SURVIVAL ANALYSIS FROM THE VIEWPOINT OF HAMPPEL'S THEORY FOR ROBUST ESTIMATION

by

Steven J. Samuels

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1163

JUNE 1978
SURVIVAL ANALYSIS FROM THE VIEWPOINT OF HAMPEL'S THEORY FOR ROBUST ESTIMATION

by

Steven J. Samuels

Department of Biostatistics
School of Public Health
University of North Carolina at Chapel Hill, NC 27413
# TABLE OF CONTENTS

**LIST OF FIGURES**

**LIST OF TABLES**

**Chapter**

1. **INTRODUCTION** .................................................. 1  
   1.1 Survival Background, Parameterization ....................... 1  
   1.2 Description of Contents .................................. 3  

2. **IDEAS FROM THE THEORY OF ROBUST ESTIMATION** ............ 5  
   2.1 Estimators as Functionals ................................. 5  
   2.2 Fisher Consistency .............................. 7  
   2.3 Robustness, the Influence Curve ...................... 9  
      The Influence Curve .................................... 10  
      Theoretical Considerations ............................ 13  
      Use of the I.C. with Outliers .......................... 15  
      Assessing Robustness with the I.C. .................... 15  
      Interpretation of the I.C. in Regression ............... 16  

3. **EXTENSION OF HAMPEL'S FRAMEWORK TO SURVIVAL DATA** .... 19  
   3.1 Introduction ............................................. 19  
   3.2 The Random Censorship Model with Covariates ............ 20  
   3.3 Related Work ............................................ 24  
   3.4 Independence of Censoring and Survival .................. 26  
   3.5 Limitations of Hampel's Approach ....................... 28  

4. **THE ONE PARAMETER EXPONENTIAL MODEL** ..................... 30  
   4.1 The Model ................................................. 30  
      The Estimator ........................................... 30  
      Functional Form ....................................... 31  
      Fisher Consistency ................................... 31  
   4.2 The Influence Curve ..................................... 32  
      Interpretation ......................................... 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3 The Limiting Variance and Distribution of $\hat{\lambda}_n$</td>
<td>35</td>
</tr>
<tr>
<td>4.4 Other Functions of $\lambda$</td>
<td>36</td>
</tr>
<tr>
<td>5. NONPARAMETRIC ESTIMATION OF AN ARBITRARY SURVIVAL FUNCTION</td>
<td>39</td>
</tr>
<tr>
<td>5.1 Note on Rank Estimators</td>
<td>39</td>
</tr>
<tr>
<td>5.2 The Product Limit and Empirical Hazard Estimators</td>
<td>39</td>
</tr>
<tr>
<td>5.3 Functional Form: Fisher Consistency Under Random Censoring</td>
<td>41</td>
</tr>
<tr>
<td>5.4 The Influence Curve</td>
<td>42</td>
</tr>
<tr>
<td>5.5 Discussion</td>
<td>45</td>
</tr>
<tr>
<td>5.6 The Squared I.C.</td>
<td>47</td>
</tr>
<tr>
<td>6. AN EXPONENTIAL REGRESSION MODEL</td>
<td>52</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>52</td>
</tr>
<tr>
<td>6.2 Likelihood Equations</td>
<td>52</td>
</tr>
<tr>
<td>6.3 The Estimator as a Von Mises Functional, Fisher Consistency</td>
<td>54</td>
</tr>
<tr>
<td>6.4 The Influence Curve</td>
<td>56</td>
</tr>
<tr>
<td>6.5 The Squared I.C. under Random Censorship</td>
<td>58</td>
</tr>
<tr>
<td>6.6 Discussion</td>
<td>59</td>
</tr>
<tr>
<td>7. COX'S ESTIMATOR</td>
<td>62</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>62</td>
</tr>
<tr>
<td>7.2 The Proportional Hazards Model, Time Dependent Covariates</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>64</td>
</tr>
<tr>
<td>7.3 Cox's Likelihood</td>
<td>65</td>
</tr>
<tr>
<td>7.4 Solvability of the Likelihood Equations</td>
<td>69</td>
</tr>
<tr>
<td>7.5 The Estimator as a Von Mises Functional</td>
<td>70</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.6   Fisher Consistency.</td>
<td>73</td>
</tr>
<tr>
<td>7.7   The Influence Curve</td>
<td>76</td>
</tr>
<tr>
<td>7.8   The Squared I.C. under Random Censorship.</td>
<td>84</td>
</tr>
<tr>
<td>7.9   The Empirical I.C. and Numerical Studies of Cox's Estimate.</td>
<td>91</td>
</tr>
<tr>
<td>7.10  The Simulation Study.</td>
<td>105</td>
</tr>
<tr>
<td>7.11  Conclusions and Recommendations</td>
<td>111</td>
</tr>
<tr>
<td>8.    ESTIMATION OF THE UNDERLYING DISTRIBUTION IN COX'S MODEL.</td>
<td>113</td>
</tr>
<tr>
<td>8.1   Introduction</td>
<td>113</td>
</tr>
<tr>
<td>8.2   Cox's Approach</td>
<td>113</td>
</tr>
<tr>
<td>8.3   The Approaches of Kalbfleisch and Prentice</td>
<td>116</td>
</tr>
<tr>
<td>8.4   The Likelihoods of Oakes and Breslow</td>
<td>119</td>
</tr>
<tr>
<td>8.5   Functional Form; Fisher Consistency of Breslow's Estimator</td>
<td>121</td>
</tr>
<tr>
<td>8.6   The Influence Curve</td>
<td>123</td>
</tr>
<tr>
<td>8.7   Estimation of $\tilde{G}(s</td>
<td>\tilde{z})$</td>
</tr>
<tr>
<td>8.8   Conjectured Limiting Distribution for Breslow's Estimator</td>
<td>127</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>129</td>
</tr>
</tbody>
</table>

Appendices

A1. APPENDIX TO CHAPTER FIVE. 133
A2. APPENDIX TO CHAPTER SEVEN 139
A3. APPENDIX TO CHAPTER EIGHT 142
A4. COMPUTER LISTING OF SAMPLES FOR CHAPTER SEVEN PLOTS. 144
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.9.1.</td>
<td>Plots of the empirical influence curve for Cox's estimate against $t$: $n=49, \beta_n=0.36$</td>
<td>97</td>
</tr>
<tr>
<td>7.9.2.</td>
<td>Plots of the normalized difference in Cox estimates against $t$: $n=49, \beta_n=0.36$</td>
<td>98</td>
</tr>
<tr>
<td>7.9.3.</td>
<td>Plots of the empirical influence curve for Cox's estimate against $t$: $n=49, \beta_n=1.36$</td>
<td>99</td>
</tr>
<tr>
<td>7.9.4.</td>
<td>Plots of the normalized difference in Cox estimates against $t$: $n=49, \beta_n=1.36$</td>
<td>100</td>
</tr>
<tr>
<td>7.9.5.</td>
<td>Plots of the empirical influence curve for Cox's estimate against $z$: $n=49, \beta_n=0.36$</td>
<td>101</td>
</tr>
<tr>
<td>7.9.6.</td>
<td>Plots of the normalized difference in Cox estimates against $z$: $n=49, \beta_n=0.36$</td>
<td>102</td>
</tr>
<tr>
<td>7.9.7.</td>
<td>Plots of the empirical influence curve for Cox's estimate against $z$: $n=49, \beta_n=1.36$</td>
<td>103</td>
</tr>
<tr>
<td>7.9.8.</td>
<td>Plots of the normalized difference in Cox estimates against $z$: $n=49, \beta_n=1.36$</td>
<td>104</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.10.1.</td>
<td>Percentiles for the Regression Coefficient Data</td>
<td>108</td>
</tr>
<tr>
<td>7.10.2.</td>
<td>Percentiles for the Studentized W Statistic</td>
<td>109</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I am pleased to acknowledge the teaching and help of my advisor, Professor Lloyd Fisher. The comments of Professor Norman Breslow led to improvements in the final draft, while a conversation with Ms. Anne York was very helpful at an early stage of the work. I want to express special appreciation to Professors O. Dale Williams and C. E. Davis, for making possible the completion of this study. Dr. Geoff Gordon wrote the computer programs, and Mrs. Gay Hinnant did the superb typing. To all these I am grateful. Finally, I want to thank my parents, for their unfailing encouragement over the years.

This work was supported, in part, by U.S.P.H.S. Grant 5-T01-GM-1269 and, in part, by U.S. National Heart, Lung, and Blood Institute Contract NIH-NHLBI-71-2243 from the National Institutes of Health.
1. INTRODUCTION

Clinical trials, industrial life testing, and longitudinal studies of human or animal populations often require the statistical analysis of censored survival data. (An elementary account can be found in the book by Gross and Clark (1976). A review article with many references is Breslow (1975). Farewell and Prentice (1977) provide advanced parametric methods.) In this study we are concerned with the robustness of some estimators encountered in survival analysis. In particular we extend to the censored survival problem a formal framework, due to Hampel (1968, 1974), for the study of robustness. By robustness, we mean, loosely, insensitivity of an estimator to deviant data and departures from assumptions. For concreteness we will speak mostly in terms of human follow-up studies; but the results may apply to other fields.

1.1. Survival Background, Parameterization

In survival analysis, interest centers on time to an inevitable response, which we term "death" or "failure." The time to failure we call "failure time" or "survival time." (No distinction is intended.) Failure time is modeled as a non-negative random variable $S$.

Suppose we follow a sample of $n$ persons, each with a potential failure time $S_i$ ($i=1, \ldots, n$). Censoring arises when $S_i$ is unobservable. This can happen for a variety of reasons. For example, the study may end before all $n$ people fail; a person may drop out or move away. (A different kind of censoring, more common in industrial life testing, is progressive censoring. A progressive censoring scheme is one in which the experimenter deliberately removes objects from study.)
Because of censoring, we observe for the $i$-th person in the sample:

$$t_i = \text{time observed without failure and a censoring indicator } \delta_i;$$

$$\delta_i = \begin{cases} 
1 & t_i = S_i \quad \text{(uncensored)} \\
0 & t_i < S_i \quad \text{(censored)}.
\end{cases}$$

In most studies, known factors such as age, treatment group, sex, and extent of disease may affect survival time. We express these as a $p \times 1$ vector of covariates $\mathbf{z}_i$, with $\mathbf{z}_i^T = (z_{i1}, z_{i2}, \ldots, z_{ip})$. The data for the experiment then consist of $(t_i, \delta_i)$ or $(t_i, \delta_i, \mathbf{z}_i)$, $i=1, \ldots, n$.

Throughout this paper, lower case "t" denotes a time variate; upper case "T" denotes matrix or vector transpose. Unfortunately, both symbols may appear in the same expression, e.g. 6.1.1.

The distribution of the failure time random variable $S$ can be described by

(1.1.1) The cumulative distribution function

$$G(t) = P(S \leq t);$$

(1.1.2) The density function

$$g(t) = \frac{d}{dt}G(t);$$

(1.1.3) The survival function

$$\bar{G}(t) = 1 - G(t) = P(S > t);$$

(1.1.4) The hazard function

$$\lambda(t) = \frac{g(t) / \bar{G}(t)}{g(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq S < t + \Delta t | S > t)}{\Delta t};$$

(1.1.5) The integrated or cumulative hazard

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log e(\bar{G}(t)).$$

When failure time depends on covariates $\mathbf{z}$, we write (1)-(5) above conditioned on the value of $\mathbf{z}$: $\bar{G}(t | \mathbf{z}) = P(S > t | \mathbf{z}); \lambda(t | \mathbf{z}) = g(t | \mathbf{z}) / (\bar{G}(t | \mathbf{z}))$ etc.
Inference from survival data requires the choice of models and tests of hypotheses about parameters. These are important topics, but in this thesis we concentrate on estimators and their properties.

1.2. Description of Contents

Chapter 2 sketches the main results from Hampel's robustness theory. A central idea is to define estimators based on \( x_1, x_2, \ldots, x_n \) as functionals of the empirical distribution function \( \tilde{F}_n : \tilde{\theta} = \tilde{w}(\tilde{F}_n) \). Such estimators are called von Mises functionals. An estimator is Fisher consistent if \( \tilde{w}(\tilde{F}_0) \equiv \tilde{\theta} \), where \( \tilde{F}_0 \) is the underlying distribution of the \( X \)'s. The influence curve (I.C.) is introduced. The I.C. is essentially a derivative of the functional \( \tilde{w}(\tilde{F}) \) and shows the sensitivity of \( \tilde{w}(\tilde{F}) \) to local changes in the underlying model. An empirical version of the I.C. shows how the estimator reacts to perturbations in the data. Related theory is discussed. In particular, if \( \sqrt{n}\tilde{\theta} \) is asymptotically normal, the asymptotic variance \( \sigma^2 \) is often related to the I.C. by

\[
(1.2.1) \quad \sigma^2 = \int I(C)^2(x) \, dF(x),
\]

or by a multivariate version of (1.2.1) if \( \tilde{\theta} \) is a vector.

The general method for applying Hampel's ideas to censored survival data is outlined in Chapter 3. The idea is to write \( X = (t, \delta) \) or \( X = (t, \delta, z) \) and define the empirical distribution function \( \tilde{F}_n(x) \) accordingly. We also need to model the \( X \)'s as random variables with a common underlying distribution \( F(x) \). One such distribution is proposed, termed a random censorship model with covariates. Limitations of Hampel's approach are discussed, and some related work in the survival literature is reviewed.

Each of the remaining chapters is devoted to a particular estimation problem and associated estimator. Chapter 4 considers the simple
exponential model \( \lambda(t) = \lambda \) and the maximum likelihood estimator (M.L.E.) for the model. In Chapter 5, a nonparametric estimator for an arbitrary survival function \( \tilde{G}(t) \) is examined. Chapter 6 studies an exponential regression model \( \lambda(t|z) = \exp(\beta^T z) \lambda_0 \) and the corresponding M.L.E.'s.

Each of the estimators in these chapters is shown to be a von Mises functional and, under random censorship, is proved Fisher consistent. The influence curve is derived, and, again under random censorship, the relationship 1.2.1 is proved.

Chapters 7 and 8 are devoted to a major topic of this thesis--the proportional hazards (P.H.) model of D. R. Cox (1972). In this model
\[
\lambda(t|z) = \exp(\beta^T z) \lambda_0(t),
\]
where \( \lambda_0(t) \) is the hazard function of an arbitrary unspecified underlying distribution. Cox's estimator for \( \beta \) is examined in detail in Chapter 7. In Chapter 8 we study an estimator (due to Breslow 1972a,b) for the underlying distribution in the P.H. model.
2. IDEAS FROM THE THEORY OF ROBUST ESTIMATION

This chapter introduces those ideas from the theory of robust estimation which underlie the succeeding chapters. The formal theory originated in Hampel's doctoral dissertation (Hampel, 1968). For illuminating reviews of robust statistics, including Hampel's theory, see Huber (1972) and Hampel (1973, 1974). There is also a brief introduction in Cox and Hinkley (1974, Section 9.4). Hampel began by characterizing estimators as functionals on the space of probability distributions.

2.1. Estimators as Functionals

The statistical problem is to estimate a $p \times 1$ parameter vector $\theta \in \Theta$. At hand are observations $x_1, x_2, \ldots, x_n$, possibly multivariate, with empirical distribution function $F_n(x)$.

Definition: An estimator $\hat{\theta}_n$ for $\theta$ is said to be a von Mises functional if $\hat{\theta}_n$ can be written as a functional of $F_n$, that is if

$$\hat{\theta}_n = w(F_n),$$

where the functional $w(\cdot)$, defined on the space of probability measures and taking values in $\Theta$, does not depend on $n$ (von Mises, 1947; Hampel, 1968, 1974; Huber, 1972).

Examples

(1) The arithmetic mean can be written

$$\bar{X}_n = \int x \, dF_n(x) = w(F_n),$$

where

$$w(F) = \int x \, dF(x).$$
(2) The M estimator for a location parameter (Huber, 1964, 1972) is implicitly defined by

\[ \sum_{i=1}^{n} \psi(X_i - \hat{\theta}) = 0, \]

for an odd function \( \psi(\cdot) \). (A unique solution is not guaranteed.) We can divide both sides of (2.1.2) by \( n \). Then \( \hat{\theta}_n = w(F_n) \), where \( w(F_n) \) is implicitly defined by

\[ \int \psi(X - w(F)) \, dF(x) = 0. \]

The sample mean in Example 1 is a special case, with \( \psi(x) = x \).

(3) Regression. In multiple regression, the observations consist of a (univariate) response \( y_i \) and a \( 1 \times p \) vector of covariates \( z_i^T \), \( i = 1, \ldots, n \), where \( z_i \) may include indicators. Define \( \tilde{x}_i = (y_i, z_i) \), \( i = 1, \ldots, n \), and let \( F_n(x) \) be the empirical distribution function of the \( X \)'s. Let \( Z(n \times p) \) be the design matrix, with \( z_i^T \), the \( i \)-th row, and let \( y^T = (y_1, \ldots, y_n) \).

In the model \( E(y_i \mid z_i) = z_i^T \beta \), the least squares estimator \( \hat{\beta}_n \) is defined by the normal equations:

\[ Z\hat{\beta} = y, \]

or

\[ Z^T(y - Z\hat{\beta}) = 0. \]

We can rewrite the normal equations as a sum over the sample:

\[ \sum_{i=1}^{n} z_i(y_i - z_i^T \hat{\beta}) = 0. \]

Now divide both sides of (2.1.4) by \( n \) to find:

\[ \frac{1}{n} \sum_{i=1}^{n} z_i(y_i - z_i^T \hat{\beta}) = 0. \]
Therefore \( \hat{\beta}_n = w(F_n) \), where \( w(F) \) is defined by

\[
(2.1.5) \quad \int_{\mathcal{Z}} \int_{\mathcal{Z}} z(\mathcal{y} - \mathcal{z})^T w(F) \, dF(\mathcal{y}, \mathcal{z}) = 0.
\]

and \( \mathcal{Z} \) is the domain of the \( \mathcal{z}'s \). In this case, we can of course explicitly solve (2.1.5) for \( \sim w(F) \):

\[
\sim w(F) = \left[ \int_{\mathcal{Z}} \mathcal{z}^T dF(\mathcal{y}, \mathcal{z}) \right]^{-1} \int_{\mathcal{Z}} \mathcal{y} dF(\mathcal{y}, \mathcal{z}).
\]

The reader can easily show that this gives the usual least squares solution when \( F = F_n \).

An estimator that is not exactly a von Mises functional itself may be nearly equal to such a functional. The sample median \( M_n \), for example, depends for its definition on whether the sample size is even or odd; but in large samples

\[
(2.1.6) \quad M_n \sim F_n^{-1}(\frac{1}{2}).
\]

The right hand side (hereafter abbreviated r.h.s.) of (2.1.6) is a von Mises functional \( w(F_n) \), defined by

\[
(2.1.7) \quad w(F) = F_n^{-1}(\frac{1}{2}).
\]

2.2. **Fisher Consistency**

Suppose \( w(F) \) is a von Mises functional for estimating \( \theta \in \Theta \). Now assume that the observations \( x_1, x_2, \ldots, x_n \) are independent random variables, with common cumulative distribution function \( F_\theta(x) \).

**Definition:** The estimator \( w(F) \) is Fisher consistent (F.C.) (Hampel, 1974) iff

\[
\approx w(\theta) \equiv \theta, \text{ identically in } \theta.
\]
Remarks

1. The virtue of a Fisher consistent estimator is that it estimates the right quantity if faced with the "true" distribution.

2. The definition given above is the one used by Hampel (1968, 1974) and others. Rao (1965, 5c.1) requires in addition that \( w(\cdot) \) be a weakly continuous functional over the space of probability distributions.

3. Fisher consistency does not refer to a limiting operation. Fisher consistency and ordinary weak consistency (convergence in probability) cannot be related without regularity conditions on \( w(\cdot) \) and the space of probability distributions.

4. The parameter \( \theta \) need not completely identify the distribution \( F_0 \).

Examples

1. Consider the family of distributions \( F_\mu \) with finite first moment \( \mu \). Then the functional (2.1.1) evaluated at \( F_\mu \) gives

\[
w(F_\mu) = \int x \, dF_\mu(x) = \mu.
\]

Therefore the arithmetic mean is F.C.

2. In the regression setup of Example 3 in Section 2.1, let the covariates \( z \) be drawn from a probability space \( Z \) with distribution function \( K(z) \). Conditional on \( z \), draw the response \( y \) from a distribution \( G_\beta(y|z) \) such that

\[
E_y(y|z) = \int y \, dG_\beta(y|z) = z^T \beta.
\]

Then the joint distribution of \( (y, z) \) is

\[
dF_\beta(y, z) = dG_\beta(y|z) \, dK(z).
\]

If this distribution is substituted in the normal equations, (2.1.5), one can easily show that \( \hat{\theta} = \beta \) is a solution. (Details are left to the reader).
Therefore, the least squares estimator \( w(F) \) defined by (2.1.5) is, under this model, Fisher consistent.

2.3. Robustness

Hampel chose to study an estimator \( w(F) \) in terms of its functional properties. For example, the natural requirement that \( w(F) \) estimate the "right thing" is translated, in functional terms, into Fisher consistency. "Robustness" is identified, in part, with stability of \( w(F) \) to changes in \( F \).

As outlined by Huber (1972) and Hampel (1974), there are three aspects to this stability: continuity of the estimator, its "breakdown point," and sensitivity to infinitesimal changes in \( F \).

Continuity is the requirement that small changes in \( F \) (errors, contamination, distortion) lead to only small changes in the estimator \( w(F) \). Formally, a metric is introduced onto the space of probability distributions, and continuity of \( w(F) \) in this metric is required. The Prokhorov metric is often convenient for studying robustness, though other metrics can be used (Prokhorov, 1956; Hampel, 1968, 1974; Martin, 1974). Among familiar estimators, the sample mean defined by (2.1.1) is nowhere continuous in the space of distributions.

The breakdown point is, roughly, the smallest percentage of observations that can be grossly in error (at \( \pm \infty \), say) before the estimator itself becomes unbounded. For example, the breakdown point of the mean is zero; that of the median is 50 percent. Again, there is a formal definition in terms of the Prokhorov metric.

The Prokhorov metric proves difficult to work with in the survival setup of the next chapter. Therefore, we will not formally study continuity and the breakdown point in this thesis. Instead we concentrate on the third aspect of robustness--stability of \( w(F) \) in the face of local
or infinitesimal changes in $F$. This leads naturally to consideration of a "derivative" of $w(F)$ at $F$ and consequently to the influence curve.

**The Influence Curve**

The influence curve (I.C.) was introduced by Hampel (1968). Among published discussions are an expository article by Hampel (1974) and sections of Huber (1972) and of Cox and Hinkley (1974, Sec. 9.4).

Most applications of the I.C. have been to estimation of univariate location and scale (Andrews, et al., 1972). For multivariate applications see Mallows (1974) (regression and correlation) and Devlin, Gnanadesikan, and Kettenring (1975) (correlation).

We first define a general first order von Mises derivative. Let $F(x)$ and $K(x)$ be probability measures on a complete, separable metric space. Suppose $w(\cdot)$ is a vector valued von Mises Functional (estimator). Form the $\varepsilon$-mixture distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon K$. For sufficiently nice $F$, the estimator $w(\cdot)$ has a von Mises derivative at $F$:

$$
\left[ \frac{d}{d\varepsilon} w(F_\varepsilon) \right]_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{w(F_\varepsilon) - w(F)}{\varepsilon} = \int_{\mathcal{X}} IC(x;w,F) \, dK(x),
$$


In particular, we can find $IC(x;w,F)$ by letting $K(x) = I_x$, the distribution function which puts mass one at $x$.

**Definition 2.3.1.** The influence curve of $w$ with respect to $F$ defined (pointwise) at $x$ is given by:
\[ IC(x; \hat{w}, F) = \left[ \frac{d}{d\epsilon} \hat{w}(F) \right]_{\epsilon=0} \]

(2.3.2)

\[ w[(1-\epsilon)F + \epsilon I_x] - \hat{w}(F) = \lim_{\epsilon \to 0} \frac{w[(1-\epsilon)F + \epsilon I_x] - \hat{w}(F)}{\epsilon} \]

where \( F_\epsilon = (1-\epsilon)F + \epsilon I_x \).

**Example 2.3.1 (means)**

Let \( \hat{w}(F) = E_F(b(x)) \)

\[ = \int b(u) \, dF(u), \]

for some function \( b(x) \). Then, contaminating \( F \) with the point \( x^* \), we have

\[ w(F_\epsilon) = \int b(u) \, d[(1-\epsilon)F + \epsilon I_{x^*}](u) \]

\[ = (1-\epsilon) \int b(u) \, dF(u) + \epsilon b(x^*) \]

\[ = w(F) + \epsilon [b(x^*) - w(F)]. \]

And

\[ \frac{w(F_\epsilon) - \hat{w}(F)}{\epsilon} = \hat{w}(F) - b(x) \quad (\forall \epsilon > 0). \]

Therefore

(2.3.3) \[ IC(x^*; \hat{w}, F) = b(x^*) - E_F(b(x)). \]

When \( F = F_n \), the empirical I.C. of the sample mean \( b(x) = x \) is

(2.3.4) \[ IC(x; \hat{w}, F_n) = x - \bar{X}_n. \]

**Remarks**

1. If \( \hat{w}(F) \) is a \( p \times 1 \) vector valued estimator, \( \hat{w}(F)^T = (\hat{w}_1(F), \hat{w}_2(F), \ldots, \hat{w}_p(F)) \),

the \( IC(x, \hat{w}, F) \) is also a \( p \times 1 \) vector. The \( k \)-th coordinate of the IC is
That is, $IC_k$ is just the I.C. of $w_k$.

2. Let us note for future use that $e=0$ corresponds to the "central case":

$$F_0 = F$$

and

$$w(F_0) = w(F).$$

3. How does the I.C. express the "influence" of an observation on an estimator? For one answer, we consider the empirical I.C. for a sample with d.f. $F_n$. Let $w(F_n)$ be the estimator based on $F_n$ and $x$ a possible observation (not necessarily in the sample). Form $F_n, e = (1-e)F_n + e I_x$.

A first order Taylor's series expansion about $w(F_n)$ using (2.3.2) gives

$$w(F_n) - w(F) = e IC(x; w, F) + O(e^2).\ldots$$

Now suppose $e = 1/(n+1)$. Then each of the observations: $x_i (i=1, \ldots, n)$, which in $F_n$ had mass $1/n$, has in $F_n, 1/(n+1)$ mass

$$(1-e)/n = 1/(n+1).$$

The "new" observation $x$ has mass $e = 1/(n+1)$.

Therefore, the mixture distribution $F_{n+1}/(n+1)$ is the distribution of the sample of size $n+1$, $F^x_{n+1}$, say, consisting of $x_1, \ldots, x_n$ and $x$.

The estimator $w(F_{n+1}/(n+1)) = w(F^x_{n+1})$ is the estimator based on the augmented sample, and the expansion (2.3.5) becomes

$$w(F^x_{n+1}) - w(F) = \frac{IC(x; w, F)}{n+1} + O(1/(n+1)^2).$$

Approximately

$$w(F^x_{n+1}) - w(F) \approx IC(x; w, F_n).$$

In other words, the empirical I.C. measures the normalized change in the estimator caused by the addition of a point $x$ to the sample.
For the case of the sample mean \( \bar{w}(F) = \frac{\bar{x}}{n} \), one can easily show
\[
\frac{x_{n+1} - \bar{x}}{n+1} = \frac{\bar{x} - x_n}{n} = \frac{x_n - x}{n+1}.
\]
That is, the relationship (2.3.8) is exact. In other cases, where \( w(F) \) is not linear in \( F \), (2.3.8) may provide only a qualitative approximation. (The approximation will be examined in Chapter 7 for the case of Cox's estimator.)

4. If (2.3.1) holds, then
\[
\sum_{i=1}^{n} IC(x_i;w,F) \cdot dF(x) = 0.
\]
In particular, when \( F = F_n \)
\[
\sum_{i=1}^{n} IC(x_i;w,F) = 0.
\]

For the estimators examined in this study, we will not prove the existence of the general von Mises derivative satisfying (2.3.1). But equations (2.3.9) and (2.3.10) will be checked for the I.C.'s derived in the chapters to come.

**Theoretical Considerations**

When the general von Mises derivative (2.3.1) exists, another Taylor series expansion may hold.
\[
\sum_{i=1}^{n} IC(x_i;w,F) \cdot dK(x) = \sum_{i=1}^{n} IC(x_i;w,F) dK(x) + o(\epsilon^2).
\]
This leads to a proof of the asymptotic normality of \( \hat{\theta}_n = w(F) \) (Cox and Hinkley, 1974).

Take \( \epsilon = 1/n \) and let \( K(x) = nF(x) - (n-1) F(x) \). Then
\[
F_n = (1-\epsilon) F + \epsilon K \quad \text{and}
\]
\[
\int IC(x;w,F) \cdot dK(x) = \int IC(x;w,F) \cdot dF_n.
\]
using (2.3.9).
Then by (2.3.11)

\[(2.3.12) \quad w(F_n) - w(F) = (1/n) \sum_{i=1}^{n} IC(x_i; w, F) + O((1/n)^2).\]

Note that the I.C. with respect to \( F \), not \( F_n \), appears on the r.h.s. of \( (2.3.12) \). Here \( IC(x_i; w, F) \) expresses the "influence" of \( x_i \) on \( w(F) \) in a way different from \( (2.3.7) \). The summands \( IC(x_i; w, F) \) are i.i.d. random variables with mean zero (by \( (2.3.9) \) and variance-covariance matrix

\[ A(w, F) = E_F \left( IC(x_i; w, F) IC(x_i; w, F)^T \right) \]

\[(2.3.13) \quad = \int IC(x_i; w, F) IC(x_i; w, F)^T dF(x). \]

Appeal to the central limit theorem leads to the conclusion that

\[(2.3.14) \quad \sqrt{n}(w(F_n) - w(F)) \xrightarrow{d} N(0, A(w, F)). \]

The conditions under which \( (2.3.14) \) holds have been discussed by von Mises (1947), Filippova (1961), and Miller and Sen (1972). The conditions are so heavy that in practice, asymptotic normality is usually proved by other means.

When asymptotic normality has been proved by other theory for estimates in succeeding chapters, we will check that \( A(w, F) \) agrees with the limiting covariance matrix.

A sample approximation to \( A(w, F) \) is (for \( p=1 \))

\[(2.3.15) \quad A(w, F_n) = 1/n \sum_{i=1}^{n} IC^2(x_i; w, F_n). \]

This estimate of \( A(w, F) \) is discussed by Cox and Hinkley (1975) and by Mallows (1974).

For one estimator studied in this thesis, the Breslow estimate of the underlying distribution in the Cox model, Chapter 8, no limiting results are known. In this case we will conjecture that asymptotic normality holds,
with limiting variance \( A(w, F) \) given by (2.3.13) and estimated (before some simplification) by \( A(w, F_n) \).

**Use of the I.C. with Outliers**

Devlin, Gnanadesikan, and Kettenring (1975) advocated the empirical I.C. as an indicator of the influence of each data point \( x_i = (y_{i1}, y_{i2}) \) on the sample correlation coefficient \( \hat{\rho}_n \). They found the complicated exact expression for \( \hat{\rho}_n - \hat{\rho}_{n-1} \) and showed

\[
\begin{align*}
(2.3.16) \quad n(\hat{\rho}_n - \hat{\rho}_{n-1}) & \leq IC(x_i; \hat{\rho}_n, F_n).
\end{align*}
\]

Devlin et al. recommended plots of the empirical \( IC(x_i; \hat{\rho}_n, F_n) \) contours on a \((y_1, y_2)\) plane to detect outliers.

For estimators of dimension \( p > 2 \) functions like

\[
Q_i = IC(x_i; \hat{\rho}_n, F_n)^T IC(x_i; \hat{\rho}_n, F_n)
\]

might with similar reasoning show the effect of a point \( x_i \) on the estimate \( \hat{\theta}_n = w(F_n) \).

Unfortunately, equation (2.3.8) suggests that the empirical \( IC(x_i) \) better estimates the effect of adding another point at \( x_i \) than of deleting the point already there. If \( x_i \) is an outlier originally it may have already distorted \( \hat{\theta}_n = w(F_n) \). The effect of adding the same outlier to the sample a second time will not, in general, be as great as the effect of deleting the outlier altogether. In practice, Devlin et al. found the approximation (2.3.16) using the empirical I.C. satisfactory, but this may not be so in other problems.

**Assessing Robustness with the I.C.**

The I.C. helps to assess robustness of estimators against two kinds of disturbances (Hampel, 1974). The first kind is the "throwing in" of bad data points—contamination, outliers. A bounded I.C. indicates
protection against such errors. (Hampel calls the \( \sup_x |IC(x; w, F)| \) the "gross error sensitivity.")

A second kind of disturbance comes from "wiggling" observations: rounding, grouping, local shifts. A measure of robustness to "wiggling" is the "local shift sensitivity":

\[
\lambda = \sup_{x \neq y} \frac{|IC(x) - IC(y)|}{|x - y|}.
\]

This will be infinite if the I.C. jumps as a function of \( x \). In such cases, the estimator will jump with infinitesimal shifts in the data. This is true, for example, of the Cox estimator in Chapter 7.

**Interpretation of the I.C. in Regression**

Three of the chapters in this thesis discuss regression problems.

Chapter 6, on an exponential model, and Chapters 7 and 8, on the proportional hazards model of Cox. In regression, the influence of each observation comes not only from the (random) response variable \( y \) but also from the covariables \( z \).

Let us illustrate by solving for the I.C. of the least squares estimator \( \hat{\beta} \) defined by (2.1.5). Suppose we want to find the I.C. at the point \( x^* = (y^*, z^*) \). We write

\[
F_{\hat{\beta}} = (1-\tau)F + \tau F_{x^*}
\]

and substitute \( F_{\hat{\beta}} \) for \( F \) in (2.1.5): The equation defining \( \hat{\beta}(\epsilon) = w(F_{\hat{\beta}}) \) is then

\[0 = \int z(y - z^T \hat{\beta}(\epsilon)) dF_{\hat{\beta}}(y, z) = \int z(y - z^T \hat{\beta}(\epsilon)) dF(y, z) = \int [z^* |y^* - z^*^T \hat{\beta}(\epsilon)| - z(y - z^T \hat{\beta}(\epsilon))] dF(y, z)\]

(2.3.18)
Now differentiate (2.3.18) w.r.t. \( \varepsilon \) and evaluate at \( \varepsilon = 0 \). This gives, (using 2.3.9),

\[
0 = -\left[ z z^T dF(y, z) \right] \left[ \frac{d}{d\varepsilon} \beta(\varepsilon) \right]_{\varepsilon=0}
\]

(2.3.19)

\[
+ z^T \left[ y - z^T \beta(0) \right].
\]

Or

\[
\text{IC}(y^*, z^*; \beta, F) = \left[ \frac{d}{d\varepsilon} \beta(\varepsilon) \right]_{\varepsilon=0}
\]

(2.3.20)

\[
= \left[ \int_{-\infty}^{\infty} z z^T dF(y, z) \right]^{-1} z^T \left[ y - z^T \beta(0) \right]
\]

when \( F = F_n \), \( \beta(0) = \beta(F_n) \), the l.s. estimate based on \( F_n \). The empirical I.C. is therefore

(2.3.21)

\[
\text{IC}(y^*, z^*; \beta, F_n) = n \left( z z^T \right)^{-1} z^T \left( y - z^T \beta_n \right).
\]

This implies, that

(2.3.22)

\[
\beta_{n+1}^T (y^*, z^*) - \beta_n^T \approx \left( \frac{n}{n+1} \right) \left( z z^T \right)^{-1} z^T \left( y^* - z^T \beta_n^T \right)
\]

(2.3.23)

\[
\approx (z z^T)^{-1} z r^*
\]

where \( r^* = y^* - z^T \beta_n^T \).

We see that the point \((y^*, z^*)\) influences the estimate \( \beta_n^T \) in two ways. One, as expected, is by means of the discrepancy between \( y^* \) and its predicted value \( z^T \beta_n^T \); this discrepancy is measured by the residual \( r^* = y^* - z^T \beta_n^T \). The second way is by means of the magnitude of the independent variables \( z^* \). Even if the residual is slight, extreme values of \( z^* \) may strongly affect the estimate. Hampel (1973) calls this effect of \( z^* \) the "influence of position in factor space."
Errors in the data may give rise to large residuals, but a (relatively) large $z^\ast$ may not be an error at all. We may nonetheless want to make an estimator robust against the influence of $z$'s far from the data (Hampel, 1973; Huber, 1973). This influence can stand out strongly in estimators that are otherwise robust (e.g. the Cox estimator, based on ranks, in Chapter 7).
3. EXTENSION OF HAMPEL'S FRAMEWORK TO SURVIVAL DATA

3.1. Introduction

In this chapter we show how Hampel's ideas can be applied to censored survival data with covariates. In the last part of the chapter, some other work on the robustness of survival estimators is reviewed.

Application of Hampel's ideas first requires definition of an empirical distribution function $F_n$. Recall from Chapter 1 that in a survival setup, the data consist of an observation $(t_i, \delta_i)$ or $(t_i, \delta_i, Z_i)$ for each patient: $t_i$ is the observed time on study, $\delta_i$ is a censoring indicator (0 = censored, 1 = uncensored); and $Z_i$ is a vector of covariates.

We therefore find it natural to make the following definitions:

Definition 3.1.1. The observations in the survival problem consist of vectors $x_1, x_2, \ldots, x_n$, where

$$x_i = (t_i, \delta_i, Z_i)$$

if covariates are present; and

$$x_i = (t_i, \delta_i)$$

if covariates are not present. To simplify notation the wavy underscore is omitted from the $x$'s, though not from other vectors, throughout this dissertation.

Definition 3.1.2. The empirical distribution function $F_n(x)$ based on the sample $x_1, x_2, \ldots, x_n$ is that distribution function which puts mass $1/n$ at each observed $x_i (i = 1, \ldots, n)$.

With $F_n$ defined, we can study survival estimators as von Mises functionals $w(F_n)$. 
3.2. The Random Censorship Model with Covariates

Application of Hampel's theory also requires that the observations $x_1, x_2, \ldots, x_n$ be modeled as i.i.d. random variables with distribution function $F_\theta$. (Here $\theta$ is the parameter to be estimated.) This allows consideration of Fisher consistency and properties of the influence curve with regard to an underlying distribution.

This requirement poses a problem: In survival analysis, the only distribution specified is $\tilde{G}_\theta(s|z)$ (or $\tilde{G}_\theta(s)$, a special case), the distribution of the failure time random variable $S$; the only parameters of interest are parameters of $\tilde{G}_\theta$. The observations $x = (t, \delta, z)$ depend, however, not only on $\tilde{G}_\theta$ but on the occurrence of censoring and on the covariates $z$. The mechanisms generating the censoring and the covariate are arbitrary nuisance mechanisms, which may not be random at all. The data can therefore arise in many ways.

To apply Hampel's ideas, we define a distribution $F(x)$ in which the censoring and covariates are random but otherwise unspecified. This distribution is an extension of the model of random censorship (Gilbert, 1962; Breslow, 1970; Breslow and Crowley, 1974; Crowley, 1973). We may therefore call this new model the random censorship model with covariates (and where there is no misunderstanding, the random censorship model, for short).

**Definition 3.2.1.** An observation $x_i = (t_i, \delta_i, z_i)$ from the random censorship model is generated in the following steps:

1. The $z_i$ are a random sample from a distribution with d.f. $K(z)$ on domain $Z$. We may assume that the distribution has a density $k(z)dz = dK(z)$ (defined suitably when $Z$ contains both continuous and discrete elements).

2. Obtain the failure time $S_i$ from $G(s|z_i)$ with density $g(s|z_i)$, where $z_i$ was generated as in (1).
3. Independent of step (2) pick the censoring time $C_i$ from a distribution $H(c|z_i) = P(C < c|z_i)$, with density $dH(c|z_i) = h(c|z_i)$, and with $\bar{H}(c|z_i) = 1 - H(c|z_i)$. Again $z_i$ is from step (1). In other words, $C_i$ is, conditional on $z_i$, independent of $S_i$.

4. Form $t_i = \min(S_i, C_i)$ and $\delta_i = I[t_i = S_i]$ (where $I[A]$ is the indicator of the event $A$).

5. Then $x_i = (t_i, \delta_i, z_i)$ and the resulting distribution function is $F(x)$.

If there are no covariates, the process consists of picking $S_i$ and $C_i$ independently from $G(s)$ and $H(c)$ respectively, and of forming $(t_i, \delta_i)$ as in steps (4) and (5).

We can calculate expressions for the joint distribution $f(t, \delta, z)$ as follows. For the moment, let upper case "T" denote the random variable, while lower case "t" denotes a particular value.

$$P(T < t, \delta=1|z) = P(S < t, S < C|z)$$

$$= \int_0^t \bar{H}(y|z)g(y|z)dy.$$ 

The conditional density is found by differentiating the last expression with respect to $t$:

$$(3.2.1) \quad f(t, \delta=1|z) = \frac{dF(t, \delta=1|z)}{dt}.$$ 

We will denote this $f(t, \delta=1|z) = \tilde{H}(t|z)g(t|z)$. 

Then

$$f(t, 1, z) = f(t, \delta=1|z)k(z)$$

$$= \tilde{H}(t|z)g(t|z)k(z).$$
Similarly,
\[
f(t, 0, z) = \bar{G}(t | z) h(t | z) k(z),
\]
where, again \( f(t, 0, z) = \frac{\partial F(t, 0, z)}{\partial t} \) is short for \( f(t, 0, z) \).

The total density, with respect to the product of Lebesgue and counting measure, is
\[
(3.2.3) \quad f(t, \delta, z) = [\bar{H}(t | z) g(t | z)]^\delta [\bar{G}(t | z) h(t | z) 2^{1-\delta}] k(z)
\]
\[
(3.2.4) \quad = [g(t | z)^\delta \bar{G}(t | z)^{1-\delta}] [h(t | z)^\delta \bar{H}(t | z)^{1-\delta}] k(z).
\]

In estimation problems, the parameters of interest are parameters of \( G \) and \( g \). The factorization of \( f(t, \delta, z) \) is noted only implicitly and only the first term
\[
(3.2.5) \quad g(t | z)^\delta \bar{G}(t | z)^{1-\delta}
\]
appears in the likelihood.

The symbols \( F \) and \( f \) will designate marginal, conditional, and unconditional distributions of \((t, \delta, z)\). Thus we write
\[
(3.2.6) \quad f(t, l) = \frac{dF(t, l)}{dt} = \int f(t, l, z) dK(z),
\]
for the marginal density of observed failure times. We will always assume enough regularity so that integrations in \( t, z \), or in \( C, S, \) and \( z \) can be performed in any order.

Derivations, when there are no covariates, are easily made.

Remarks

1. This random censorship model with covariates was inspired by earlier work of Breslow (1970). Breslow studied a generalization of the Kruskal-Wallis test for comparing survival distributions of \( K \) populations. The long run proportion of observations from the \( j \)-th population was
\[
\lambda_j (\sum_j \lambda_j = 1);
\]
and a different censoring distribution was allowed for each population.
Define K 0-1 variables $z_j$. Let $\text{Prob}(z_j=1, z_i=0, i \neq j) = \lambda_j$.

Then the K sample problem becomes a special case of random censorship with covariates.

2. By allowing a different censoring distribution for each population, Breslow implicitly weakened the usual assumption of survival analysis that censoring and survival be unconditionally independent. In extension to regression problems, the current model requires only that censoring and failure time be independent conditional on the value of $z$'s. This further weakens the usual assumption.

Suppose, for example, that younger patients are more likely than older patients to move away during the course of a clinical trial. Then short censoring times are apt to be associated with long (unobserved) survival times, since both are associated with younger ages. But, conditional on age at entry, censoring and survival may be independent. The inclusion of age as a covariate in such a case allows the inference to proceed.

3. Not all censoring schemes are covered by this model. Especially excluded are progressive schemes in which censorship depends on events in the course of the trial. Many statistical procedures are valid for such schemes; the only requirement is that failure times of the censored observations not be affected by censoring. We cannot, however, study such procedures by means of the random censorship model.

4. To what extent can covariate vectors $z_i$ be regarded in practice as i.i.d. random variables? Some covariates are truly random. For example, treatment assignment may be the outcome of a randomization process. Or, the patients may represent some sampled population, entry into the trial being a "random" process.
At the other extreme are covariates which are fixed. The need for an intercept in a regression model, for example, may require a covariate \( z_{ij} = 1 \). We may think of this covariate value as being sampled with probability one.

Between the extremes of truly random and truly fixed covariates are others, in which the sampled distribution may be hard to describe at best: patient assignment schemes, "balancing" of patients within treatments, biased or arbitrary recruitment and selection. In such cases (which are common), we would only conjecture that the covariates behave like random variables with some regularity conditions.

5. In Cox's (1972) regression model, the covariate values may change with time. Such covariates may also be considered in a random censorship model. A discussion is deferred until Chapter Seven, where the study of Cox's model takes place.

3.3. Related Work

In this section, some related work on robustness in the survival literature is outlined. These references do not constitute a complete survey; they are intended only to alert the reader to other approaches.

Fisher and Kanarek (1974a) studied the extent to which choice of incorrect model affected inference in survival analysis. They studied four exponential regression models, with the distribution depending on one covariate \( z \):

1. \( \lambda_1(z) = \Lambda_0 + \Lambda_1 z \); Byar, et al. (1974), Kanarek (1973).
2. \( \lambda_2(z) = 1/(\beta_0 + \beta_1 z) \), Feigl and Zelen (1965).
3. \( \lambda_3(z) = \exp(C_0 + C_1 z) \), Glasser (1967).
4. \( \lambda_4(z) = -\frac{1}{t_0} \ln \left\{ \exp(D_0 + D_1 z) \right\}, \frac{\exp(D_0 + D_1 z)}{1 + \exp(D_0 + D_1 z)} \); Trueitt, et al. (1967).
Model 4 is an exponential model which gives the usual multiple logistic probability of survival past $t_0$.

A distance between models $i$ and $j$ was defined by the average (over $z$) of the supremum norm:

$$d(\lambda_i, \lambda_j) = E_z \left( \sup_{t \geq 0} \left| e^{-\lambda_i(z)t} - e^{-\lambda_j(z)t} \right| \right).$$

The covariate $z$ was given a uniform $[-1,1]$ distribution; censoring was not considered.

When model $i$ is known to be the true model, the parameters must still be estimated. Here maximum likelihood estimation was used. The distance of the estimated model from the true model was defined as:

$$d(\hat{\lambda}_i, \lambda_i) = E_{\hat{\lambda}_i} \left( E_z \left( \sup_{t \geq 0} \left| e^{-\lambda_i(z)t} - e^{-\hat{\lambda}_i(z)t} \right| \right) \right).$$

The outer expectation was taken with respect to the asymptotic normal distribution of $\hat{\lambda}_i$.

Examination of (3.3.1) and (3.3.2) showed how the loss of precision from incorrect model choice compared to the loss from the variability of estimation. (It might have been informative to compute $d(\lambda_j, \hat{\lambda}_i)$, but this was not done.)

The results showed models 1, 3, and 4 were very close

$$d(\lambda_i, \lambda_j) \leq \frac{1}{3} d(\lambda_i, \hat{\lambda}_i),$$

except when $z$ had a strong effect or when sample size was large. On the other hand, model 2, the Peitl-Zelen model, was considerably different from the others.

Prentice (1974, 1975) and Farewell and Prentice (1977) embedded a number of well known parametric models (Weibull, log normal, logistic, gamma) in a single parametric family. In this family, the log of failure time is modeled as a linear combination of the covariates:

$$\log S(z) = \alpha + \beta^T z + \omega.$$
Here $\beta$ is a scale parameter and $w$ is an error random variable. According to the generality required, $w$ follows either the distribution of the log of a gamma R.V. (one parameter) or the log of an $F$ random variable (two parameters). Censoring is taken into account; likelihood methods provide tests and estimates. Prentice and Farewell show that inference about $\beta$ may be more robust to spurious response times, if the model (3.3.3) is fitted.

**Independence of Censoring and Survival**

Most procedures in survival analysis depend upon some kind of assumption of the independence of censoring and survival. (In the random censorship model, we require independence conditional on $z$.) The assumption enables one to factor likelihoods, so that estimation of the survival function $\tilde{G}$ can proceed.

In many applications, the assumption of independence is suspect. Patients who drop out of clinical trials may be more sickly than those who stay in, or they may be healthier. For example, patients who drop out because of drug side effects may be constitutionally different from those who do not suffer such side effects. Or, patients who are healthier may be better able to move away from the city in which a study is located.

What are the consequences for survival analysis? This question has been studied by several authors. Peterson (1975) and Tsiatis (1975) in a competing risks set-up proved the following: that any underlying distribution $F(t,\delta)$ (or $F(t,\delta,z)$) of the observed data may arise from an infinite number of survival distributions $\tilde{G}$, if dependent censoring is possible. Therefore $\tilde{G}$ is not identifiable.

Peterson studied nonparametric estimation of $\tilde{G}$ via the Kaplan-Meier estimate (described in Chapter Four). He obtained sharp upper and lower
bounds for the estimate. The lower bound is equivalent to designating all censored observations as failures immediately after censoring. The upper bound is found by treating all censored observations as if they never fail. For some data sets (in which there is little censoring) the bounds can be narrow; with heavy censoring they can be wide.

Fisher and Kanarek (1974b) also studied the problem of estimating a single survival function $\tilde{G}$, when censoring and failure time are dependent. They presented a model with two kinds of censoring. The first kind was "administrative" censoring caused by end of study; such censoring should be independent of survival. The second censoring was loss to follow-up due to dropping out. In the model, such loss to follow-up affected failure time through a non-estimable scaling parameter, $0 \leq \alpha \leq \infty$. Large values of $\alpha (>1)$ corresponded to poor survival, following censoring; small values ($<1$) corresponded to better survival. The extremes of $\alpha=0$ and $\alpha=\infty$ corresponded to Peterson's upper and lower bounds, respectively, while $\alpha=1$ was equivalent to independence. By varying $\alpha$ and finding the M.L. estimate of $\tilde{G}$ for each value, an investigator could assess the robustness of his estimates to the independence assumption.

Fisher and Kanarek also discussed ways of using auxiliary information to test the assumption of independence. Suppose that at time $t_j$, $N_j$ people are being followed and $K_j$ of these are lost to follow-up (in some small interval following $t_j$). If the assumption of independence holds, any set of $K_j$ people of the $N_j$ at risk might, with equal probability, have been lost. If the assumption is not true the covariate values for the $K_j$ who were lost should be somewhat separated from those who were not lost. These facts are exploited to give tests of the independence assumption.

The applicability of these results in regression problems is not clear. Certainly one can apply the trick of failing censored observations
at the time of censoring or of letting them live forever. This should be easy to do and to program on a computer. Perhaps an analogue of Fisher and Kanarek's (1974b) $\alpha$-technique can be extended to the regression problem; this might provide a more realistic assessment than the extreme bounds (corresponding to $\alpha=0$ and $\alpha=\infty$).

3.5. **Limitations of Hampel's Approach**

First consider the estimation problem when there are no covariates. We would like to study the continuity of $\theta=w(F)$ with respect to changes in $G$, the survival function. We are restricted to studying continuity with respect to changes in $F$, the argument of the functional $w$. Results quoted in the last section show that the same $F$ may correspond to many different survival distributions $G$, if censoring and survival are not independent. This nonidentifiability property makes a general study of continuity extremely difficult.

One approach is the following: Assume that random censorship holds (thus ensuring independence of failure and censoring times). In addition, assume that the censoring distribution $H(c)$ is fixed. The resulting distribution $F$ of $X=(t,\delta)$ should be a 1-1 functional of $G$: $F=q(G)$, say. The estimator is then also a functional of $G$, defined by $\eta(G) = w(q(G))$. The formal study of the functional $\eta=wq$ can then proceed; we shall not, however, do so.

With covariates, the situation is more complicated. We cannot define the estimator as a functional $\theta=\eta(G)$, because there is no single underlying $G$. Instead there is a different distribution $G(\cdot|z)$ for each value of the covariates $z$. In regression problems, therefore, the functional approach is restricted to study of $\theta=w(F)$. 
The restriction of Hampel's ideas to estimation is another limitation. Some extensions to test statistics are possible, but we will not pursue the subject in this study.
4. THE ONE-PARAMETER EXPONENTIAL MODEL

4.1. The Model

We start with a simple parametric model: the one-parameter exponential. In this model, the hazard, density, and survivor functions are:

\[ \lambda(t) = \lambda \quad ; \]

(4.1.1)

\[ q_\lambda(t) = \lambda e^{-\lambda t} \quad ; \quad \tilde{G}_\lambda(t) = e^{-\lambda t} . \]

Actually, the model 4.1.1 is a special case of the Glasser multiple regression exponential model, which will be studied in Chapter Six. We study 4.1.1 for two reasons: (1) the simple exponential model is often fitted in its own right to survival data; and (2) the calculations and results serve as a good warm-up to the more complex models in the chapters to come.

The Estimator

The goal is to estimate \( \lambda \), and the usual method is by maximum likelihood. Each observation \( (t_i, \delta_i) \) contributes to the likelihood a term.

\[ \delta_i \quad \lambda \quad \delta_i \quad 1-\delta_i \quad q_\lambda(t_i) \quad \tilde{G}_\lambda(t_i) \]

(4.1.2)

\[ = \lambda \quad i \quad e^{-\lambda t_i} . \]

(See 3.2.5).

The log likelihood is

\[ \ell(\lambda) = \sum_{i=1}^{n} \ell_i(\lambda) = (\ln \lambda) \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} t_i . \]
The estimating equation is therefore

\[ \sum_{i=1}^{n} \frac{d}{d\lambda} U_i(\lambda) = \frac{d}{d\lambda} \lambda(\lambda) = \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i = 0, \]

where \( U_i(\lambda) = \frac{\delta_i}{\lambda} - t_i \) is the efficient score for a single observation.

Equation 4.1.3 is solved by

\[ \hat{\lambda}_n = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}. \]

\( \hat{\lambda}_n \) is the ratio of the observed number of failures to the total exposure time in the sample.

**Functional Form**

Let \( F_n \) be the sample distribution function of \((t,\delta)\). \( \hat{\lambda} \) can be written

\[ \hat{\lambda}_n = \frac{\left(\sum_{i=1}^{n} \delta_i\right)}{\left(\sum_{i=1}^{n} t_i\right)} \]

\[ = \frac{\int \delta \, dF_n(t)}{\int t \, dF_n(t)}, \]

\[ = A(F_n)/B(F_n), \text{ say,} \]

where \( A(F) = E_F(\delta) \), \( B(F) = E_F(t) \). Therefore \( \hat{\lambda}_n = w(F_n) \), with

\[ w(F) = A(F)/B(F). \]

**Fisher Consistency**

Let \( H(c) \) be an arbitrary distribution of censoring time. With the model 4.1.1 for failure times, a random censorship model (without covariates) is easily defined, as in Section 3.2. Let \( F_\lambda(t,\delta) \) be the resulting distribution of the data. Fisher consistency will hold, if

\[ w(F_\lambda) = A(F_\lambda)/B(F_\lambda) = \lambda. \]
We show this by evaluating $A(F_\lambda)$ and $B(F_\lambda)$, which are the expectations (over $F_\lambda$) of $\delta$ and $t$, respectively:

$$A(F_\lambda) = E_{F_\lambda}(\delta) = P\{S < C\}$$

(4.1.8)

$$= \int_0^\infty \tilde{H}(t)g_\lambda(t)dt = \lambda \int_0^\infty \tilde{H}(t)e^{-\lambda t}dt.$$  

(4.1.9)

$$B(F_\lambda) = E_{F_\lambda}(t) = \int_0^\infty (1-F_\lambda(t))dt. \quad \text{(Rao, 1965, 2b2.1)}$$

$$1-F_\lambda(t) = P\{S > t\}P\{C > t\}, \text{ by independence},$$

(4.1.10)

$$= \tilde{G}_\lambda(t) \tilde{H}(t)$$

$$= e^{-\lambda t} \tilde{H}(t).$$

Therefore,

$$B(F_\lambda) = \int_0^\infty e^{-\lambda t} \tilde{H}(t)dt,$$

(4.1.11)

and

$$A(F_\lambda)/B(F_\lambda) = \lambda.$$  

4.2. The Influence Curve

Suppose we contaminate a distribution $F$ with a point $x = (t, \delta)$. We write the contaminated distribution as $F_\varepsilon = (1-\varepsilon)F + \varepsilon I_{\{x\}}$, where $I_{\{x\}}$ is the c.d.f. which puts mass 1 at $x$.

The estimator of the exponential hazard based on $F_\varepsilon$ is:

$$w(F_\varepsilon) = A(F_\varepsilon)/B(F_\varepsilon),$$

(4.2.1)

or, to simplify notation:

$$w(\varepsilon) = A(\varepsilon)/B(\varepsilon).$$

(4.2.2)
The influence curve is given by:

\[(4.2.3) \quad \text{IC}(x; w, F) = \frac{d}{dc} w(c) \bigg|_{c=0} = \frac{B(0)A'(0) - A(0)B'(0)}{B'(0)},\]

where \(A'(0) = \frac{d}{dc} A(c) \bigg|_{c=0}\) etc. and \(\Lambda(0) = A(F)\). The derivatives \(A'(0)\) and \(B'(0)\) are easy to evaluate from (2.3.3), since \(A(F)\) and \(B(F)\) are means:

\[(4.2.4) \quad A'(0) = \delta - A(F)\]

\[(4.2.4) \quad B'(0) = t - B(F).\]

Simplifying (4.2.3), we find

\[\text{IC}(t, \delta; w, F) = \frac{\delta - [(A(F)/B(F))t]}{B(F)}\]

\[(4.2.5) \quad \text{IC}(t, \delta; w, F) = \frac{\delta - w(F)t}{B(F)}\]

The empirical I.C. is found by substituting \(\frac{F}{n}\) for \(F\) in (4.2.5):

\[(4.2.6) \quad \text{IC}(t, \delta; w, F_n) = \frac{\delta - \lambda_n t}{B(F_n)} = \frac{\delta - \hat{\lambda} t}{t_n}.\]

Check on Computation

By (2.3.9), an I.C. should satisfy \(E_F(\text{IC}(x; w, F)) = 0\). We can easily show this for the I.C. (4.2.5). By definition, we have for any \(F\),

\[E_F(\delta) = A(F), \quad E_F(t) = B(F).\]

By (4.2.5)

\[E_F(\text{IC}(t, \delta; w, F)) = E_F\left[\frac{\delta - (A(F)/B(F))t}{B(F)}\right]\]

\[(4.2.7) \quad = \left(E_F(\delta) - \left(A(F)/B(F)\right) E_F(t)\right)/B(F)\]

\[= (A(F) - A(F))/B(F) = 0.\]

The argument above holds for \(F = F_n\), so that

\[\sum_{i=1}^{n} \text{IC}(x_i; w, F_n) = 0,\]

which is (2.3.10).
Interpretation

The influence curves for the estimated hazard rate are fairly easy to interpret. We go into some detail, because the principles will apply in later chapters.

How will the addition of a new data point \( x = (t, \delta) \) affect an estimate \( \hat{\lambda} \) based on a sample of size \( n \)? For an answer, let us first examine the formula 4.1.5 for \( \hat{\lambda} \),

\[
\hat{\lambda} = \frac{(1/n) \sum_{i=1}^{n} \delta_i}{(1/n) \sum_{i=1}^{n} t_i}.
\]

We see that \( \delta = 1 \), an observed failure, will increase the estimate, if the corresponding \( t \) is less than or equal to \( \hat{t}_n \). Similarly, \( \delta = 0 \) and \( t > \hat{t}_n \) will result in \( \hat{\lambda}_{n+1} < \hat{\lambda}_n \). The change in other cases depends on the relative changes in \( (1/n) \sum \delta_i \) and \( (1/n) \sum t_i \). Intuitively, a relatively large \( t \) ought to decrease the estimator, even if \( \delta = 1 \).

Let us now see how the empirical I.C. clarifies matters. Recall this is

\[
IC(x; w, \hat{\lambda}_n) = \frac{\delta - \hat{\lambda}_n t}{\hat{t}_n}.
\]

From (2.3.7), we regard the I.C. as a (first order) approximation to \( (n+1)(\hat{\lambda}_{n+1} - \hat{\lambda}_n) \). To an approximation, then, we expect the estimate to decrease if \( \delta/t < \hat{\lambda}_n \), to increase if \( \delta/t > \hat{\lambda}_n \) and to show no change if \( \delta/t = \hat{\lambda}_n \). Note that the first inequality is not entirely correct. An observation \( (t, \delta) \) with \( \delta = 0 \) can still result in a decreased estimate if \( t \) is small enough.

Another shortcoming of the approximation is also apparent. For large enough \( t \), the approximation implies that \( \hat{\lambda}_{n+1} \) can decrease below zero, but \( \lambda > 0 \).
Another point of interest is this: We can regard $\delta$ as the observed number of failures (0) in $x=(t,\delta)$ and $\lambda_n t$ as the expected number of failures (E) in time $t$. Then the empirical I.C. is proportional to $0-E$, a kind of "residual" of the observation $(t,\delta)$. We therefore see that the I.C. gives at least a qualitative indication of the influence of $(t,\delta)$, when $(t,\delta)$ is added to the sample.

4.3. The Limiting Variance and Distribution of $\hat{\lambda}_n$

We now compare the limiting distribution of $\hat{\lambda}_n$ from standard likelihood theory to that suggested by the theory of von Mises derivatives.

Suppose that the data arise from a random censorship model $F_\lambda$, with survival given by (4.1.1). The information in a single observation $(t,\delta)$ is

$$I(\lambda) = E_{F_\lambda} U_1^2(\lambda) = E_{F_\lambda} \left(\frac{\delta}{\lambda} - t\right)^2 = -E_{F_\lambda}\left(\frac{d}{d\lambda} U_1(\lambda)\right)$$

(4.3.1)

$$= -E_{F_\lambda} \left(\frac{\delta}{\lambda^2}\right) = A(F_\lambda)/\lambda^2.$$

If the ordinary regularity conditions apply to the density $f_\lambda(t,\delta)$, then we have the standard result (Rao, 1965, 5f).

(4.3.2)

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{L} N(0, I(x)^{-1})$$

where

(4.3.3)

$$I(\lambda)^{-1} = \lambda^2/A(F_\lambda).$$

The regularity conditions apply to the total random censorship model, which includes the distribution of the censoring times $C_i (i=1,\ldots,n)$. This poses no real problem, as one can first condition on $C_i$ before taking expectations. See Cox and Hinkley (1975, 4.8).

The result (4.3.2) can also be obtained directly from the formula

$$\lambda_n = \Sigma \delta_i / n t_i.$$ The asymptotic normality of $n^2 (\Sigma \delta_i / n, \Sigma t_i / n)$ is proved by a
multivariate central limit theorem. The limiting distribution of \( n^{-1}(\hat{\lambda} - \lambda) \)
is then obtained by the "δ method" (Rao, 1965, 6a.2). Details are left to the reader.

According to the theory of von Mises derivative

\[
\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, E_{\lambda}^2(\text{IC}^2(x; w, F)).
\]

We now show that the limiting variances of (4.3.4) and (4.3.3) agree.

The I.C. (4.2.5) can be written

\[
\text{IC}(x; w, F_{\lambda}) = \frac{\delta - \lambda t}{B(F_{\lambda})}
\]

(4.3.5)

\[
= \frac{\lambda}{B(F_{\lambda})} \left( \frac{\delta - \lambda t}{\lambda} \right)
\]

\[
= \frac{\lambda}{B(F_{\lambda})} U_1(\lambda),
\]

where \( U_1(\lambda) \) is the efficient score defined at 4.1.3.

Therefore

\[
E_F(\text{IC}^2(x; w, F)) = \frac{\lambda^2}{B^2(F_{\lambda})} E(U_1^2(\lambda))
\]

(4.3.6)

\[
= \frac{\lambda^2}{B^2(F_{\lambda})} \bar{I}(\lambda) = \frac{\lambda^2}{B^2(F_{\lambda})} \frac{A(F_{\lambda})}{\lambda^2}
\]

\[
= \frac{A(F_{\lambda})}{\lambda^2} \bar{I}(\lambda)^{-1}.
\]

Therefore the two variances are identical.

4.4. Other Functions of \( \lambda \)

Suppose we are interested not only in estimating \( \lambda \) but in estimating a parameter \( \theta \) which is a 1-1 function of \( \lambda \): \( \theta = q(\lambda) \). The maximum likelihood estimate of \( \theta \) is then \( \hat{\theta}_n = q(\hat{\lambda}_n) \). We now show that the results for \( \hat{\lambda}_n \) extend in a straightforward way to \( \hat{\theta}_n \).

First, it is easy to see that \( \hat{\theta} \) is a von Mises functional defined by \( r(F_{\lambda}) = q(\theta) = q(w(F_{\lambda})) \). This functional is Fisher consistent:

\[
r(F_{\lambda}) = q(w(F_{\lambda})) = q(\lambda) = \theta.
\]
The influence curve for \( r(F) \) follows directly from that of \( w(F) \). Write \( r(\varepsilon) = q(w(\varepsilon)) \). Provided that \( q(\cdot) \) is differentiable, we have

\[
IC(x;r,F) = \frac{d}{d\varepsilon} r(\varepsilon) \bigg|_{\varepsilon=0}
\]

(4.4.1) \[
= q'(w(F)) \left( \frac{d}{d\varepsilon} w(\varepsilon) \bigg|_{\varepsilon=0} \right)
\]

= \( q'(w(F)) \ IC(x;w,F) \).

If the theory of von Mises derivatives is applied to \( \theta \), then

\[
\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0,E_F(IC^2(x;r,F)))
\]

where

(4.4.2) \[
E_F(IC^2(x;r,F)) = [q'(w(F))]^2 E_F(IC^2(x;w,F)).
\]

This is exactly the same result obtained from applying the "\( \delta \) method" to the function \( q(\lambda) \).

Example 4.4.1. Let \( \theta = 1/\lambda \), the mean of the exponential distribution (4.1.1). Here \( \hat{\theta} = 1/\hat{\lambda} = \Sigma t_i/\Sigma \delta_i \). Formally \( q(\lambda) = \lambda^{-1} \), and \( r(F) = w(F)^{-1} \). By (4.4.1)

\[
IC((t,\delta);r,F) = -\lambda^{-2} IC((t,\delta);w,F)
\]

(4.4.3) \[
= A(F)^{-1} (t-\theta\delta).
\]

As expected, a large observation time \( t \) will probably increase the estimate of \( \theta \). In form, the IC (4.4.2) for \( \hat{\theta} \) is unbounded as a function of \( t \). In practice, censoring will usually impose an upper limit, \( C_{\text{max}} \) say, on the observation times. Deviant survival times greater than \( C_{\text{max}} \) are "moved back" to \( C_{\text{max}} \) and have influence curve no greater than \( A(F)^{-1} C_{\text{max}} \). If only one failure time is observed in a sample of size \( n \), the rest being censored, \( \hat{\theta} \) will remain bounded. Therefore, the estimate does not break down easily, unlike the mean for uncensored observations.

Nonetheless, the bound \( A(F)^{-1} C_{\text{max}} \) may still be large, if \( C_{\text{max}} \) is. And, as Hampel (1968, p. 89) remarks in a similar context, the estimate \( \hat{\theta} \)
may be sensitive to distribution of failure times near $C_{\text{max}}$.

**Example 4.4.2.** Let $\theta = \log \lambda$, a location parameter for log failure time. The influence curve is easily found to be

$$IC(x; w, F) = IC(x; w, F)/\lambda$$

$$= \frac{\delta}{A(F)} - \frac{t}{B(F)}.$$
5. NONPARAMETRIC ESTIMATION OF AN ARBITRARY SURVIVAL FUNCTION

5.1. Note on Rank Estimators

In this chapter and in Chapters Seven and Eight, we consider estimators based on the rank ordering of observations. Some comment is needed to show how the rank representation is based on the sample $F_n$.

Let the data be $(t_i, \delta_i)$ or $(t_i, \delta_i, z_i)$, as $(i = 1, \ldots, n)$. The marginal distribution of $t_i$ in the sample, is

\[ F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I[t_i < t]. \]

If $t(i)$ is the $i$-th largest observation, then 5.1.1 means that

\[ F_n(t(i)) = (i-1)/n \]

and that

\[ 1 - F_n(t(i)) = (n-i+1)/n. \]

Here $F_n$ is defined to be continuous from the right, following Breslow and Crowley (1974).

Note that the rank of $t_i$ is usually defined by

\[ R_i = \sum_{j=1}^{n} I[t_j < t_i]. \]

Therefore

\[ 1 - F_n(t(i)) = (n-R_i+1)/n, \]

a fact used repeatedly.

5.2. The Product Limit and Empirical Hazard Estimators

When a parametric model is not specified, a nonparametric estimate of $G(s)$ is called for. A well known estimator is the product limit (P.L.)
estimate, discovered by Kaplan and Meir (1958). If there are no ties
among the failure times, the P.L. estimate is defined by

\[
\bar{G}^\text{PL}_n(s) = \prod_{i: t_i < s} \left( \frac{n-R_i}{n-R_i+1} \right) \delta_i
\]

(5.2.1)

\[
= \prod_{i: t_i < s} \left( 1 - \frac{1}{n-R_i+1} \right) \delta_i
\]

This is a step function, jumping at observed failure times and
constant between. The P.L. estimator has been studied by Efron (1967) and
Breslow and Crowley (1974), with extensions to multiple decrement models
by Peterson (1975) and Aalen (1976).

Peterson (1975) has succeeded in writing 5.2.1 as a functional of
\( F_n \). His representation is difficult to work with, so we consider instead
a nearly equivalent estimator of \( \bar{G}(s) \). This is the empirical hazard
estimator (Grenander, 1956; Altschuler, 1970; Breslow and Crowley, 1974):

(5.2.2) \[
\bar{G}^e_n(s) = \exp(-\Lambda^e_n(s))
\]

where

(5.2.3) \[
\bar{G}^e_n(s) = \sum_{i: t_i < s, \delta_i = 1} \left( \frac{1}{n-R_i+1} \right)
\]

estimates the cumulative hazard \( \Lambda(s) \). (Again the absence of ties is
assumed.) If the largest observation, \( t_{\text{max}} \), is a failure time, then
\( \Lambda^e_n(s) = +\infty \), for \( s > t_{\text{max}} \).

The empirical hazard estimate can be derived as the maximum likeli-
hood estimate of \( \bar{G}(s) \), assuming a constant hazard function between observed
failure times. Breslow and Crowley prove the following lemma relating
\( \Lambda^e_n(s) \) and \(-\log e^\{\bar{G}^\text{PL}_n(s)\}\).
Lemma: Let \( n(t) \) equal the number of individuals at risk of failure at \( t \).

Then,

\[
0 < -\log \left( \frac{n - \Pi_n(s)}{n \cdot n(s)} \right) < \frac{n(s)}{n \cdot n(s)}.
\]

Proof: Breslow and Crowley, 1974, Lemma 1.

The estimates are therefore close and are asymptotically equivalent.

5.3. Functional Form: Fisher Consistency under Random Censoring

We choose to work primarily with \( \Lambda_n(s) \); conclusions about \( \Lambda_n(s) \) will follow immediately.

From 5.1.5 and 5.2.3, we see that

\[
\Lambda_n(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i I_{t_i < s}}{1 - F_{n,t_i}}
\]

(5.3.1)

\[
= \int_0^s (1 - F_{n,t})^{-1} dF_n(t,1).
\]

This representation was first given by Breslow and Crowley (1974, eq. 7.6).

For general \( F \), we define the functional

\[
\Lambda(s,F) = \int_0^s (1 - F(t))^{-1} dF(t,1).
\]

Then

\[
\Lambda_n(s) = \Lambda(s,F_n).
\]

Fisher consistency under random censorship now follows, as Breslow and Crowley showed. For, suppose \( F \) is a random censorship distribution.

By 3.2.2

\[
dF(t,1) = \tilde{H}(t)g(t)dt.
\]

Also,

\[
1 - F(t) = \tilde{H}(t)\tilde{G}(t).
\]
Therefore

\[
\Lambda(s, F) = \int_0^s \frac{g(t)}{G(t)} \, \lambda(t) \, dt
\]

\[
= \Lambda(s).
\]

The estimator of \( \tilde{G}(s) \) in functional form is

\[
\tilde{G}_n(s) = \exp(-\Lambda(s, F_n)),
\]

and this is also Fisher consistent:

\[
\exp(-\Lambda(s, F)) = \tilde{G}(s).
\]

Breslow and Crowley actually proved the weak convergence of \( \Lambda(s, F_n) \) to \( \Lambda(s) \). This follows from continuity of \( \Lambda(s, F) \) in the supremum norm and from the convergence in the norm of \( F_n \) to \( F \).

5.4. The Influence Curve

We now find the I.C. of \( \Lambda(s, F) \). Write the contaminating point as \( x^* = (t^*, \delta^*) \), where the asterisks are temporarily added for clarity. Let

\[
F_\varepsilon = (1-\varepsilon)F + \varepsilon I_x, \quad x^* = F + \varepsilon I_x(1, F).
\]

The estimator based on \( F_\varepsilon \) of the cumulative hazard, is \( \Lambda(s, F_\varepsilon) \), and the I.C. will be given by

\[
(5.4.1) \quad \text{I.C.}(x^*; \Lambda(s, F)) = \left. \frac{d}{d\varepsilon} \Lambda(s, F_\varepsilon) \right|_{\varepsilon=0}.
\]

From 5.3.2

\[
(5.4.2) \quad \Lambda(s, F_\varepsilon) = \int I_{[y<s]} \delta(1-F_\varepsilon(y))^{-1} dF(y, \delta).
\]
We expand this:

\[ \Lambda(s, F_\varepsilon) = \int I_{[y \leq s]} \delta(1-F_\varepsilon(y))^{-1} dF(y, \delta) \]

\[ + \varepsilon \int I_{[t^* \leq s]} \delta^*(1-F_\varepsilon(t^*))^{-1} dF(y, \delta) \]

\[ - \varepsilon \int I_{[y \leq s]} \delta(1-F_\varepsilon(y))^{-1} dF(y, \delta). \]

Let us assume that we can differentiate with respect to \( \varepsilon \) under the integral sign in 5.4.3. Then

\[ I_{C((t^*, \delta^*); \Lambda(s, F_\varepsilon))} = \frac{d}{d\varepsilon} \Lambda(s, F_\varepsilon) \bigg|_{\varepsilon=0} \]

\[ = \int I_{[y \leq s]} \delta \left( \frac{d}{d\varepsilon} (1-F_\varepsilon(y))^{-1} \bigg|_{\varepsilon=0} \right) dF(y, \delta) \]

\[ + \int I_{[t^* \leq s]} \delta^*(1-F_\varepsilon(t^*))^{-1} \Lambda(s, F_\varepsilon), \]

since

\[ F_0 = F. \]

We must evaluate \( \frac{d}{d\varepsilon} (1-F_\varepsilon(y))^{-1} \):

\[ \frac{d}{d\varepsilon} (1-F_\varepsilon(y))^{-1} \bigg|_{\varepsilon=0} = -(1-F(y))^{-2} \frac{d}{d\varepsilon} (1-F_\varepsilon(y)) \bigg|_{\varepsilon=0} \]

\[ = +(1-F(y))^{-2} \frac{d}{d\varepsilon} F(y) \bigg|_{\varepsilon=0}. \]

The last derivative is

\[ \frac{d}{d\varepsilon} \left[ F(y) + \varepsilon (I_{[y>t^*]} - F(y)) \right] = I_{[y>t^*]} - F(y). \]

The indicator in 5.4.6 is the distribution function with mass 1 at \( t^* \), evaluated at \( y \).

\[ I_{t^*}(y) = \begin{cases} 0 & y \leq t^* \\ 1 & y > t^* \end{cases}, \]

or

\[ I_{t^*}(y) = I_{[y>t^*]}. \]
Incorporating 5.4.6 and 5.4.5 into 5.4.4, we find,

\[
\mathcal{I}C((t^*, \delta^*); \Lambda(s,F)) = \int_0^s (1-F(y))^{-2} I_{[y>t^*]} dF(y,1) \\
- \int_0^s (1-F(y))^{-2} F(y) dF(y,1) \\
+ \int_{[t^*<s]} \delta^* (1-F(t^*))^{-1} \\
- \Lambda(s,F).
\]

(5.4.9)

This can be simplified, if we add and subtract from the first two terms the following expression:

\[
\min(s,t^*) D = \int_0^s (1-F(y))^{-2} dF(y,1)
\]

(5.4.10)

\[
= \int_0^s (1-F(y))^{-2} I_{[y\leq t^*]} dF(y,1).
\]

Then the sum of the first two terms in 5.4.9 is

\[
\int_0^s (1-F(y))^{-2} (I_{[y\leq t^*]} + I_{[y>t^*]} - F(y)) dF(y,1) - D
\]

(5.4.11)

\[
= \int_0^s (1-F(y))^{-2} (1-F(y)) dF(y,1) - D
\]

\[
= \Lambda(s,F) - D.
\]

Therefore, we can rewrite the I.C. 5.4.9, dropping the asterisks, as

\[
\mathcal{I}C((t, \delta); \Lambda(s,F)) = -\int_0^{\min(t,s)} (1-F(y))^{-2} dF(y,1)
\]

(5.4.12)

\[
+ \delta I_{[t\leq s]} (1-F(t))^{-1}.
\]
The empirical I.C. of $\Lambda(s,F_n)$ is found by substituting $F$ for $F_n$ in 5.4.12:

$$IC((t,\delta);\Lambda(s,F_n)) = -\frac{1}{n} \sum_{i=1, t_i \leq \text{min}(s,t)} n_i (1-F_n(t_i))^{-2}$$

(5.4.13)

$$+ \delta \prod_{t \leq s} (1-F_n(t))^{-1}.$$

In Appendix A1, it is shown that $E_F(IC(x;\Lambda(s,F))) = 0$, for differentiable $F$, and that the empirical I.C. sums to zero over the sample. Note that 5.4.12 and 5.4.13 will be undefined for any $t$ such that $F(t) = 1$. We therefore agree to define the estimators only for $s$ such that $F(s) < 1$.

The influence curve for

$$G(s,F) = \exp{-\Lambda(s,F)}$$

(5.4.14)

the estimator of $\tilde{G}$ based on $F$, follows easily from 5.4.12:

$$IC(x;G(s,F)) = \frac{d}{d\epsilon} \exp{-\Lambda(s,F_{\epsilon})} \bigg|_{\epsilon=0}$$

(5.4.15)

$$= -G(s,F) \frac{d}{d\epsilon} \Lambda(s,F_{\epsilon}) \bigg|_{\epsilon=0}$$

$$= -G(s,F) \prod_{t \leq s} (1-F(t))^{-1}.$$

In particular, when $F=F_n$, we obtain the empirical I.C. of $G_n(s)$:

$$IC((t,\delta);G(s,F_n))$$

(5.4.16)

$$= \prod_{t \leq s} (1-F_n(t))^{-1} + \frac{1}{n} \sum_{i=1, t_i \leq \text{min}(s,t)} (1-F_n(t_i))^{-2}.$$

5.5. Discussion

Let us examine the empirical I.C. for $G_n(s)$ (5.4.16) to see how a new point $(t,\delta)$ affects the estimator. We consider separately the cases $\delta=0$ and $\delta=1$. 
Case 1: $\delta = 0$.

When $(t, 0)$ is added to the sample, the I.C. is

\[(5.5.1) \quad \bar{G}_n^e(s) \frac{1}{n} \left\{ \sum_{i: \delta_i = 1, t_i < \min(s, t)} (1 - F_n(t_i))^{-2} \right\} \]

For $t$ less than the smallest sample failure time, the I.C. is zero, implying no change in the estimate. This is also apparent by reference to 5.2.2 and 5.2.3.

Suppose there are $m$ failure times, ordered $t(1) < t(2) < \ldots < t(m)$. Then as $t$ moves past $t(j)$, for $t(j) < s$, the I.C. is increased by

\[(5.5.2) \quad \bar{G}_n^e(s) \frac{1}{n} (1 - F_n(t(j)))^{-2} \]

Each term 5.5.2 is larger than the previous ones, suggesting that larger censored observations have a stronger influence than smaller ones. This larger influence is clearly related to the diminishing number at risk, expressed by increasing $F_n(t)$.

As soon as $t > s$, the I.C. does not add any more terms but instead remains constant. This constant influence indicates that $\bar{G}_n^e(s)$ is not affected badly by outliers greater than $s$.

Case 2: $\delta = 1$.

When a failure time $(t, 1)$ is added to the sample, the picture changes dramatically. Now there is a negative contribution to the I.C. from

\[(5.5.3) \quad -\bar{G}_n^e(s) I_{t \leq s} (1 - F_n(t))^{-1} \]

For $t < t(1)$, the negative contribution of 5.5.3 constitutes the entire I.C. As $T$ moves past the sample failure times 5.5.3 grows larger in absolute value, but its contribution is offset by 5.5.1. The tradeoff continues until $t > s$; at that point the contribution of 5.5.3, which had
been increasing, becomes zero. Thereafter the observation has the same influence as a censored time greater than \( s \).

The lesson in all this is fairly clear and expected. Failure times tend to decrease the estimated probability of survival. Late failure times and censored times are influential because the risk sets have diminished in size.

5.6. The Squared I.C.

Suppose \( 0 < s < u \). Define the functional

\[
\Lambda(s, u, F) = \begin{bmatrix} \Lambda(s, F) \\ \Lambda(u, F) \end{bmatrix}.
\]

Then

\[
\Lambda(s, u, F) = \begin{bmatrix} \Lambda^e_n(s) \\ \Lambda^e_n(u) \end{bmatrix} = \Lambda^e(s, u), \text{ say.}
\]

For arbitrary \( F \), the influence curve of 5.6.1 is

\[
\text{IC}(x; \Lambda(s, u, F)) = \begin{bmatrix} \text{IC}(x; \Lambda(s, F)) \\ \text{IC}(x; \Lambda(u, F)) \end{bmatrix}
\]

Now assume that the observations \( x_i = (t_i, \delta_i), i = 1, \ldots, n \) are distributed according to \( F(t, \delta) \), not necessarily a random censorship distribution. According to the theory of von Mises derivatives (2.3.14)

\[
\sqrt{n}(\Lambda^e_n(s, u) - \Lambda(s, u, F)) \xrightarrow{d} N(0, \Sigma(F))
\]

where

\[
\Sigma(F) = E_F\{\text{IC}(x; \Lambda(s, u, F))\text{IC}(x; \Lambda(s, u, F))^T\}.
\]
Explicitly, the elements of $a(F)$ are

$$a_{11} = E_F \{ IC^2(x; \Lambda(s,F)) \}$$

(5.6.6)

$$a_{22} = E_F \{ IC^2(x; \Lambda(u,F)) \}$$

$$a_{12} = a_{21} = E_F \{ IC(x; \Lambda(s,F)) IC(x; \Lambda(u,F)) \}.$$ 

Breslow and Crowley (1974) provide a rigorous proof of the limiting normality of $\sqrt{n}(\Lambda^e_n(s,u) - \Lambda(s,u,F))$. Their proof (which does not depend on the von Mises theory) holds if (a) $F(t) < 1$ if $t < \infty$ and (b) $F(t)$ and $F(t,1)$ are continuous. The covariance matrix of the limiting normal process is

$$\begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix}$$

where for $s \leq u$, $w_{11} = w_{12} = w_{21} = W(s)$, defined by

$$W(s) = \int_0^s (1 - F(t))^{-2} dF(t,1),$$

and $w_{22} = W(u)$. If a random censoring model holds, condition (b) above is replaced by (b') $G$ and $H$ are continuous. In this case, the covariance function $w(s)$ becomes

$$W(s)^* = \int_0^s (1 - F(T))^{-1} G(T)^{-1} dG(T).$$

For the result of Breslow and Crowley to agree with the theory of von Mises derivatives, we must show $A(F) = W$, or for $s \leq u$

(5.6.8)

$$E_F \{ IC^2(x; \Lambda(s,F)) \} = W(s)$$

(5.6.9)

$$E_F \{ IC(x; \Lambda(s,F)) IC(x; \Lambda(u,F)) \} = W(s).$$

We now prove (5.6.8), assuming that $F$ is differentiable. The proof of (5.6.9) is similar and is therefore omitted.
We start by expanding the IC (5.4.12)

\[ \text{IC}((t,\delta),\Lambda(s,F)) = - \int_0^{\min(t,s)} (1-F(y))^{-2} \, dF(y,1) \]

\[ + \delta I_{[t<s]} (1-F(t))^{-1} \]

\[ = -I_{[t<s]} \int_0^t (1-F(y))^{-2} \, dF(y,1) \]

\[ -I_{[t>s]} \int_0^s (1-F(y))^{-2} \, dF(y,1) \]

\[ + \delta I_{[t<s]} (1-F(t))^{-1} \]

\[ = -I_{[t<s]} B(t) \]

\[ -I_{[t>s]} B(s) \]

\[ + \delta I_{[t<s]} (1-F(t))^{-1} \]

where

\[ B(t) = \int_0^t (1-F(y))^{-2} \, dF(y,1). \]

Squaring the IC, we have

\[ \text{IC}^2((t,\delta);\Lambda(s,F)) = I_{[t<s]}^2 B^2(t) + I_{[t<s]}^2 \delta^2 (1-F(t))^{-2} \]

\[ + I_{[t>s]}^2 B^2(s) - 2 \delta I_{[t<s]} I_{[t<s]} B(t)(1-F(t))^{-1} \]

\[ = -2 I_{[t<s]} I_{[t>s]} B(t)B(s) \]

\[ - 2 \delta I_{[t>s]} I_{[t<s]} B(s)(1-F(t))^{-1}. \]

Because \( I_{[t>s]} I_{[t<s]} \equiv 0 \), the last two cross products are zero. The squared I.C. therefore reduces to:
\[
\text{IC}^2((t,\delta);\Lambda(s,F)) = \int_{t<s} B^2(t) + \int_{t<s} \delta(1-F(t))^{-2} \]
\[ + \int_{t>s} B^2(s) - 2 \delta \int_{t<s} B(t)(1-F(t))^{-1}. \tag{5.6.14} \]

Taking the expectation of these terms, we find
\[
\int \text{IC}^2((t,\delta);\Lambda(s,F)) \ dF(t,\delta) = \int_0^s B^2(t) \ dF(t) + \int_0^s (1-F(t))^{-2} \ dF(t,1) \]
\[ + B^2(s)(1-F(s)) - 2 \int_0^s B(t)(1-F(t))^{-1} \ dF(t,1). \tag{5.6.15} \]

The second integral on the r.h.s. of 5.6.15 can be recognized as \( W(s) \). Let the sum of the remaining terms be denoted \( Q(s) \). Then 5.5.8 will hold if \( Q(s) = 0 \ \forall \ s > 0 \).

\[
Q(s) = \int_0^s B^2(t) \ dF(t) + B^2(s) - B(s)F(s) \]
\[ - 2 \int_0^s B(t)(1-F(t))^{-1} \ dF(t,1). \tag{5.6.16} \]

Now differentiate both sides of 5.6.16 with respect to \( s \). Because \( B(0) = 0 \), we may neglect terms multiplied by \( B(0) \).

Then
\[
Q'(s) = B^2(s) \frac{dF(s)}{ds} + 2B(s)B'(s) - B^2(s) \frac{dF(s)}{ds} - 2B(s)B'(s)F(s) - 2B(s)(1-F(s))^{-1} \frac{dF(s,1)}{ds} \]
\[ = 2B(s)\left[B'(s)(1-F(s)) - (1-F(s))^{-1} \frac{dF(s,1)}{ds}\right] \tag{5.6.17} \]

Because \( B'(s) = (1-F(s))^{-2} \frac{dF(s,1)}{ds} \), the term in brackets is zero, and \( Q'(s) = 0 \ \forall \ s > 0 \).
This implies $Q(s)$ is constant. With the immediate side condition $Q(0) = 0$, we must have $Q(s) \equiv 0$, proving 5.6.8.

The asymptotic normality of $\sqrt{n}(\bar{G}_n^e(s) - \bar{G}(s))$ follows from application of the δ-method to $\bar{G}_n^e(s) = \exp(-A_n^e(s))$. The limiting variance of $\sqrt{n} \bar{G}_n^e(s)$ is given by

$$(5.6.19) \quad V(s) = (1-F(s))W(s).$$

As shown in 4.4, this is identical to $E_{\bar{F}} \{I_C^2(x; G(s,F))\}$. 


6. AN EXPONENTIAL REGRESSION MODEL

6.1. Introduction

In this chapter we encounter the first of the two regression models we will study. This model is exponential, with hazard function depending on the covariates but not on time.

(6.1.1) \[ \lambda^\beta(t|z) = \exp(\beta^T z). \]

Here \( \beta \) is a \( p \times 1 \) vector of regression coefficients, including intercepts. (Recall upper case "T" denotes a transpose.)

Model (6.1.1) was proposed by Feigl and Zelen (1965), who gave the likelihood equations for uncensored data with one intercept and one continuous covariate. Glasser (1967) wrote down the likelihood equations for censored data with several intercepts. This work was in turn extended by Breslow (1972a, 1974) to multiple covariates. Prentice (1973) treated the problem from the viewpoint of structural inference.

The likelihood equations below are slightly different from those appearing in earlier work. In an analysis of covariance with parallel lines, the estimated treatment effects (intercepts) can be written in terms of the other estimated parameters. The likelihood equations and the observed information matrix are thereby reduced in dimension and changed in form. In this chapter, we retain full generality and do not separate treatment intercepts from other coefficients.

6.2. Likelihood Equations

The density and survival functions corresponding to the hazard 6.1.1 are
(6.2.1) \[ q_{\beta}(t|z) = \exp(\beta^T z) \exp(-\exp(\beta^T z)t) \]

and

(6.2.2) \[ \tilde{g}_{\beta}(t|z) = \exp(-\exp(\beta^T z)t), \]

respectively.

From 3.2.5 the contribution of a point \( x_i = (t_i, \delta_i, z_i) \) to the log likelihood is

\[
\ell(\beta; x_i) = \log \left[ \frac{q_{\beta}(t_i|z_i)^{\delta_i} \tilde{g}_{\beta}(t_i|z_i)^{1-\delta_i}}{\tilde{g}(\delta_i, x_i)} \right]
\]

(6.2.3)

\[
= \beta^T z_i \delta_i - \exp(\beta^T z_i t_i).
\]

The efficient score for the observation is then (utilizing vector notation)

\[
U(\beta; x_i) = \frac{\partial}{\partial \beta} \ell(\beta; x_i)
\]

(6.2.4)

\[
= z_i (\delta_i - \exp(\beta^T z_i t_i)).
\]

Based on the sample \( x_1, x_2, \ldots, x_n \), the M.L.E. is the solution to the

\[
U(\hat{\beta}; x_i) = \sum_{i=1}^{n} z_i (\delta_i - \exp(\beta^T z_i t_i)) = 0.
\]

(6.2.5)

The equations are usually solved by a Newton Raphson procedure. The observed second derivative of \( \ell(\beta; x_i) \) is

\[
C(\beta, x_i) = \frac{\partial^2}{\partial \beta \partial \beta} U(\beta, x_i)
\]

(6.2.6)

\[
= -z_i z_i^T \exp(\beta^T z_i t_i).
\]

The observed average information matrix is therefore,

\[
\hat{I}_{\text{avg}} = \frac{1}{n} \sum_{i=1}^{n} C(\beta, x_i) = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \exp(\beta^T z_i t_i).
\]
The actual average information is:

$$E(U(\beta; x_i) U(\beta, x_i)^T)$$

(6.2.8) 

$$I_{\text{avg}}(\beta) = E(z_i z_i^T \exp(\beta^T z_i) t_i),$$

if the $z$'s are regarded as random variables, or

(6.2.9) 

$$I_{\text{avg}}(\beta) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \exp(\beta^T z_i) E(t_i | z_i)$$

in general.

According to standard likelihood theory for nonidentically distributed variables (Cox and Hinkley, 1975, Sec. 9.2).

(6.2.10) 

$$\sqrt{n}(\beta - \beta_0) \overset{d}{\to} N(0, I^{-1}_{\text{avg}}(\beta)),$$

assuming the model is true. In practice $I^{-1}_{\text{avg}}(\beta)$ is estimated by $I^{-1}_{\text{avg}}(\hat{\beta}_n)$.

6.3. **The estimator as a von Mises Functional, Fisher Consistency**

We now show that $\hat{\beta}_n$ is implicitly defined as a functional of $F_n$.

The proof is similar to that in the multiple regression example in 2.1; we simply divide the likelihood equations 6.2.5 by $n$:

(6.3.1) 

$$\frac{1}{n} U_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} z_i (\delta_i - \exp(\beta^T z_i) t_i) = 0$$

or

(6.3.2) 

$$\int z(\delta - \exp(\beta^T z) t) dF_n(t, \delta, z) = 0.$$

For the formal approach, we define

(6.3.3) 

$$U(\beta, F) = \int z(\delta - \exp(\beta^T z) t) dF$$

where $\beta$ is $p \times 1$ and $F$ is an arbitrary distribution function.

Then the Glasser estimator based on $F$ is defined to be that value of $\beta$, such that
assuming a unique solution exists. For the solution to 6.3.4, write

\[ \hat{\beta}(F) = \beta(F) \]

where \( \hat{\beta}(\cdot) \) is a functional. Then \( \beta(F) \) is defined by

\[ U(\hat{\beta}(F), F) = \int z(\delta - \exp(\hat{\beta}(F)^T z)) t \, dF = 0. \]

The finite sample M.L.E. is just \( \hat{\beta}_n = \beta(F_n) \), and the likelihood equations 6.3.2 are

\[ U(\hat{\beta}_n, F_n) = 0. \]

Suppose now that the random censorship model with covariates (see 3.2) governs the distribution of the data \( x = (t, \delta, z) \). Failure time is assumed to follow the exponential regression model 6.1.1, and we label the resulting distribution \( F_{\beta} \).

By the definition 6.3.6 of the estimator \( \hat{\beta}(F) \), Fisher Consistency will hold if

\[ U(\hat{\beta}, F_{\beta}) = \int z(\delta - \exp(\hat{\beta}^T z)) t \, dF_{\beta} = 0. \]

But 6.3.8 is just the condition

\[ E_{F_{\beta}}(U(\hat{\beta}, F_{\beta})) = 0, \]

which follows from a standard result for regular likelihood problems (Cox and Hinkley, 1975, 4.8).

We will also find it helpful to define the average information matrix 6.2.8 as a functional of \( F \):

\[ I_{\text{avg}}(\hat{\beta}, F) = \int z \, z^T \exp(\hat{\beta}^T z) t \, dF. \]

Then the reader can easily check that the observed sample information matrix \( I_{\text{avg}}(\hat{\beta}_n) \) (6.2.7) is, in functional form, written \( I_{\text{avg}}(\hat{\beta}(F_n), F_n) \).
6.4. The Influence Curve

We contaminate an arbitrary distribution \( F \) with a point \( x^* = (t^*, \delta^*, z^*) \)
and write

\[
F_\varepsilon = (1-\varepsilon)F + \varepsilon I_{x^*}.
\]

Let \( \beta(F_\varepsilon) \) be the regression estimator based on \( F_\varepsilon \). By 6.3.6, \( \beta(F_\varepsilon) \)
is defined by the relation

\[
U(\beta(F_\varepsilon), F_\varepsilon) = 0.
\]

For notational convenience, we write

\[
\beta(\varepsilon) = \beta(F_\varepsilon),
\]

Then

\[
\beta(0) = \beta(F_0) = \beta(F).
\]

Recall that the influence curve of the estimate is given by

\[
IC(x^*; \beta, F) = \left. \frac{d}{d\varepsilon} \beta(\varepsilon) \right|_{\varepsilon=0},
\]

where the differentiation is coordinate-by-coordinate (2.3.5). We will find the I.C. by implicit differentiation of the estimating equation 6.4.1.

Expanding \( F_\varepsilon \), we find

\[
U(\beta(\varepsilon), F_\varepsilon) = \int z \{ \delta - \exp(z^T \beta(\varepsilon)) \} dF_\varepsilon
\]

\[
= \int z \delta \exp(z^T \beta(\varepsilon)) t dF
\]

\[
+ \varepsilon \int z \{ \delta \exp(z^T \beta(\varepsilon)) t \}
\]

\[
- \varepsilon \int z \{ \delta \exp(z^T \beta(\varepsilon)) t \} dF.
\]
Or

\[ 0 = U(\beta(\varepsilon), F) \]

(6.4.5)

\[ = U(\beta(\varepsilon), F) + \varepsilon [z^* \{ \delta^* - \exp(z^T \beta(\varepsilon) t^*) \} - U(\beta(0), F). \]

The first term on the r.h.s. can be evaluated by vector calculus if we differentiate under the integral sign.

\[
\frac{d}{d\varepsilon} U(\beta(\varepsilon), F) \bigg|_{\varepsilon=0} = \frac{d}{d\varepsilon} \left[ \int z(\delta^* - \exp(z^T \beta(\varepsilon) t) \right] dF
\]

(6.4.6)

\[
= - \left[ z(\frac{d}{d\varepsilon} \exp(z^T \beta(\varepsilon) t) \right]_{\varepsilon=0} dF
\]

\[
= - \left[ z z^T \exp(z^T \beta(0) t) dF \frac{d}{d\varepsilon} \beta(\varepsilon) \bigg|_{\varepsilon=0} \right]
\]

\[
= - I_{\text{avg}}(\beta(0), F) \text{ IC}(x^*; \beta, F)
\]

The third term on the r.h.s. of (6.4.5) is \( U(\beta(0), F) = 0 \) by definition of \( \beta(0) \) 6.4.1.

Thus

\[ 0 = - I_{\text{avg}}(\beta(F), F) \text{ IC}(x^*; \beta, F) \]

(6.4.7)

\[ + z^* \{ \delta^* - \exp(z^T \beta(F) t^*) \}. \]

Now we drop the asterisks from \( x^* = (t^*, \delta^*, z^*) \), and solve (6.4.7) for the I.C.

\[
\text{IC}((t, \delta, z); \beta, F)
\]

(6.4.8)

\[ \text{p} \times 1
\]

\[ = I_{\text{avg}}^{-1}(\beta(F), F) \{ z(\delta - \exp(z^T \beta(F) t) \},
\]

\[ \text{p} \times \text{p} \]

\[ \text{p} \times 1
\]

where a generalized inverse may be used.
We recognize the I.C. as

\[(6.4.9) \quad \text{IC}(x; \beta, F) = I_{\text{avg}}^{-1}(\beta, F) U(\beta; x)\]

where \(U(\beta, x)\) is the efficient score 6.2.4 based on \(x = (T, \delta, \bar{z})\).

The empirical I.C. follows by substituting \(F\) for \(F_n\):

\[(6.4.10) \quad \text{IC}((t, \delta, \bar{z}); \beta, F_n) = \hat{I}_{\text{avg}}^{-1}(\beta, F_n) \{z(\delta - \exp(\bar{z}^T \beta) t)\}.\]

We can easily show that the basic property 2.3.9 holds for the I.C.

6.4.6. That is,

\[(2.3.9) \quad \int \text{IC}(x; \beta, F) dF = 0,\]

for arbitrary \(F\). For, integrating 6.4.6 over \(F\), we find

\[(6.4.11) \quad = \hat{I}_{\text{avg}}^{-1}(\beta, F) \{z(\delta - \exp(\bar{z}^T \beta) t)\}.\]

The proof applies to \(F = F_n\), so that also

\[(6.4.12) \quad \sum_{i=1}^{n} \text{IC}(x; \beta, F) = 0.\]

6.5. The Squared I.C. Under Random Censorship

Suppose that the model 6.1.1 is true and that random censorship holds.

The asymptotic covariance matrix of \(\hat{\beta}_n\) given by likelihood theory (6.2.10) is \(\hat{I}_{\text{avg}}^{-1}(\beta, F)\). From the theory of von Mises derivatives (2.3.14), we therefore expect:
But by 6.4.9 the l.h.s. is equal to

\[
\left. I^{-1}_{\text{avg}}(\beta, F_{\beta}) \right| E_F \left\{ (\beta, x) \right\} = I^{-1}_{\text{avg}}(\beta, F_{\beta})
\]

(6.5.2)

= \left. I^{-1}_{\text{avg}}(\beta, F_{\beta}) \right| I^{-1}_{\text{avg}}(\beta, F_{\beta}) \left. I^{-1}_{\text{avg}}(\beta, F_{\beta}) \right|

= \left. I^{-1}_{\text{avg}}(\beta, F_{\beta}) \right|

6.6. Discussion

Recall from 2.3.8 that for \( x = (t, \delta, z) \),

(6.6.1)

\[
(n+1) \left[ \hat{\beta}^{+x}_{n+1} - \hat{\beta}^{+x}_{n} \right] = IC((t, \delta, z); \beta, F_{\beta})
\]

where \( \hat{\beta}^{+x}_{n+1} \) is the estimated based on \( x_1, x_2, \ldots, x_n \) and \( x \). From this representation, we can draw some conclusions about the way \( \hat{\beta} \) is affected by individual data points:

1. The I.C. is continuous as a function of \( t \), for \( t < C_{\text{max}} \), the maximum possible observation time. The estimator therefore depends (for fixed \( z \)) strongly on the value of \( t \).

2. The term

\[
R = \{ \delta - \exp(z^T \beta) t \}
\]

multiplies all entries in the I.C.

For any random censorship model \( F \), it can be shown that

\[
E_F(\delta) = E_F(\Lambda(t | z)),
\]

where \( \Lambda(t | z) \) is the cumulative hazard 1.1.5. of the failure time. In the exponential model, \( \Lambda(t | z) = \exp(z^T \beta) t \); therefore

\[
E_F(\delta) = E_F(\exp(z^T \beta) t).
\]

Now, \( \delta \) can be considered the observed number of failures (0) in the observation \( (t, \delta, z) \), and \( \exp(z^T \beta) t \) may be regarded as the expected
number of failures (E) in the observation. Therefore

\[ R = \delta - \exp(z^T \hat{\beta}) t \]

= Observed - Expected = O-E,

is a kind of residual associated with the observation. (This is the viewpoint of Peto and Peto, 1972, p. 193.)

A negative residual large in absolute value can arise when \( \exp(z^T \hat{\beta}) >> 1 \) (implying early failure) and instead a large t is observed. A large positive residual can occur when a small predicted hazard function \( \exp(z^T \hat{\beta}) \) is coupled with an early observed failure. Though the residuals sum to zero over the sample (6.4.12), their distribution will be highly skewed.

3. The residual O-E is multiplied by z. Therefore the influence of a contaminating observation depends on the relative size of z, apart from the size of the residual. This phenomenon was encountered in the multiple regression example at 2.3.19.

4. Note the effect of \( I^{-1}_{\text{avg}} (\hat{\beta},F) \) in 6.6.1. An extreme value in one of the coordinates of z (say the k-th) affects not only \( z_k \) but the other coefficients as well. The strength of the influence of \( z_k \) on \( \hat{\beta}_j \) depends on the size of the \((k,j)\)-th element of \( I^{-1}_{\text{avg}} (\hat{\beta},F) \). This element is an estimate of the asymptotic covariance of \( \sqrt{n} \hat{\beta}_j \) and \( \sqrt{n} \hat{\beta}_k \).

5. With the interpretation of the residual in remark 2, the I.C. 6.4.10 shows a remarkable correspondence to that for the least squares estimator in multiple linear regression (2.3.21). In both

\[ I_C(x;\hat{\beta},F) = A z R, \]

where \( A \) is a \( p \times p \) matrix proportional to the estimated covariance matrix of \( \hat{\beta} \); \( z \) is the \( p \times 1 \) vector of covariates for the contaminating data point; and R is a scalar residual, appropriately defined. The I.C.'s in both
cases depend continuously on the response \((y \text{ or } t)\) and on \(z\). We conclude that the exponential estimator will, like its least squares counterpart, be sensitive to outliers in the data.
7. COX'S ESTIMATOR

7.1. Introduction

In many applications prognostic factors may vary with time.

Example 7.1.1. If a patient undergoes a series of treatment, define a covariate \( z(t) \), indicating the treatment current at \( t \). Or \( z(t) \) may indicate the whole sequence of treatments up to \( t \).

Example 7.1.2. The covariate might be the current value of some biological variate: white blood count, number of cigarettes smoked/day, tumor size, cholesterol, and so on.

Where such variables act through cumulative exposure, one might define \( z(t) \) as the "effective dose" at \( t \). Such a dose might be measured as a weighted average of past exposure levels.

With such time dependent variables, D. R. Cox (1972) introduced the proportional hazards (P.H.) model:

\[
\lambda_\beta(t|z(t)) = \exp(B^Tz(t))\lambda_0(t).
\]

Here \( \lambda_0(t) \) is an arbitrary unspecified underlying hazard function. Cox introduced likelihood methods for estimating \( \beta \), and in this chapter we consider Cox's estimator. Estimation of the underlying distribution is of interest in its own right and some estimates will be examined in Chapter 8.

In the next section, we further develop the model 7.1.1., including an extension of random censorship. Section 7.3 presents the likelihood equations for estimating \( \beta \). In 7.4 we briefly state some conditions under which the likelihood equations have no finite solution. Section 7.5 contains the demonstration that the estimator \( \hat{\beta}_n \) is a von Mises functional. In 7.6
Fisher consistency under random censorship is proved when the covariates do not depend on time. The influence curve of \( \hat{\beta} \) and related theory are contained in Sections 7.7 and 7.8. The chapter closes with some numerical and Monte Carlo studies of the effects of outliers.

7.2. The P.H. Model

Let us define the covariate \( z \) as a function of time.

Definition 7.2.1. A time dependent covariate \( z(p \times 1) \) is defined by

\[
z = \{z(t) : t \geq 0\},
\]

where \( z(t) \) is a real function of \( t \). We assume \( z \) is an element of a (measurable) space \( Z \).

With time dependent \( z \), define the integrated hazard from 7.1.1.

\[
\Lambda_{\beta}(t|z) = \int_{0}^{t} \lambda_{\beta}(y|z) \, dy
\]

(7.2.1)

\[
= \int_{0}^{t} \exp(\beta^T_z(y)) \lambda_0(y) \, dy.
\]

The survival function is

\[
\tilde{G}_{\beta}(t|z) = \exp\{-\Lambda_{\beta}(t|z)\}.
\]

(7.2.2)

If \( z \) does not depend on time, this simplifies to:

\[
\tilde{G}_{\beta}(t|z) = \exp\{-\exp(\beta^T_z z) \Lambda(t)\}
\]

(7.2.3)

\[
= \tilde{G}(t) \exp(\beta^T_z z)
\]

where

\[
\Lambda(t) = \int_{0}^{t} \lambda_0(y) \, dy \text{ and } \tilde{G}(t) = \exp\{-\Lambda(t)\},
\]

(7.2.4)

are the underlying integrated hazard and survival functions.
Again, with time dependent \( z \), the density in the P.H. model is

\[
g_\beta(t|z) = \frac{-d}{dt} \bar{g}_\beta(t|z) = \lambda_\beta(t|z(t)) \bar{g}_\beta(t|z).
\]

With \( z \) a function of time, we can again define each observation as \( x_i = (t_i, \delta_i, z_i) \), \( i = 1, \ldots, n \).

The fact that \( z \) is a function of time creates a host of potential problems for the results in this chapter: definition of \( F(t, \delta, z) \), measurability, the existence of expectations, and so on. We adopt the following compromise: (1) All proofs will be valid, with conditions stated, when \( z \) is not a function of time; (2) when \( z \) is a function of time, we regard our methods as heuristic, the manipulations as formal only.

As examples of the "heuristic" method, we "define" the sample empirical d.f. \( F_n(t, \delta, z) \) to be that distribution function which puts mass \( \frac{1}{n} \) at each observed sample point \( x_i = (t_i, \delta_i, z_i) \). We also will "define" arbitrary distributions \( F(t, \delta, z) \) and write down expectations over these distributions.

**Random Censorship**

Our formal manipulations with time dependent \( z \) will not be extended to the random censorship model. For example, our proof of Fisher consistency for \( \hat{\beta} \) will be carried out only for time independent covariates; the expectation of the squared I.C. will be found under the same conditions.

Let us, however, briefly indicate how the random censorship model could be extended to the case of time dependent covariates. We define a Borel Field \( B \) on the space \( Z \) of the functions, \( \tilde{z} = \{z(t)\} \). Let \( P(\cdot) \) be a probability measure on \( B \), so that \((Z, B, P)\) is a probability space. With a suitable metric on \( B \), let \( K(\tilde{z}) \) be a distribution function corresponding to \( P \). Then observations \( x = (t, \delta, \tilde{z}) \) are formed in the usual way: Choose \( \tilde{z} \) according to \( K \) (or \( P \)); contingent on \( \tilde{z} \), choose failure and censoring times.
S and C; let \( t = \min(S,C) \); and set \( \delta \).

This model has promise, but we will not pursue it any further. To summarize the strategy for this chapter, formal manipulations of \( F(t,\delta,z) \) for time dependent covariables will be carried out only in Section 7.5 (showing \( \hat{\beta} \) is a von Mises functional) and in Section 7.7 (finding the influence curve). In other sections—involving random censorship—the proofs involving \( F(t,\delta,z) \) will be confined to variates which do not depend on time.

7.3. Cox's Likelihood

Suppose the sample consists of \( (t_i,\delta_i,z_i) \), \( i = 1, \ldots, n \), where now \( z_i = \{z_i(t) \mid 0 < t < t_i \} \). (We do not observe \( z_i(t) \), if \( t > t_i \)).

Definition 7.3.1. The risk set \( R(t) \) at time \( t \) is the set of labels for individuals still at risk of failing (observed) at time \( t \). That is:

\[
R(t) = \{i \mid t_i > t \}.
\]

Suppose now there are \( m \) distinct failure times in the sample, at \( t(1) < t(2) < \ldots < t(m) \). By convention the \( t(j) \), \( j = 1, \ldots, m \) are failure times, a subset of the \( t_i \), \( i = 1, \ldots, n \), the observation times. The label of the \( (j) \)-th failure is \( (j) \).

In what follows, we assume no ties among failure times. Modifications to Cox's likelihood in case of ties have been suggested by Cox (1972), Peto and Peto (1972b) and Efron (1977), Kalbfleisch and Prentice (1973, and Breslow (1975); Breslow's approach will be reviewed in Chapter Eight, in conjunction with his estimate of the underlying distribution.

Assuming no ties, Cox argued as follows: With the P.H. model 7.1.1 and conditional on (a) the risk set \( R(t(j)) \) and (b) the fact that a failure occurs at \( t(j) \), the probability that the observed failure was \( (t(j),z(j)) \) is the ratio of hazards:
The reasoning is this. Let \( E_k \) be the event \{Individual k fails at \( t(j) \); the others in \( R(t(j)) \) survive\}. The probability density of \( E_k \) is, by independence, 7.1.1 and 7.2.5.

\[
 f(E_k) = g(\beta(t(j)|z_k) \prod_{l \in R(t(j)) \atop l \neq k} \bar{g}(\beta(t(j)|z_k))
\]

\[
 = \lambda(\beta(t(j)|z_k(t(j))) \prod_{l \in R(t(j))} \bar{g}(\beta(t(j)|z_k))
\]

\[
 = \exp(z_k(t(j))^T \beta) \lambda(\beta(t(j)|z_k(t(j))) \prod_{l \in R(t(j))} \bar{g}(\beta(t(j)|z_k)).
\]  

The events \( \{E_k : k \in R(t(j))\} \) are mutually exclusive, so that the probability that a single failure occurs at \( t(j) \), conditional on \( R(t(j)) \), is

\[
 \sum_{k \in R(t(j))} f(E_k).
\]

Therefore, the probability that the failed individual at \( t(j) \) has label \( (j) \), conditional on \( R(t(j)) \) is

\[
 \frac{f(E(j))}{\sum_{k \in R(t(j))} f(E_k)},
\]

and this reduces immediately to \( p(t(j), \beta) \). Notice that both numerator and denominator in (7.3.3) are probability densities conditional on \( R(t(j)) \).

Multiplication of the \( p(t(j), \beta) \) leads to Cox's likelihood for \( \beta \):

\[
 L(\beta) = \prod_{j=1}^{m} \left( \sum_{k \in R(t(j))} \frac{\exp(z_k(t(j))^T \beta)}{\exp(z_k(t(j)))} \right).
\]
A notable fact about this likelihood is that it depends on the observed failure times in the sample only through their ranks.

The likelihood (7.3.5) was a subject of controversy when it first appeared. Cox (1972) originally called it a "conditional" likelihood, but this was incorrect: the conditioning events were not identical at each failure time \( t_{(j)} \). Kalbfleisch and Prentice derived (7.3.5) as the marginal likelihood of \( \beta \), based on the ranks of the failure and censoring times, when there were no ties nor time dependent covariates. Breslow (1972, 1974, 1975) derived (7.3.5) as an unconditional likelihood for \( \hat{\beta} \), assuming \( \lambda_0(t) \) jumped at observed failures and was constant between. (We repeat Breslow's argument in connection with his estimate of \( \hat{\lambda}_0(t) \) in Chapter 8.) Finally Cox in a fundamental paper "Partial Likelihood" (1975) showed the 7.3.5 can be treated as an ordinary likelihood, for purposes of inference.

Let us write the efficient score at \( t_{(j)} \):

\[
\mathbf{u}_{(j)}(\beta) = \frac{d \log p(t_{(j)}, \beta)}{d \beta} \\
= \mathbf{z}_{(j)}(t_{(j)}) \mathbf{u}(t_{(j)}, \beta).
\]

The quantity \( \mathbf{u}(t_{(j)}) \) is

\[
\mathbf{u}(t_{(j)}) = \sum_{k \in R(t_{(j)})} \frac{z_k(t_{(j)}) \exp(z_k(t_{(j)}) T \beta)}{\sum_{k \in R(t_{(j)})} \exp(z_k(t_{(j)}) T \beta)}.
\]

Note that \( \mathbf{u}(t_{(j)}, \beta) \) is an "exponentially weighted" average of the \( z(t_{(j)}) \) in the risk set at \( t_{(j)} \).

The equation for estimating \( \beta \) is:

\[
\sum_{j=1}^{m} \mathbf{u}_{(j)}(\beta) = 0.
\]
Or, by writing $\hat{\beta}_n$, for the solution, we define $\hat{\beta}_n$ by the equation

$$
(7.3.9) \quad \sum_{j=1}^{m} \{z(j)\dot{t}(j) - \mu(t(j), \hat{\beta}_n)\} = 0.
$$

A solution to (7.3.9) may not exist, as discussed in the next section.

The observed p$x$p information matrix is

$$
(7.3.10) \quad \hat{I}(\beta) = \sum_{j=1}^{m} C(t(j), \beta),
$$

where

$$
C(t(j), \beta) = \frac{\partial^2}{\partial \beta^2} u_j(\beta).
$$

That is, $C(t(j), \beta)$ is the variance-covariance matrix of the $z$'s in the risk set at $t(j)$, under the scheme of exponentially weighted sampling.

In the 1975 paper, Cox shows that under a wide variety of censoring schemes, $E(u_j(\beta)) = 0$ and $\text{var}(u_j(\beta)) = -E(C(t(j), \beta)) = i_j$, say. Write

$$
(7.3.12) \quad \hat{I}_{\text{avg}}(\beta) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \Sigma_i.
$$

The sum $\Sigma_i$ has $m$ terms; the divisor $n$ is needed for work with $F_n$. If $m/n = Q$, say, then $I_{\text{avg}}(\beta) = Q(\lim_{m \to \infty} \frac{1}{m} \Sigma_i)$. Under regularity conditions, a C.L.T. for dependent R.V.'s applies to $n^{-1/2} \Sigma u_j(\beta)$, and

$$
(7.3.13) \quad n^{1/2} (\beta - \beta) \sim N(0, I_{\text{avg}}(\beta)).
$$

$I_{\text{avg}}(\beta)$ will be consistently estimated by

$$
(7.3.14) \quad \hat{I}_{\text{avg}}(\hat{\beta}_n) = \frac{1}{n} \sum_{j=1}^{m} C(t(j), \hat{\beta}_n),
$$

providing for large sample inference.
We can, therefore, base large sample inference on the fact that

\[
\hat{\beta} \sim \mathcal{N}(\beta, \frac{1}{n} \hat{I}^{-1}) - \mathcal{N}(\beta, I_p) \sim \mathcal{N}(\beta, I_p).
\]

7.4. Solvability of the Likelihood Equations

In this section, we briefly outline some conditions under which the estimating equations 7.3.9 have no finite solution. For simplicity, assume the \(z\)'s do not depend on time. Then the likelihood equation 7.3.9 is

\[
\sum_{j=1}^{m} (\hat{z}(j) - \hat{\mu}(t(j), \beta)) = 0.
\]

There will be no finite solution if, for some coordinate of \(z\), say the \(q\)-th, and for all \(\beta\),

\[
z_q(j) \geq (\leq) \, \hat{\mu}_q(t(j), \beta), \quad j = 1, \ldots, m
\]

with at least one strict inequality. In words, we require \(z_q\) for the observed failure at \(t(j)\) to be greater than (less than) or equal to the exponentially weighted mean of the \(z\)'s in the risk set at \(t(j)\).

Let us see how this can arise in practice. Suppose \(z_q\) is a 0-1 indicator of group membership. Then 7.4.2 will hold if all of the 1's are failed or censored before the first 0 fails. For, suppose the first 0-failure takes place at \(t(\ell)\). Then, \(z_q(1) = z_q(2) = \cdots = z_q(\ell-1) = 1\).

But the risk sets \(\{R(t(j)): j = 1, \ldots, \ell-1\}\) all contain some 0's. Therefore, \(\hat{\mu}_q(t(j), \beta) \leq 1, \quad j = 1, \ldots, \ell-1\), \(-\infty < \beta < \infty\), precisely because \(\hat{\mu}_q(t(j), \beta)\) is an average of the \(z\). As a result,

\[
z_q(j) - \hat{\mu}_q(t(j), \beta) > 0, \quad j = 1, \ldots, \ell-1.
\]

But if there are no 1's left in the risk set at \(t(\ell)\), then

\[
z_q(j) - \hat{\mu}_q(t(j), \beta) \leq 0, \quad \ell \leq j \leq m.
\]
Together, 7.4.3 and 7.4.1 imply 7.4.2, and Cox's likelihood has no finite solution in this case.

Note that if there are any 1's left in the risk set at \( t(\bar{l}) \), then \( z(q(j)) < \mu_q(\tau(j), \bar{\beta}) \) for at least one \( j \geq 1 \) and for any \( \bar{\beta} \). The estimating equations 7.4.1 may then have a finite solution, illustrating a peculiar role of censoring.

With no censoring, condition 7.4.2 will hold in general if, for some \( q, 1 \leq q \leq p \),

\[
(7.4.5) \quad z_q(1) > z_q(2) > \cdots > z_q(n),
\]

with at least one strict inequality. (The direction of the inequalities 7.4.5 may of course be reversed.)

7.5. The Estimator as a von Mises functional

With time dependent \( z \)'s \( F_n \) is again defined to be the distributions with mass \( 1/n \) at \( (t_i, \delta_i, z_i) \), is \( i=1, \ldots, n \). Note that \( z_i(t) \) is only defined if \( t_i > t \); otherwise we may set \( z_i(t) = 0 \) if \( t_i < t \). Our strategy is to write sums in 7.3.7 and 7.3.9 over the whole sample and divide by \( n \).

The statement "\( k \in R(t) \)" about risk set measurement is converted to

"\( I_{[t_k > t]} = 1 \).

We begin by writing the exponentially weighted mean \( \mu(t, \bar{\beta}) \) as a functional of \( F_n \). Let \( \bar{\theta} \) be a \( p \times 1 \) vector. Then

\[
(7.5.1) \quad \mu(t, \bar{\theta}) = \frac{1}{n} \sum_{j=1}^{n} I_{[t_j > t]} z_j(t) \exp(z_j(t)^T \bar{\theta})
\]

The division of numerator and denominator by \( n \) has converted the sums to averages over the sample. We make the dependence on \( F_n \) explicit by redefining \( \mu \):
where (letting \( y \) be a dummy for time):

\[
A(t, \theta, F) = \int \left[ I_{[y > t]} \right]_{\sim} z(t) \exp(z(t)^T \theta) \, dF(y, \sim).
\]

(7.5.4) \[ B(t, \theta, F) = \int I_{[y > t]} \exp(z(t)^T \theta) \, dF(y, \sim). \]

\( F(y, \sim) \) is the marginal distribution of \((y, \sim)\), and integration is over \([0, \infty) \times \mathbb{Z}\) where \( \mathbb{Z} \) is the domain of the covariates.

Note that 7.5.3 is formally equivalent to

\[
A(t, \theta, F) = E_{(Y, \sim)} \left\{ I_{[y > t]} z(t) \exp(z(t)^T \theta) \right\}
\]

(7.5.5) \[ = E_{\sim} \{ P(Y > t | z) \, z(t) \exp(z(t)^T \theta) \}. \]

Here \( t \) is a fixed time point, while \( Y \) and \( z(t) \) are random quantities. We must assume the existence of the expectation on the r.h.s. of 7.5.5 to avoid measurability problems. Similar remarks apply to 7.5.4 and other expressions in this chapter.

There is a connection here with the work of Efron (1977) on the efficiency of Cox's likelihood, Efron makes his calculations on the basis of the fixed sample \( z_1, z_2, \ldots, z_n \). The randomness in the risk sets \( R(t) \) is in his work due only to censorship and failure of the sample items. For example, let \( P_i(t) \) be the probability that sample item \( i \) is in \( R(t) \). Then Efron (c.f. his equation 3.18) works with quantities like the following:
\[
\mathbb{E}\left[ \sum_{i \in \mathcal{R}(t)} z_i(t) \exp(z_i(t)^T \theta) \right]
\]

(7.5.7)

\[
= \sum_{i=1}^{n} P_i(t) z_i(t) \exp(z_i(t)^T \theta).
\]

Now, the probability \( P_i(t) \) is \( P(Y_i > t | z_i) \). Therefore each term \( P_i(t) z_i(t) \exp(z_i(t)^T \theta) \) summed in 7.5.7 can be identified with the quantity whose expectation is taken in 7.5.6.

The approach of this chapter explicitly recognizes the randomness of the \( z \)'s, while the approach of Efron does not. However, asymptotic arguments with Efron's work require some sort of long run distribution for the \( z \)'s. Indeed, in the two sample problem presented by Efron as an example (\( z_i = 0 \) or \( z_i = 1 \)), the two samples are present in fixed proportions \( p \) and \( 1-p \).

Using 7.5.2 we can write the estimating equation 7.3.9 for \( \hat{\beta} \) over the whole sample and divide by \( n \).

(7.5.8)

\[
\frac{1}{n} U(\beta) = \frac{1}{n} \sum_{i=1}^{n} \delta(z_i(t_i) - \mu(t_i, \beta, F_n)) = 0.
\]

For an arbitrary distribution \( F \), let us define a functional \( \hat{\beta}(F) \) implicitly by:

(7.5.9)

\[
\Delta(\hat{\beta}(F), F) = \int \delta(\hat{\beta}(t, F) - \mu(t, \hat{\beta}(F), F)) \ dF(t, \delta, z) = 0.
\]

Then by 7.5.5, Cox's estimator is a von Mises functional, with \( \hat{\beta}_n = \hat{\beta}(F_n) \).

Note that existence and uniqueness of a solution to 7.5.9 are not guaranteed. Nevertheless, we will continue to speak of "the" Cox estimator defined by 7.5.9.
7.6. **Fisher Consistency**

Suppose now that the covariates \( z \) do not depend on time. Let the observations \( x = (t, \delta, z) \) be from a random censorship model (Definition 3.2.1), with failure time \( S \) from a P.H. distribution 7.1.1:

\[
\lambda_B(s|z) = \exp(\beta^T z)\lambda_0(s).
\]

Label the resulting random censorship distribution \( F_B(t, \delta, z) \).

In this section we show that Cox's estimator is Fisher consistent under the random censoring. The equation defining the estimator for arbitrary \( F \) is 7.5.6:

\[
(7.6.1) \quad U(B(F), F) = \int \delta(z - \mu(t, B(F), F)) dF(t, \delta, z) = 0.
\]

Fisher consistency will hold if \( U(B, F_B) = 0 \), that is if:

\[
(7.6.2) \quad \int \delta(z - \mu(t, B, F_B)) dF_B(t, \delta, z) = 0.
\]

Let us rewrite 7.6.2 as

\[
(7.6.3) \quad \int_{\mathcal{Z}} \int_{[0,0]} dF_B(t, \delta, z) = \int_{\mathcal{Z}} \int_{[0,0]} \mu(t, B, F_B) dF_B(t, \delta, z),
\]

making explicit the region of integration as \( \mathcal{Z} \times [0,0] \) and \( \delta=1 \). We will evaluate both sides of 7.6.3 and show them to be equal.

We assume enough regularity in the distribution of \( F_B(t, 1, z) \) so that a density function (with respect to Lebesgue and counting measures) exists (3.2.2):

\[
(7.6.4) \quad dF_B(t, 1, z) = f_B(t, 1, z) \, dz \, dt
\]

\[
= g_B(t|z) \, h(t|z) \, k(z) \, dz \, dt.
\]

We also assume that \( z f_B(t, 1, z) \) and \( \mu(t, B, f_B) \) are integrable on \( \mathcal{Z} \times [0,0] \), so that Fubini's theorem may be invoked.
We begin by writing
\[ \mu(t, \beta, F_\beta) = \frac{A(t, \beta, F_\beta)}{B(t, \beta, F_\beta)} \]
and evaluating \( A(t, \beta, F_\beta) \) and \( B(t, \beta, F_\beta) \).

By 7.5.3,
\[ A(t, \beta, F_\beta) = \int \int I_{[y \geq t]} z \exp(z^T \beta) \, dF_\beta(y,z). \]  

By definition, we have
\[ dF_\beta(y,z) = f_\beta(y|z) k(z) \, dy \, dz. \]  

When 7.6.6 is substituted into 7.6.5, let us assume that the resulting function is integrable. We apply Fubini's theorem, finding
\[ A(t, \beta, F_\beta) = \int \int z \exp(z^T \beta) k(z) \left[ \int_{y \geq t} f_\beta(y|z) \, dy \right] \, dz \]
\[ = \int z \exp(z^T \beta) k(z) \, dF_\beta(t|z) \, dz, \]
\[ = \int z \exp(z^T \beta) k(z) \, \tilde{H}(t|z) \, \tilde{G}_\beta(t|z) \, dz. \]

The last equation holds because
\[ \tilde{F}_\beta(t|z) = P_\beta(T \geq t|z) \]
\[ = P(C \geq t|z) P_\beta(S \geq t|z) \]
\[ = \tilde{H}(t|z) \, \tilde{G}_\beta(t|z), \]
by the conditional independence of \( S \) and \( C \), given \( z \), in the random censorship model.

In a similar fashion, we can show
\[ B(t, \beta, F_\beta) = \int \exp(z^T \beta) k(z) \, \tilde{H}(t|z) \, \tilde{G}_\beta(t|z) \, dz. \]
We now write the r.h.s. of 7.6.3 as

\[
\text{R.H.S.} = \int \int \frac{A(t, \beta, \beta)}{B(t, \beta, \beta)} \frac{dF_\beta(t, 1, z)}{dF_\beta} \] 

(7.6.10)

using Fubini's theorem.

By 7.6.4, this is

\[
\text{R.H.S.} = \int \int \frac{A(t, \beta, \beta)}{B(t, \beta, \beta)} \left[ \int g_\beta(t | z) H(t | z) k(z) \, dz \right] dt, 
\]

(7.6.11)

\[
= \int \int \frac{A(t, \beta, \beta)}{B(t, \beta, \beta)} \left[ \int f(t | z) \tilde{H}(t | z) \, dz \right] dt, 
\]

(7.6.12)

\[
= \int \int \frac{A(t, \beta, \beta)}{B(t, \beta, \beta)} \lambda_0(t) B(t, \beta, \beta) \, dt, \text{ using 7.6.9,} 
\]

(7.6.13)

\[
= \int \int A(t, \beta, \beta) \lambda_0(t) \, dt. 
\]

(7.6.14)

We further evaluate 7.6.14 using 7.6.7

\[
\text{R.H.S.} = \int \int \int z \exp(z T \beta) k(z) H(t | z) \tilde{g}_\beta(t | z) \lambda_0(t) \, dz \, dt. 
\]

\[
= \int \int \int z \lambda_0(t) B(t, \beta, \beta) \, dt \, dz, 
\]

\[
= \int \int \int z \, dF_\beta(t, 1, z), \text{ by 7.6.4} 
\]

\[
= \text{L.H.S. of 7.6.3.} 
\]

This proves Fisher consistency.
Remark: The reader is invited to try a "proof" of Fisher consistency with
time dependent z's, along the lines of this section. Surprisingly, if the
formal manipulations are carried out with Fubini's theorem--such a "proof"
will go through.

7.7. The Influence Curve

In this section, we derive the I.C. for Cox's estimator, allowing
the covariates to depend on time. The strategy for finding the I.C. is
similar to that in Chapter 6. We contaminate an arbitrary F with
\[ x^* = (t^*, \delta^*, z^*) \]
the resulting distribution is \[ F_\varepsilon = (1-\varepsilon)F + \varepsilon I_x^* \]. The
Cox estimator based on \[ F_\varepsilon \] is defined by \[ \tilde{U}(\beta(F_\varepsilon), F) = 0 \] (7.5.6). This
equation will be differentiated with respect to \( \varepsilon \) in order to find
\[ \frac{d}{d\varepsilon} \tilde{U}(F_\varepsilon) \bigg|_{\varepsilon=0} \]
the influence curve.

Before we proceed, it will be helpful to restate the definitions of
some functionals needed in the sequel and to define some new functionals.
To avoid confusion with the contaminating point \( (t^*, \delta^*, z^*) \), let \( y \) and \( u \)
be variables for time.

From 7.5.2, 7.5.3, and 7.5.4, we have the exponentially weighted
mean at \( y \):
\[ \mu(y, \theta, F) = \frac{A(y, \theta, F)}{B(y, \theta, F)} \]
where
\[ A(y, \theta, F) = \int_{[u>y]} z(y) \exp(z(y)^T\theta) \, dF(u,z) \]
and
\[ B(y, \theta, F) = \int_{[u>y]} \exp(z(y)^T\theta) \, dF(u,z). \]
Define the matrix \( D(y, \theta, F) \) by

\[
(7.7.1) \quad D(y, \theta, F) = \int_{\{u \geq y\}} z(y) z^\top(y) \exp(z(y)^\top \theta) \, dF(u, z).
\]

Then let

\[
(7.7.2) \quad C(y, \theta, F) = \frac{D(y, \theta, F)}{B(y, \theta, F)} - \mu(y, \theta, F) \mu(y, \theta, F)^\top.
\]

The reader may verify that \( C(y, \theta, F) \) is just the exponentially weighted covariance matrix of the \( z \)'s at \( y \), defined at 7.3.11.

Finally let

\[
(7.7.3) \quad I_{\text{avg}}(\theta, F) = \int \delta C(y, \theta, F) \, dF(y, \delta)
\]

\[
= \int C(y, \theta, F) \, dF(y, 1)
\]

Note that \( I_{\text{avg}}(\theta, F) \) is analogous to the average information matrix \( I_{\text{avg}}(\beta) \) defined at 7.3.12. The relationship will be discussed in 7.8.

To ease the notational burden, we adopt the convention

\[
(7.7.4) \quad \beta(\epsilon) = \beta(F_\epsilon).
\]

Then

\[
\beta(0) = \beta(F).
\]

The defining equation for \( \beta(\epsilon) \) is

\[
(7.7.5) \quad U(\beta(\epsilon), F_\epsilon) = \int \delta(z(y) - \mu(y, \beta(\epsilon), F_\epsilon)) \, dF_\epsilon = 0.
\]

The I.C. will be

\[
\frac{d}{d\epsilon} \beta(\epsilon) \bigg|_{\epsilon=0} = \beta'(0).
\]

We now expand 7.7.5.
\[ u(\tilde{z}(\varepsilon),\tilde{F}_\varepsilon) = (1-\varepsilon) \int \delta(z(y) - \tilde{u}(y,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon)) \, dF(y,\delta,\bar{z}) \]

(7.7.6)

\[ + \varepsilon \delta^*(z^*(t^*) - \tilde{u}(t^*,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon)) = 0. \]

Or

\[ q = (1-\varepsilon) \, u(\tilde{z}(\varepsilon),\tilde{F}) + \varepsilon[\delta^*(z^*(t^*) - \tilde{u}(t^*,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon))]. \]

We now differentiate 7.7.6 coordinate by coordinate with respect to \( \varepsilon \)
and evaluate at \( \varepsilon = 0 \).

\[ 0 = \frac{\partial}{\partial \varepsilon} \, u(\tilde{z}(\varepsilon),\tilde{F}_\varepsilon) \bigg|_{\varepsilon = 0} = \frac{\partial}{\partial \varepsilon} \, u(\tilde{z}(\varepsilon),\tilde{F}) \bigg|_{\varepsilon = 0} \]

(7.7.7)

\[ + \delta^*(z^*(t^*) - \tilde{u}(t^*,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon)) = u(\tilde{z}(0),\tilde{F}) \]

\[ = \frac{\partial}{\partial \varepsilon} \, u(\tilde{z}(\varepsilon),\tilde{F}) \bigg|_{\varepsilon = 0} + \delta^*(z^*(t^*) - \tilde{u}(t^*,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon)), \]

because \( u(\tilde{z}(0),\tilde{F}) = 0 \) by 7.5.6.

Now

\[ \frac{\partial}{\partial \varepsilon} \, u(\tilde{z}(\varepsilon),\tilde{F}) \bigg|_{\varepsilon = 0} = \frac{\partial}{\partial \varepsilon} \int (z(y) - \tilde{u}(y,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon)) \, dF(y,\bar{z}) \bigg|_{\varepsilon = 0} \]

(7.7.8)

\[ = -\frac{\partial}{\partial \varepsilon} \int \tilde{u}(y,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon) \, dF(y,\bar{z}) \bigg|_{\varepsilon = 0}. \]

We assume sufficient regularity in 7.7.8 to allow differentiation under
the integral sign. Sufficient conditions are given, for example, in
Theorem 17.3e of Fulks (1969). Assume:

a. \( \int \tilde{u}(y,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon) \, dF(y,\bar{z}) \) and \( \int \frac{\partial}{\partial \varepsilon} \tilde{u}(y,\tilde{z}(\varepsilon),\tilde{F}_\varepsilon) \, dF(y,\bar{z}) \)
are continuous with respect to \( y \) and \( \varepsilon \).

b. There is an \( \varepsilon' \) for which \( \int \tilde{u}(y,\tilde{z}(\varepsilon'),\tilde{F}_{\varepsilon'}) \, dF(y,\bar{z}) \) converges
(true here for \( \varepsilon' = 0 \)).
With these assumptions,

\[
\frac{\partial}{\partial \varepsilon} \left[ \mu(y, \beta(\varepsilon), F_{\varepsilon}) \right] \, dF(y, 1) = - \int \left[ \frac{\partial}{\partial \varepsilon} \mu(y, \beta(\varepsilon), F_{\varepsilon}) \right]_{\varepsilon=0} \, dF(y, 1).
\]

The inner derivative is:

\[
\frac{\partial}{\partial \varepsilon} \mu(y, \beta(\varepsilon), F_{\varepsilon}) \bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \left[ A(y, \beta(\varepsilon), F_{\varepsilon}) \right]_{\varepsilon=0}
\]

\[
= \left\{ B(y, \beta(0), F) \left[ \frac{\partial}{\partial \varepsilon} A(y, \beta(\varepsilon), F_{\varepsilon}) \right]_{\varepsilon=0} \right\} \left( \frac{1}{\beta^2(y, \beta(0), F)} \right)
\]

\[
A(y, \beta(0), F) \left[ \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F_{\varepsilon}) \right]_{\varepsilon=0} \left( \frac{1}{\beta^2(y, \beta(0), F)} \right)
\]

We require

\[
\frac{\partial}{\partial \varepsilon} \left. A(y, \beta(\varepsilon), F_{\varepsilon}) \right|_{\varepsilon=0} \quad \text{and} \quad \frac{\partial}{\partial \varepsilon} \left. B(y, \beta(\varepsilon), F_{\varepsilon}) \right|_{\varepsilon=0}.
\]

We find the first of these expressions by writing out \(A(y, \beta(\varepsilon), F_{\varepsilon})\):

\[
A(y, \beta(\varepsilon), F_{\varepsilon}) = \int_{[u \geq y]} z(y) \exp(z(y) \mathbb{1}_{\beta(\varepsilon)}) \, dF_{\varepsilon}(u, z)
\]

\[
A(y, \beta(\varepsilon), F_{\varepsilon}) = A(y, \beta(\varepsilon), F) + \varepsilon \left[ \int_{[t \geq y]} z^*(y) \exp(z^*(y) \mathbb{1}_{\beta(\varepsilon)}) - A(y, \beta(\varepsilon), F) \right].
\]

Here

\[
A(y, \beta(\varepsilon), I_{x^*}) = \varepsilon \int_{[t \geq y]} z^*(y) \exp(z^*(y) \mathbb{1}_{\beta(\varepsilon)}).
\]

Then, differentiating 7.7.12,

\[
\frac{\partial}{\partial \varepsilon} \left. A(y, \beta(\varepsilon), F_{\varepsilon}) \right|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \left. A(y, \beta(\varepsilon), F) \right|_{\varepsilon=0}
\]

\[
+ \int_{[t \geq y]} z^*(y) \exp(z^*(y) \mathbb{1}_{\beta(\varepsilon)}) - A(t, \beta(0), F).
\]
The first term of 7.7.14 is

\[ \frac{\partial}{\partial \varepsilon} A(y, \beta(\varepsilon), F) \bigg|_{\varepsilon=0} \]

(7.7.15)

\[ = \frac{\partial}{\partial \varepsilon} \left[ \int _{u>y} z(y) \exp(z(y)^T \beta(\varepsilon)) \, dF(u,z) \right] \bigg|_{\varepsilon=0}. \]

Again let us assume that we can differentiate with respect to \( \varepsilon \) under the integral sign in 7.7.15. (Recall that the regularity conditions in Theorem 17.3e of Fulks (1969) will apply only when \( z \) does not depend on time.) Thus, by the chain rule of multivariable calculus, we have

\[ \frac{\partial}{\partial \varepsilon} A(y, \beta(\varepsilon), F) \quad (p \times 1) \]

\[ = \int _{u>y} z(y) \left[ \frac{\partial}{\partial \varepsilon} \exp(z(y)^T \beta(\varepsilon)) \bigg|_{\varepsilon=0} \right] \, dF(u,z) \]

(7.7.16)

\[ = \int _{u>y} z(y) z(y)^T \exp(z(y)^T \beta(0)) \left[ \frac{d}{d \varepsilon} \beta(\varepsilon) \bigg|_{\varepsilon=0} \right] \, dF(u,z) \]

\[ = D(y, \beta(0), F) \beta'(0), \quad p \times p \]

where \( D(y, \beta(0), F) \) was defined at 7.7.1. Note that \( \beta'(0) \) is the influence curve we are seeking.

Therefore, collecting terms from 7.7.16 and 7.7.14, we have

\[ \frac{\partial}{\partial \varepsilon} A(y, \beta(\varepsilon), F) \bigg|_{p \times 1} \bigg|_{\varepsilon=0} \]

\[ = D(y, \beta(0), F) \beta'(0) + I_{[t^* \geq y]} z^*(y) \exp(z^*(y)^T \beta(0)) \quad p \times p \quad p \times 1 \]

\[ - A(y, \beta(0), F). \quad p \times 1 \]
In a similar fashion, we can find:

\[ \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F_\varepsilon); \]

\[ (7.7.18) \quad \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F_\varepsilon) \bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F) \bigg|_{\varepsilon=0} + I_{[t^* > y]} \exp(z^*(y) \beta(0)) - B(y, \beta(0), F); \]

parallel to 7.7.12 - 7.7.14.

The first term on the r.h.s. of 7.7.18 is

\[ \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F) \bigg|_{\varepsilon=0} \]

\[ = \frac{\partial}{\partial \varepsilon} \int I_{[u > y]} \exp(z(y) \beta(\varepsilon)) dF(u, z) \bigg|_{\varepsilon=0} \]

\[ (7.7.19) \]

\[ = \int I_{[u > y]} z(y)^T \exp(z(y) \beta(0)) \left[ \frac{\partial}{\partial \varepsilon} \beta(\varepsilon) \bigg|_{\varepsilon=0} \right] dF(u, z) \]

\[ = \Lambda(y, \beta(0), F)^T \beta'(0), \]

assuming we can differentiate under the integral sign. Again \( \beta'(0) \) is the I.C.

Therefore

\[ \frac{\partial}{\partial \varepsilon} B(y, \beta(\varepsilon), F_\varepsilon) \bigg|_{\varepsilon=0} = \Lambda(y, \beta(0), F)^T \beta'(0) \]

\[ (7.7.20) \]

\[ + I_{[t^* > y]} \exp(z^*(y) \beta(0)) - B(y, \beta(0), F). \]

We now simplify notation by writing
We also write

$$\beta(0) = \beta,$$

suppressing the dependence on $F$.

We now substitute from 7.7.20 and 7.7.17 into 7.7.11:

$$\frac{\partial}{\partial \epsilon} \mu(y, \beta(\epsilon), F) \bigg|_{\epsilon = 0}$$

$$= [B(y)D(y)\beta'(0) + B(y)I [t^*_\gamma] z^*_y(y) \exp(z^*_y(y)T\beta) - B(y)A(y)$$

$$- A(y)A(y)T\beta'(0) - A(y)I [t^*_\gamma] \exp(z^*_y(y)T\beta)$$

$$+ A(y)B(y)] \frac{1}{B^2(y)}]$$

$$= \begin{bmatrix} D(y) \\ B(y) \end{bmatrix} - \frac{A(y)A(y)}{B(y)B(y)} \beta'(0)$$

$$+ I [t^*_\gamma] \exp(z^*_y(y)T\beta) \begin{bmatrix} z^*_y(y) - \frac{A(y)}{B^2(y)} \end{bmatrix}$$

$$= \zeta(y)\beta'(0) + I [t^*_\gamma] \exp(z^*_y(y)T\beta) \begin{bmatrix} z^*_y(y) - \frac{\mu(y)}{B(y)} \end{bmatrix}.$$
\[
\frac{\partial}{\partial \varepsilon} \mathcal{U}(\varepsilon; \beta, F) \bigg|_{\varepsilon = 0} = - \int_0^\infty C(y) \, dF(y, 1) \beta'(0)
\]
(7.7.25)

\[
\begin{align*}
&- \int_{[t^* \geq y]} \exp(z^*(y)^T \beta)(z^*(y) - \mu(y)) \, dF(y, 1) / B(y) \\
&+ \delta^*(z^*(t^*) - \mu(t^*)) \end{align*}
\]
(7.7.26)

We substitute 7.7.26 in turn back into 7.7.7:

\[
0 = -I_{\text{avg}}(\beta, F) \beta'(0) - \int_0^{t^*} \exp(z^*(y)^T \beta)(z^*(y) - \mu(y)) \, dF(y, 1) / B(y)
\]
(7.7.27)

We can now solve for \( \beta'(0) = IC(x^*; \beta, F) \).

The result is a

**Theorem 7.7.1**

Let Cox's estimator \( \beta(F) \) be defined by 7.5.6, and let \( I_{\text{avg}}(\beta(F), F)^{-1} \) be an inverse of \( I_{\text{avg}}(\beta(F), F) \). (A generalized inverse will do.) Then the influence curve for Cox's estimator at \( x^* = (t^*, \delta^*, z^*) \) is

\[
IC((t^*, \delta^*, z^*); \beta, F)
\]

\[
= I_{\text{avg}}(\beta(F), F)^{-1} \left\{ - \int_0^{t^*} \exp(z^*(y)^T \beta)(z^*(y) - \mu(y)) \, dF(y, 1) / B(y)
\]
\[
+ \delta^*(z^*(t^*) - \mu(t^*)) \right\}.
\]

In Appendix A2 it is shown that \( E_{\tilde{F}}(IC(x^*; \beta, F) = 0 \). The empirical I.C. and robustness are discussed in 7.9.
7.8. The Squared I.C. Under Random Censorship

According to the theory of von Mises derivatives, we expect

\[
\sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} N(0, \Sigma(F))
\]  

where

\[
\Sigma(F) = E_F[I_C(x; \beta, F) I_C(x; \beta, F)^T].
\]

Suppose now that the P.H. model holds in the random censorship setup of Section 7.6. (In that setup, recall that the covariates do not depend on time.) As in 7.6, call the resulting distribution \( F_\beta \).

The main result of this section is Theorem 7.8.1, which states:

\[
E_{F_\beta}[I_C(x; \beta, F) I_C(x; \beta, F)^T] = I_{avg}(\beta, F_\beta)^{-1}.
\]

This implies \( \Sigma(F_\beta) = I_{avg}(\beta, F_\beta)^{-1} \).

How does this compare to the asymptotic covariance matrix from Cox's (1975) partial likelihood? That theory, as quoted in 7.3, gave a covariance matrix of \( I_{avg}(\beta)^{-1} \), where \( I_{avg}(\beta) \) was defined at 7.3.12.

Writing 7.3.12 in functional form, we see that

\[
I_{avg}(\beta) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \delta_i C(t_i, \beta, F_\beta)
\]

\[
= \lim_{n \to \infty} E_{F_\beta}[I_{avg}(\beta, F_\beta)].
\]

The identity of \( I_{avg}(\beta) \) and \( I_{avg}(\beta, F_\beta) \) should follow from the mild conditions of Cox's theory; but no proof will be given.

We begin the proof of the theorem with two closely related lemmas. These were implicitly proved in Section 7.6. Lemma 7.8.1 will be of use in Chapter 8.
Assume the P.H. model 7.1.1 for failure time:

\[ \lambda(t|z) = \exp(\beta^T z) \lambda_0(t), \]

with a random censorship distribution as in Section 7.6. With the integrability conditions of Section 7.6, then

**Lemma 7.8.1.** \( dF_\beta(t,1) = \lambda_0(t) B(t,\beta, F_\beta) dt. \)

**Lemma 7.8.2.** \( dF_\beta(t,1|z) = \lambda_0(t) \exp(z^T \beta) \tilde{F}_\beta(t|z) dt. \)

**Proofs:**

By 7.6.4 and 7.2.5

\[
\begin{align*}
    dF_\beta(t,1,z) &= g_\beta(t|z) \tilde{H}(t|z) k(z) dz dt \\
    &= \lambda_0(t) \exp(z^T \beta) \tilde{G}_\beta(t|z) \tilde{H}(t|z) k(z) dz dt.
\end{align*}
\]

Then

\[
\begin{align*}
    dF_\beta(t,1) &= \int dF_\beta(t,1,z) \\
    &= \lambda_0(t) B(t,\beta, F_\beta) dt,
\end{align*}
\]

by 7.6.9. This proves Lemma 7.8.1.

Also,

\[
\begin{align*}
    dF_\beta(t,1|z) &= \frac{dF_\beta(t,1,z)}{k(z) dz} \\
    &= \lambda_0(t) \exp(z^T \beta) \tilde{F}_\beta(t|z) dt,
\end{align*}
\]

by 7.6.8, proving the second lemma.

**Theorem 7.8.1.** With the assumptions of the previous lemmas

\[
E_\beta(\mathbf{IC}(x;\beta, F_\beta) \mathbf{IC}(x;\beta, F_\beta)^T) = I_{\text{avg}}(\beta, F_\beta)^{-1}.
\]
Proof:

\[ I^{\gamma}((t, \delta, z; \beta, F_\beta) \sim) \]

\[ = I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ \int_{0}^{t} \exp \left( z^T \beta \right) (z-\mu(y)) dF(y,1) + \delta(z-\mu(t)) \right\} \]

(7.8.9)

\[ = I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ -w(t,z) + \delta(z-\mu(t)) \right\}, \text{say.} \]

Then,

\[ I^{\gamma}((t, \delta, z; \beta, F_\beta) \sim) I^{\gamma}((t, \delta, z; \beta, F_\beta)^T \sim) \]

\[ = I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ w(t,z) w(t,z)^T - 2\delta w(t,z) (z-\mu(t))^T + \delta(z-\mu(t))(z-\mu(t))^T \right\} I^{\gamma} \left( \beta, F_\beta \right)^{-1} \]

(7.8.10)

\[ = I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ M(t,z) - N(t,\delta,z) + \delta(z-\mu(t))(z-\mu(t))^T \right\} I^{\gamma} \left( \beta, F_\beta \right)^{-1}, \]

where

\[ M(t,z) = w(t,z) w(t,z)^T \]

and

(7.8.11)

\[ N(t,\delta,z) = 2\delta w(t,z) (z-\mu(t))^T. \]

Then

\[ E_F (IC((t, \delta, z; \beta, F_\beta) \sim) I^{\gamma}((t, \delta, z; \beta, F_\beta)^T \sim)) \]

(7.8.12)

\[ = I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ E_F (M(t,z) - N(t,z)) \right\} I^{\gamma} \left( \beta, F_\beta \right)^{-1} \]

\[ + I^{\gamma} \left( \beta, F_\beta \right)^{-1} \left\{ E_F (\delta(z-\mu(t))(z-\mu(t))^T \right\} I^{\gamma} \left( \beta, F_\beta \right)^{-1}. \]
We complete the proof in two parts. In Part I, we show that
\[ E_\beta (M(t,z) - N(t,\delta,z)) = 0. \]
In the second part, we show that
\[ E_\beta \{ \delta(z-\overline{\mu}(t))(z-\overline{\mu}(t))^T \} = I_{avg} (\beta, \overline{F}_\beta). \]
The theorem follows immediately.

Part I.

To show: \[ E_\beta (M(t,z) - N(t,\delta,z)) = 0. \]

The following step is crucial. Define the \( p \times p \) matrix
\[ M_1(t,z) = \frac{\partial}{\partial t} M(t,z), \]
where the differentiation is coordinate by coordinate. Then

\[
(7.8.13) \quad M_1(t,z) = 2w(t,z) \exp(\beta^T(z-\overline{\mu}(t))^T) \frac{d \overline{F}_\beta(t,1)}{dt}.
\]

Comparing \( M_1(t,z) \) to \( N(t,\delta,z) \) defined at 7.8.10, we see that

\[
(7.8.14) \quad N(t,\delta,z) = \frac{\delta M_1(t,z) \overline{B}(t)}{\exp(\beta^T(z-\overline{\mu}(t))^T) \frac{d \overline{F}_\beta(t,1)}{dt}}.
\]

By Lemma 7.8.1, this is

\[
(7.8.15) \quad N(t,\delta,z) = \frac{\delta M_1(t,z) \overline{B}(t)}{\exp(\beta^T(z-\overline{\mu}(t))^T) \lambda_0(t)}.
\]

Then

\[
(7.8.16) \quad E_\beta (N(t,\delta,z)) = \int_0^\infty \int_0^\infty M_1(t,z) \overline{F}_\beta(t,1,z) \overline{B}(t) \lambda_0(t) \exp(\beta^T(z-\overline{\mu}(t))^T) \lambda_0(t) \]
\[
= \int_0^\infty \int_0^\infty M_1(t,z) \overline{F}_\beta(t,1,z) \lambda_0(t) \exp(\beta^T(z-\overline{\mu}(t))^T) \lambda_0(t) \]
by Lemma 7.8.2.

We also evaluate $E_{\beta_{\infty}}(M(t,z))$:

$$E_{\beta_{\infty}}(M(t,z)) = \int_0^\infty \int_{\mathbb{Z}} M(t,z) \, d\bar{F}_\beta(t|z) \, dK(z)$$

(7.8.17)

$$= -\int_0^\infty \int_{\mathbb{Z}} M(t,z) \, d\bar{F}_\beta(t|z) \, dK(z).$$

Therefore

$$E_{\beta_{\infty}}(M(t,z) - N(t,o,z))$$

(7.8.18)

$$= -\int_0^\infty \int_{\mathbb{Z}} \left[ M(t,z) \, d\bar{F}_\beta(t|z) + M_{-1}(t,z) \, \bar{F}_\beta(t|z) \right] \, dt \, dK(z).$$

The elementary rule for differentiation of products reveals

$$M(t,z) \left( \frac{\partial}{\partial t} \bar{F}_\beta(t|z) \right) + M_{-1}(t,z) \, \bar{F}_\beta(t|z)$$

(7.8.19)

$$= \frac{\partial}{\partial t} \left[ M(t,z) \, \bar{F}_\beta(t|z) \right].$$

Let us assume that for every $z$, the elements of the matrix $M(\infty,z)$ are finite. This will be true if, for instance, the covariates are bounded. Or, more realistically, the assumption will hold if there is an upper limit of observation, $C_{\max}$, so that $\tilde{H}(C_{\max}|z) = 0$.

With the assumption, the inner integral in 7.8.18 can be evaluated by the fundamental theorem of calculus and 7.8.19:
Therefore, the entire integral 7.8.18 is zero, and

\[
E_{\beta} (M(t,z) - N(t,\delta,z)) = 0.
\]

**Part II.**

We must show that

\[
(7.8.20) \quad E_{\beta} (\delta(z-\mu(t)) (z-\mu(t))^T) = I_{\text{avg}} (\beta, F_{\beta}).
\]

It is easier to start from the r.h.s, defined at 7.7.3:

\[
I_{\text{avg}} (\beta, F_{\beta}) = \int_{0}^{\infty} C(t,\beta, F_{\beta}) \, dF_{\beta}(t, l).
\]

Recall from the remark at 7.7.2 that \(C(t,\beta, F_{\beta})\) is the exponentially weighted covariance matrix of the \(z\)'s in \(R(t)\); we can, therefore, write it as:

\[
(7.8.21) \quad C(t,\beta, F_{\beta}) = \int_{\mathbb{R}^t} \exp \left( -\frac{T}{h(t)} \right) (z-\mu(t)) (z-\mu(t))^T \, dF_{\beta}(y, z).
\]

And, assuming we can apply Fubini's Theorem:

\[
I_{\text{avg}} (\beta, F_{\beta})
\]

\[
= \int_{0}^{\infty} \left[ \int_{\mathbb{R}^t} \exp \left( -\frac{T}{h(t)} \right) (z-\mu(t)) (z-\mu(t))^T \, dF_{\beta}(y, z) \right] \, dF_{\beta}(t, l)
\]
\[
I_{\text{avg}}(\beta, F_\beta) = \int_0^\infty \int_0^\infty \exp(z_\beta (z_\beta(t)) T_dF_\beta(t,1, z) dz \wedge dF_\beta(t,1).
\]

(7.8.22)

\[
= \int_0^\infty \int_0^\infty \exp(z_\beta (z_\beta(t)) T_dF_\beta(t|z) k(z) dz \wedge dF_\beta(t,1).
\]

With Lemmas 7.8.1 and 7.8.2, we can write

\[
\exp(z_\beta T_dF_\beta(t|z) k(z) dz \wedge dF_\beta(t,1)) = \exp(z_\beta T_dF_\beta(t|z) k(z) dz \wedge \lambda_0(t) B(t) dt.
\]

(7.8.23)

\[
= dF_\beta(t,1|z) k(z) dz \wedge B(t),
\]

\[
= dF_\beta(t,1,z) B(t).
\]

Substituting the last expression into 7.8.22, we find that B(t) cancels in numerator and denominator, and

\[
I_{\text{avg}}(\beta, F_\beta) = \int_0^\infty \int_0^\infty (z_\beta(t)) T_dF_\beta(t,1, z)
\]

\[
= E_{F_\beta}(\delta(z_\beta(t)) (z_\beta(t)) T),
\]

as we were to show.

The theorem now follows, since by 7.8.12,

\[
E_{F_\beta}(IC(x; \beta, F_\beta) IC(x; \beta, F_\beta)^T) = I_{\text{avg}}(\beta, F_\beta)^{-1} I_{\text{avg}}(\beta, F_\beta) I_{\text{avg}}(\beta, F_\beta)^{-1}
\]

\[
= I_{\text{avg}}(\beta, F_\beta)^{-1}.
\]

\[\square\]
Suppose $\hat{\beta}$ is Cox's estimate for a sample of size $n$, with failures at $y_1 < y_2 \ldots < y_m$. The empirical version of the I.C. 7.7.28, is evaluated at a point $x = (t, \delta, z)$ (z time dependent) is

$$\text{IC}(\hat{\beta}, F_n) = \text{IC}(\hat{\beta}, F_n)_{(t, \delta, z)}$$

(7.9.1)

$$= \left[ \frac{1}{n} \sum \exp \left( z(y_k) \beta \right)(y_k - \mu(y_k)) \right]_{(t, \delta, z)}$$

$$+ \delta(z(t) - \mu(t, \hat{\beta})),$$

Suppose the point $x = (t, \delta, z)$ is now actually added to the original sample. Let $\hat{\beta}_{n+1}$ denote the estimate based on the augmented sample. Then by 2.3.8 we expect that

$$\text{IC}(x; \hat{\beta}, F_n) \approx (n+1)(\hat{\beta}_{n+1} - \hat{\beta}).$$

(7.9.2)

In this section, we shall discuss robustness of $\hat{\beta}$ by examination of the empirical I.C. As a check, computations of both sides of 7.9.2 have been carried out for actual samples with a scalar covariate. Two samples of size $n=49$ were drawn.

The observations for each sample were created in the following way. A scalar covariate $z$ was generated from a uniform $[-1,1]$ distribution. Conditional on $z$, a failure time $S$ was drawn from an exponential distribution with hazard $e^{\beta z}$. For the first sample, $\beta=0$, and for the second $\beta=1$. A censoring time $C$ was drawn from a $U(0,4)$ distribution. From $z, S,$ and $C$, the observation $x = (t, \delta, z)$ was formed. A similar procedure was used in the Monte Carlo study described in the next section.

The two resulting samples are listed in Appendix A4 for reference.

The resulting estimates of $\hat{\beta}$ were: $\hat{\beta}_{49} = 0.36$ and $\hat{\beta}_{49} = 1.36$. The
first sample thus provides a "small" estimated $\hat{\beta}$ and the second a "large" one, as desired.

We begin our examination of the empirical I.C. by observing that it is the sum of two parts, both multiplied by $\frac{1}{\text{avg } \beta_1}$. The first part is

$$Q_1(t, z) = -\sum_{k:y_k \leq t} \exp(z(y_k) \beta_k) \left( z(y_k) - \mu(y_k) \right) .$$

The second part is

$$Q_2(t, \delta, z) = \delta(z(t) - \mu(t, \hat{\beta}_n)).$$

In the augmented sample, the added observation is present in those risk sets $R((k)$, such that $y_k \leq t$. This presence is represented by $Q_1$.

On the other hand, $Q_2$ can contribute to the I.C. only if the added observation is a failure ($\delta = 1$). In this case a new risk set at $t$ is formed. This may not be one of the original $m$ risk sets. As a convention, we agree that if $t$ is larger than all $n$ sample times, then $\mu(t, \hat{\beta}_n) = z(t)$. In this case there is therefore no contribution from $Q_2$. Indeed in the augmented sample, with $t$ the largest observation, we must have

$$z(t) - \mu(t, \hat{\beta}^+_{n+1}, F_{n+1}) = 0,$$

since only $z$ is in $R(t)$. Reference to the original likelihood equations (7.3.9) shows that the contribution of $x$ then does not depend on $\delta$.

We now consider in detail how an observation $x = (t, \delta, z)$ influences $\hat{\beta}_n$. To do so, it seems useful to consider two cases separately. In the first, we assume $z$ and $\delta$ are fixed, but $t$ moves. In the second case $t$ and $\delta$ are held constant, but $z$ is allowed to vary.
Case 1: Fix \( z \) and \( \delta \); vary \( t \).

For fixed \( z = \{z(y): y>0\} \), the term \( Q_1(t, z) \) is a step function in \( t \), taking on \( m+1 \) possible values. For \( t<y_1 \), \( Q_1(t, z) = 0 \). A term is added to the sum for each failure \( y_k \) with \( y_k<t \). For \( t>y_m \), \( Q_1(t, z) \) is constant. \( Q_1 \) is therefore bounded.

The term \( Q_2(t, \delta, z) \) is also bounded. For time dependent \( z \)'s, \( z(t) \),
and \( \mu(t, \hat{z}, \hat{n}) \) may vary continuously with \( t \). Therefore \( Q_2 \) is not, in general, limited to a finite number of values, as \( Q_1 \) is. For \( z \)'s which are not time dependent, \( Q_2 \) takes on at most \( n+1 \) values; there is a value for each original observation \( t_i \), \( i = 1, \ldots, n \), and zero, as agreed, for \( t > t_n \), the largest sample value.

From these observations we may conclude that the I.C. is discontinuous as a function of \( t \) and is bounded. The same conclusions should apply to the quantity \( 50(\hat{\beta}^{+X}_{50} - \hat{\beta}_{49}) \). Let us turn to the sample data to check these conclusions.

For the first sample, \( \hat{\beta}_{49} = 0.36 \). The largest observation was a failure at \( t = 3.69 \). In Figure 9.2.1 the IC(\( t, \delta, z \)) is plotted for the four combinations resulting from \( z = \{-1, +1\} \) and \( \delta = \{0, 1\} \). Thus the curve labeled "1" is the combination \( (z = -1, \delta = 0) \); the curve labeled "2" is \( (z = -1, \delta = 1) \) and so on. The horizontal axis is plotted from \( t = 0 \) to \( t = 4 \).

The corresponding normalized differences are \( 50(\hat{\beta}^{+X}_{50} - \hat{\beta}_{49}) \). These are designated "DIFF" and are plotted in Figure 7.9.2. The I.C. and DIFF were plotted and were both calculated at increments of 0.10 in \( t \).

We can see that the I.C. plots are very similar in shape to the corresponding DIFF plots. The I.C. exaggerates the true influence at values of \( t \). Both the I.C. and DIFF curves are discontinuous in \( t \), as
The curves at \( z = +1 \) decrease with increasing \( t \). This is to be expected. As \( t \) increases, we expect the estimated hazard to decrease. For \( z = +1 \), this hazard is \( \hat{e}^\beta \); therefore \( \hat{\beta} \) should increase. Also at \( z = +1 \), an added failure (\( \delta = 1 \)) should result in a larger estimated hazard than an added censored time (\( \delta = 0 \)), no matter what the value of \( t \) is. This is precisely the behavior shown by the I.C. and DIFF.

For \( z = -1 \), the estimated hazard is \( \hat{e}^\beta \). By symmetry we can account for the increase in \( \hat{\beta} \) with increasing \( t \). For either value of \( z ( -1 \) or +1), the value of \( \delta \) does not affect the estimate of \( t > 3.69 \), the maximum sample observation. In this sample, the largest change is \( \hat{\beta}^x_{50} \) occurs in the case \( (z = 1, \delta = 1) \). The re-estimated values of \( \hat{\beta}^x_{50} \) range from 0.42 at \( t = 0 \) to 0.11 for \( t = 3.69 \).

Now let us turn to the second sample, which resulted in \( \hat{\beta}^x_{49} = 1.36 \). In this sample, the maximum observation was a censoring time at \( t = 3.99 \). The four empirical influence curves for \( z = \{ -1, +1 \} \) and \( \delta = \{ 0, 1 \} \) are plotted in Figure 7.9.3. The DIFF curves for the normalized change in \( \hat{\beta} \) are plotted in Figure 7.9.4.

Again the I.C. curves strongly mimic the shape of the DIFF curves. But the influence curve also badly overestimates the large changes in the estimator for \( z = +1 \). The discontinuous effect of the risk sets is also apparent here. Indeed, of 27 failures, only two occurred after \( t = 1.5 \).

The curves for \( z = +1 \) show much more of a change with \( t \) than do the curves for \( z = -1 \). The original data give, for \( z = +1 \), an estimated mean failure time of \( \bar{E}(S) = e^{-1.36} = 0.25 \). Larger observed times dramatically decrease the estimate (from \( \hat{\beta}^x_{50} = 1.42 \) at \( t = 0.0 \) and \( \delta = 1 \) to \( \hat{\beta}^x_{50} = 0.95 \) for \( t = 4.0 \) and \( \delta = 1 \)).
From these examples, we may conclude that the influence of $t$ on $\hat{\beta}_n$ is indeed bounded. A virtue of the Cox estimator is that only the ranks and not the values of $t$ enter into the estimate. Therefore any monotone transformation of the original sample times leaves the values of the I.C. unchanged. The most that an outlier can do is shift from being largest observation to smallest. The actual value does not matter.

The discontinuities in the I.C. and normalized difference curves have been noted. These discontinuities mean that $\hat{\beta}$ is sensitive to local shifts in the data—the "wiggling" mentioned in Section 2.3.

**Case 2:** Fix $t$ and $\delta$; vary $z$.

We have seen that the I.C. and normalized change in the estimate are both bounded, as functions of $t$. As a function of $z$, the I.C. is continuous and unbounded. There are terms linear in $z$ and also there are exponential terms: $\exp(\hat{\beta}'z(y_k))$. If the actual change $(\hat{\beta}_{n+1}^+ - \hat{\beta}_n^-)$ depends exponentially on $z$, this would pose a serious problem. For outliers in $z$ could potentially cause large changes in $\hat{\beta}$. This phenomenon of "influence" of position of the independent variables was encountered in the cases of multiple least squares regression (Section 2.3) and of exponential survival regression (Chapter Six).

It is therefore of interest to look at some examples.

Plots of the I.C. and of $\text{DIFF} = (n+1)(\hat{\beta}_{n+1}^+ - \hat{\beta}_n^-)$, as functions of $z$, were made from the two samples of $n = 49$ described earlier. Four combinations of $(t, \delta)$ were considered. In each, $z$ ranges from $-5$ to $+5$ in increments of $0.2$; recall that in the original samples $z$ is in the interval $[-1,1]$. The four combinations of $(t, \delta)$ are: (1) $t=0, \delta=1$; (2) $t=1, \delta=0$; (3) $t=1, \delta=1$; and (4) $t=4$. By our earlier discussion, the curves for $t=4$ will be the same for $\delta=0$ and for $\delta=1$. Also, an added point with $t=0$ and $\delta=0$ has no influence on the estimate. Therefore the
points chosen cover the range of possible \((t, \hat{\theta})\) values.

For the first sample \((\hat{\theta} = 0.36)\) the I.C.'s are plotted in Figure 7.9.5 and the corresponding DIFF curves are plotted in Figure 7.9.6. For the I.C. all curves except that at \(t=0\) show the effect of the exponential term. Indeed the curve at \(t = 4.0\) has a value of -401.0 at \(z=5\).

When we turn to the actual DIFF plots, we notice that the predicted exponential declines in \(\hat{\beta}\) do not take place in the curves 2, 3, and 4. These curves seem to level off as \(z\) becomes larger; indeed this is true of curve 1 also.

Nonetheless, the I.C. mimics the qualitative behavior of all four DIFF curves. For example, both curves for situation 1 are approximately linear. It is interesting that the DIFF curves are not necessarily monotone in \(z\), for curves 2, 1, and 4. The I.C. captures this behavior, at least for curves 3 and 4.

Plots of the I.C.'s from the second sample \((\hat{\theta} = 1.36)\) are shown in Figure 7.9.7. With the larger value of \(\hat{\theta}\), the exponential fall-off in curves 2, 3, and 4 is more precipitous than before. The corresponding DIFF curves are plotted in Figure 7.9.8. Again the actual fall-off in \(\hat{\beta}_{50}\) is not nearly so severe as the I.C.'s predict.

The I.C. therefore has exaggerated in both samples the extreme effects of outliers in \(z\). But comparison of the I.C. and DIFF curves show close agreement for smaller values of \(z\). And, in both samples, the changes in \(\hat{\beta}_{50}\) caused by outliers in \(z\) are more extreme than the changes caused by the variation in \(t\).

Remarks

We close this section with a few miscellaneous remarks on the I.C. A small Monte Carlo simulation is presented in Section 7.10. Conclusions and recommendations will be made in Section 7.11.
Fig. 7.9.1. Plots of the empirical influence curve for Cox's estimate against $t$: $n=49$, $\hat{\theta}_n=0.36$.

The cases are: 1: $(z=1, \delta=0)$ 2: $(z=-1, \delta=1)$ 3: $(z=1, \delta=0)$ 4: $(z=1, \delta=1)$. 
Fig. 7.9.2. Plots of the normalized difference in Cox estimates against $t$: $n=49$, $\hat{\beta}_n=0.36$.

The cases are: 1: $(z=-1, \delta=0)$ 2: $(z=-1, \delta=1)$ 3: $(z=1, \delta=0)$ 4: $(z=1, \delta=1)$. 
Fig. 7.9.3. Plots of the empirical influence curve for Cox's estimate against $t$: $n = 49$, $\hat{\beta} = 1.36$.
The cases are: 1: $(z = -1, \delta = 0)$ 2: $(z = -1, \delta = 1)$ 3: $(z = 1, \delta = 0)$ 4: $(z = 1, \delta = 1)$. 

\[ t \]

\[ IC \]

\[ -16 \]

\[ -24 \]

\[ -32 \]

\[ -40 \]

\[ -48 \]

\[ -56 \]

\[ 0 \]

\[ 1 \]

\[ 2 \]

\[ 3 \]

\[ 4 \]
Fig. 7.9.4. Plots of the normalized difference in Cox estimates against $t$: $n=49$, $\hat{\beta}_n = 1.36$. The cases are: 1: $(z=-1, \delta=0)$ 2: $(z=-1, \delta=1)$ 3: $(z=1, \delta=0)$ 4: $(z=1, \delta=1)$. 
Fig. 7.9.5. Plots of the empirical influence curve for Cox's estimate against $z$: $n=49$, $\hat{\beta}_n = 0.36$.
The cases are: 1: $(t=0, \delta=1)$ 2: $(t=1, \delta=0)$ 3: $(t=1, \delta=1)$ 4: $(t=4)$. 
Fig. 7.9.6. Plots of the normalized difference in Cox estimates against $z$: $n=49$, $\hat{\beta}_n = 0.36$.

The cases are: 1: $(t=0, \delta=1)$ 2: $(t=1, \delta=0)$ 3: $(t=1, \delta=1)$ 4: $(t=4)$. 
Fig. 7.9.7. Plots of the empirical influence curve for Cox's estimate against $z$: $n=49$, $\hat{\beta}_n=1.36$. The cases are: 1: $(t=0, \delta=1)$ 2: $(t=1, \delta=0)$ 3: $(t=1, \delta=1)$ 4: $(t=4)$. 
Fig. 7.9.8. Plots of the normalized difference in Cox estimates against $z$: $n=49$, $\hat{\beta}_n=1.36$.
The cases are: 1: $(t=0, \delta=1)$ 2: $(t=1, \delta=0)$ 3: $(t=1, \delta=1)$ 4: $(t=4)$. 
1. The discussion in this section has been mainly in terms of a scalar covariate. The sensitivity of Cox's estimator to outliers will not be simple to describe when the z's are multivariate.

One cause of this complexity is multiplication by $\tilde{I}_{\text{avg}}(\hat{\beta})$ in 7.9.1. This means that an outlier in only one coordinate of $\tilde{z}$ can affect the estimates of coefficients for the other coordinates. This phenomenon is also encountered in the exponential survival model (Section 6.6, Remark 4) and in least squares regression.

2. The I.C. (7.9.1) shows why time dependent covariables are natural when used with Cox's model. For example, the $I_C((t, \delta, z), \hat{\beta}, F)$ is zero if $z(y) = \hat{u}(y)$, for every $y > 0$.

3. Most of the discussion in this section has centered around the empirical I.C. The $I_C(x; \hat{\beta}, F)$ evaluated at a true long run distribution $F$ serves as the limit of the empirical I.C. (Mallows, 1974). No attempt has been made here to evaluate the I.C.: for a theoretical distribution. Still, some observations can be made.

The behavior of the I.C. as a function of $z$ should be analogous to the behavior of the empirical I.C. As a function of $t$, the I.C. should be smoother. Discontinuities in the censoring pattern and in values of the time dependent z's may still produce discontinuities in the I.C. itself.

7.10. The Simulation Study

The last section showed the effects of outliers in single samples. To provide a probabilistic look at the problem of outliers, a simulation study was carried out. The simulation was designed to show the influence outliers have in a random censorship setting.
Five models were studied.

1. BO: no outliers, $\beta = 0$.

   Each observation $(t, \delta, z)$ was picked according to a random censorship model. In this model, the covariate $z$ was picked from a $U[-1,1]$ distribution. Conditional on $z$, the failure time $S$ was chosen from an exponential distribution with hazard $e^{\beta z}$. For model BO, $\beta$ is zero, so the hazard is equal to one. Finally a censoring time $S$ was generated from a $U[0,4]$ distribution. The endpoint 4 was chosen to achieve approximately a 25% censoring rate at $z=0$. To complete the observation, $t$ is set to the minimum of $S$ on $C$ and $\delta=0$. This process is repeated 100 times to form a sample.

2. Bl: no outliers, $\beta = 1$.

   Model Bl is the same as BO, except that the failure times $S$ are exponential with hazard $e^1$.

3. BO/T4:

   In model BO/T4, the basic setup is similar to BO. But each observation has a 10% chance of being $t=4.0$, $\delta=0$, and $z=5$. In BO/T4, we expect the estimates to be decreased by the 10% contamination.

4. Bl/T4:

   Bl/T4 is model Bl contaminated with 10% outliers with $t=4.0$, $\delta=0$, and $z=5$.

5. BO/T0:

   BO/T0 is model BO contaminated with approximately 10% outliers at $t=0$, $\delta=1$, $z=5$. In contrast to BO/T4, the contamination is expected to increase the estimate of $\beta$.

For each of these models, 1000 samples of size $n=100$ were formed. The estimated regression coefficient in Cox's model was found with its
estimated variance by Newton Raphson iteration. In about six cases, a solution to the likelihood equations could not be found, and replacement samples were required. An attempt was made to produce estimates from a Bl/T0 model analogous to the BO/T0 model above. However, a solution to the likelihood equation could not be found for over 10% of the samples generated; therefore a Bl/T0 case is not included.

Computations were carried out in single precision arithmetic on an IBM 370/65 at the Triangle Universities Computation Center, Research Triangle Park, North Carolina. The basic random numbers generator was a Tausworthe generator of uniform [0,1] numbers, described by J. Whittlesey, Communications of the A.C.M. 11 (1968). The same uniform numbers were reused for each model. Therefore the covariate and censoring times are identical for samples in different models, if the observation is not an outlier.

Percentile points for the estimated regression coefficients appear in Table 7.10.1. The tabled percentiles are: 1 2.5 5 10 25 50 75 90 95 92.5 99.

In addition to regression coefficients, a Student-type statistic was calculated for each sample. For the j-th sample let \( \hat{\beta}_j \) be the estimated regression coefficient and \( \hat{\nu}_j \) its estimated variance. If the uncontaminated portion of the sample was generated with the true beta equal to zero, define

\[
W_j = \frac{\hat{\beta}_j}{\hat{\nu}_j}. \tag{7.10.1}
\]

For those samples for models Bl and Bl/T4,

\[
W_j = \frac{\hat{\beta}_j - 1}{\hat{\nu}_j}. 
\]
Since outliers also affect the estimated variances, the \( W \) distributions need not show the same patterns as the \( \hat{\beta} \) distributions. Percentiles for the \( W \)'s are presented in Table 7.10.2. All the percentiles were calculated according to an empirical distribution function definition. For comparative purposes, the percentage points from a \( N(0,1) \) distribution are also tabled with the \( W \)'s.

**Table 7.10.1**

**Percentiles for the Regression Coefficient Beta**

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>( B_0 )</th>
<th>( B_0/T4 )</th>
<th>( B_0/T0 )</th>
<th>( B_1 )</th>
<th>( B_1/T4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>-0.51</td>
<td>-0.73</td>
<td>0.45</td>
<td>0.48</td>
<td>-0.39</td>
</tr>
<tr>
<td>2.5</td>
<td>-0.45</td>
<td>-0.67</td>
<td>0.53</td>
<td>0.58</td>
<td>-0.37</td>
</tr>
<tr>
<td>5.0</td>
<td>-0.34</td>
<td>-0.62</td>
<td>0.59</td>
<td>0.64</td>
<td>-0.35</td>
</tr>
<tr>
<td>10.0</td>
<td>-0.26</td>
<td>-0.58</td>
<td>0.63</td>
<td>0.71</td>
<td>-0.32</td>
</tr>
<tr>
<td>25.0</td>
<td>-0.15</td>
<td>-0.52</td>
<td>0.69</td>
<td>0.85</td>
<td>-0.28</td>
</tr>
<tr>
<td>50.0</td>
<td>-0.01</td>
<td>-0.45</td>
<td>0.75</td>
<td>0.99</td>
<td>-0.23</td>
</tr>
<tr>
<td>75.0</td>
<td>0.13</td>
<td>-0.39</td>
<td>0.82</td>
<td>1.15</td>
<td>-0.18</td>
</tr>
<tr>
<td>90.0</td>
<td>0.25</td>
<td>-0.34</td>
<td>0.88</td>
<td>1.29</td>
<td>-0.12</td>
</tr>
<tr>
<td>95.0</td>
<td>0.32</td>
<td>-0.29</td>
<td>0.91</td>
<td>1.37</td>
<td>-0.09</td>
</tr>
<tr>
<td>97.5</td>
<td>0.38</td>
<td>-0.26</td>
<td>0.94</td>
<td>1.76</td>
<td>-0.05</td>
</tr>
<tr>
<td>99.0</td>
<td>0.46</td>
<td>-0.22</td>
<td>0.98</td>
<td>1.56</td>
<td>0.00</td>
</tr>
</tbody>
</table>
TABLE 7.10.2
PERCENTILES FOR THE STUDENTIZED W STATISTIC

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>BO</th>
<th>BO/T4</th>
<th>BO/T0</th>
<th>Bl</th>
<th>Bl/T4</th>
<th>N(0,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>-2.44</td>
<td>-4.76</td>
<td>2.72</td>
<td>-2.40</td>
<td>-15.76</td>
<td>-2.33</td>
</tr>
<tr>
<td>2.5</td>
<td>-2.19</td>
<td>-4.57</td>
<td>3.58</td>
<td>-1.94</td>
<td>-15.35</td>
<td>-1.96</td>
</tr>
<tr>
<td>5.0</td>
<td>-1.70</td>
<td>-4.43</td>
<td>4.17</td>
<td>-1.66</td>
<td>-15.15</td>
<td>-1.64</td>
</tr>
<tr>
<td>10.0</td>
<td>-1.25</td>
<td>-4.28</td>
<td>4.80</td>
<td>-1.36</td>
<td>-14.89</td>
<td>-1.28</td>
</tr>
<tr>
<td>25.0</td>
<td>-0.69</td>
<td>-3.98</td>
<td>5.50</td>
<td>-0.71</td>
<td>-14.30</td>
<td>-0.67</td>
</tr>
<tr>
<td>50.0</td>
<td>-0.04</td>
<td>-3.62</td>
<td>6.10</td>
<td>-0.07</td>
<td>-13.66</td>
<td>0.00</td>
</tr>
<tr>
<td>75.0</td>
<td>0.64</td>
<td>-3.24</td>
<td>6.51</td>
<td>0.62</td>
<td>-12.93</td>
<td>0.67</td>
</tr>
<tr>
<td>90.0</td>
<td>1.20</td>
<td>-2.81</td>
<td>6.79</td>
<td>1.21</td>
<td>-12.28</td>
<td>1.28</td>
</tr>
<tr>
<td>95.0</td>
<td>1.54</td>
<td>-2.45</td>
<td>6.91</td>
<td>1.54</td>
<td>-11.87</td>
<td>1.64</td>
</tr>
<tr>
<td>97.5</td>
<td>1.83</td>
<td>-2.27</td>
<td>6.99</td>
<td>1.88</td>
<td>-11.42</td>
<td>1.96</td>
</tr>
<tr>
<td>99.0</td>
<td>2.18</td>
<td>-1.98</td>
<td>7.06</td>
<td>2.19</td>
<td>-10.79</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Here are some short comments on the results for each model.

1. Model BO

   Both Beta and W show a slight negative bias. In addition the left tails are slightly more stretched than the right tails.

2. Model BO/T4

   The distribution of Beta is shifted to the left, as expected: the shift is about -0.44 at the 50-th percentile and is greater in the right tail than in the left. The resulting distribution is much more tightly bunched than the BO distribution. Similar remarks apply to the W distribution.
3. Model B0/T0

Both the Beta and W distribution show the expected shifts to the right. The W distribution has a longer right tail than left tail. The Beta distribution shows the same tendency to a lesser extent.

4. Model B1

For this no-outlier model, the Beta percentiles show a slight tendency to a longer right tail. The W distribution, on the other hand, shows a slightly more stretched left tail. This was a characteristic of the B0 W distribution. The similarity may be due to use of the same random numbers.

5. Model B1/T4

Both the Beta and W distributions show a large shift to the left from the B1 distribution. For Beta, the shift at the 50-th percentile is about -1.22. The greater magnitude of this shift compared to that from B0 to B0/T4 is expected from the plots in Section 7.9 of the normalized changes in \( \hat{\beta} \): the influence of outliers at \( t=4 \) and \( z=5 \) was stronger in the sample with \( \hat{\beta}_{49} = 1.36 \) than in the sample with \( \hat{\beta}_{49} = 0.36 \). The right tail is slightly longer than the left, and the distribution is much tighter than the B1 case with no outliers.

Both tails of the W distribution are heavier than the corresponding \( N(0,1) \) distribution. In addition the right tail is more stretched than the left tail. The middle part of the distribution is symmetric around the median; the interquartile range is 1.37, very close to the theoretical 1.35.

Perhaps the results for the outlier models are not surprising, since the outliers are unusually bad. Ten percent gross errors is not unusual in practice. (Hampel, 1973).
7.11. **Conclusions and Recommendations**

From the results of the last two sections, some conclusions can be drawn about the robustness of Cox's estimator:

1. The estimator is robust to outliers in the observed sample times. This robustness reflects the nonparametric rank approach of the Cox procedure.

2. On the other hand, the estimator is relatively more sensitive to outliers in the covariates. The influential covariates are likely to be found at the extremes of the sample ranges. In this sensitivity to covariate position, the Cox estimator is like parametric regression models.

3. Infinitesimal shifts in sample times can cause abrupt changes in the estimated coefficients. The influential local shifts are those which move an observation in or out of a risk set. These shifts may result in practice from errors, rounding, and grouping.

The user of Cox's estimate can take some practical action to protect the analysis against outliers in the covariates:

1. Identification of extreme values is essential.

2. Once extremes in the covariates are identified, a number of steps can be taken. (Hampel, 1973). Impossible values should be rejected. On the other hand, the extreme values may provide most of the variation in a particular covariate; in such cases it may be necessary to keep and trust the values.

3. As a middle way, Hampel suggests moving doubtful outlying \( z \)'s towards the rest of the data. This might be done by applying a transformation.

4. At some point, the statistician may choose to delete observations with outliers to see how Cox's estimate is affected. It may be the case
that the influence is not large.

5. The approach described above deals with extremes in the z's on a coordinate by coordinate basis. Assessment of the total influence of each observation would be helpful. To do this one might calculate the empirical influence curve. The use of the empirical I.C. to detect outliers was not dealt with in this chapter, but it is a topic worthy of investigation. For further discussion, see Chapter Two. As a substitute for calculating the I.C., observations could be ranked in order of their estimated hazards.

There seems to be no simple remedy for the sensitivity of the estimator to local shifts in the observation times. To check on this sensitivity, the estimation procedure might be run with different scalings of time. Observations measured in days might, for example, be grouped in two-day or weekly units.

Another approach is to decrease all censoring times by a fixed amount, or--equivalently--to increase all failure times. This tactic should alter some of the risk sets sufficiently to provide an additional check on the sensitivity to local shifts.

If one finds that a sample estimate is sensitive to perturbations or outliers, the question to ask is: how sensitive? One measure may be the change in the estimate in comparison to the size of its estimated standard error. Of course the standard error, too, may be disturbed.

Of poor comfort is the fact that the influence curve is related to a change in estimates multiplied by sample size. This fact implies that the influence of any particular observations diminishes as sample size grows. Unfortunately, the number of bad data points is likely to increase at the same time.
8. ESTIMATION OF THE UNDERLYING DISTRIBUTION IN COX'S MODEL

8.1. Introduction

Recall that in the P.H. model (7.1.1)

$$\lambda(t|z) = \exp(\beta^T z)\lambda_0(t).$$

Cox's likelihood for $\hat{\beta}$ did not require knowledge of the underlying survival distribution $\bar{G}(t)$, with hazard $\lambda_0(t)$. Nonparametric estimation of $\bar{G}(t)$ is of obvious interest in its own right, and several approaches have been proposed. We assume that the covariates do not depend on time.

We will describe briefly the estimation schemes of Cox (1972) and of Kalbfleisch and Prentice (1973). A simple proof of the inconsistency of Cox's method is given.

The main topic of this chapter is an estimator for $\bar{G}(t)$ (in two versions) due to Breslow (1972a,b) and considered in a slightly different form by Oakes (1972).

8.2. Cox's Approach

Cox (1972) chose to model the underlying distribution as discrete, with mass at the observed failure times $t(1), t(2), \ldots, t(m)$. Let $\pi_k(z)$ be the conditional probability of failure at $t(k)$, for a patient with covariates $z$ who has survived up to $t(k)-0$. In the underlying distribution, $z=0$, and the underlying conditional probability is $\pi_k(0)$. To estimate $\pi_k(z)$, Cox postulated a logistic relationship:
\[(8.2.1) \quad \pi_k(z) = \frac{\exp(\beta^T z)}{1 - \pi_k(z)} \quad \pi_k(0) = \frac{1 - \pi_k(0)}{1 - \pi_k(0)} \quad (k=1, \ldots, m)\]

The previously derived value of $\hat{\beta}$ was inserted for $\beta$ in (8.2.1), and separate maximum likelihood estimation for $\pi_k(0)$, was carried at each observed failure point.

It is natural to regard the discrete probabilities $\pi_k(z)$ defined by (8.2.1) as approximations to probabilities in continuous time. Doing so leads to Cox's estimator for $\hat{G}_0(t)$

\[(8.2.2) \quad \hat{G}(t) = \prod_{k: t_k < t} (1 - \pi_k(0))\]

Unfortunately, the logistic discrete model (8.2.1) is not compatible with any continuous time model. This inconsistency was noted by Breslow (1972a) and by Kalbfleisch and Prentice (1973). We give a (new) simple proof here.

Suppose that $t_1$ and $t_2$ ($t_1 < t_2$) are successive failure times. In continuous time, we would write the conditional probability of failure at $t_1$ as

\[(8.2.3) \quad \pi(z) = \Pr(t_1 < s < t_2 \mid s > t_1; z),\]

with $\pi(z)$ related to $\pi(0)$ by (8.2.1).

Now introduce a new failure time $t'$, with $t_1 < t' < t_2$.

Define

\[(8.2.4) \quad \pi_1(z) = \Pr(t_1 < s < t' \mid s > t_1; z); \quad \pi_2(z) = \Pr(t' < s < t_2 \mid s > t_1; z);\]

probabilities corresponding to $\pi(z)$. We will show that the logistic relationship (8.2.1) cannot hold for $\pi(z)$, $\pi_1(z)$, and $\pi_2(z)$ simultaneously.
We start with two equivalent equations relating $\pi(z)$ to $\pi_1(z)$ and $\pi_2(z)$:

\begin{align}
(8.2.5) \quad \pi(z) &= \pi_1(z) + (1-\pi_1(z))\pi_2(z) \\
(8.2.6) \quad 1-\pi(z) &= (1-\pi_1(z))(1-\pi_2(z)).
\end{align}

(Equation 8.2.5 is obtained from the fact that failure in \([t_1,t_2)\) requires failure in \([t_1,t')\) or, conditional on survival in \([t_1,t')\), failure in \([t',t_2)\). Equation 8.2.6 is obtained by subtraction or by noting that survival of the entire interval \([t_1,t_2)\) requires survival of the two subintervals.)

Therefore

\begin{align}
(8.2.7) \quad \frac{\pi(z)}{1-\pi(z)} &= \frac{\pi_1(z)}{(1-\pi_1(z))(1-\pi_2(z))} + \frac{\pi_2(z)}{1-\pi_2(z)}.
\end{align}

In particular, for $z=0$

\begin{align}
(8.2.8) \quad \frac{\pi(0)}{1-\pi(0)} &= \frac{\pi_1(0)}{(1-\pi_1(0))(1-\pi_2(0))} + \frac{\pi_2(0)}{1-\pi_2(0)}.
\end{align}

Now apply the logistic relationship (8.1.1) to $\pi(z)$:

\begin{align}
(8.2.9) \quad \frac{\pi(z)}{1-\pi(z)} &= \exp(\beta^Tz) \cdot \frac{\pi(0)}{1-\pi(0)}.
\end{align}

Substituting for $\pi(0)/1-\pi(0)$ from 8.2.8 into 8.2.9 we find:

\begin{align}
(8.2.10) \quad \frac{\pi(z)}{1-\pi(z)} &= \exp(\beta^Tz) \left\{ \frac{\pi_1(z)}{(1-\pi_1(z))(1-\pi_2(z))} + \frac{\pi_2(z)}{1-\pi_2(z)} \right\}.
\end{align}
By (8.2.1) we also have

\[
(8.2.11) \quad \frac{\pi_1(z)}{1-\pi_1(z)} = \exp(\beta^T z) \frac{\pi_1(0)}{1-\pi_1(0)}
\]

and

\[
\frac{\pi_2(z)}{1-\pi_2(z)} = \exp(\beta^T z) \frac{\pi_2(0)}{1-\pi_2(0)}
\]

We can substitute from 8.2.11 into 8.2.7 and find another expression for \(\pi(z)/1-\pi(z)\):

\[
(8.2.12) \quad \frac{\pi(z)}{1-\pi(z)} = \exp(\beta^T z) \frac{\pi_1(0)}{1-\pi_1(0)} \left( \frac{1}{1-\pi_2(0)} \right) + \exp(\beta^T z) \frac{\pi_2(0)}{1-\pi_2(0)}
\]

If we equate the right hand sides of 8.2.10 and 8.2.12, we conclude that

\[
(8.2.13) \quad \pi_2(z) = \pi_2(0).
\]

But 8.2.13 implies that \(\pi_2(z)\) does not depend on \(z\), a clear contradiction.

We should note that Cox used the discrete model 8.2.1 not only for estimation of \(\bar{G}\) but also for estimation of \(\bar{G}\) when tied failure times were present. This estimate, as shown by Kalbfleisch and Prentice, was inconsistent.

8.3. The Approaches of Kalbfleisch and Prentice

Kalbfleisch and Prentice (1973) proposed a discrete model which is compatible with the P.H. model in continuous time. The time axis is again partitioned at the observed failure times \(t_{(k)} \), \(k = 1, \ldots, m\). The model specifies
where \( \pi_k(z) \) is the conditional probability of failure at \( t(k) \). To simplify notation, let us follow Kalbfleisch and Prentice by writing

\[
\alpha_k = 1 - \pi_k(0).
\]

The probability that a person with covariates \( z \) fails at \( t(k) \) is:

\[
(8.3.3) \quad \exp(B^T z) \prod_{j=0}^{k-1} \alpha_j \exp(B^T z) = (1 - \alpha_k \exp(B^T z)) \prod_{j=0}^{k-1} \alpha_j \exp(B^T z).
\]

The probability that a person with covariate \( z \) survives past \( t(k) \) is, in the discrete model,

\[
(8.3.4) \quad \prod_{j=0}^{k-1} \alpha_j \exp(B^T z).
\]

Expression (8.3.3) is therefore the contribution to the likelihood of a person failing at \( t(k) \). Expression (8.3.4) is the contribution of a person censored in \([t(k), t(k+1)]\). The use of (8.3.4) is equivalent to moving a censoring time in \([t(k), t(k+1)]\) back to \( t(k) + 0 \).

For a fixed value of \( B \), the likelihood equation for \( \alpha_k \) is:

\[
(8.3.5) \quad \sum_{i \in F(k)} \frac{\exp(B^T z_i)}{(1 - \alpha_k \exp(B^T z_i))} = \prod_{j \in R(k)} \exp(B^T z_j),
\]

where \( F(k) \) is the set of individuals failing at \( t(k) \). (Therefore ties are allowed.) The previously derived estimate \( \hat{B} \) is inserted into (8.3.5), which is then solved for \( \hat{\alpha}_k = 1 - \pi_k(0) \). (Iteration is necessary if there is more than one failure at \( t(k) \).) The estimate of the survivor function is given by (8.2.2), resulting in a step function.

Kalbfleisch and Prentice also proposed a method of estimating \( \tilde{G}_0(t) \) by a connected series of straight lines. For fixed intervals
I₁ = [y₀, y₁), I₂ = [y₁, y₂), ..., Iᵣ = [yᵣ₋₁, yᵣ), the hazard function was approximated by a step function \( \lambda_0(t) = \lambda_\ell \) for \( t \in \mathcal{I}_\ell \).

Let \( \mathcal{R}_\ell \) be the risk set at the fixed point \( y_\ell \), and let \( \mathcal{R}_{\ell-1}/\mathcal{R}_\ell \) be the set of individuals failed or censored in \( I_\ell \). Of these \( m_\ell \) are observed failures in \( I_\ell \).

Then, for \( \beta = \hat{\beta} \), (previously derived), the likelihood function of \( \lambda_1, \lambda_2, ..., \lambda_r \) is proportional to

\[
\prod_{\ell=1}^{r} \lambda_\ell^{m_\ell} \exp\left\{ -\lambda_\ell \left( \sum_{t_i \in \mathcal{R}_{\ell-1}/\mathcal{R}_\ell} (t_i - y_{\ell-1}) \exp(\hat{\beta}^T z_i) + \sum_{k \in \mathcal{R}_\ell} (y_\ell - y_{\ell-1}) \exp(\hat{\beta}^T z_k) \right) \right\}
\]

(8.3.6)

\[
= \prod_{\ell=1}^{r} \lambda_\ell^{m_\ell} \exp\left\{ -\lambda_\ell Q_\ell \right\}, \text{ say.}
\]

The maximum likelihood estimator of \( \lambda_\ell \) is easily found to be

(8.3.7)

\[
\hat{\lambda}_\ell = m_\ell/Q_\ell.
\]

A continuous estimate of the underlying survival function is:

\[
\hat{G}(t) = \exp\{ -\hat{\Lambda}_0(t) \},
\]

where

\[
\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u) du
\]

(8.3.8)

\[
= \sum_{\ell=1}^{r} \hat{\lambda}_\ell (y_\ell - y_{\ell-1}) + (t - y_{J-1}) \hat{\lambda}_J , \quad t \in I_J.
\]

This estimate was suggested by the estimators of Oakes (1972) and Breslow (1972a,b), which are considered in the next section.

We note that if \( \hat{\lambda}_\ell = m_\ell/Q_\ell \) is inserted into (8.3.6), the resulting likelihood is a function of \( \hat{\beta} \) alone. This likelihood can in turn be maximized with respect to \( \hat{\beta} \), without recourse to a previously derived
estimate. Holford (1976) gives details of this approach. The resulting estimator $\hat{\beta}$ is extremely dependent on the stepwise exponential assumption, on the intervals chosen, and on the exact times of failure and censoring within those intervals. Holford's estimator $\hat{\beta}$ is therefore not recommended, especially as the much more robust estimator of Cox is available.

8.4. The Likelihoods of Oakes and Breslow

Oakes (1972) and Breslow (1972a,b) both approximated the underlying hazard by a step function constant between observed failure times. Suppose there are failure times: $t(1), t(2), \ldots, t(m)$. Then, the model is

\[(8.4.1) \quad \lambda_0(t) = \lambda_k \quad t_{k-1} \leq t < t_k.\]

Oakes' estimator is essentially that of equation (8.3.7), where now there are $m$ random intervals. Oakes' estimate therefore incorporates knowledge of exact censoring times between failures; a previously derived rank estimator of $\hat{\beta}$ is required.

Breslow, on the other hand, followed the practice of Kalbfleisch and Prentice (1973) with regard to the likelihood (8.3.5): he moved censoring times in $[t(k), t_{k+1})$ back to $t_k$. The resulting likelihood provided for simultaneous estimation of the $\hat{\lambda}'s$ and $\hat{\beta}$. We develop Breslow's estimators, following the technical report, (Breslow, 1972a) in which they were first derived. See also Crowley and Hu (1977).

With constant hazard between failures, the probability that a person with covariates $z$ fails at $t_k$ is:

\[(8.4.2) \quad \lambda^T_k e^{\beta^T z} \exp\left( - e^{\beta^T z} \sum_{l=1}^{k} \lambda_l (t_l - t_{l-1}) \right).\]

With Kalbfleisch and Prentice's simplification of the censoring times, the probability of being censored at $t_k$ is:
The likelihood for the data is therefore

\[
\exp\left\{ -e^{-e^{-\sum_{k=1}^{m} \lambda_k z_k (t_k - t_{k-1})}} \right\}.
\]

where \( S_k \) is the sum of the covariates for those who fail at \( t_k \), and

\( R_k / R_{k+1} \) is the set of individuals failing or censored in \([t_k, t_{k+1})\).

[This is the same as (8.3.6), with the adjustment to censoring times.]

The log likelihood is

\[
L(\lambda_1, \lambda_2, \ldots, \lambda_m, \beta) = \sum_{k=1}^{m} \ln m_k + \beta^T S_k - \sum_{j \in R_k / R_{k+1}} \sum_{\ell=1}^{k} \lambda_{\ell} (t_{\ell} - t_{\ell-1})
\]

where

\[
\sum_{\ell=1}^{k} \lambda_{\ell} (t_{\ell} - t_{\ell-1}) = \lambda_k (t_k - t_{k-1}) \sum_{j \in R_k} e^{-\beta^T z_j}.
\]

The likelihood equation for \( \lambda_k \) is therefore

\[
\lambda_k = \frac{m_k}{\ln m_k} \left( \frac{t_k - t_{k-1}}{\sum_{j \in R_k} e^{-\beta^T z_j}} \right) \exp(\beta^T z_j),
\]

When these values of \( \lambda_k \) are substituted into the log-likelihood (8.4.5), the resulting likelihood equation for \( \beta \) is

\[
\hat{\beta}(\lambda) = \sum_{k=1}^{m} \left[ S_k - m_k \hat{\mu}(t_k) \right] = 0,
\]

where \( \hat{\mu}(t_k) \) is the exponentially weighted mean (7.3.7).

When there are no ties, \((m_k = 1)\) the likelihood equation (8.4.7) is identical to Cox's original likelihood (7.3.9). In any case, ties do not make the estimation of \( \hat{\beta} \) or the \( \hat{\lambda}_k \)'s overly complex, as is the case with other approaches.

The estimator \( \hat{\beta} \) is inserted into (8.4.6) yielding
This estimator is also a first order approximation to the solution of Kalbfleisch and Prentice's equation (8.3.5).

For the remainder of the chapter, we consider in detail properties of estimates of $\widehat{G}(t)$ based on Breslow's estimator (8.4.8). Both Oakes and Breslow called their estimates "maximum likelihood." But this is not strictly correct, as Kalbfleisch and Prentice (1973) remark: the model is not chosen independent of the data; the parameters depend on the sample; and usual likelihood inference does not apply. We will therefore consider estimators based on (8.4.8) under the general random censorship model and will disregard the original piecewise model (8.4.1).

8.5. Functional Form; Fisher Consistency of Breslow's Estimate

From the estimated hazard rate $\hat{\lambda}(t)$ (8.4.8), two different estimators of $\widehat{G}(t)$ can be defined. The first treats the quantities $\hat{\lambda}_k(t_{(k)}-t_{(k-1)})$ as conditional probabilities of failure in $(t_{(k-1)}, t_{(k)})$; thus

$$G_n^{(1)}(t) = \prod_{k: t_{(k)} < t} \left(1 - \frac{m_k}{\sum_{j \in R_k} \exp(\beta^T z_j)}\right)$$

(8.5.1)

This version therefore generalizes the product limit estimator (5.2.1), (Breslow, 1974).

The second estimator treats the quantities $\hat{\lambda}_k(t_{(k)}-t_{(k-1)})$ as estimates of $\int_{t_{(k-1)}}^{t_{(k)}} \lambda(u) du$:

$$G_n^{(2)}(s) = \exp\left\{- \sum_{k: t_{(k)} < s} \left(\frac{m_k}{\sum_{j \in R_k} \exp(\beta^T z_j)}\right)\right\}$$

(8.5.2)

(Breslow, 1975). $G_n^{(2)}(\cdot)$ generalizes the empirical cumulative hazard
process estimator (5.2.2).

Just as the product limit and empirical hazard process estimators are close, so we expect \( \hat{G}_n^{(1)} \) and \( \hat{G}_n^{(2)} \) to be close in practice. \( \hat{G}_n^{(2)} \) is easier to write as a von Mises functional, so it will be studied in detail.

From now on let us assume that there are no ties among the observed times in the data \( m_k \equiv 1, \forall k \). Then, the solution \( \hat{\beta} = \hat{\beta}(F_n) \) to Breslow's likelihood is a von Mises functional, Fisher consistent, with a known I.C., on the basis of the last chapter. These facts make the proofs of the next theorems easy.

**Theorem (8.5.1).** \( \hat{G}_n^{(2)}(s) \) is a von Mises functional.

**Proof:** With \( \hat{\beta} = \hat{\beta}(F_n) \), we can write \( \hat{G}_n^{(2)}(s) \) as

\[
\hat{G}_n^{(2)}(s) = \exp\left\{ - \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{2} \sum_{j=1}^{n} \mathbb{I}\{t_j > t_i\} \exp(\hat{\beta}_i z_j) \right] \right\}
\]

(8.5.3)

\[
= \exp\left\{ - \int_0^s dF_n(t,1)/B(t,\hat{\beta}(F_n),F_n) \right\}
\]

where

\[
B(t,\hat{\beta}(F_n),F_n) = \int \mathbb{I}\{y > t\} \exp(z^T \hat{\beta}(F_n)) dF_n(y,z)
\]

was defined at 7.2.

Therefore \( \hat{G}_n^{(2)}(s) = \bar{G}(s,F_n) \), a von Mises functional with

(8.5.4)

\[
\bar{G}(s,F) = \exp\left\{ - \int_0^s dF(t,1)/B(t,\hat{\beta}(F),F) \right\}.
\]

**Theorem (8.5.2).** Assume that the observations \( (t,\delta,\zeta) \) are generated from a random censorship model \( F_{\beta} \) with covariates, in which the P.H. model (7.1.1) holds. Assume the regularity conditions required in Section 7.6, so that \( \hat{\beta} \) is Fisher consistent. Then \( \hat{G}_n^{(2)}(s) \) is Fisher consistent.
Proof:

By Lemma 7.8.1,

$$\frac{dF_\beta(t,1)}{dt} = \lambda_0(t) B(t, \beta(F), \text{F}_\beta).$$

Therefore,

$$\tilde{G}(s, F_\beta) = \exp\left\{- \int_0^s \frac{dF_\beta(t,1)}{B(t, \beta(F), \text{F}_\beta)} \right\}$$

(8.5.5)

$$= \exp\left\{- \int_0^s \lambda_0(t) \, dt \right\}$$

$$= \tilde{G}(s).$$

8.6. The Influence Curve

The influence curve for \(\tilde{G}(s)\) follows from results in Chapter 7.

First we let the added point by \(x^* = (t^*, \delta^*, z^*)\), and write

\(F_\varepsilon = (1-\varepsilon)F + \varepsilon I [x^*]\). The estimate for \(\tilde{G}(s)\) based on \(F_\varepsilon\) is:

$$\tilde{G}(s, F_\varepsilon) = \exp\left\{- \int_0^s \frac{dF_\varepsilon(t,1)/B(t, \beta(\varepsilon), F_\varepsilon)}{B(t, \beta(F), \text{F}_\beta)} \right\}$$

where following our previous practice, we write \(\beta(\varepsilon) = \beta(F_\varepsilon)\) and \(\hat{F} = \hat{F}(0) = \hat{F}(F)\).

The I.C. is given by

$$IC(x^*; \tilde{G}(s), F) = \frac{d}{d\varepsilon} \tilde{G}(s, F_\varepsilon) \bigg|_{\varepsilon=0}$$

(8.6.1)

$$= \tilde{G}(s, F) \left[ \frac{d}{d\varepsilon} \int_0^s \frac{dF_\varepsilon(t,1)/B(t, \beta(\varepsilon), F_\varepsilon)}{B(t, \beta(F), \text{F}_\beta)} \right]_{\varepsilon=0} \tilde{G}(s, F).$$

Let us expand the integral in (8.6.1) and differentiate each term:
\[ \int_0^S \frac{dF(t,1)}{B(t,\beta(\epsilon),F_\epsilon)} = \int_0^S \frac{dF(t,1)}{B(t,\beta(\epsilon),F_\epsilon)} \]  
(8.6.2)

\[ + \epsilon \left[ \frac{\delta^* I[t^*<s]}{B(t^*,\beta(\epsilon),F_\epsilon)} - \frac{\int_0^S dF(t,1)}{B(t,\beta(\epsilon),F_\epsilon)} \right]. \]

We differentiate the first term on the r.h.s. of (8.6.2) under the integral sign:

\[ \frac{d}{d\epsilon} \left[ \int_0^S \frac{dF(t,1)}{B(t,\beta(\epsilon),F_\epsilon)} \right] = - \left[ \frac{\int_0^S \frac{d}{d\epsilon} B(t,\beta(\epsilon),F_\epsilon)}{B^2(t,\beta,F)} \right] dF(t,1) \]  
(8.6.3)

By 7.7.20, this is

\[ - \left[ \int_0^S \frac{\mu(t,\beta,F) dF(t,1)}{B(t,\beta,F)} \right] I_{C(x;\beta,F)} \]  
(8.6.4)

where \( I_{C(x;\beta,F)} \) is the I.C. of \( \beta \) (7.7.28). The derivative of the second term in (8.6.2) is simply

\[ \frac{d}{d\epsilon} \left[ \epsilon \left[ \frac{\delta^* I[t^*<s]}{B(t^*,\beta(\epsilon),F_\epsilon)} - \frac{\int_0^S dF(t,1)}{B(t,\beta(\epsilon),F_\epsilon)} \right] \right] \]  
(8.6.5)

\[ = \frac{\delta^* I[t^*<s]}{B(t^*,\beta,F)} - \int_0^S \frac{dF(t,1)}{B(t,\beta,F)}. \]
Adding (8.6.4) and (8.6.5), we find that the I.C. of $\tilde{G}(2)(s)$ at $x^* = (t^*, \delta^*, z^*)$ is $IC(x^*; \tilde{G}(2)(s), F) =$

$$
\tilde{G}(s, F) \left\{ -\delta^* I_{t^* < s} + \int_0^{\min(t^*, s)} \frac{\exp(z^* T_\beta) dF(t^*, l)}{B^2(t, \hat{\beta}, F)} \right. 
$$

(8.6.6)

$$
+ \left[ \int_0^s \frac{u(t, \hat{\beta}, F) dF(t, l)}{B(t, \hat{\beta}, F)} \right] I_{C(X^*, F)} 
$$

8.7. Estimation of $\tilde{G}(s \mid z)$

A statistician may wish to estimate not $\tilde{G}(s)$ (corresponding to $z=0$) but

$$
\tilde{G}(s \mid z) = \tilde{G}(s) \exp(\beta_T z) 
$$

The obvious estimate, based on a sample of size $n$ is

(8.7.1) $\tilde{G}^{(2)}_n(s \mid z) = [\tilde{G}_n^{(2)}(s)] \exp(\hat{\beta}_n T z)$

The properties of $\tilde{G}^{(2)}_n(s \mid z)$ follow in a straightforward way from those of $\tilde{G}_n^{(2)}(s)$ and $\hat{\beta}_n$. Thus $\tilde{G}^{(2)}_n(s \mid z)$ is a von Mises functional and is Fisher consistent under random censorship.

The influence curve of $\tilde{G}^{(2)}(s \mid z)$, based on arbitrary $F$, is easily derived. Let us write $\tilde{G}^\varepsilon(s \mid z)$ for the estimate based on $F_\varepsilon = (1-\varepsilon)F + \varepsilon I_{x^*}$ temporarily dropping the superscript $^{(2)}$ or

$$
\tilde{G}^\varepsilon(s \mid z) = [\tilde{G}(s)] \exp(\beta(\varepsilon)_T z),
$$

where $\tilde{G}^\varepsilon$ and $\beta(\varepsilon)$ are functionals based on $F_\varepsilon$. Then the I.C. of $\tilde{G}^{(2)}(s \mid z)$ is
\[
\text{IC}((t^*, \delta^*, z^*); \hat{G}(s|z), F) = \frac{d}{d\varepsilon} \hat{G}(s|z) \bigg|_{\varepsilon = 0} \\
= \frac{d}{d\varepsilon} \exp(\log \hat{G}(s) \exp(\beta(Tz))) \bigg|_{\varepsilon = 0}
\]

(8.7.2)

\[
= \hat{G}(s|z) \{ \log \hat{G}(s) \exp(\beta(Tz)) \text{IC}(x^*; \beta, F) \}_{Tz} \\
+ \exp(\beta(Tz)) \text{IC}(x^*; \hat{G}(s), F) \\
\frac{1}{\hat{G}(s)}
\]

This can be simplified by using 8.6.6 and the fact that

\[
\log \hat{G}(s) = -\int_0^s \frac{dF(t, l)}{B(t, \beta, F)}.
\]

(8.7.3)

\[
\text{IC}((t^*, \delta^*, z^*); \hat{G}(s|z), F)
\]

\[
= \exp(\beta(Tz) \hat{G}(s|z) \left\{ -\int_0^s \frac{(z-\mu(t))dF(t, l)}{B(t)} \right\} \text{IC}(x^*; \beta, F)
\]

\[
- \delta^* I_{[t^* < s]} + \exp(\beta(Tz^*) \frac{\min(t^*, s)}{B(t^*)}) \left\{ \frac{dF(t, l)}{B(t)} \right\}
\]

In Appendix A3, we show that \( EF(\text{IC}(x^*; \hat{G}(s|z), F)) = 0, \forall s \geq 0. \)

The empirical I.C. follows when \( F_n \) is substituted for \( F \) in 8.7.3.

Let \( T(j), j = 1, \ldots, m \) be the observed failure times. Then

\[
\text{IC}((t^*, \delta^*, z^*); \hat{G}_n(s|z), F_n)
\]

(8.7.4)
8.8. Conjectured Limiting Distribution for Breslow's Estimate

Let \( s < u \) be two points in time and \( z \) a covariate, with Breslow's estimates \( \tilde{G}_n^{(2)}(s|z) \) and \( \tilde{G}_n^{(2)}(u|z) \). If \( F \) is the true distribution of the data, then \( \tilde{G}_n^{(2)}(s|z) \) and \( \tilde{G}_n^{(2)}(u|z) \) are, by definition Fisher consistent for some constants, \( \tilde{G}(s,F|z) \) and \( \tilde{G}(u,F|z) \), respectively. If these constants exist, they may have no meaning unless a P.H. model holds.

Now define

\[
\begin{align*}
    w_{1n} &= \sqrt{n}(\tilde{G}_n^{(2)}(s|z) - \tilde{G}(s,F|z)) \\
    w_{2n} &= \sqrt{n}(\tilde{G}_n^{(2)}(u|z) - \tilde{G}(u,F|z))
\end{align*}
\]

(8.8.1)

If the theory of von Mises derivatives is applied to the functionals \( \tilde{G}_n^{(2)}(\cdot|z) \), we would conclude the \( w_n \) has a limiting multivariate normal distribution:

\[
\begin{align*}
    w_n &\xrightarrow{L} N(0,A(F)),
\end{align*}
\]

(8.8.2)

where \( A(F) = (a_{ij}) \) is \( 2 \times 2 \), and

\[
\begin{align*}
    a_{11} &= E_F\{IC^2(x^*; \tilde{G}_n^{(2)}(s|z),F)\}, \\
    a_{12} &= a_{21} = E_F\{IC(x^*; \tilde{G}_n^{(2)}(s|z),F)IC(x^*; \tilde{G}_n^{(2)}(u|z),F)\}, \\
    a_{22} &= E_F\{IC^2(x^*; \tilde{G}_n^{(2)}(u|z),F)\}.
\end{align*}
\]

(8.8.3)

In other words, we conjecture that the estimator \( \tilde{G}_n^{(2)}(s|z) \) is, as a function of \( s \), a normal process. No rigorous proof (or disproof) of this conjecture is known. It is interesting that the von Mises theory...
suggests such a conclusion for general $F$. The topic deserves further investigation, perhaps through Monte Carlo methods.

To estimate the variance $a_{11}$ of Breslow's estimate, the empirical influence curve may be used. The proposed estimate is

\[
\hat{a}_{11} = \frac{1}{(n-1)} \sum_{i=1}^{n} \text{IC}^2(x_i^*; G(s|z), F_n),
\]

where the summands are the squared empirical I.C.'s 7.7.4, evaluated at $x_i^*, i = 1, \ldots, n$. The denominator $(n-1)$ is suggested because

\[
\sum_{i=1}^{n} \text{IC}(x_i^*; G(s|z), F_n) = 0.
\]

Again, the utility of 8.8.4 must be checked, and this will be the topic of further research.


Theorem Al.1.

For differentiable F and for the IC(\(x; \Lambda(s,F)\)) defined by 5.4.12,

(Al.1) \(M(s) = \int IC(x; \Lambda(s,F))dF = 0.\)

Proof:

Define

(Al.2) \(B(t) = \int_0^t (1-F(y))^{-2}dF(y,1).\)

Then the I.C. 5.4.12 can be written as

\[
IC((t,\delta); \Lambda(s,F)) = - I_{[t<s]} B(t) - I_{[t>s]} B(s) + \delta I_{[t<s]} (1-F(t))^{-1}.
\]

And

(Al.4) \(M(s) = E_F(IC((t,\delta); \Lambda(s,F))) = A_1(s) + A_2(s) + A_3(s)\)

where

(Al.5)
\[
A_1(s) = - \int_0^s B(t)dF(t)
\]
\[
A_2(s) = - B(s)(1-F(s))
\]
\[
A_3(s) = \int_0^s (1-F(t))^{-1}dF(t,1) = \int_0^s (1-F(t))(1-F(t))^{-2}dF(t,1).
\]
We can evaluate $A_1(s)$

\[(Al.6) \quad A_1(s) = - B(t)F(t) \left[ s^s \right] + \int_0^s F(t)B'(t) dt.\]

Now $B(0) = F(0) = 0$, and

\[(Al.7) \quad B'(t) = (1-F(t))^{-2} dF(t,1).\]

Therefore

\[(Al.8) \quad A_1(s) = - B(s)F(s) + \int_0^s F(t)(1-F(t))^{-2} dF(t,1).\]

\[(Al.9) \quad A_1(s) + A_3(s) = - B(s)F(s) + \int_0^s (1-F(t))^{-2} dF(t,1)\]

\[= - B(s)F(s) + B(s) = - A_2(s).\]

Theorem A2.2. Let $t_{i,1}$, $\delta_i$ $(i = 1, \ldots, n)$ constitute the sample used to estimate $\Lambda_n(s)$, we assume there are no ties among the $t$'s. Let the I.C. 5.4.13 be written

\[IC(t_{i,1}, \delta_i; \Lambda_n(s)) = \int_0^1 \cdot \]

for simplicity. Then $\forall s > 0$, (2.3.10) holds:

\[(Al.10) \quad \sum_{i=1}^n IC(t_{i,1}, \delta_i; \Lambda_n(s)) = 0.\]

Proof:

The proof is by induction. For convenience we reorder the sample

\[t(1) < t(2) < \ldots < t(n).\]

Then the conclusion of the theorem is

\[(Al.11) \quad \sum_{i=1}^n IC(t_{(i),1}, \delta_i(t_{(i)}; \Lambda_n(s)) = 0.\]
The proof rests on the observation that, considered as a function of

\[ IC(t(i), \delta(i); \Lambda_n(s)) \]

takes on at most \( i+1 \) distinct values, \( \forall i = 1, \ldots, n \). The possible values occur when \( s < t(1) \), \( s = t(1) \), \( s = t(2) \), \ldots, \( s = t(n) \). In fact,

\[
IC(t(i), \delta(i); \Lambda_n(s)) = \begin{cases} 
0, & s < t(1); \\
- \frac{1}{n} \sum_{j=1}^{k} \delta(i) (1-F_n(t(j)))^{-2}, & \text{for } k \text{ such that } t(k) \leq s < t(k+1) \leq t(i); \\
- \frac{1}{n} \sum_{j=1}^{i} \delta(i) (1-F_n(t(j)))^{-2} + \delta(i) (1-F_n(t(i)))^{-1}, & t(i) \leq s. 
\end{cases}
\]

(Al.12)

For example, \( IC(t(1), s(1); \Lambda_n(s)) \) takes on at most two values: for \( s < t(1) \) and for \( s \geq t(1) \).

To prove Al.11, therefore, it is sufficient to consider the cases \( s < t(1) \) and \( s = t(k), k = 1, \ldots, n \). The proof is by induction on \( k \).

**Case I:** \( s < t(1) \).

Then \( IC(t(i), \delta(i); \Lambda_n(s)) = 0 \) \( \forall i = 1, \ldots, n \), and the conclusion holds in this case.

**Case II:** \( s = t(1), \) (or \( t(1) \leq s < t(2) \)).

Recall that \( (1-F_n(t(j))) = (n-j+1)/n \). Then \( (1-F_n(t(1))) = (n-1+1)/n = 1 \).
Now
\[ \text{IC}(t^{(1)}, \delta^{(1)}; \Lambda_{n}(t^{(1)})) = -\frac{1}{n} (1-F_{n}(t^{(1)}))^{-2} + (1-F_{n}(t^{(1)}))^{-1} = -\frac{1}{n} + 1. \]

(A1.13)

\[ \text{IC}(t^{(i)}, \delta^{(i)}; \Lambda_{n}(t^{(1)})) = -\frac{1}{n} (1-F_{n}(t^{(1)}))^{-2} = -\frac{1}{n} \text{ for } i = 2, \ldots, n. \]

Therefore
\[ \sum_{i=1}^{n} \text{IC}(t^{(i)}, \delta^{(i)}; \Lambda_{n}(t^{(1)})) = -\frac{1}{n} + 1 - (n-1)/n = 0. \]

(A1.15)

Now assume that the lemma holds for \( s = t^{(k)} \) (or, equivalently, for \( t^{(k)} \leq s < t^{(k+1)} \)).

This means that the sum
\[ H_{k} = \sum_{i=1}^{n} \text{IC}(t^{(i)}, \delta^{(i)}; \Lambda_{n}(t^{(k)})) = 0. \]

Explicitly
\[ 0 = H_{k} = \sum_{i=1}^{k} \left[ \delta^{(i)} (1-F_{n}(t^{(i)}))^{-1} + (1-F_{n}(t^{(i)}))^{-1} \right] + \sum_{j=1}^{(n-k)} \left[ \frac{1}{n} \sum_{j=1}^{k} \delta^{(j)} (1-F_{n}(t^{(j)}))^{-2} \right]. \]

(A1.16)

\[ (A1.17) \]

The first two sums are contributions from the first \( k \) observations. The last term comes from the \((n-k)\) observations greater than \( t^{(k)} \), all having the same value for the I.C. at \( s = t^{(k)} \). We must show \( H_{k+1} = 0 \).

\( H_{k+1} \) may be evaluated in three parts:
\[ H_{k+1} = C_{1} + C_{2} + C_{3}. \]
1. A contribution to $H_{k+1}$ from $t(1), t(2), \ldots, t(k)$:

\[
C_1 = \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{l} \delta(j) (1 - F_{n}(t(j)))^{-2}
\]

(Al.18)

\[
+ \sum_{i=1}^{k} \delta(i) (1 - F_{n}(t(i)))^{-1}.
\]

2. Contribution from $IC(t_{(k+1)}, \delta(k+1); A_{n}(t_{(k+1)}))$:

\[
C_2 = -\frac{1}{n} \sum_{j=1}^{k+1} \delta(j) (1 - F_{n}(t(j)))^{-2} + \delta(k+1) (1 - F_{n}(t(k+1)))^{-1}
\]

(Al.19)

\[
= -\frac{1}{n} \sum_{j=1}^{k} \delta(j) (1 - F_{n}(t(j)))^{-2} - \delta(k+1) (1 - F_{n}(t(k+1)))^{-2} + \delta(k+1) (1 - F_{n}(t(k+1)))^{-1} .
\]

3. Contribution from the I.C.'s of the $n-k-1$ observations $t_{(k+2)}, t_{(k+3)}, \ldots, t_{(n)}$ larger than $t_{(k+1)}$:

\[
C_3 = -\frac{(n-k-1)}{n} \sum_{j=1}^{n} \delta(j) (1 - F_{n}(t(j)))^{-2}
\]

(Al.20)

\[
= -\frac{(n-k-1)}{n} \sum_{j=1}^{k} \delta(j) (1 - F_{n}(t(j)))^{-2} - \frac{(n-k-1)}{n} \delta(k+1) (1 - F_{n}(t(k+1)))^{-2} .
\]

Now writing $H_{k+1} = C_1 + C_2 + C_3$ and collecting terms multiplied by $\delta(k+1)$, we have:

\[
H_{k+1} = \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{l} \delta(j) (1 - F_{n}(t(j)))^{-2} + \sum_{i=1}^{k} \delta(i) (1 - F_{n}(t(i)))^{-1}
\]

(Al.21)

\[
+ (n-k) \left[ -\frac{1}{n} \sum_{j=1}^{k} \delta(j) (1 - F_{n}(t(j)))^{-2} \right]
\]

\[
+ \delta(k+1) \left\{ -\frac{(1 - F_{n}(t(k+1)))^{-2}}{n} + (1 - F_{n}(t(k+1)))^{-1} \right\}
\]

\[
- \frac{(n-k-1)}{n} (1 - F_{n}(t(k+1)))^{-2} .
\]
The first three terms add to $H_k$, which is zero by the induction hypothesis. The last term is (using $1 - F_n(t_{k+1}) = \frac{n-k}{n}$)

$$
\delta_{k+1} \left(1 - F_n(t_{k+1})\right)^{-2} \left\{ -1 + n \left(1 - F_n(t_{k+1})\right) - (n-k-1) \right\}
$$

(A1.22)

= 0.

Therefore $H_{k+1} = 0$ and the theorem is proved.
A2. APPENDIX TO CHAPTER SEVEN

In this appendix we show that the I.C. (7.7.28) of Cox's estimator satisfies 2.3.9:

\[(A2.1) \quad E_F(\text{IC}(x;\tilde{\beta},F) = 0,\]

where \(F\) is arbitrary.

We drop the asterisks from \(x = (t,\delta,z)\), where \(z = \{z(y) : y \geq 0\}\) depends on time. (As in Chapter Seven, the proof will be strictly valid only when \(z\) does not depend on time.) Again write \(\tilde{\beta} = \tilde{\beta}(F)\) for the estimator defined by 7.5.6.

Then

\[
\text{IC}((t,\delta,z);\tilde{\beta},F)
= I_{\text{avg}}(\tilde{\beta},F)^{-1}\left\{ \int_0^t \exp(\tilde{\beta}^T z(y)) (z(y) - \mu(y,\tilde{\beta},F)) \, dF(y,1) \right. \\
+ \delta(z(t)) - \mu(t,\tilde{\beta},F) }.
\]

\[(A2.2) \quad E_F(\text{IC}((t,\delta,z);\tilde{\beta},F))
= I_{\text{avg}}(\tilde{\beta},F)^{-1}\left\{ -E_F\left( \int_0^t \exp(\tilde{\beta}^T z(y)) (z(y) - \mu(y)) \, dF(y,1) \right) \\
+ E_F(\delta(z(t)) - \mu(t,\tilde{\beta},F)) \right\}.
\]

Writing out the second term in brackets, we find

\[(A2.3) \quad E_F(\delta(z(t)) - \mu(t,\tilde{\beta},F) = \int_0^t \delta(z(t)) - \mu(t,\tilde{\beta},F) \, dF(t,\delta,z) = 0,
\]

by the definition 7.5.6 of Cox's estimator \(\tilde{\beta} = \tilde{\beta}(F)\).
Then $E_E(E_C(x; \beta, F) = 0$ if

$$E_F\left( \begin{array}{l} \exp(\beta^T z) \left( z(y)-\mu(y) \right) \quad \text{if} \\ 0 \quad \text{if} \end{array} \right) \, dF(y,1) = 0,$$

or

$$Q(\beta, F) = 0, \text{ say.}$$

We have

$$Q(\beta, F) = \iint_{\mathbb{R} \times \mathbb{R}} I_{[y \leq t]} \exp(\beta^T z(y)) \left( z(y)-\mu(y) \right) \, dF(y,1) \, dF(t,z).$$

Let us assume we can interchange the order of integration in $(t, z)$ and $y$. (The conditions for Fubini's theorem will, of course, strictly apply only if $z$ does not depend on time.)

With the interchange,

$$Q(\beta, F) = \int_{\mathbb{R}} \frac{1}{B(y)} \int_{\mathbb{R}} \left[ \int_{\mathbb{R}} \exp(\beta^T z(y)) \left( z(y)-\mu(y) \right) \, dF(t,z) \right] \, dF(y,1)$$

$$= \int_{\mathbb{R}} w(y) \, dF(y,1), \text{ say.}$$

We complete the proof by showing

$$w(y) = 0, \quad y > 0.$$

$$w(y) = \frac{1}{B(y)} \int_{\mathbb{R}} \exp(\beta^T z(y)) z(y) \, dF(t,z)$$

$$- \mu(y) \int_{\mathbb{R}} \exp(\beta^T z(y)) \, dF(t,z)$$

$$= A(y) - \frac{\mu(y)}{B(y)} B(y), \text{ by 7.5.3 and 7.5.4}$$

$$= \frac{\mu(y)}{B(y)} - \frac{\mu(y)}{B(y)} = 0.$$
The proof applies to $F = F_n$, so that we also have

$$\sum_{i=1}^{w} IC(x_i, \hat{\beta}_n, F_n) = 0.$$
A3. APPENDIX TO CHAPTER EIGHT

Proposition: For an arbitrary distribution \( F \) and for \( \text{IC}(x^*; \tilde{G}^{(2)}(s), F) \) given by 8.6.6,

\[
\text{A3.1. } \quad \mathbb{E}_F\{\text{IC}(x^*; \tilde{G}^{(2)}(s), F)\} = 0, \quad \forall \ s > 0.
\]

Proof:

Let us drop the asterisk \(*\) from \( x^* \), so that \( x = (t, \delta, z) \)

Then

\[
\text{ IC}(x; \tilde{g}^{(2)}(s), F) = \tilde{g}^{(2)}(s) \{ -\delta I_{t \leq s} + \exp(\begin{bmatrix} T \\ \beta \end{bmatrix}) \} \mathbb{E}_F \frac{\min(t, s)}{B(t)} \]

\[
= \int_0^s \frac{u(y) \ dF(y, 1)}{B(y)} \text{ IC}(x; \beta, F) \}
\]

(A3.2)

By the result of A2,

\[
\mathbb{E}_F\{\text{IC}(x; \beta, F)\} = 0.
\]

Therefore the expected value of the last bracketed term in (A3.2) is zero. The proposition will hold if (ignoring the factor \( \tilde{g}(s) \))

(A3.3) \[
\mathbb{E}_F\left\{ \delta I_{t \leq s} \right\} = \mathbb{E}_F\{\exp(\begin{bmatrix} T \\ \beta \end{bmatrix}) \} \mathbb{E}_F \frac{\min(t, s)}{B(t)} \]

The l.h.s. of A3.3 can be written

(A3.4) \[
\text{L.H.S. = } \mathbb{E}_F\left\{ \delta I_{t \leq s} \right\} = \int_0^s \frac{dF(t, 1)}{B(t)} = \lambda_0(s, F) .
\]
The r.h.s. is

\[
\text{R.H.S.} = E_p \{ \exp(z^T \beta) \int_0^{\min(t,s)} \frac{dF(y,1)}{B^*(y)} \}
\]

(A3.5)

\[
= E_p \{ \exp(z^T \beta) \int_0^{\infty} \frac{I_{[y<t]} I_{[y<s]} dF(y,1)}{B^2(y)} \}
\]

\[
= \int_{\mathbb{Z}} \int_0^{\infty} \exp(z^T \beta) \{ \int_0^{\infty} \frac{I_{[y<t]} I_{[y<s]} dF(y,1)}{B^2(y)} \} dF(t,z).
\]

Here the inner integral is with respect to \( y \), the outer ones are with respect to \((t,z)\), \( \mathbb{Z} \) is the domain of \( z \). Let us interchange the order of \((t,z)\) and \( y \) integration. Then A3.6 becomes

\[
\text{R.H.S.} = \int_0^s \int_{\mathbb{Z}} \exp(z^T \beta) dF(t,z) \{ 1/B^2(y) \} dF(y,1)
\]

\[
= \int_0^s \{ 1/B^2(y) \} dF(y,1), \text{ by definition 7.5.4 for } B(y),
\]

\[
= \int_0^s \lambda_0(y) dy = \Lambda_0(s,\Gamma) = \text{L.H.S.}
\]
Sample I.  \( n=49, \hat{\beta}_n = 0.36 \). The observations are \((t, \delta, z)\).

<table>
<thead>
<tr>
<th>t</th>
<th>( \delta )</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0007</td>
<td>1</td>
<td>0.750</td>
</tr>
<tr>
<td>0.0035</td>
<td>1</td>
<td>0.994</td>
</tr>
<tr>
<td>0.0127</td>
<td>-0.333</td>
<td>0.0007</td>
</tr>
<tr>
<td>0.0302</td>
<td>0</td>
<td>0.821</td>
</tr>
<tr>
<td>0.0350</td>
<td>0</td>
<td>-0.667</td>
</tr>
<tr>
<td>0.0650</td>
<td>1</td>
<td>0.372</td>
</tr>
<tr>
<td>0.0990</td>
<td>1</td>
<td>-0.558</td>
</tr>
<tr>
<td>0.1490</td>
<td>1</td>
<td>-0.196</td>
</tr>
<tr>
<td>0.1840</td>
<td>0</td>
<td>-0.847</td>
</tr>
<tr>
<td>0.1979</td>
<td>1</td>
<td>0.589</td>
</tr>
<tr>
<td>0.2231</td>
<td>0</td>
<td>-0.880</td>
</tr>
<tr>
<td>0.2386</td>
<td>1</td>
<td>-0.010</td>
</tr>
<tr>
<td>0.2567</td>
<td>0</td>
<td>0.902</td>
</tr>
<tr>
<td>0.2577</td>
<td>1</td>
<td>-0.383</td>
</tr>
<tr>
<td>0.2655</td>
<td>0</td>
<td>0.878</td>
</tr>
<tr>
<td>0.3445</td>
<td>1</td>
<td>0.692</td>
</tr>
<tr>
<td>0.3489</td>
<td>0</td>
<td>-0.855</td>
</tr>
<tr>
<td>0.3699</td>
<td>0</td>
<td>-0.414</td>
</tr>
<tr>
<td>0.4004</td>
<td>1</td>
<td>0.472</td>
</tr>
<tr>
<td>0.4308</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>0.4558</td>
<td>1</td>
<td>-0.228</td>
</tr>
<tr>
<td>0.4878</td>
<td>1</td>
<td>0.053</td>
</tr>
<tr>
<td>0.5295</td>
<td>1</td>
<td>-0.038</td>
</tr>
<tr>
<td>0.5362</td>
<td>1</td>
<td>0.860</td>
</tr>
<tr>
<td>0.5808</td>
<td>1</td>
<td>0.251</td>
</tr>
<tr>
<td>0.6015</td>
<td>0</td>
<td>-0.360</td>
</tr>
<tr>
<td>0.7461</td>
<td>1</td>
<td>-0.316</td>
</tr>
<tr>
<td>0.7639</td>
<td>1</td>
<td>0.710</td>
</tr>
<tr>
<td>0.7801</td>
<td>0</td>
<td>0.691</td>
</tr>
<tr>
<td>0.8698</td>
<td>1</td>
<td>-0.603</td>
</tr>
<tr>
<td>0.9755</td>
<td>1</td>
<td>0.419</td>
</tr>
<tr>
<td>1.0633</td>
<td>1</td>
<td>-0.463</td>
</tr>
<tr>
<td>1.1368</td>
<td>1</td>
<td>-0.868</td>
</tr>
<tr>
<td>1.1470</td>
<td>0</td>
<td>0.893</td>
</tr>
<tr>
<td>1.2178</td>
<td>1</td>
<td>-0.728</td>
</tr>
<tr>
<td>1.3289</td>
<td>1</td>
<td>0.208</td>
</tr>
<tr>
<td>1.3437</td>
<td>1</td>
<td>0.638</td>
</tr>
<tr>
<td>1.3921</td>
<td>1</td>
<td>-0.747</td>
</tr>
<tr>
<td>1.4845</td>
<td>0</td>
<td>-0.693</td>
</tr>
<tr>
<td>1.6322</td>
<td>1</td>
<td>-0.574</td>
</tr>
<tr>
<td>1.6565</td>
<td>0</td>
<td>-0.785</td>
</tr>
<tr>
<td>1.7818</td>
<td>1</td>
<td>0.029</td>
</tr>
<tr>
<td>1.8607</td>
<td>0</td>
<td>-0.166</td>
</tr>
<tr>
<td>2.0741</td>
<td>1</td>
<td>0.970</td>
</tr>
<tr>
<td>2.2691</td>
<td>0</td>
<td>0.875</td>
</tr>
<tr>
<td>2.4268</td>
<td>1</td>
<td>-0.654</td>
</tr>
<tr>
<td>2.6624</td>
<td>1</td>
<td>0.517</td>
</tr>
<tr>
<td>2.9724</td>
<td>0</td>
<td>-0.729</td>
</tr>
<tr>
<td>3.6949</td>
<td>1</td>
<td>-0.619</td>
</tr>
</tbody>
</table>
Sample II. \( n=49, \beta_n = 1.36 \). The observations are \((t, \delta, z)\).

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \delta )</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0003</td>
<td>1</td>
<td>0.750</td>
</tr>
<tr>
<td>0.0013</td>
<td>1</td>
<td>0.994</td>
</tr>
<tr>
<td>0.0177</td>
<td>1</td>
<td>-0.333</td>
</tr>
<tr>
<td>0.0302</td>
<td>0</td>
<td>0.821</td>
</tr>
<tr>
<td>0.0350</td>
<td>0</td>
<td>-0.667</td>
</tr>
<tr>
<td>0.0448</td>
<td>1</td>
<td>0.372</td>
</tr>
<tr>
<td>0.1098</td>
<td>1</td>
<td>0.589</td>
</tr>
<tr>
<td>0.1725</td>
<td>1</td>
<td>0.692</td>
</tr>
<tr>
<td>0.1730</td>
<td>1</td>
<td>-0.558</td>
</tr>
<tr>
<td>0.1812</td>
<td>1</td>
<td>-0.196</td>
</tr>
<tr>
<td>0.1840</td>
<td>0</td>
<td>-0.847</td>
</tr>
<tr>
<td>0.2231</td>
<td>0</td>
<td>-0.880</td>
</tr>
<tr>
<td>0.2270</td>
<td>1</td>
<td>0.860</td>
</tr>
<tr>
<td>0.2411</td>
<td>1</td>
<td>-0.010</td>
</tr>
<tr>
<td>0.2497</td>
<td>1</td>
<td>0.472</td>
</tr>
<tr>
<td>0.2567</td>
<td>0</td>
<td>0.902</td>
</tr>
<tr>
<td>0.2655</td>
<td>0</td>
<td>0.878</td>
</tr>
<tr>
<td>0.3489</td>
<td>0</td>
<td>-0.855</td>
</tr>
<tr>
<td>0.3699</td>
<td>0</td>
<td>-0.414</td>
</tr>
<tr>
<td>0.3755</td>
<td>1</td>
<td>0.710</td>
</tr>
<tr>
<td>0.3780</td>
<td>1</td>
<td>-0.383</td>
</tr>
<tr>
<td>0.4308</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>0.4518</td>
<td>1</td>
<td>0.251</td>
</tr>
<tr>
<td>0.4628</td>
<td>1</td>
<td>0.053</td>
</tr>
<tr>
<td>0.5502</td>
<td>1</td>
<td>-0.038</td>
</tr>
<tr>
<td>0.5727</td>
<td>1</td>
<td>-0.228</td>
</tr>
<tr>
<td>0.5814</td>
<td>1</td>
<td>0.893</td>
</tr>
<tr>
<td>0.6015</td>
<td>0</td>
<td>-0.360</td>
</tr>
<tr>
<td>0.6415</td>
<td>1</td>
<td>0.419</td>
</tr>
<tr>
<td>0.7099</td>
<td>1</td>
<td>0.638</td>
</tr>
<tr>
<td>0.7801</td>
<td>0</td>
<td>0.691</td>
</tr>
<tr>
<td>0.7866</td>
<td>1</td>
<td>0.970</td>
</tr>
<tr>
<td>1.0234</td>
<td>1</td>
<td>-0.316</td>
</tr>
<tr>
<td>1.0789</td>
<td>1</td>
<td>0.208</td>
</tr>
<tr>
<td>1.1297</td>
<td>0</td>
<td>-0.463</td>
</tr>
<tr>
<td>1.3102</td>
<td>1</td>
<td>0.875</td>
</tr>
<tr>
<td>1.3203</td>
<td>0</td>
<td>-0.868</td>
</tr>
<tr>
<td>1.3309</td>
<td>0</td>
<td>-0.728</td>
</tr>
<tr>
<td>1.4111</td>
<td>0</td>
<td>-0.603</td>
</tr>
<tr>
<td>1.4845</td>
<td>0</td>
<td>-0.693</td>
</tr>
<tr>
<td>1.5877</td>
<td>1</td>
<td>0.517</td>
</tr>
<tr>
<td>1.6565</td>
<td>0</td>
<td>-0.785</td>
</tr>
<tr>
<td>1.7302</td>
<td>1</td>
<td>0.029</td>
</tr>
<tr>
<td>1.8607</td>
<td>0</td>
<td>-0.166</td>
</tr>
<tr>
<td>1.8611</td>
<td>0</td>
<td>-0.574</td>
</tr>
<tr>
<td>2.9376</td>
<td>1</td>
<td>-0.747</td>
</tr>
<tr>
<td>2.9724</td>
<td>0</td>
<td>-0.729</td>
</tr>
<tr>
<td>3.9241</td>
<td>0</td>
<td>-0.619</td>
</tr>
<tr>
<td>3.9954</td>
<td>0</td>
<td>-0.654</td>
</tr>
</tbody>
</table>