; Therneau and Grambsch, 1990). We found similar behavior with MB deviance residuals in these simulations.

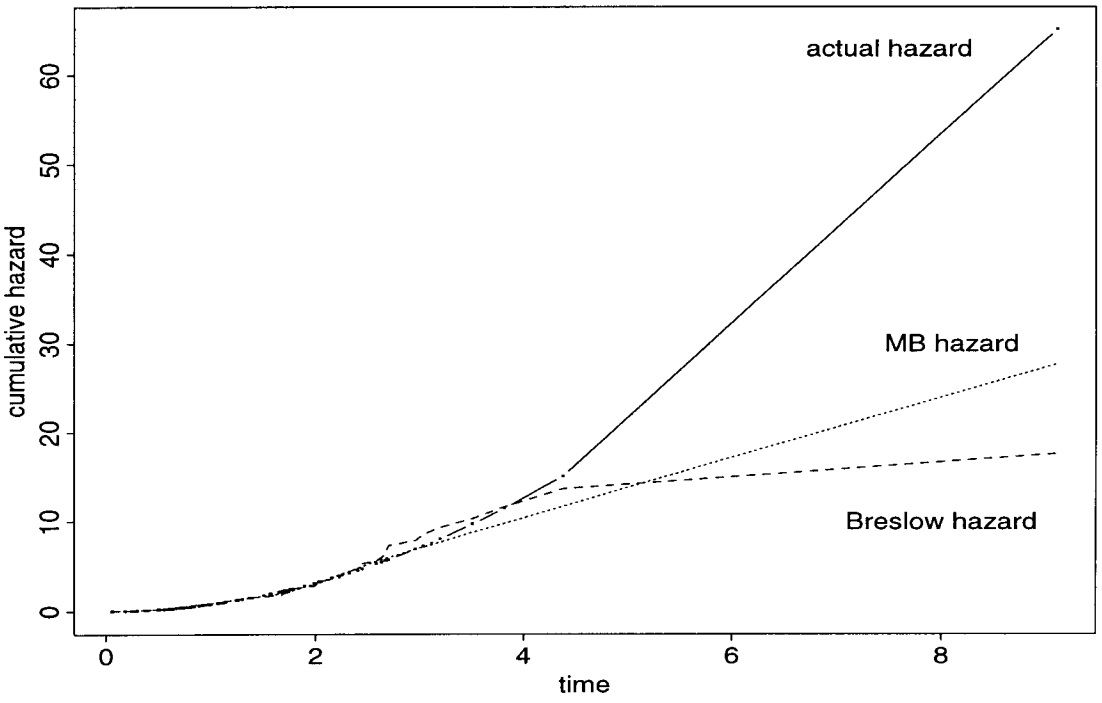

Figure 2.5 Three cumulative hazards for the example data set. The true cumulative hazard, our MB cumulative hazard based on clustering the baseline hazard, and the Breslow cumulative hazard.

On the other hand, when there is an outlier in the data, it is supposed the outlier will have a deviance residual indicating a large deviation from a standard normal distribution. Both the parametric and MB deviance residuals display this behavior to varying degrees. FH deviance residuals, however, seem to behave as independent standard normal random variates whether or not there is an outlier in the data, making their diagnostic value suspect. Clustering events, with the resulting modified hazard function, generates semi-parametric deviance residuals with increased diagnostic power.

Clustering events may have other benefits as well. Baltazar-Aban and Pena (1995) considered a semi-parametric residual similar to 1 - M_i, where M_i is defined in (3). This

residual is closely related to the deviance residual and is another tool for checking model assumptions in PH models. They similarly found those residuals displayed the distributional properties expected under model assumptions, even when the model assumptions were violated, calling their diagnostic value into question. They conjectured, as we have, that "overusing the data" to estimate $\Lambda(\cdot)$ may be the source of the problem. The Breslow hazard function, with its numerous nuisance parameters, may not be appropriate for diagnostic purposes. A hazard function with fewer nuisance parameters, but still non-parametric, may generally be more useful.

Deviance residuals of all kinds detect early events as outliers much more readily than they detect late events as outliers. This is surprising because, when the late event is an outlier of the magnitude considered here (Figure 2.1), it is obviously an unusual observation in any univariate analysis. The apparent difficulty is the influence of the late event on the cumulative hazard function when it is an outlier. Estimation of the Breslow cumulative hazard on the far right depends solely on the last event; consequently, this event has significant influence on the right extreme of the function (see Figure 2.5). Clustering only marginally reduces this problem. A residual that explicitly addresses this pattern of influence may better detect unusual observations among late events in failure time data.

Finally, there is nothing special about treating the baseline hazard as a step function, nor in the particular cluster sizes chosen. Both kernel smoothers (Staniswalis, 1989) and polynomial splines (Kooperberg et. al., 1995) have been suggested for estimating the baseline hazard. Either a more sophisticated clustering strategy or one of these other non-parametric smoothing techniques could be used to estimate martingale-based residuals, potentially leading to better results than found here.

2.8 References

BALTAZAR-ABAN, I. and PENA, E. A. (1995). Properties of hazard-based residuals and implications in model diagnostics. *Journal of the American Statistical Association* 90, 185-97.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.

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

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* 60, 267-78.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

KOOPERBERG, C., STONE, C.J., and TROUNG, Y.K. (1995). Hazard regression. *Journal of the American Statistical Association* 90, 78-94.

LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

MCCULLAGH, P. and NELDER, J. A. (1994). *Generalized Linear Models.* Cambridge: University Press.

OAKES, D. (1972). Contribution to the discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* 34, 187-220.

PIERCE, D. A. and SCHAFER, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* 81, 977-86.

STANISWALIS, J. G. (1989). The kernel estimation of a regression function in likelihood based models. *Journal of the American Statistical Association* 84, 276-83.

THERNEAU, T. M. and GRAMBSCH, P. A. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147-60.

Chapter 3

A Diagnostic Tool for Identifying the Shape of the Baseline Hazard in Survival Data

DeWayne R. Derryberry and Paul A. Murtaugh

3.1 Abstract

A new deviance residual is derived for proportional hazards survival data. It is used to identify the shape of the baseline hazard and to judge when a reasonable model has been fitted to the hazard. Some important properties of the residual are here derived; its use is demonstrated with several examples, and its effectiveness is evaluated with simulations.

3.2. Introduction

The most common models for censored survival data are the Cox (1972) model and the Weibull model. Both regression models share the proportional hazards (PH) assumption:

$$\Lambda(t,z) = \Lambda(t)\exp(z\beta), \tag{1}$$

where $\Lambda(\cdot)$ is a cumulative hazard, z is a covariate vector and β is a vector of parameters to be estimated. For Weibull regression, $\Lambda(t) = \lambda^{\rho}t^{\rho}$. For Cox regression, $\Lambda(t)$ is left unspecified, although the Breslow cumulative hazard, a non-parametric estimate, is often used in conjunction with Cox regression (Fleming and Harrington, 1991). The exponential or constant-hazards model is a special case of the Weibull, having ρ equal to one.

Although $\Lambda(t)$ is a nuisance parameter, there is often an interest in $\Lambda(t)$ itself, or more frequently in

$$\frac{d\Lambda(t)}{dt} = \lambda(t).$$

This latter quantity, often called the baseline hazard, offers considerable information about the nature of the failure mechanism. In medical applications, the baseline hazard is often conjectured to be constant or monotone increasing (Lawless, 1982, p.11;

Padgett and Wei, 1980). In Engineering, it is often argued that the baseline hazard is U-shaped (Lawless, 1982, p. 11; Grosh, 1989, p. 27).

We will present a new residual as a diagnostic tool for identifying the form of the baseline hazard. Use of the residual is demonstrated as follows: data from four types of baseline hazards are simulated and three hypothesized models are fitted, yielding residuals predicated on that hypothesized model. When the hypothesized model is correct, these residuals appear random, but when the hypothesized model is incorrect, these residuals display systematic departures from randomness. The four true models and three hypothesized models were chosen so that every hypothesized model would be appropriate in some cases and inappropriate in others.

These residuals give a local (in time) estimate of the difference between the modeled baseline hazard and the observed baseline hazard. When the hypothesized model is constant hazard, these residuals are of special significance, because the discrepancy between the model and the observed failure rates are deviations from a constant rate over time. For example, these residuals (assuming constant hazard), when plotted in temporal order, can be used to visually identify whether the baseline hazard is constant, monotone increasing, or U-shaped.

The following presentation will be three-fold: first, we will discuss the properties of the new residual. Next, we will present illustrative examples showing how the residual is used to identify the shape of the baseline hazard. Finally, we will examine via simulation the effectiveness of the new residuals.

3.3 Deviance Residuals

3.3.1 *The PH Loglikelihood*

For randomly censored survival data, with proportional hazards (1), the log-likelihood is

$$-\sum_{i=1}^{n} \Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n} \delta_i \log\{\lambda(t_i)\exp(z_i\beta)\}, \qquad (2)$$

where $t_1 \ldots t_n$ are the ordered observation times and δ_i is an indicator variable set to 1 for observations that are failures and 0 for observations that are censored. When $\Lambda(t)$ is the Breslow cumulative hazard, (2) is equivalent to the partial likelihood used in Cox regression (Breslow, 1974). The diagnostic tool we consider is a deviance residual derived from this loglikelihood, with each hypothesized model (constant hazard, Weibull, non-parametric monotone) representing a different choice of $\Lambda(t)$.

A deviance residual can be defined as the signed square root of a likelihood ratio test where the relevant hypotheses are: H_o: the current model fits at sample value i; H_a: a different model is appropriate for sample value i. For each observation in the sample

$$d_i = \pm [2\{\text{loglik(full model}_i) - \text{loglik(reduced model)}\}]^{\frac{1}{2}},$$

where sample value i has a unique value for the parameter of interest in the full model.

The usual deviance residual in the Cox model is

$$d_i = \pm (2[\Lambda(t_i)\exp(z_i\hat{\beta}) - \delta_i - \delta_i \log\{\Lambda(t_i)\exp(z_i\hat{\beta})\}])^{\frac{1}{2}}, \qquad (3)$$

where the sign is determined by the sign of $\Lambda(t_i)\exp(z_i\hat{\beta}) - \delta_i$ (Fleming and Harrington, 1991, p. 168). The covariate parameters are estimated using Cox regression, and $\Lambda(t)$ is usually estimated with the Breslow hazard. This residual, found by treating the covariate parameters as those of interest and the hazard function as a nuisance

parameter, is useful in detecting outliers (Therneau and Grambsch, 1990; Fleming and Harrington, 1991, p. 189).

The residual above was found by treating the observed times (censored or failures) as the sample values. By defining $\Delta_i = t_i - t_{i-1}$ and $t_0 = 0$, where t is the observed time, a different set on n sample values is obtained. This approach treats the baseline hazard as the parameter of interest and the covariate parameters as nuisance parameters, we can derive (see Appendix) a new deviance residual of the form

$$d_i^t = \pm (2[\Delta_i \hat{\lambda}(t_i) \sum_{k=i}^{n} \exp(z_k \hat{\beta}) - \delta_i - \delta_i \log\{\Delta_i \hat{\lambda}(t_i) \sum_{k=i}^{n} \exp(z_k \hat{\beta})\}])^{\frac{1}{2}}. \qquad (4)$$

This residual measures differences between observed and predicted failure rates on the intervals between observed times.

3.3.2 *Properties of the New Deviance Residual*

When the correct model is chosen for the baseline hazard, with certain qualifications stated below, these residuals are independent and identically distributed, from a bell-shaped (but not normal) distribution with known statistical properties.

Property i - For a data set with n independent observation times (failures or censored observations), the n residuals defined by (4) are "almost" independent. In this case, Δ_i and Δ_j, with $i \neq j$, are independent, unless j = i-1 or i+1. As is common when taking differences, adjacent Δ_i's have a negative correlation of $2^{-\frac{1}{2}}$. Because the deviance residual is a transformation of this random variable, only adjacent deviance residuals are dependent; all others are independent of each other. Hence, the n deviance residuals are "almost" independent in the sense that each one is independent of 100(n-3)/(n-1)% of the other residuals. Some additional dependence may be induced because parameters are estimated from the data and are then used to calculate the residuals.

Property ii - For identically distributed uncensored data, the quantity $\lambda_i^b = \{\Delta_i(n-i+1)\}^{-1}$ is both the observed local failure rate and the Breslow baseline hazard rate on the interval $(t_{i-1},t_i]$. It can be shown (4) has the form:

$$d_i^t = \text{sign}[\log\{\hat{\lambda}(t_i)/\lambda_i^b\}](2[\hat{\lambda}(t_i)/\lambda_i^b - 1 - \log\{\hat{\lambda}(t_i)/\lambda_i^b\}])^{\frac{1}{2}}.$$

By Taylor series expansion near $\lambda_i^b = \hat{\lambda}(t_i)$,

$$-d_i^t \approx \log(\lambda_i^b) - \log\{\hat{\lambda}(t_i)\}. \tag{5}$$

This residual is always zero when the observed failure rate equals the hypothesized failure rate, negative when the hypothesized failure rate exceeds the observed failure rate, and positive otherwise. The residual defined in (4) can be thought of as a monotone transformation and generalization of (5) that improves the overall distributional shape and extends to censored regression data.

Property iii - The residual defined by (4) closely parallels (3), the standard deviance residual. Suppose the baseline hazard rate is almost constant on each observed interval. Further, recall that the minimum of several exponential event times is exponential with a rate that is the sum of the individual rates. Then

$$h_i^t = \Delta_i\hat{\lambda}(t_i)\sum_{k=i}^{n}\exp(z_k\hat{\beta}) \;\dot{\sim}\; \text{EXPO}(1),$$

which can be extended to include censoring by using the memoryless property of exponentials (Lawless, 1982, p. 281) to

$$h_i^t = \Delta_i\hat{\lambda}(t_i)\sum_{k=i}^{n}\exp(z_k\hat{\beta}) + 1 - \delta_i \;\dot{\sim}\; \text{EXPO}(1). \tag{6}$$

Then (4) can be re-expressed as,

$$d_i^t = \text{sign}(h_i^t - 1)[2\{h_i^t - 1 - \delta_i\log(h_i^t)\}]^{\frac{1}{2}}. \tag{7}$$

Similarly, the standard deviance residual (Lawless, 1982, p. 366) has

$$h_i = \hat{\Lambda}(t_i)\exp(z_i\hat{\beta}) + 1 - \delta_i \;\dot{\sim}\; \text{EXPO}(1), \tag{8}$$

and (3) can be re-expressed as

$$d_i = \text{sign}(h_i - 1)[2\{h_i - 1 - \delta_i\log(h_i)\}]^{\frac{1}{2}}.$$

Property iv - For identically distributed uncensored data, detailed knowledge of the distribution of the new deviance residuals is possible. Using (6) and (4), or (8) and (3), both d_i^t and d_i become a transformation of a unit exponential:

$$y_i \sim \text{EXPO}(1), \text{ and } d_i = \text{sign}(y_i - 1)[2\{y_i - 1 - \log(y_i)\}]^{\frac{1}{2}}. \quad (9)$$

These are not standard normal, as has been conjectured (Therneau and Grambsch, 1990). A plot of 10,000 such random variables (Figure 3.1) shows that although they display an approximate bell shape, they are shifted away from 0. The median is $-(2[\log(2) - 1 - \log\{\log(2)\}])^{\frac{1}{2}} \approx -0.34543$.

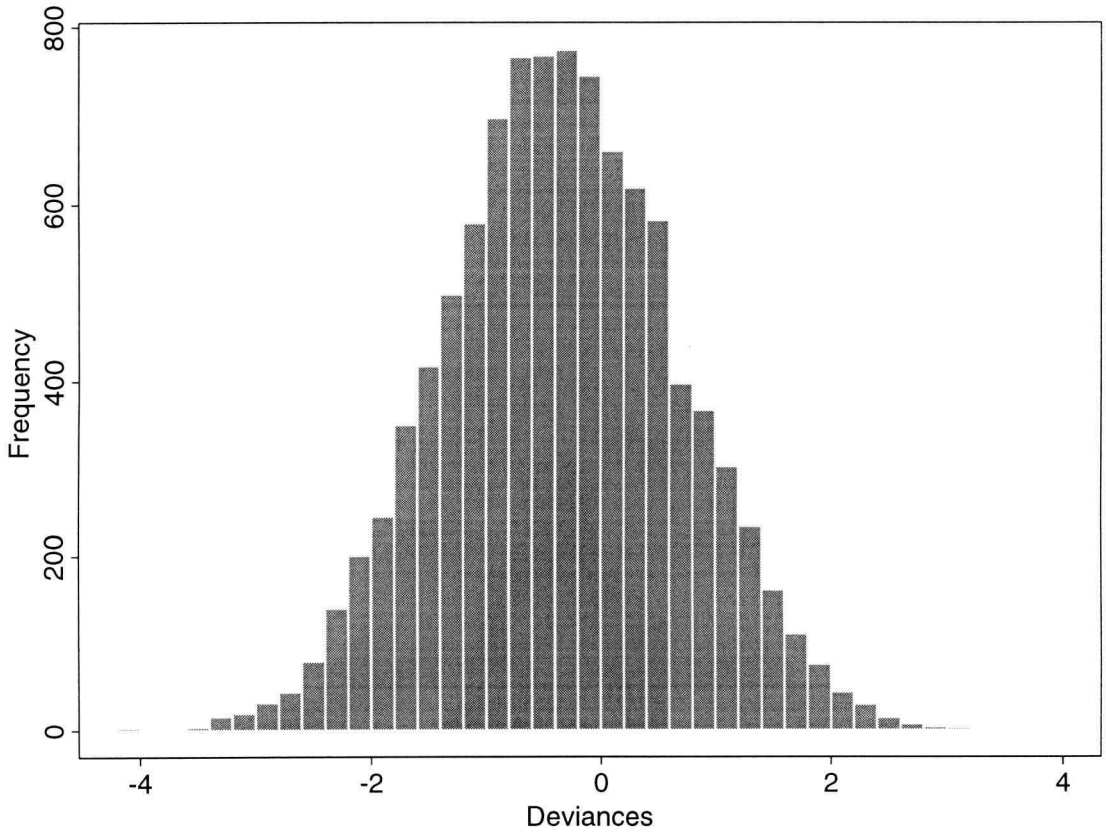

Figure 3.1 Histogram of the transformation $y = \text{sign}(e - 1)\{2(e - 1 - \log(e)\}^{\frac{1}{2}}$ applied to 10,000 unit exponential random variables.

The approximate symmetry of the distribution justifies using the median as an estimate of the mean. The expectation of the squared residual can be shown to be 2γ, where $\gamma \approx 0.5772...$ (Euler's number). This distribution, then, is nearly symmetric with a mean of about -0.34543 and a variance of about 1.03508. For the data from the histogram, the sample mean was -0.352, the sample median was -0.3688, and the sample variance was 1.0537.

When the negatives of these residuals are plotted in temporal order, discrepancies between hypothesized failure rates and observed failure rates emerge. When the correct model is selected, the residuals are "almost" independent and identically distributed following the transformed distribution in (9). In this case, the residuals should look random when plotted in temporal order. When an incorrect model is selected, trends and outliers appear in the residuals. Outliers indicate intervals where observed failure rates differ greatly from predicted failure rates. Using (9), one can estimate quantiles, making identifying and interpreting outliers straightforward.

When the hypothesized model is constant hazard, patterns in the negative residuals indicate the form of the true baseline hazard. Systematic trends have a simple interpretation. For example, a U-shaped pattern in the negative residuals is evidence of a U-shaped baseline hazard.

3.4 Examples of Diagnostic Plots

3.4.1 *Examples with Known Baseline Hazard*

Below are examples of residuals from various models fitted to simulated data with a known baseline hazard.

In each of Figures 3.2-3.5, 100 failure times from one of the following baseline hazards were simulated:

Baseline hazard	Constant	Weibull	Monotone	U-shaped
$\lambda(t) =$	1	$2\{\Gamma(1.5)\}^2 t$	$.25 + t^2$	$.45 + .9(t - 1)^2$

The characteristics we identify in each figure are typical for data sets generated in this manner.

Residuals were found subject to three fits: constant, Weibull, and non-parametric monotone increasing baseline hazards. Parameter estimates were found by maximum likelihood in the Weibull and constant-hazards cases. For a non-parametric monotone increasing hazard, a Cox regression estimate of the regression parameter was found; then a monotone step function was found for the baseline hazard by using an algorithm suggested by the proofs given by Chung and Ching (1994).

In each of Figures 3.2-3.5, four plots appear: a scatterplot and three residual plots. The plot in the upper left corner is a scatterplot of log (hazard) versus log (time). A linear relation indicates a Weibull model, and no pattern indicates a constant hazard. Unfortunately, any other systematic pattern is difficult to interpret. The straight line and step function superimposed on this plot are the Weibull and non-parametric estimates of log (hazard), respectively.

This plot, though useful, has several limitations: it is unclear how to extend this plot to include censored data, how much deviation from linearity is required before we

claim the data are not Weibull, and how to identify other models by using this plot or some modification of it.

The other three graphs in each of Figures 3.2-3.5 are of residuals from fitting a constant hazard (upper right), Weibull hazard (lower left), and non-parametric monotone hazard (lower right) to the data. When the correct model is chosen, the residual plots appear random; when an incorrect model is chosen, trends and outliers appear in the residuals.

The three horizontal lines are quantiles of 0.01, 0.50, and 0.99 (i.e. about half the residuals should be above the center line, about 1 in 100 above the upper line, and about 1 in 100 below the lower line) based on (9), and are useful for detecting outliers. The S-Plus function "lowess" (Venables and Ripley, 1994, section 10.1) was applied to the residuals in each plot to help identify trends.

In Figure 3.2, the true baseline hazard is constant. The scatterplot of log (hazard) versus log (time) suggests a single, constant hazard rate. The step function (non-parametric fit) and straight line (Weibull fit) overlaid on the scatterplot both approximate a constant hazard rate. The residual plot for a constant hazard seems random: all three residual plots fit the data well and have similar lowess-smoothed residuals. Although there is an outlier, it persists in all three models. Because all three fitted models are reasonable, constant hazard is the parsimonious choice.

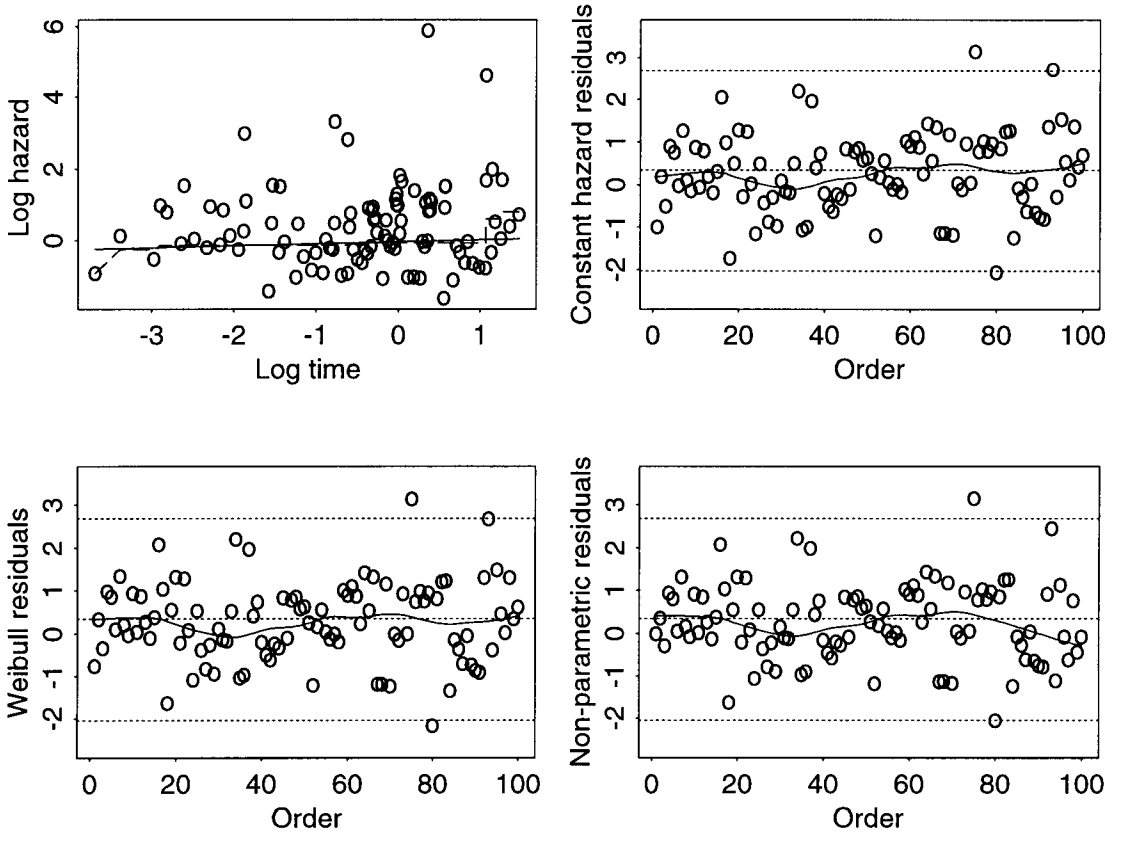

Figure 3.2 For a sample data set of 100 failure times with a constant baseline hazard, the upper left plot is a scatterplot of log(time) versus log(hazard). The straight line and step function superimposed on the scatterplot are the Weibull and non-parametric estimates of log(hazard). The remaining three graphs are of residuals, plotted in temporal order, from a constant-hazards fit (upper right), a Weibull fit (lower left), and a non-parametric monotone fit (lower right). There are 0.01, 0.50, and 0.99 quantiles and a lowess-smoothed curve overlaid on each residual plot.

In Figure 3.3, with data generated from a Weibull baseline hazard, the relation between log (time) and log (hazard) appears linear. The non-parametric (step-function) and parametric (straight line) fits are similar. The residual plot for constant hazard has a trend in the residuals -- evidence that a constant hazard is inadequate and that a better fit is monotone increasing. The earliest interval has an extremely small residual,

suggesting a failure rate far less than expected under constant hazards. Three other early residuals are borderline outliers in the same direction. The next two plots show an appropriate fit -- the outliers are gone and the lowess-smoothed curve is relatively flat. As expected, the Weibull and non-parametric monotone models are both adequate for the data. Parsimony suggests the simpler Weibull model.

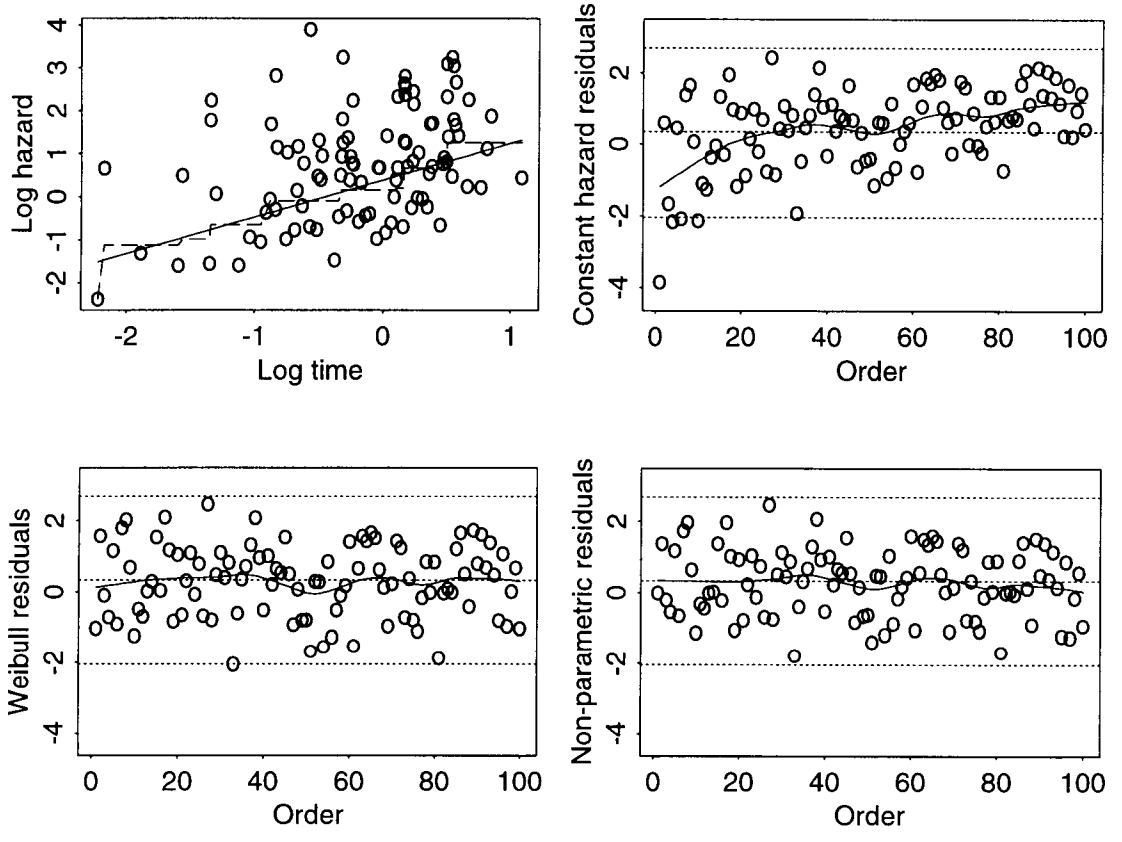

Figure 3.3 For a sample data set of 100 failure times with a Weibull baseline hazard, the four plots are as in Figure 3.2.

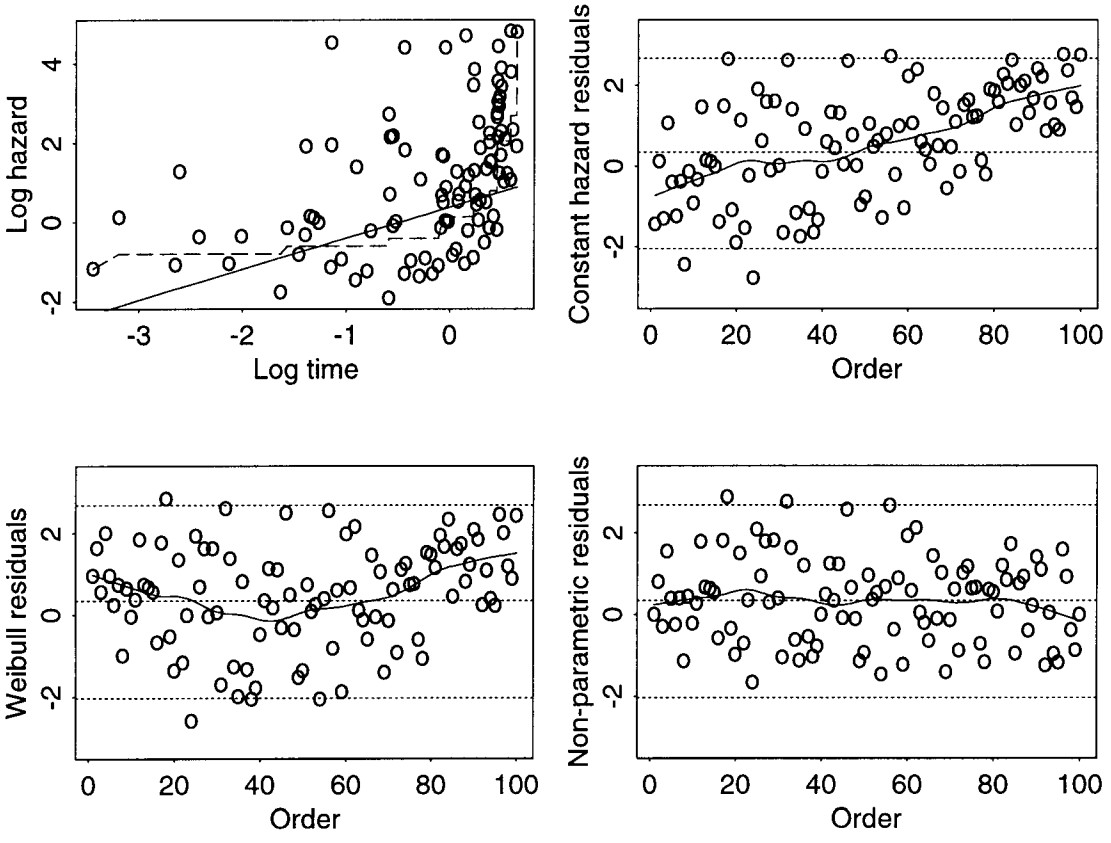

Figure 3.4 For a sample data set of 100 failure times with a monotone (but not Weibull) baseline hazard, the four plots are as in Figure 3.2.

Figure 3.4 has data simulated with a non-Weibull monotone baseline hazard. The scatterplot suggests deviation from a Weibull fit. Comparing the two fits overlaid on the scatterplot, we see that the straight line (Weibull fit) is inadequate. The step function (non-parametric fit), which does fit the scatterplot, appears almost L-shaped. Again, the plot of residuals from a constant fit indicates a monotone increasing hazard. Two negative outliers are associated with early time intervals, and several mild positive outliers are associated with late time intervals, which is evidence that the constant-hazard model overestimates the early failure rates and underestimates later failure rates.

The residuals from a Weibull fit are also inadequate (as expected); a slight dip persists in the middle of the residuals (the lowess-smoother accentuates this), and a large negative outlier is still present. Although the non-parametric fit does have two mild outliers in the first half of the data, it has the flattest lowess-smoothed curve and the mildest outliers, and it seems a reasonable model.

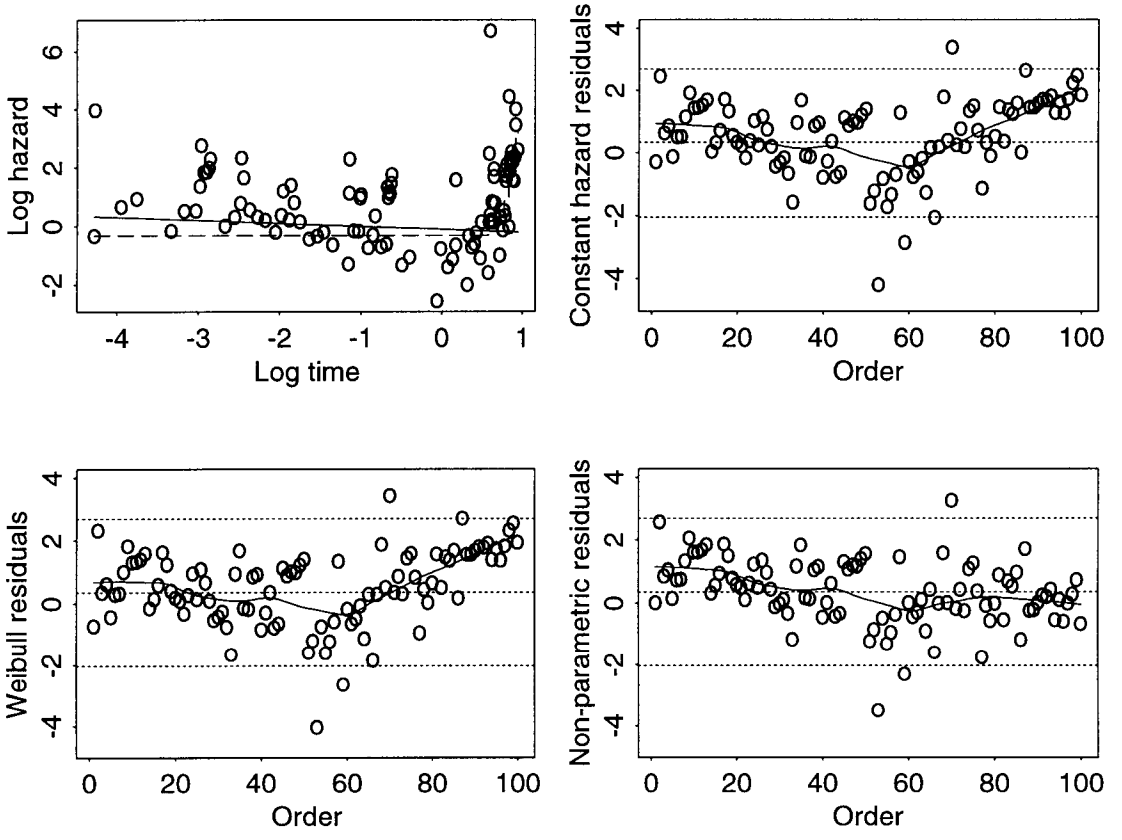

Figure 3.5 For a sample data set of 100 failure times with a U-shaped baseline hazard, the four plots are as in Figure 3.2.

A U-shaped baseline hazard was used to generate the data in Figure 3.5. The scatterplot is uninterpretable, except that a Weibull fit is out of the question. The large negative outliers in the middle of each residual plot approximate the minimum of the true baseline hazard. The residual plot predicated on a constant hazard is indicative of a U-shaped hazard.

As expected, both monotone models poorly fit the data from a U-shaped hazard. The residuals from both the Weibull and non-parametric fit still contain systematic departures from randomness. The non-parametric fit, being more flexible, displays a better fit then the Weibull. Nevertheless, the non-parametric residuals display several severe outliers and some trend in the lowess-smoothed curve.

For each of the above examples, examination of trends and outliers in the residual plots help identify the model we knew to be correct.

3.4.2 *An Example from Lifetime Testing*

Lawless (1982, p. 189) presents data on 40 specimens of cable insulation. A voltage stress test was performed with the following results.

Insulation type	Failure times
I	32.0, 35.4, 36.2, 39.8, 42.1, 43.3, 45.5, 46.0, 46.2, 46.4, 46.5, 46.8, 47.3, 47.4*, 47.6, 49.2, 50.4, 50.9, 52.4, 56.3
II	39.4, 45.3, 49.3*, 49.4, 51.3, 52.0, 53.2, 53.3*, 54.9, 55.5, 57.1, 57.2, 57.5, 59.2, 61.0, 62.4, 63.8, 64.3, 67.3, 67.7

* - Ties were broken by adding 0.10 to some of the original failure times.

There was no censoring. The data are claimed by Lawless to follow a Weibull distribution of the form:

$$\lambda(t,z) = \rho\lambda^{\rho}t^{\rho-1}\exp(z\beta),$$

where z is and indicator for insulation type. If the Weibull model is appropriate, then $H_o:\beta = 0$ is the test of whether insulation types differ.

In Figure 3.6, the first plot is the usual scatterplot of log(hazard) versus log(time). The straight line based on a Weibull maximum likelihood fit characterizes the relation. Nevertheless, the new residuals will be presented as an alternative analysis that is at least as illuminating but more general.

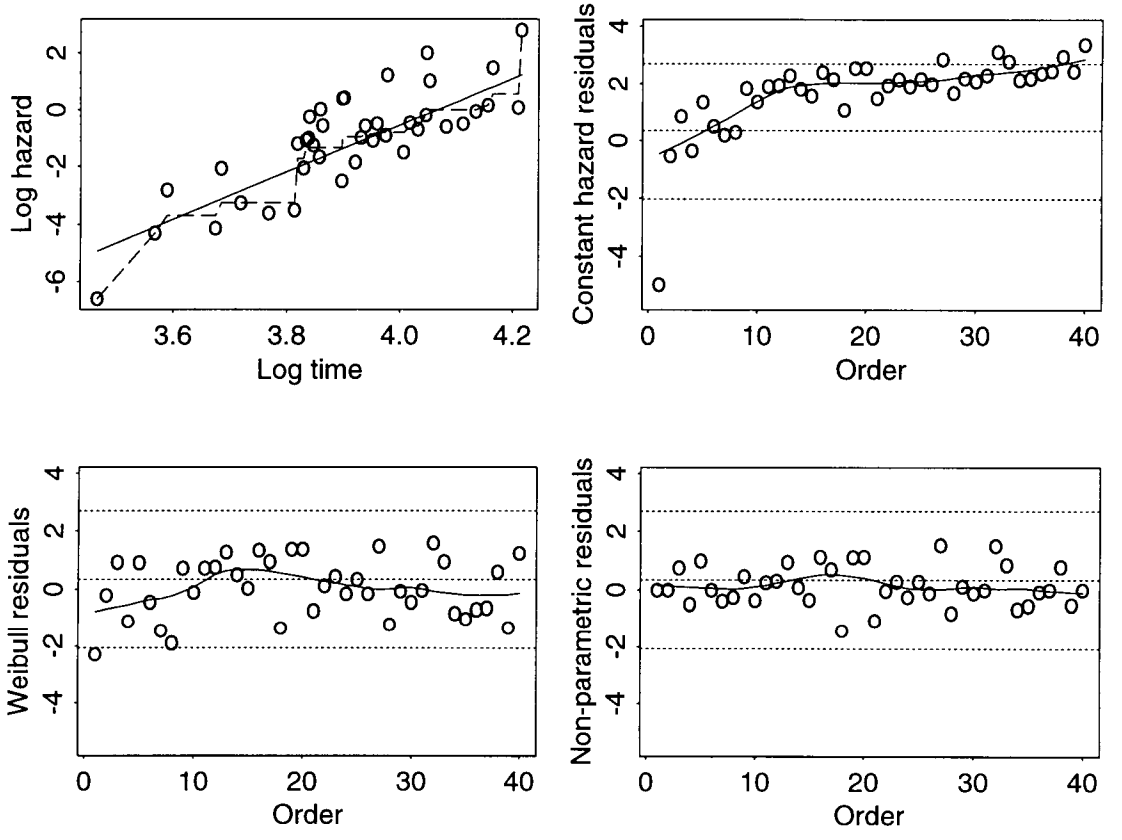

Figure 3.6 For the insulation type, voltage test data (Lawless 1982, p. 189), the four plots are as in Figure 3.2.

The residual plot predicated on constant hazard, in Figure 3.6, gives evidence of a monotone increasing hazard, suggested both by the trend in the lowess-smoothed residuals and by the many outliers in the data. The large negative deviation associated with the first time interval suggests this interval has a much smaller observed failure rate than is consistent with constant hazard, while several positive outliers at much later time segments indicate failure rates much higher than predicted, assuming constant hazard. A runs test performed on these residuals yielded a p-value of 0.004, indicating the inadequacy of a constant-hazards model.

The Weibull model residuals appear random (runs test p-value = 0.30), except perhaps for the slight outlier associated with the initial time interval and a slight monotone trend in the lowess curve for early time segments. The non-parametric fit has neither outliers nor any indication of trend (runs test p-value = 0.646). More experience is required with these residuals before we can say whether the Weibull model is sufficient, or whether the non-parametric fit is needed.

3.5 A Simulation

The presentation of the previous section is inherently anecdotal and subjective. An objective assessment of the residual plots as diagnostic tools was undertaken by simulating data with the four different baseline hazards. Data were simulated both with and without censoring; with and without a regression covariate; and at two sample sizes. With censoring, rates were about 10-15%, and censoring times were from an exponential distribution. All three models (constant, Weibull, and monotone baseline hazard) were fitted to each data set. A runs test (Daniel, 1990) was performed on each set of residuals to assess the appropriateness of the assumed model.

The first expectation was that the runs test has a rejection rate of about 5% when the null hypothesis is true, and a higher rate when the null hypothesis is false. Even for small sample sizes (Table 3.1, sample size = 50), this was generally the case. The lone exception occurred when the fitted model was non-parametric and the true model was Weibull or monotone. In these cases, rejection rates were high (a 95% confidence interval of 0.064 to 0.08).

Table 3.1 Each entry is a rejection rate (at the 0.05 level) for 500 simulated data sets of sample size 50; the rates in italics should be 0.05 (the null hypothesis is true), and the rates in bold should exceed 0.05 (the null hypothesis is false).

		Actual model			
Data type	Fitted model	Constant	Weibull	Monotone	U-shaped
Identically distributed	constant	*0.050*	**0.236**	**0.434**	**0.320**
	Weibull	*0.056*	*0.046*	**0.082**	**0.282**
	monotone	*0.050*	*0.078*	*0.098*	**0.136**
Censored					
	constant	*0.060*	**0.246**	**0.362**	**0.300**
	Weibull	*0.050*	*0.058*	**0.080**	**0.274**
	monotone	*0.046*	*0.064*	*0.060*	**0.134**
Regression					
	constant	*0.038*	**0.334**	**0.374**	**0.318**
	Weibull	*0.038*	*0.076*	**0.068**	**0.304**
	monotone	*0.060*	*0.068*	*0.066*	**0.140**
Censored, regression	constant	*0.042*	**0.316**	**0.458**	**0.286**
	Weibull	*0.040*	*0.052*	**0.120**	**0.268**
	monotone	*0.032*	*0.064*	*0.080*	**0.148**

Next, we considered whether rejection rates were similar in the four cases: identically distributed data, censored data, regression data, and censored regression data. Although the properties of the residuals were derived under the assumption of identically distributed, uncensored data, we found these residuals useful for censored regression data as well. For each of the twelve possible pairings of true and hypothesized models, rejection rates were of comparable magnitude in each of the four cases. For example, when a constant hazard is fitted to data with a Weibull hazard, the rejection rates were 0.236, 0.246, 0.334, and 0.316 for the identically distributed, censored, regression, and censored-regression cases, respectively.

Table 3.2 Each entry is a rejection rate (at the 0.05 level) for 500 simulated data sets of sample size 100; the rates in italics should be 0.05 (the null hypothesis is true), and the rates in bold should exceed 0.05 (the null hypothesis is false)

		Actual model			
Data type	Fitted model	Constant	Weibull	Monotone	U-shaped
Identically distributed	constant	*0.042*	**0.352**	**0.660**	**0.556**
	Weibull	*0.050*	*0.056*	**0.058**	**0.550**
	monotone	*0.044*	*0.106*	*0.086*	**0.228**
Censored					
	constant	*0.056*	**0.428**	**0.662**	**0.486**
	Weibull	*0.054*	*0.042*	**0.126**	**0.486**
	monotone	*0.044*	*0.068*	*0.102*	**0.202**
Regression					
	constant	*0.050*	**0.502**	**0.754**	**0.498**
	Weibull	*0.052*	*0.054*	**0.070**	**0.508**
	monotone	*0.056*	*0.068*	*0.094*	**0.184**
Censored, regression	constant	*0.044*	**0.478**	**0.656**	**0.466**
	Weibull	*0.040*	*0.047*	**0.130**	**0.442**
	monotone	*0.052*	*0.074*	*0.104*	**0.240**

Increased sample sizes produced better results. In general, we expect rejection rates in Table 3.2 (sample size = 100) to be closer to 5% when the null hypothesis is true (all italicized rejection rates), and we expect substantial increases in rejection rates when the null hypothesis is false (all rejection rates in bold).

For all three fitted models when the true model is constant hazard, and for the Weibull fit when the true model is Weibull, rejection rates are generally closer to 0.05 in Table 3.2 than in Table 3.1. For example, the range of rejection rates in Table 3.1, over all the above cases, was 0.032 to 0.076, but in Table 3.2 the range was 0.040 to 0.056. The lone exception is, again, the non-parametric fit when the true model is either Weibull or monotone. In these cases the rejection rates are still too high.

When rejection rates should be greater than 5% (i.e., when the null hypothesis is false), the rejection rates, with one minor exception, were much higher in Table 3.2 than in Table 3.1. In several cases, rejection rates are nearly twice as high in Table 3.2 than in Table 3.1.

When a Weibull model is fitted to data with a more general monotone hazard, little difference appears in the rejection rates between Table 3.1 and Table 3.2. These data sets may just be too small, or the runs test too general to detect subtle differences between a Weibull hazard and a different shaped, but still monotone hazard.

In Table 3.2 (as in Table 3.1), differences in the rejection rates were not practically significant for the four cases (identically distributed, regression, censored, and censored-regression data), so the residuals seem equally useful in all four cases.

The non-parametric monotone fit is problematic. The non-parametric residuals seemed to perform well in practice (for example, see Figures 3.2-3.6), but they did not behave as expected when a runs test was applied to them. The persistence of rejection rates greater than 5% when the null hypothesis is true requires further examination. Although we were not able to find a satisfactory explanation, two conjectures come to

mind. Because the fit is non-parametric, a degree of over-fitting of the data may be present. Over-fitting almost certainly accounts for the low rejection rates for a U-shaped hazard (compared to parametric models), but it might also account for the high rejection rates for monotone hazards (over-fitting might cause too many runs in a set of residuals). A second possibility relates to the dependence among the residuals. Although this lack of independence did not seem to be a problem in general, it may interact with the non-parametric fit in some subtle way.

3.6 Discussion

The deviance residuals defined here are effective diagnostic tools for detecting and modeling the form of the baseline hazard. A temporally ordered plot of these residuals, with hypothesized constant hazard, indicates which sorts of models to consider. Any trend in this plot describes the shape of the correct hazard. Subsequent plots with new hypothesized models aid model selection because the residual plot will appear random (especially without trend or outliers) when the correct model is used.

The residuals seem effective with mild censoring and in a simple regression setting, as well as in the identically distributed, uncensored case.

These plots might be more useful then the simulations suggest. The runs test cannot, nor can any statistical test, simultaneously detect all departures from randomness. Statisticians rely on judgment rather than tests when examining residuals. We think that an experienced statistician would correctly reject as non-random many residual plots that the runs test does not reject without incurring a higher rate of type I error. In Figures 3.2-3.6, for example, the impression from each residual plot agreed with the p-value reported from a runs test with one exception. In the upper right corner of Figure 3.2, the residuals have a pronounced monotone trend, although the p-value

from a runs test was 0.51. If expert opinion could have been used in our simulations, instead of a runs test, the results in Tables 3.1 and 3.2 might have been stronger.

This paper has limited the fitted models to two widely known models (constant and Weibull hazard) and to a non-parametric model (monotone increasing hazard) taken directly from the work of Chung and Ching (1994). These models are sufficient to illustrate the diagnostic power of the residual, but they are hardly exhaustive. The residuals we have derived are useful for any PH model (2) as long as some parametric or non-parametric method exists for estimating the baseline hazard.

In particular, it seems attractive to have a method of examining the appropriateness of non-parametric models. For example, it seems possible to fit a non-parametric U-shaped baseline hazard in a manner similar to the methods used by Chung and Ching (1994) for a monotone hazard. Given the use of U-shaped hazards in engineering, and the failure of a good parametric model to emerge, developing this algorithm seems important.

Many non-parametric methods provide consistent estimators of the baseline hazard. These estimators may be as simple as smoothing the Breslow baseline hazard (Kooperberg, Stone and Troung, 1995; Staniswalis, 1989). Such models often have smoothing parameters or other constants that must be selected by the user. Residuals could play an important role both in evaluating the quality of estimators and in the selection of smoothing parameters or related constants.

3.7 References

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.

CHUNG, D. and CHING, M. N. (1994). An isotonic estimator of the baseline hazard function in Cox regression model under order restrictions. *Statistics and Probability Letters* 21, 223-228.

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

DANIEL, W. W. (1990). *Applied Nonparametric Statistics*. Boston: PWS-KENT.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

GROSH, D. L. (1989). *A Primer of Reliability Theory*. New York: Wiley.

KOOPERBERG, C., STONE, C. J., and TROUNG, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* 90, 78-94.

LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

PADGETT, W. J. and WEI, L. J. (1980). Maximum likelihood estimation of a distribution function with increasing failure rate based on censored observations. *Biometrika* 67, 470-4.

PIERCE, D. A. and SCHAFER, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* 81, 977-86.

STANISWALIS, J. G. (1989). The kernel estimation of a regression function in likelihood based models. *Journal of the American Statistical Association* 84, 276-83.

THERNEAU, T. M. and GRAMBSCH, P. A. (1990). Martingale-based residuals for survival models. *Biometrika* 77, 147-60.

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

3.8 Appendix

A deviance residual for a sample value i can be defined as the signed square root of a likelihood ratio test for H_o: the current model fits at sample value i; H_a: a different model is appropriate for sample i. The residual has the form

$$d_i = \text{sign}(e_i - \text{predicted } e_i)\{2(\text{loglik[full model}_i\text{]} - \text{loglik[reduced model]})\}^{\frac{1}{2}},$$

where e_i denotes sample value i. We take the sample value to be the observed failure rate on the interval $(t_{i-1},t_i]$ between observed times. For the reduced model, we must find $\hat{\beta}_c$, $\hat{\Lambda}(t)_c$ and $\hat{\lambda}(t)_c$ that maximize

$$\text{L(reduced model)} = -\sum_{i=1}^{n}\Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n}\delta_i\log\{\lambda(t_i)\exp(z_i\beta)\}.$$

Using $\Lambda(t_i) \approx \sum_{j=1}^{i}\lambda(t_j)\Delta_j$, the baseline hazard is replaced by a step function. The order of summation is also reversed to give

$$\text{L(reduced model)} \approx -\sum_{i=1}^{n}\{\hat{\lambda}(t_i)_c\Delta_i\sum_{k=i}^{n}\exp(z_k\hat{\beta}_c)\} + \sum_{i=1}^{n}\delta_i\log\{\hat{\lambda}(t_i)_c\} + \sum_{i=1}^{n}\delta_i z_i\hat{\beta}_c.$$

The full model with respect to sample value i is maximized with the relation defined by the hypothesized model holding for all $\lambda(t_j)$ except j = i, but the maximum likelihood estimate of $\lambda(t_i)$ is chosen without any model constraints.

$$\text{L(full model}_i) \approx -\sum_{i\neq j}\{\hat{\lambda}(t_j)_s\Delta_j\sum_{k=j}^{n}\exp(z_k\hat{\beta}_s)\} + \sum_{i\neq j}\delta_j\log\{\hat{\lambda}(t_j)_s\}$$

$$+\sum_{i\neq j}^{j}\delta_j z_j\hat{\beta}_s - \hat{\lambda}(t_i)_s\Delta_i\sum_{k=i}^{n}\exp(z_k\hat{\beta}_s) + \delta_i\log\{\hat{\lambda}(t_i)_s\} + \delta_i z_i\hat{\beta}_s.$$

Note that for sample value i,

$$\hat{\lambda}(t_i)_s = \lambda_i^b = \delta_i\{\Delta_i\sum_{k=i}^{n}\exp(z_k\hat{\beta})\}^{-1}. \qquad (10)$$

Exact deviance residuals cannot be calculated without computing n distinct likelihood ratio tests. These calculation costs are often considered prohibitive, and $\hat{\beta}_c \approx \hat{\beta}_s \approx \hat{\beta}_{cox}$ and $\hat{\lambda}(t_j)_c \approx \hat{\lambda}(t_j)_s$ for $i \neq j$ are all assumed. Using these assumptions it is possible to approximate the exact deviance residual using the usual parameter estimates. The current model and Cox estimates have already been calculated as a by-

product of the statistical analysis. We have used the Cox estimate of β and the current model estimate of $\hat{\lambda}(t)$ with subscripts dropped. In this case,

$$d_i^2 = 2\{L(\text{full model}_i) - L(\text{reduced model})\}$$

so,

$$d_i^2 \approx 2[\{\hat{\lambda}(t_i) - \lambda_i^b\}\Delta_i \sum_{k=i}^{n} \exp(z_k\hat{\beta}) + \delta_i \log\{\lambda_i^b/\hat{\lambda}(t_i)\}].$$

Using (10) and noting the "δ_i" inside the log function is redundant,

$$d_i^2 \approx 2[\hat{\lambda}(t_i)\Delta_i \sum_{k=i}^{n} \exp(z_k\hat{\beta}) - \delta_i - \delta_i \log\{\hat{\lambda}(t_i)\Delta_i \sum_{k=i}^{n} \exp(z_k\hat{\beta})\}].$$

The sign of d_i must still be determined. The quantity $\log\{\lambda_i^b\} - \log\{\hat{\lambda}(t_i)\}$ is the difference between an observed failure rate and a predicted failure rate. When $|\log\{\lambda_i^b\} - \log\{\hat{\lambda}(t_i)\}| = 0$; then $d_i^2 = 0$. Further, d_i^2 increases as the magnitude $|\log\{\lambda_i^b\} - \log\{\hat{\lambda}(t_i)\}|$ increases. Hence the signed deviance residual is

$$d_i = \text{sign}[\log\{\lambda_i^b/\hat{\lambda}(t_i)\}](2[\hat{\lambda}(t_i)\Delta_i \sum_{k=i}^{n} \exp(z_k\hat{\beta}) - \delta_i - \delta_i \log\{\hat{\lambda}(t_i)\Delta_i \sum_{k=i}^{n} \exp(z_k\hat{\beta})\}])^{\frac{1}{2}}.$$

Chapter 4

Analysis of Split-Plot Censored Survival Data Using BLUP Estimators

DeWayne R. Derryberry and Paul A. Murtaugh

4.1 Abstract

Designed experiments in the health sciences sometimes involve blocking factors and split-plot designs, called "random effects models" in the linear models literature. Cox regression cannot be used directly to analyze data with these complex structures because the event times are not independent. We extend the Cox model to include random effects using a hierarchical Bayesian model. We describe and implement an algorithm for testing and estimating fixed effects in such cases. This algorithm produces BLUP-type predictors for the random effects. We then evaluate the algorithm for simulated split-plot data, and a data set is analyzed.

4.2 A Random Effects Model

The Cox (1972) regression model is widely used for analysis of censored survival data. As originally developed by Cox, this model is appropriate for data with independent observed times, although data sets in which the observations are not independent are common. For example, a medical study may require application of any one of several treatments to individual mice, but those many mice may come from just a few litters. If mice from the same litter are either robust or frail as a group, litter effects may overwhelm the treatment effects being studied.

In such cases, the experiment should be designed (if possible) so that all treatments are included in each litter, and a statistical tool should be available that recognizes a "litter effect" and makes treatment comparisons "within a litter." Consistent with the linear models literature, we consider the litter effect to be a random effect and the litter to be a blocking factor.

The Cox model can be extended to include a random effect as follows:

$$\Lambda(t_{ij}| z_{ij1}, ...z_{ijs}, \theta_j) = \Lambda(t_{ij})\exp(\sum_{f=1}^{s} z_{ijf}\beta_f + \theta_j) \quad (1)$$

$$\theta_j \sim N(0, \sigma^2), \quad (2)$$

where f = 1.. s fixed effects; i = 1...r_j observations in block j; and j = 1...b blocks. The observed (failure or censoring) time is t_{ij}; z_{ij1}, ...z_{ijs} are the covariates associated with observation ij; θ_j is the unobserved random effect for block j; and σ^2 is a variance common to the random effects. The cumulative hazard is $\Lambda(\cdot)$, and the unknown parameters associated with fixed effects are $\beta_1...\beta_s$.

Equation (1) is the proportional hazards assumption upon which Cox regression is based, and (2) is a distribution on an unobserved random effect. This is a proportional hazards mixed model (PH-MM) because there are both fixed and random effects.

4.3 Bayesian Analysis of the PH-MM Model

The exposition that follows flows easily using Bayesian language, but the results are available in a form more palatable to frequentists (Robinson, 1991; Searle, Casella, and McCulloch, Chapter 9, 1992). Let

$$\mathcal{L} = -\sum_{i=1}^{r}\sum_{j=1}^{b}\Lambda(t_{ij})\exp(\sum_{f=1}^{s} z_{if}\beta_f + \theta_j) + \sum_{i=1}^{r}\sum_{j=1}^{b}\delta_{ij}\log[\lambda(t_{ij})] + \sum_{i=1}^{r}\sum_{j=1}^{b}\delta_{ij}(\sum_{f=1}^{s} z_{if}\beta_f + \theta_j)$$

and

$$\mathcal{P} = b \cdot \log(\sigma) + [2\sigma^2]^{-1}\sum_{j=1}^{b}\theta_j^2,$$

where δ_{ij} indicates whether the observation is a failure, and $\lambda(t) = \partial\Lambda / \partial t$. Then $\mathcal{H} = \mathcal{L} - \mathcal{P}$ is a Bayesian posterior joint density in the observed values t_{ij}, δ_{ij} and the unobserved θ_j. $\mathcal{L}$ is equivalent to the partial loglikelihood of Cox when we estimate $\Lambda(\cdot)$ using Breslow's cumulative hazard (Breslow, 1974). $\mathcal{H}$ was first introduced in the Cox regression setting by McGilchrist and Aisbett (1991) as a penalized partial

loglikelihood. In this section we will discuss this expression from a Bayesian perspective, and in the next from the perspective of penalized loglikelihood.

Expressions similar to $\mathcal{H}$, but with different loglikelihoods for $\mathcal{L}$, are common in the literature. Henderson (1950) introduced $\mathcal{L}$-$\mathcal{P}$ when $\mathcal{L}$ is a normal loglikelihood, and he noted that it is not a true loglikelihood. Lee and Nelder (1996) discuss $\mathcal{H}$ when $\mathcal{L}$ is a generalized linear model and for a more general $\mathcal{P}$, referring to $\mathcal{H}$ as an *h*-loglikelihood. Correspondingly, we will refer to $\mathcal{H}$ as an *h*-partial loglikelihood.

As with linear models ($\mathcal{L}$ a normal loglikelihood), three strategies are possible for parameter estimation and hypothesis testing (Searle et al., Chapter 9, 1992):

i - If an improper prior, $\beta \sim \text{UNIF}(-\infty, \infty)$, is placed on the parameters associated with fixed effects, and (2) is used as a prior on θ_j, and these elements are integrated out, we get a loglikelihood in σ. This is one derivation of residual maximum likelihood (REML) estimation in linear models.

ii - If (2) is used as a prior for θ_j and these unobserved values are integrated out, we get a joint loglikelihood in β and σ. This leads to the usual maximum loglikelihood (ML) estimation procedure.

iii- A third approach is to maximize β and θ_j directly for some reasonable choice of σ. This approach, dubbed the BLUP (best linear unbiased predictor) approach by some authors, was used by Henderson (1950) for linear models and generates the mixed model equations. This approach was used by Lee and Nelder (1996) for generalized linear models, and applied to Cox regression by McGilchrist and Aisbett (1991). Of the three approaches, only this one avoids integration.

For our purposes, this third approach is especially useful for two reasons: first, because it produces estimates of the random effects (which we will use as residuals for some of the data analysis that follows), and second, because there is a simple implementation of the approach in any statistical programming language that has Cox

regression, "while" loops, and an offset function (for example, S-PLUS). In addition, we avoid matrix inversion.

For linear models this procedure produces best linear unbiased estimators for the fixed effects (BLUEs) and best linear unbiased predictors (BLUPs) for the random effects (Robinson; 1991). Lee and Nelder (1996), considering generalized linear models, found this approach to produce asymptotically best estimators of the fixed and random effects and concluded that the h-loglikelihood is a reasonable surrogate for a true loglikelihood when performing hypothesis tests for the fixed effects.

Our approach is similar to those of McGilchrist (1993) and McGilchrist and Aisbett (1991), but differs in three respects: our specific algorithm is easily coded in standard software, our choice of penalty weight avoids the excessive shrinkage of parameter estimates they reported, and we construct approximate likelihood ratio tests instead of Wald-type tests for the fixed effects.

Since point estimates of random effects have come to be called BLUP's, we will refer to this as the BLUP approach, although this approach is known to produce Best Linear Unbiased Predictors only for linear models.

The rest of this paper is composed of three parts. In the next section we consider the h-partial loglikelihood as a penalized partial loglikelihood and derive some results based on optimization theory. Secondly, we present an algorithm that computes BLUP estimates by embedding Cox regression in a "while" loop. The resulting BLUP estimators are evaluated via simulations. Finally, we use the BLUP approach to analyze a data set from censored survival data with repeated measurements per subject. This data set has fixed effects both between subject and within subject, and so represents a split-plot structure.

4.4 Penalized Partial Loglikelihood

When random effects are estimated as if they were fixed effects, the estimates are too large in absolute value. We know from (2) that the random effects have a mean of zero; this acts as prior information that requires the estimates be shrunken toward the origin compared to fixed effects (Robinson, 1991). A minimal requirement of any estimation procedure is that it should produce this shrinkage. The *h*-partial loglikelihood can be treated as a partial loglikelihood $\mathcal{L}$ and a penalty function $\mathcal{P}$ with a penalty weight of $[2\sigma^2]^{-1}$. In the Appendix we show that the random effects are indeed always shrunk towards the origin when the penalty function $\mathcal{P}$, with any positive penalty weight, is subtracted from $\mathcal{L}$ to form $\mathcal{H}$.

In spite of this shrinkage, the fixed effects are often close in value to the random effects estimates. Suppose $\sum_i \delta_{ik} > 0$ for a block (litter, subject), and that there is at least one failure in that block. Then setting derivatives to zero yields

$$\sum_i \delta_{ik}[\sum_i \Lambda(t_{ik})\exp(z_i\beta)]^{-1} = \exp(\theta_k) \tag{3}$$

for block k in the fixed effects model, and

$$\sum_i \delta_{ik}[\sum_i \Lambda(t_{ik})\exp(z_i\beta)]^{-1} = \exp(\theta_k) + \theta_k[\sigma^2 \sum_i \Lambda(t_{ik})\exp(z_i\beta)]^{-1} \tag{4}$$

for block k in the random effects model. If either the number of observations in the block becomes large [causing $\sum \Lambda(t_{ik})\exp(z_i\beta)$ to become large], or σ^2 is large, the difference between the estimated random and fixed effect is small. From (3), we see that an effect θ_k is not finite when $\sum \delta_{ik} = 0$.

The value σ should not be estimated as part of the optimization process with β and θ_j. Maximizing $\mathcal{H}$, either with the reparametrization $u = \log(\sigma)$ or by using the maximizing condition $\sigma^2 = b^{-1}\sum \theta_j^2$, we see that $\mathcal{H}$ goes to positive infinity as σ goes to zero. A procedure that estimates σ must be chosen carefully to avoid excessive shrinkage of the random effects. We use the fixed effects estimates for initial estimates

of the random effects and σ, and retain this fixed value of σ, denoted σ_{wgt}, throughout the maximization process.

4.5 The Algorithm

The estimation of parameters requires that we solve the system of equations:

$$\frac{\partial \mathcal{H}}{\partial \beta_f} = 0;\ f = 1..s \text{ fixed effects and } \frac{\partial \mathcal{H}}{\partial \theta_j} = 0;\ j = 1..b \text{ random effects.} \qquad (5)$$

Hypothesis tests for the fixed effects are then formed by either

$$T_f = 2\{\mathcal{H}(\hat{\beta}, \sigma^2_{wgt}, \hat{\theta}) - \mathcal{H}(\hat{\beta}_o, \sigma^2_{wgt}, \hat{\theta}_o)\} \quad \text{or}$$

$$T_a = T_f + 2\{A(\hat{\beta}) - A(\hat{\beta}_o)\}$$

$$\text{where} \qquad A(\beta^*) = \tfrac{1}{2}\sum \log\left(\frac{\partial^2 \mathcal{H}}{\partial \theta_k^2}\Big|_{\beta=\beta^*} \right).$$

The fixed value σ^2_{wgt} is used to form the penalty weight $[2\ \sigma^2_{wgt}]^{-1}$. $\hat{\beta}$ and $\hat{\theta}$ are the solutions to (5) under the alternative hypothesis; and $\hat{\beta}_o$ and $\hat{\theta}_o$ are the solutions to (5) under the null hypothesis. T_f is an (unadjusted) h-partial loglikelihood ratio test, and T_a is an adjusted h-partial loglikelihood ratio test. Both T_f and T_a have been suggested as reasonable hypothesis tests in the generalized linear models setting (Lee and Nelder, 1996), and we will consider them both.

A naive application of Newton's method to solve (5) would require repeated inversion of a b+p rank matrix, where p is the number of fixed effects and b is the number of blocks. We will present a method that requires no matrix inversion. The maximization of $\mathcal{H}$ can be decomposed into two steps: finding estimates of θ given values of β and $\Lambda(\cdot)$, and finding values of β and $\Lambda(\cdot)$ given values of θ. We should also maintain the constraint $\sum \theta_j = 0$.

Our algorithm consists of the following steps:

Step 0: Get initial estimates of the β, $\Lambda(\cdot)$, θ, and σ^2_{wgt}.

Step 1: Use Newton's method to estimate θ, given estimates of β and $\Lambda(\cdot)$, based on condition (4).

Step 2: Center the random effects so that $\sum\theta_j = 0$.

Step 3: Use Cox regression with the offset command to estimate β and $\Lambda(\cdot)$ for the current θ.

Repeat steps 1, 2, and 3 until convergence.

Step 1 appears to require matrix inversion, but the resulting matrix of second derivatives is diagonal. Let

$$g_i = \frac{\partial \mathcal{H}}{\partial \theta_i} \quad \text{and} \quad q_i = -[\frac{\partial^2 \mathcal{H}}{\partial \theta_i^2}]^{-1},$$

and let g be the vector of g_i's and let D be a diagonal matrix with diagonal elements q_i; then the Newton's method estimate in step 1 is Dg. Compared to the usual matrix inversion, this results in a faster algorithm with fewer convergence difficulties. Step 2 is possible because the mean of the random effects is not identifiable, so centering the random effects just alters the baseline hazard by a multiplicative factor. Newton's method is required in step 3, but this is embedded in professionally coded software and is hidden to the user.

4.6 Initial Values

In principle, initial values can be found by solving the Cox regression with the random effects assumed fixed. Two problems arise, however, especially when there are many blocks and few observations per block. If all observations in a block are censored, then the related fixed effect is not finite (discussed above). Further, the large number of

parameters relative to the number of observations sometimes causes convergence problems for Cox regression.

Initial values were found as follows:

Step1: Perform Cox regression on the fixed effects only.

Step 2: For each block, form the quantities $\sum\delta_{ik}$ and $\sum\Lambda(t_{ik})\exp(\sum z_{ikf}\beta_f)$.

Use (3) to get an initial estimate of each θ_k. If any $\sum\delta_{ik} = 0$, replace $\sum\delta_{ik}$ with

$(1- \alpha)\sum\delta_{ik} + \alpha \cdot \text{mean}(\delta)$, where α is some small positive number.

Step 3: Create a covariate vector with a value of θ_k found in step 2 assigned to each observation in block k, for all blocks.

Step 4: Perform Cox regression with the original covariate z and a new covariate vector created in step 3. Update θ_k by multiplying the current value by the estimated regression coefficient from this Cox regression.

This algorithm generates the initial values of β and θ for the previous algorithm and the penalty weight, $[2\ \sigma^2_{wgt}]^{-1}$, to be used throughout the maximization process.

This algorithm only produces approximate solutions to the fixed effects model. Because these are only initial values for the previous algorithm, we are not concerned with exact solutions. Step 2 addresses the problem of unbounded fixed effects when all observations in a block are censored. Steps 3 and 4 reduce the dimension of the parameter space for Cox regression from p + b down to p + 1. This avoids the convergence problems mentioned above.

The best choice for a penalty weight would presumably be the unknown true population variance. Because we are essentially using fixed effects in place of random effects at this point, we expect that this estimate of variance is usually too large, and the shrinkage of the random effects too little -- a result different from what others have found (McGilchrist and Aisbett,1991).

The penalty weight, $[2\ \sigma^2_{wgt}]^{-1}$, is not changed during the optimization process. Only after final estimates of β and θ are found is a final estimate of variance made: $\hat{\sigma}^2 = [b-1]^{-1}\sum\hat{\theta}_j^2$, where $\hat{\theta}_j$ are the final estimated random effects.

4.7 Evaluating the Algorithm via Simulation

In a split-plot design there is at least one treatment (covariate) that varies from block to block and at least one treatment (covariate) that varies within a block. For example, we might give subjects two different medications for an illness at two different times, but we might also be interested in the gender of the subject. In this case the medication varies within subject , but gender varies between subjects. Estimated random effects are especially useful for split-plot designs, as they are the residuals for blocks. An unusually large estimated random effect suggests an outlier, i.e. a block (subject, litter, etc.) that exhibits an unusual response to the treatment (or a possible error in the data).

Several data sets used as examples in survival analysis share common characteristics: a split-plot structure, with few subjects (10 to 50), two or three observations per subject, and mild censoring. The design is not completely balanced -- the number of observations per subject varies slightly. Three such data sets are the skin graft data from Chapter 8 of Kalbfleisch and Prentice (1980), the catheter data used by McGilchrist and Aisbett (1991), and the HIV data used by Lipsitz and Parzen (1996). We evaluated our algorithm by simulating data with these characteristics.

There were two observations simulated within each block, a treatment and a control for some factor W. Half the blocks received the treatment and the other half the control for some factor B. In terms of the model specified in (1) and (2), we let $r_j = 2$ (two observations per block) , $s = 2$ (two fixed effects, one between subjects and the other

within subjects), $z_{ijf} = 0$ or 1 (an indicator that distinguishes treatments from controls), b = 12 or 36 (the number of blocks), and $\sigma^2 = 2$. There was mild censoring (about 10%).

The failure times and censoring times were exponential. Each fixed effect had two possible values: 0 or -log(2). Simulations were run for all four possible combinations of fixed effects values. The initial value of σ^2_{wgt}, the estimated random effects, the estimated fixed effects, and the final estimate of σ^2 were all taken from a full model where both fixed effects were estimated. For hypothesis tests of the fixed effects, reduced models were fitted without the effect of interest.

From Table 4.1 we see that the adjusted and unadjusted (in parentheses) *h*-partial loglikelihood produced nearly identical rejection rates. Subsequent discussion will focus on the rejection rates based on an adjusted *h*-partial loglikelihood only.

There are few surprises in the rejection rates presented in Table 4.1. In the cases where the rejection rates should be 5% (in italics), they are very close, with two exceptions. The rejection rate for the 36-block case when there is a within-block effect, but no between-block effect, seems low (2.8%). The rejection rate is also too high (7.6%) in one case. The sample sizes for the within-block comparison were 24 and 72, and the sample sizes for the between-block comparison were 12 and 36.

Although rejection rates of about 5% when the null hypothesis is true are not unexpected, there are several reasons this result might not have occurred. First, the claim that any loglikelihood ratio test is approximately χ^2 is based on asymptotic theory, while our sample sizes are relatively small, especially for the between-block effect. Secondly, the adjusted *h*-loglikelihood ratio test is only an approximation of a loglikelihood ratio test (Lee and Nelder, 1996). Finally, our test, which is an *h*-partial loglikelihood ratio test, is the extension of this approach to a new setting rather than a special case of the work of Lee and Nelder.

Table 4.1. Hypothesis testing for fixed effects. The rejection rates are for an adjusted *h*-partial loglikelihood ratio test (rejection rates for the *h*-partial loglikelihood ratio test, unadjusted, are in parentheses) where the full model includes both parameters and the reduced model excludes a parameter. Rejection rates when the null hypothesis (null) is true are in italics and should be 0.05, and those in bold are for when the alternative hypothesis (altn) is true and should be greater than 0.05. All rejection rates are based on 500 simulated samples.

β_W	β_B	H_0: $\beta_W = 0$	H_0: $\beta_B = 0$
		Rejection Rates	
12 blocks			
null	null	*0.076 (0.080)*	*0.048 (0.048)*
null	altn	*0.052 (0.052)*	**0.106 (0.100)**
altn	null	**0.262 (0.268)**	*0.056 (0.054)*
altn	altn	**0.244(0.244)**	**0.104 (0.098)**
36 blocks			
null	null	*0.046 (0.046)*	*0.052 (0.050)*
null	altn	*0.050 (0.050)*	**0.222 (0.208)**
altn	null	**0.650 (0.652)**	*0.028 (0.028)*
altn	altn	**0.610 (0.610)**	**0.240 (0.234)**

When the alternative hypothesis is true, the rejection rates in Table 4.1 should be greater than 5%. Further, the hypothesis tests should be more powerful when the sample size is increased, and the test based on comparisons within a block should be more powerful than the test based on comparisons between blocks.

For the test based on within-block comparisons, H_0: $\beta_W = 0$, the rejection rates increased from about 25% to about 63% as the sample size tripled. For the between-block comparison, H_0: $\beta_B = 0$, the rates increased from about 10% to about 22%. So, increased sample sizes produce dramatic increases in the rejection rates.

These results also indicate that within-block comparisons are more precise than between-block comparisons. In all cases, the same alternative hypothesis was used

(β = -log2). For the smaller sample size, the within-block comparison had a rejection rate of 25%, while the between-block comparison had a rejection rate of 10%. In the data with larger sample sizes, the within-block comparison yielded a 63% rejection rate, while the between-block comparison yielded a 22% rejection rate. This is the reason blocking is used in designed experiments.

Efficient hypothesis tests with correct significance levels suggest the BLUP approach is useful, but it is also desirable to obtain reasonable point estimates of the fixed and random effects.

Table 4.2 Estimated fixed effects. Below are median, 0.05, and 0.95 quantiles for the estimated fixed effects parameters for 500 simulated samples. When the null hypothesis is true (null) the numbers are in italics, and they should be centered around zero. When the alternative hypothesis is true (altn), the numbers are in bold, and they should be centered around -log2 ($\approx$ -0.693).

β_W	β_B	$\hat{\beta}_W$			$\hat{\beta}_B$		
		lower	median	upper	lower	median	upper
12 blocks							
null	null	*-1.137*	*-0.004*	*1.343*	*-2.037*	*0.009*	*2.151*
null	altn	*-1.168*	*-0.010*	*1.150*	**-3.003**	**-0.817**	**1.121**
altn	null	**-2.294**	**-0.920**	**-0.266**	*-2.088*	*0.132*	*2.279*
altn	altn	**-2.621**	**-0.901**	**0.228**	**-3.114**	**-0.889**	**1.088**
36 blocks							
null	null	*-0.592*	*0.007*	*0.631*	*-1.108*	*-0.019*	*1.047*
null	altn	*-0.646*	*0.000*	*0.618*	**-1.874**	**-0.764**	**0.375**
altn	null	**-1.617**	**-0.895**	**-0.208**	*-1.010*	*-0.009*	*1.017*
altn	altn	**-1.552**	**-0.870**	**-0.228**	**-1.996**	**-0.802**	**0.335**

The earlier hypothesis test results indicate that within-block tests are more precise than between-block tests, and that larger sample sizes yield more precise tests.

Correspondingly, we expect the narrowest ranges for our parameter estimates in those cases where hypothesis tests are most precise. This principle explains the relative widths of the ranges for the fixed effects estimates in Table 4.2. The estimates for β_B and for β_W have similar medians, but β_B has a wider range, i.e., the betweenn-block comparison is less precise than the within-block comparison. Further, increasing sample sizes increases precision, so parameter estimates using 36 blocks have similar medians to those using 12 blocks, but narrower ranges.

It is harder to generalize about the median value of the parameter estimates. When the correct value is zero, the median value is close to zero in all cases. When the correct value is -log(2), the median values are consistently too large in magnitude, although there does appear to be slight improvement with increased sample size.

It is sometimes useful to estimate the random effects as well. The comparisons between the estimated random effects and the actual random effects in each simulated data set were made in two ways: we hope both that the variance of the estimated random effects be comparable to that of the actual random effects, and that the estimated random effect closely matches the actual random effect for each block.

Table 4.3 summarizes the estimation of random effects. The first three columns characterize the range of the sample correlation between the estimated random effects and actual random effects. The last three columns summarize the ranges of the ratio of the sample variance of the estimated random effects to the sample variance of the actual random effects.

The BLUP approach, as implemented in this paper, consistently overestimates the sample variance of the random effects. The ratio of our sample variance to the correct sample variance has a wide range for small sample sizes (about 0.459 to about 6.002) and a much smaller range for larger samples sizes (about 0.896 to 3.062), but the

median value is always about 1.60, suggesting a median inflation of our sample variance of about 60%, independent of sample size.

For each data set, the sample correlation between the estimated and actual random effects was calculated. The range of the sample correlation is contained in the first three columns of Table 4.3. The correlation between the actual and estimated random effects is wide ranging for the smaller sample sizes (0.554 to 0.931), but has a much smaller range for larger sample sizes (0.730 to 0.903). The median correlation increases with sample size, from 0.81 to 0.84.

Table 4.3 Estimated random effects. The first three columns characterize the range of the sample correlation between the estimated random effects and the simulated random effects; the second set of columns characterizes the ratio of sample variance for the estimated random effects to the sample variance for the actual random effects. The median, 0.05, and 0.95 quantiles for are given in each case. Each row was based on 500 simulated samples.

β_W	β_B	r			$\hat{\sigma}^2/s^2$		
		lower	median	upper	lower	median	upper
12 blocks							
null	null	0.545	0.816	0.934	0.465	1.626	5.400
null	altn	0.558	0.811	0.930	0.452	1.562	6.278
altn	null	0.565	0.815	0.933	0.464	1.621	6.761
altn	altn	0.549	0.806	0.928	0.453	1.537	5.568
36 blocks							
null	null	0.740	0.844	0.909	0.846	1.650	2.936
null	altn	0.735	0.838	0.901	0.937	1.635	3.067
altn	null	0.735	0.838	0.901	0.909	1.648	3.095
altn	altn	0.709	0.827	0.901	0.891	1.600	3.149

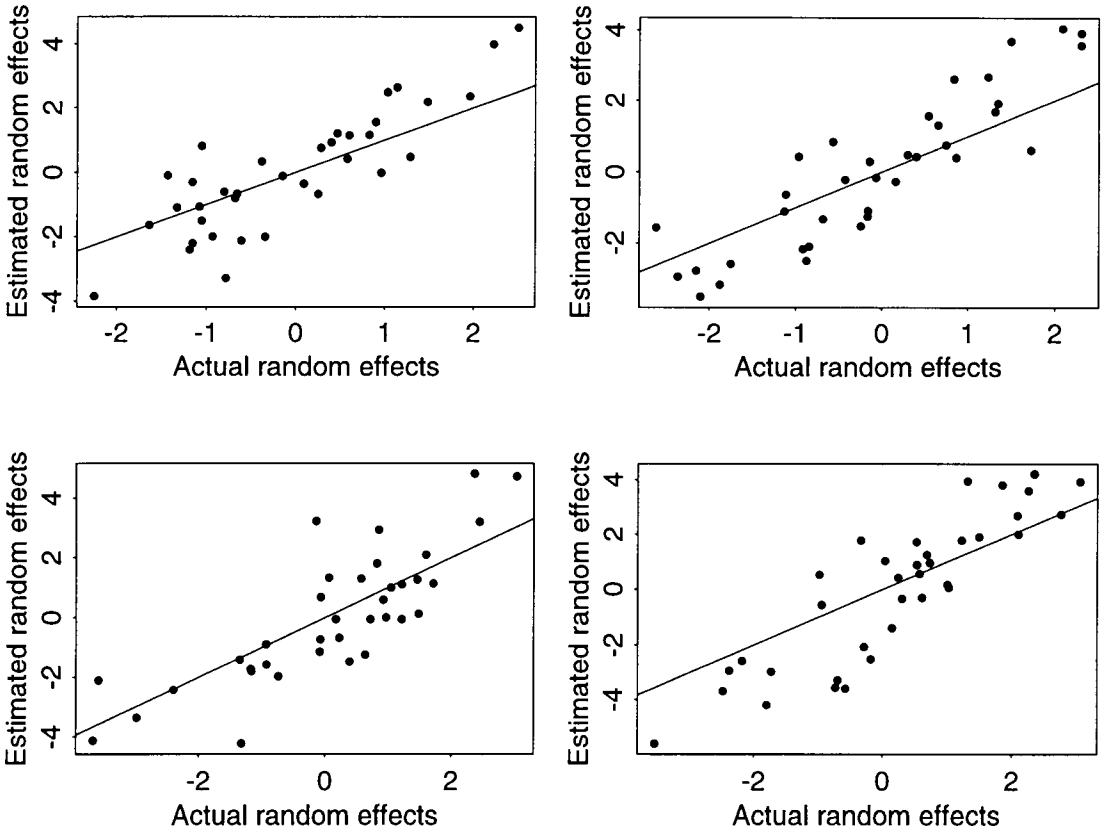

Figure 4.1. Four scatterplots of actual and estimated random effects for data sets with 36 blocks. In each scatterplot a 45° line has been added.

Figure 4.1 contains scatterplots of actual versus estimated random effects for four simulated data sets and shows the close correspondence between actual and estimated values. Although our method overestimates the overall variance between subjects, Figure 4.1 and the range of the sample correlations suggest that the random effects are being correctly estimated except for a scaling factor: the largest estimated random effects are associated with the largest actual random effects, estimated random effects near zero are associated with actual random effects near zero, and large negative

estimated random effects are associated with large negative actual random effects. It is this characteristic that makes them useful for outlier detection.

Previous authors found that the BLUP method yields estimates of the fixed and random effects that are shrunk too much (McGilchrist and Aisbett, 1991). Because of the way the initial estimate of variance is obtained, our implementation induces substantially less shrinkage. Neither our implementation nor the previous efforts induce the correct amount of shrinkage, and the bias resulting from our approach inflates parameter estimates. This bias seems relatively minor for parameter estimates of the fixed effects, but sizable for the sample variance. Strangely, the bias does not affect hypothesis tests.

4.8 An Example

Lipsitz and Parzen (1996) discuss a data set in which 36 HIV positive subjects were given one of three treatments (a placebo, or a low or high dosage of the drug ribavirin). Each subject had blood samples taken at 4, 8, and 12 weeks. The failure times are the times until the HIV virus can be detected in each blood sample; failure times are inversely related to disease severity (see Table 4.4).

The authors used Cox regression, treating time (weeks 4, 8, and 12) as a covariate. This analysis gives a reasonable point estimate of the parameters of interest but does not take the correlation between repeated measures into account. They solved this problem by using a jackknife estimate of the standard error of the parameter estimates. The authors modeled the low and high dosage of ribavirin with two indicator variables.

Table 4.4. The data for 36 HIV positive subjects, taken from Lipsitz and Parzen. For weeks 4, 8, and 12 the numbers are the days until the HIV virus was detected in the blood sample. Censored times are marked (*).

Treatment	Subject	Week 4	Week 8	Week 12	estimated random effect
Placebo	1	9	6	6	0.47
	2	4	5	10	0.87
	3	6	7	6	1.07
	4	10	--	21*	-1.35
	5	15	8	--	-0.62
	6	3	--	6	1.95
	7	4	7	3	1.38
	8	9	12	12	-0.38
	9	9	19*	19*	-1.42
	10	6	5	6	1.50
	11	9	--	18	-0.60
	12	9	20*	17*	-1.82
Low dose	13	6	4	5	2.97
	14	16	17	21*	-0.30
	15	31	19*	21*	-2.02
	16	27*	19*	--	-1.68
	17	7	16	23*	-0.05
	18	28*	7	19*	-0.82
	19	28*	3	16	-0.23
	20	15	12	16	0.52
	21	18	21*	22	-0.58
	22	8	4	7	2.15
	23	4	21*	7	0.07
High dose	24	21	9	8*	-0.67
	25	13	7	21*	-0.44
	26	16	6	20	-0.18
	27	3	8	6	1.66
	28	21	--	25*	-1.36
	29	7	19	3	-0.02
	30	11	13	21*	-0.59
	31	27*	18*	9	-1.54
	32	14	14	6	0.09
	33	8	11	15	0.34
	34	8	4	7	1.50
	35	8	3	9	1.30
	36	19*	10	17	-1.20

An alternative approach is suggested by applied linear models. The blood samples for a specific subject are repeated measures, and we will assume these measurements are independent except that they share a common random effect. Since there is a within-subject time effect and a between-subject treatment effect, this is a split-plot design (Mead, Chapter 14, 1988). We treated time as a categorical variable.

Several hypothesis tests were performed in the course of analyzing the data. All the fitted models are displayed in Table 4.5. The initial estimate of σ^2 was taken from the model labeled "Wk", but the initial estimates from the other models are similar and we assume our results do not depend on this choice.

Table 4.5. The models used in the analysis of the HIV data. The model label is given as well as the factors in the model, the number of regression parameters needed, and the initial estimated sample variance.

Model	Factors	Parameters	Initial $\hat{\sigma}^2$
I-a	low dosage, high dosage, time, low dosage × time.	6	2.284
I-b	low dosage, high dosage, time, high dosage × time.	6	2.351
Wk	low dosage, high dosage, time	4	2.297
Tr	low dosage, high dosage	2	2.177
Ld	low dosage, time	3	2.459
Hd	high dosage, time	3	2.730

Two interaction models (I-a and I-b) were used to test for any interaction effects. There was no evidence of interaction, and these will not be discussed further.

From Table 4.6, we see that our results are similar to the results of Lipsitz and Parzen. The parameter estimates are close, but our p-values are smaller. In each case, the conclusion is the same. Low doses of ribavirin are effective, but high doses are not.

Table 4.6. (a) An analysis of the HIV data in a split-plot format. (b) Comparison of the Lipsitz and Parzen parameter estimates and hypothesis test to our results. Lipsitz and Parzen computed a jackknife standard error and their p-values are based on the approximate normality of a Wald test using this standard error. Our p-values are based on the h-partial loglikelihood being approximately χ^2 with the degrees of freedom listed.

(a)

	full model	reduced model	χ^2_{df}	df	p-value
Between subject					
Low dosage	Wk	Hd	8.12	1	0.004
High dosage	Wk	Ld	3.10	1	0.078
Within subject					
Time	Wk	Tr	3.90	2	0.142

(b)

	Lipsitz & Parzen		BLUP	
	parameter estimates	p-value	parameter estimates	p-value
Low dosage	-0.8636	0.031	-0.727	0.004
High dosage	-0.4024	0.222	-0.370	0.078

Our p-values are smaller that those of Lipsitz and Parzen and there are at least two possible sources of the difference: First, the underlying models presumed in the two analyses are quite different, and may not always give identical results. Second, it is possible that our model, which is arguably "more parametric," might be noticeably more

efficient for this small data set, as there are only 36 observations for the comparisons involving ribavirin dosage.

The estimated random effects for the subjects are presented in Figure 4.2. Subject 13 in Table 4.4 may be an outlier. This subject is in the "best" treatment group, but has three of the worst failure times in the data set. This is noteworthy, and was made obvious by our estimation method, but it does not affect the analysis above. Deleting this subject and re-analyzing the data gave almost identical results for all the hypothesis tests and parameter estimates.

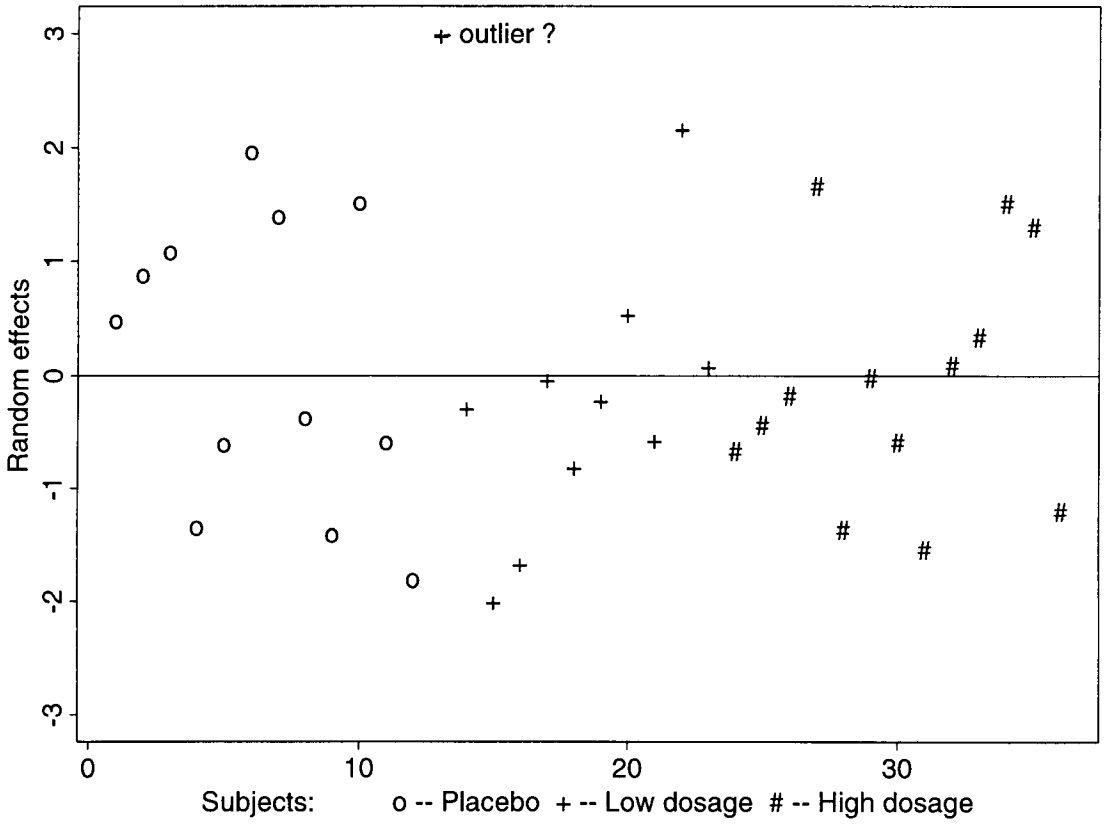

Figure 4.2. Estimated random effects for each subject for the HIV data set, with subjects ordered as in Table 4.4.

The fact that all the predicted random effects may be too large in magnitude does not affect outlier detection, because here we are only interested in whether any one is large compared to the others (scaling is irrelevant).

4.9 Conclusions

With the BLUP estimation algorithm presented here, it is possible to analyze survival data from experiments with blocking factors, including many randomized block designs and split-plots. Our algorithm can easily be implemented in statistical languages (i.e., S-PLUS) as an extension of Cox regression. This method allows for hypothesis tests for the fixed effects, estimates of the fixed effects, and the estimates of random effects (up to a scaling factor), which are useful for detection of blocks (subjects, litters, etc.) that are potential outliers at the between-subject level.

Our parameter estimates were slightly inflated in magnitude, almost certainly because the initial estimate of σ^2 is too large. It is possible that a better method of finding initial values, or a wiser method of selecting σ^2_{wgt}, would lead to better parameter estimates. Nevertheless, the hypothesis tests for fixed effects performed well, the estimated random effects appeared useful for detecting outliers, and the estimated fixed effects were reasonable, although slightly inflated.

4.10 References

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.

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

HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* 21, 309-310.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B* 58, 619-678.

LIPSITZ, S. R. and PARZEN, M. (1996). A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics* 52, 291-298.

McGILCHRIST, C. A. and AISBETT, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* 47, 461-466.

McGILCHRIST, C. A. (1993). REML estimation for survival models with frailty. *Biometrics* 49, 221-225.

MEAD, R. (1988). *The Design of Experiments.* Cambridge: Cambridge University Press.

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science.* 6, 15-51.

SEARLE, S. R., CASELLA, G. and McCULLOCH, C. E. (1992). *Variance Components.* New York: Wiley.

4.11 Appendix

Part 1 - The estimated random effects are always shrunk toward the origin when estimated using a penalty function compared to the random effects estimated as if they were fixed, i.e., $\sum\theta_{FE}^2 > \sum\theta_{RE}^2$.

Fixed effects are estimated by maximization of $f(\beta, \theta)$, and random effects are estimated by maximization of $f(\beta, \theta) - w(\sum\theta_j^2)$. Denote fixed effects solutions as (β_{FE},θ_{FE}) and random effects solutions as (β_{RE},θ_{RE}); then $f(\beta_{FE}, \theta_{FE}) > f(\beta_{RE}, \theta_{RE})$ and $f(\beta_{RE}, \theta_{RE}) - w(\sum\theta_{RE,j}^2) > f(\beta_{FE}, \theta_{FE}) - w(\sum\theta_{FE,j}^2)$. Add both sides of the preceding inequalities and drop common terms yields $-w(\sum\theta_{RE,j}^2) > -w(\sum\theta_{FE,j}^2)$. If we let $w = [2\sigma^2]^{-1} > 0$, then $\sum\theta_{FE,j}^2 > \sum\theta_{RE,j}^2$.

Part - 2. If common estimates of $\Lambda(\cdot)$ *and* β *are given and there is at least one failure in each block, random effects estimates are shrunk individually compared to the fixed effects estimates for that block.*

Suppose $\sum_i \delta_{ik} > 0$ for each block (litter, subject), i.e., that there is at least one failure in each block; let $y_k = \sum_i \delta_{ik}[\sum_i \Lambda(t_{ik})\exp(\sum_f z_{if}\beta_f)]^{-1}$ and $c_k = [\sigma^2\sum_i \Lambda(t_{ik})\exp(\sum_f z_{if}\beta_f)]^{-1}$ be a fixed non-negative constant for each block k. Then setting derivatives to zero will yield $y_k = \exp(\theta_{FE,k})$ (1) for each block k in the fixed effects model and

$y_k - c_k\theta_{RE,k} = \exp(\theta_{RE,k})$ (2) for the random effects model.

For both (1) and (2), $\frac{dy_k}{d\theta_k} > 0$ and y_k is continuous in θ_k. Also $\theta_k = 0$ when $y_k = 1$ in both cases, so $\theta_{FE,k}$ and $\theta_{RE,k}$ always agree in sign. If $\theta_k < 0$, then $0 > \theta_{RE,k} > \theta_{FE,k}$. Similarly, when $\theta_k > 0$, $0 < \theta_{RE,k} < \theta_{FE,k}$. So $\theta_{RE,k} = \theta_{FE,k} = 0$ when $\sum_i \delta_{ik} = \sum_i \Lambda(t_{ik})\exp(\sum_f z_{if}\beta_f)$; otherwise $|\theta_{RE,k}| < |\theta_{FE,k}|$.

The above demonstrates that for any common estimate of $\Lambda(\cdot)$ and β, the random effects estimates are shrunk individually, compared to the fixed effects estimates for that block. In practice, there may be exceptions because the fixed and random effects models generate slightly different estimates of $\Lambda(\cdot)$ and β.

Chapter 5
5. Conclusions

5.1 The Incompletely Specified PH Loglikelihood

We have, shown in a variety of settings, that an incompletely specified proportional hazards loglikelihood,

$$-\sum_{i=1}^{n} \Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n} \delta_i \log\{\lambda(t_i)\exp(z_i\beta)\}, \tag{1}$$

is more useful than the partial loglikelihood,

$$\sum_{i=1}^{n} \delta_i [z_i\beta - \log\{\sum_{j=i}^{n} \exp(z_j\beta)\}], \tag{2}$$

originally introduced by Cox. The loglikelihood (1) is incomplete in that it is not a loglikelihood (or partial loglikelihood) until $\Lambda(\cdot)$ and $\lambda(\cdot)$ are specified.

In principle, many of the results we found could have been derived using only the partial loglikelihood. Nevertheless, many results were more easily obtained using (1). It was particularly beneficial to use (1) instead of (2) when forming partial derivatives that have intuitive interpretations.

The main difference between the partial loglikelihood and the incompletely specified PH loglikelihood, is that the hazard function is completely removed from the partial loglikelihood. Deviance or martingale residuals, however, require that a cumulative hazard be specified. Deriving residuals using the partial loglikelihood involves the rather artificial process of removing the nuisance parameter at one step, only to re-introduce it a few steps later. Nor is it clear that the Breslow cumulative hazard, implicit in the partial loglikelihood, is always the best choice.

5.2 Implications for Further Study

In Chapter 2, we found that the deviance residuals used in Cox regression fail to detect certain obvious outliers. Although it was easy to identify the problem, a solution is difficult. Our partial solution was to cluster neighboring observations to form a smoothed baseline hazard. For early outliers, the assumption of a locally constant baseline hazard was too crude and introduced a bias in the deviance residual that depends on whether the true underlying hazard is monotone increasing or decreasing. This bias disappears with larger sample sizes, but rather slowly. Late outliers, on the other hand, are influential in the calculation of the estimated parameters $\Lambda(\cdot)$ and β, introducing a bias that carries over to the deviance residual itself. As with the early outlier, the bias disappears very slowly with increased sample size.

The first problem, associated with early outliers, could probably be solved by using a kernel smoother or lowess smoother to smooth the baseline hazard, which would give the benefits of clustering while producing a smoother function. The latter problem, associated with late outliers, might be solved by sequentially deleting each observation when computing the deviance residual for that observation.

In Chapter 3, we found deviance residuals that reveal the shape of the underlying baseline hazard. Ignoring censored data, these are equivalent to:

$$d_i = \text{sign}(h_i - 1)\{2[h_i - 1 - \log(h_i)]\}^{\frac{1}{2}}, \tag{3}$$

where h_i is the ratio of the observed and predicted (maximum likelihood) local baseline hazard rate. When the predicted failure rate is constant, this suggests a one-to-one mapping between the deviance residuals and the baseline hazard function. An efficient way to find a smoothed baseline hazard would be to compute deviance residuals, smooth the deviance residuals, then numerically invert (3) back to a baseline hazard.

5.3 Constrained Optimization

There are several other areas where use of (1) in place of (2) might be useful. In Chapter 3, a non-parametric monotone baseline hazard was found. For ordered observed times $t_1 \ldots t_n$, the optimization problem we solved was:

Maximize $$-\sum_{i=1}^{n} \Lambda(t_i)\exp(z_i\beta) + \sum_{i=1}^{n} \delta_i \log\{\lambda_i \exp(z_i\beta)\}$$

subject to $$\Lambda(t_i) = \sum_{j=1}^{n} \lambda_j \Delta_j \quad \text{and} \quad \lambda_m \leq \lambda_k \quad \text{for all } t_m < t_k.$$

This general class of optimization problems produces a non-parametric step function for the baseline hazard, but it can be used to place other constraints on the baseline hazard. Two obvious examples are a U-shaped baseline hazard (common in reliability applications) and a baseline hazard smoothed to limit variation.

5.4 Hierarchical Models

In Chapter 4, we developed a Bayesian hierarchical model to create the random effects model needed for a split-plot design. Many interesting designed experiments include random effects or repeated measures, so the approach we used should be generally applicable in these cases.

Bayesian methods should directly apply to (1) because it is a loglikelihood. A typical case in which (1) is far more useful then (2) is when the prior information is on the hazard function, so the hazard must be explicitly modeled.

A hierarchical model that is different from the random effects model is the measurement error model, in this model covariates are random rather than fixed. This kind of problem can be attacked using (1).

5.5 Grouped and Tied Data

Grouped data and data with ties are quite common in survival analysis. Methods for dealing with such data are "ad hoc". When a loglikelihood is available, tied and grouped data present no special problems. Using (1), it is possible to directly derive the correct loglikelihood in these cases.

5.6 Summary

In many applications of the proportional hazards semi-parametric model a cumulative hazard must be specified, e. g., when we compute median failure times, compute residuals, or plot a baseline hazard. It is artificial to remove the cumulative hazard, derive formulas, and then re-introduce a cumulative hazard. But this is required when using (2) in place of (1). Avoiding this awkwardness produces derivations that are simpler and often intuitively meaningful.

Using (2) in place of (1) also presupposes that one particular non-parametric hazard is best. There is no unique, correct non-parametric hazard function; but using (2) implicitly commits the analyst to the Breslow cumulative hazard. Using (1), the analyst can specify the cumulative hazard as she wishes.

BIBLIOGRAPHY

BALTAZAR-ABAN, I. and PENA, E. A. (1995). Properties of hazard-based residuals and implications in model diagnostics. *Journal of the American Statistical Association* 90, 185-97.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.

CHUNG, D. and CHING, M. N. (1994). An isotonic estimator of the baseline hazard function in Cox regression model under order restrictions. *Statistics and Probability Letters* 21, 223-228.

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

DANIEL, W. W. (1990). *Applied Nonparametric Statistics*. Boston: PWS-KENT.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

GROSH, D. L. (1989). *A Primer of Reliability Theory*. New York: Wiley.

HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* 21, 309-310.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* 60, 267-78.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

KOOPERBERG, C., STONE, C.J., and TROUNG, Y.K. (1995). Hazard regression. *Journal of the American Statistical Association* 90, 78-94.

LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B* 58, 619-678.

LIPSITZ, S. R. and PARZEN, M. (1996). A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics* 52, 291-298.

BIBLIOGRAPHY (Continued)

McCULLAGH, P. and NELDER, J. A. (1994). *Generalized Linear Models.* Cambridge: University Press.

McGILCHRIST, C. A. and AISBETT, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* 47, 461-466.

McGILCHRIST, C. A. (1993). REML estimation for survival models with frailty. *Biometrics* 49, 221-225.

MEAD, R. (1988). *The Design of Experiments.* Cambridge: Cambridge University Press.

OAKES, D. (1972). Contribution to the discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* 34, 187-220.

PADGETT, W. J. and WEI, L. J. (1980). Maximum likelihood estimation of a distribution function with increasing failure rate based on censored observations. *Biometrika* 67, 470-4.

PIERCE, D. A. and SCHAFER, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* 81, 977-86.

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science.* 6, 15-51.

SEARLE, S. R., CASELLA, G. and McCULLOCH, C. E. (1992). *Variance Components.* New York: Wiley.

STANISWALIS, J. G. (1989). The kernel estimation of a regression function in likelihood based models. *Journal of the American Statistical Association* 84, 276-83.

THERNEAU, T. M. and GRAMBSCH, P. A. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147-60.

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus.* New York: Springer-Verlag.