APPLIED SURVIVAL ANALYSIS

REGRESSION MODELING OF TIME TO EVENT DATA

DAVID W. HOSMER, JR.

STANLEY LEMESHOW

Wiley Series in Probability and Statistics

Applied Survival Analysis

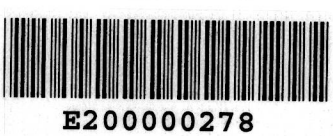
Regression Modeling of Time to Event Data

DAVID W. HOSMER, Jr.

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts



A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ♾

Library of Congress Cataloging in Publication Data:

Hosmer, David W.
Applied survival analysis : regression modeling of time to event data / David W. Hosmer, Jr., Stanley Lemeshow
p. cm. — (Wiley series in probability and statistics)
Includes bibliographical references and indexes.
ISBN 0-471-15410-5 (cloth : alk. paper)
1. Medicine—Research—Statistical methods. 2. Medical sciences—Statistical methods—Computer programs. 3. Regression analysis—Data processing. 4. Prognosis—Statistical methods. 5. Logistic distribution. I. Lemeshow, Stanley. II. Title. III. Series.
R853.S7H67 1998
610'.7'27—dc21 98-27511

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Applied Survival Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS
TEXTS AND REFERENCES SECTION

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *Vic Barnett, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, David G. Kendall, David W. Scott, Bernard W. Silverman, Adrian F. M. Smith, Jozef L. Teugels; Ralph A. Bradley, Emeritus, J. Stuart Hunter, Emeritus*

A complete list of the titles in this series appears at the end of this volume.

To Trina, Wylie, Tri,
and the memory of my parents
D. W. H.

To Elaine, Jenny, Adina, Steven,
my mother, Jack, and the memory of my father & Marisha
S. L.

Preface

The study of events involving an element of time has a long and important history in statistical research and practice. Examples chronicling the mortality experience of human populations date from the 1700s [see Hald (1990)]. Recent advances in methods and statistical software have placed a seemingly bewildering array of techniques at the fingertips of the data analyst. It is difficult to find either a subject matter or a statistical journal that does not have at least one paper devoted to use or development of these methods.

In spite of the importance and widespread use of these methods there is a paucity of material providing an introduction to the analysis of time to event data. A course dealing with this subject tends to be more advanced and often is the third or fourth methods course taken by a student. As such, the student typically has a strong background in linear regression methods and usually some experience with logistic regression. Yet most texts fail to capitalize on this statistical and experiential background. The approach is either highly mathematical or does not emphasize regression model building. The goal of this book is to provide a focused text on regression modeling for the time to event data typically encountered in health related studies. For this text we assume the reader has had a course in linear regression at the level of Kleinbaum, Kupper, Muller and Nizam (1998) and one in logistic regression at the level of Hosmer and Lemeshow (1989). Emphasis is placed on the modeling of data and the interpretation of the results. Crucial to this is an understanding of the nature of the "incomplete" or "censored" data encountered. Understanding the censoring mechanism is important as it may influence model selection and interpretation. Yet, once understood and accounted for, censoring is often just another technical detail handled by the computer software allowing emphasis to return to model building, assessment of model fit and assumptions and interpretation of the results.

The increase in the use of statistical methods for time to event data is directly re-

lated to their incorporation into major and minor (specialized) statistical software packages. To a large extent there are no major differences in the capabilities of the various software packages. When a particular approach is available in a limited number of packages it will be noted in this text. In general, analyses have been performed in STATA [Stata Corp. (1997)]. This easy to use package combines reasonably good graphics and excellent analysis routines, is fast, is compatible across Macintosh, Windows and UNIX platforms and interacts well with Microsoft Word 6.0. Other major statistical packages employed at various points during the preparation of this text include BMDP [BMDP Statistical Software (1992)], SAS [SAS Institute Inc. (1989)] and S-PLUS [S-Plus Statistical Sciences (1993)].

This text was prepared in camera ready format using Microsoft Word 6.0.1 on a Power Macintosh platform. Mathematical equations and symbols were built using Math Type 3.5 [Math Type: Mathematical Equation Editor (1997)]. When necessary, graphics were enhanced and modified using MacDraw.

Early on in the preparation of the text we made a decision that data sets used in the text would be made available to readers via the World Wide Web rather than on a diskette distributed with the text. The ftp site at John Wiley & Sons, Inc. for the data in this text is ftp://ftp.wiley.com/public/sci_tech_med/survival. In addition, the data may also be found, by permission of John Wiley & Sons Inc., in the archive of statistical data sets maintained at the University of Massachusetts at Internet address http://www-unix.oit.umass.edu/~statdata in the survival analysis section. Another advantage to having a text web site is that it provides a convenient medium for conveying to readers text changes after publication. In particular, as errata become known to us they will be added to an errata section of the text's web site at John Wiley & Sons, Inc. Another use that we envision for the web is the addition, over time, of new data sets to the statistical data set archive at the University of Massachusetts.

As in any project with the scope and magnitude of this text, there are many who have contributed directly or indirectly to its content and style and we feel quite fortunate to be able to acknowledge the contributions of others. One of us (DWH) would like to express special thanks to a friend and colleague, Petter Laake, Head of the Section of Medical Statistics at the University of Oslo, for arranging for a Senior Scientist Visiting Fellowship from the Research Council of Norway that supported a sabbatical leave visit to the Section in Oslo during the winter of 1997. We would like to thank Odd Aalen for reading and commenting on several sections of the text. His advice was most helpful in preparing the material on frailty and additive models in Chapter 9. While in Oslo, and afterwards, Ørnulf Borgan was especially helpful in clarifying some of the details of the counting process approach and graciously shared some, at that time, unpublished research of his and his student, J. K. Grønnesby. Thoughtful and careful commentary by outside reviewers, in particular Daniel Commenges, of the UFR de Santé Publique at the University of Bordeaux II, improved the content and quality of the text.

We are grateful to colleagues in our Department who have contributed to the development of this book. These include Drs. Jane McCusker, Anne Stoddard and Carol Bigelow for the use and insights into the data from the Project IMPACT Study

and Janelle Klar and Elizabeth K. Donohoe for their extraordinarily careful reading of the manuscript and editorial suggestions.

DAVID W. HOSMER, JR.
STANLEY LEMESHOW

Amherst, Massachusetts
August, 1998

Contents

CHAPTER 1

Introduction to Regression Modeling of Survival Data

1.1 INTRODUCTION

Regression modeling of the relationship between an outcome variable and independent predictor variable(s) is commonly employed in virtually all fields. The popularity of this approach is due to the fact that biologically plausible models may be easily fit, evaluated and interpreted. Statistically, the specification of a model requires choosing both systematic and error components. The choice of the systematic component involves an assessment of the relationship between an "average" of the outcome variable and the independent variable(s). This may be guided by an exploratory analysis of the current data and/or past experience. The choice of an error component involves specifying the statistical distribution of what remains to be explained after the model is fit (i.e., the residuals).

In an applied setting, the task of model selection is, to a large extent, based on the goals of the analysis and on the measurement scale of the outcome variable. For example, a clinician may wish to model the relationship between a measure of nutritional status (e.g., caloric intake) and various demographic and physical characteristics of the child such as gender, socio-economic status, height and weight, among children between the ages of two and six seen in the clinics of a large health maintenance organization (HMO). A good place to start would be to use a model with a linear systematic component and normally distributed errors, the usual linear regression model. Suppose instead that the clinician decides to convert the nutrition data into a dichotomous variable that indicated whether the child's diet met specified intake criteria (1 =

yes and 0 = no). If we assume the goal of this analysis is to estimate the "effect" of the various factors via an odds-ratio, then the logistic regression model would be a good choice. The logistic regression model has a systematic component that is linear in the log-odds and has binomial/Bernoulli distributed errors. There are many issues involved in the fitting, refinement, evaluation and interpretation of each of these models. However, the clinician would follow the same basic modeling paradigm in each scenario.

This basic modeling paradigm is commonly used in texts taking a data-based approach to either linear or logistic regression [e.g., Kleinbaum, Kupper, Muller and Nizam (1998) and Hosmer and Lemeshow (1989)]. We use it in this text to motivate our discussion of the similarities and differences between the linear (and the logistic) regression model and regression models appropriate for survival data. In this spirit we begin with an example.

Example

A large HMO wishes to evaluate the survival time of its HIV+ members using a follow-up study. Subjects were enrolled in the study from January 1, 1989 to December 31, 1991. The study ended on December 31, 1995. After a confirmed diagnosis of HIV, members were followed until death due to AIDS or AIDS-related complications, until the end of the study or until the subject was lost to follow-up. We assume that there were no deaths due to other causes (e.g., auto accident). The primary outcome variable of interest is survival time after a confirmed diagnosis of HIV. Since subjects entered the study at different times over a 3-year period, the maximum possible follow-up time is different for each study participant. Possible predictors of survival time were collected at enrollment into the study. Data listed in Table 1.1 for 100 subjects are: TIME: the follow-up time is the number of months between the entry date (ENT DATE) and the end date (END DATE), AGE: the age of the subject at the start of follow-up (in years), DRUG: history of prior IV drug use (1 = Yes, 0 = No), and CENSOR: vital status at the end of the study (1 = Death due to AIDS, 0 = Lost to follow-up or alive).[1] Of many possible covariates, age and prior drug use

[1] Although it may seem odd that if the subject's time to failure is *not* censored the subject receives a "1" for this variable, this is the convention followed in the literature and will be followed throughout this text as well.

were chosen for their potential clinical relevance as well as for statistical purposes to illustrate techniques for continuous and nominal scale predictor variables.

One of the most important differences between the outcome variables modeled via linear and logistic regression analyses and the time variable in the current example is the fact that we may only observe the survival time partially. The variable TIME listed in Table 1.1 actually records two different things. For those subjects who died, it is the outcome variable of interest, the actual survival time. However, for subjects who were alive at the end of the study, or for subjects who were lost, TIME indicates the length of follow-up (which is a partial or incomplete observation of survival time). These incomplete observations are referred to as being *censored.* For example, subject 1 died from AIDS 5 months after being seen in the HMO clinic (CENSOR = 1) while subject 2 was not known to have died from AIDS at the conclusion of the study and had been followed for 6 months (CENSOR = 0). It is possible for a subject to have entered the study 6 months before the end or he/she could have entered the study much earlier, eventually becoming lost to follow-up as a result of moving, failing to return to the clinic or some other reason. For the time being we do not differentiate between these possibilities and consider only the two states: dead (as a result of AIDS) and not known to be dead.

The main goal for a statistical analysis of these data is to fit a model that will yield biologically plausible and interpretable estimates of the effect of age and drug use on survival time, for HIV+ patients. Before beginning any statistical modeling, we should perform a thorough univariate analysis of the data to obtain a clear sense of the distributional characteristics of our outcome variable as well as all possible predictor variables. The fact that some of our observations of the outcome variable, survival time, are incomplete is a problem for conventional univariate statistics such as the mean, standard deviation, median, etc. If we ignore the censoring and treat the censored observations as if they were measurements of survival time, then the resulting sample statistics are not estimators of the respective parameters of the survival time distribution. They are estimators of parameters of a combination of the survival time distribution and a second distribution that depends on survival time as well as statistical assumptions about the censoring mechanism. For example, the average of TIME for subjects 1 and 2 in Table 1.1 is 5.5 months. The number 5.5 months is not an estimate of the mean length of survival. We can say the mean survival is estimated to be *at least* 5.5 months. But how can we appropriately use the fact that the survival time

Table 1.1 Study Entry and Ending Dates, Survival Time (Time), Age, History of IV Drug Use (Drug) and Vital Status (Censor) at Conclusion of Study

ID	Ent Date	End Date	Time	Age	Drug	Censor	ID	Ent Date	End Date	Time	Age	Drug	Censor
1	15May90	14Oct90	5	46	0	1	51	11Nov89	10Feb91	15	33	0	1
2	19Sep89	20Mar90	6	35	1	0	52	1Oct90	31Oct90	1	31	0	1
3	21Apr91	20Dec91	8	30	1	1	53	20Mar90	18Jan91	10	33	0	1
4	3Jan91	4Apr91	3	30	1	1	54	30Jul90	29Aug90	1	50	1	1
5	18Sep89	19Jul91	22	36	0	1	55	17Jul89	14Feb90	7	36	1	1
6	18Mar91	17Apr91	1	32	1	0	56	10Nov90	9Feb91	3	30	1	1
7	11Nov89	11Jun90	7	36	1	1	57	5Mar89	4Jun89	3	42	1	1
8	25Nov89	25Aug90	9	31	1	1	58	2Mar91	1May91	2	32	1	1
9	11Feb91	13May91	3	48	0	1	59	11Sep89	11May92	32	34	0	1
10	11Aug89	11Aug90	12	47	0	1	60	12Sep89	12Dec89	3	38	1	1
11	11Apr90	10Jun90	2	28	1	0	61	8Apr90	6Feb91	10	33	0	0
12	11May91	10May92	12	34	0	1	62	20Apr89	20Mar90	11	39	1	1
13	17Jan89	16Feb89	1	44	1	1	63	31Jan91	2May91	3	39	1	1
14	16Feb91	17May92	15	32	1	1	64	15Sep89	15Apr90	7	33	1	1
15	9Apr91	6Feb94	34	36	0	1	65	7Dec91	7May92	5	34	1	1
16	9Mar91	8Apr91	1	36	0	1	66	4Mar90	1Oct92	31	34	0	1
17	3Aug90	2Dec90	4	54	0	1	67	20Apr89	19Sep89	5	46	1	1
18	10Jun90	8Jan92	19	35	0	0	68	16Jun89	15Apr94	58	22	0	1
19	12Jun91	11Sep91	3	44	1	0	69	1Oct90	31Oct90	1	44	1	1
20	7Jan91	8Mar91	2	38	0	1	70	1Feb91	3May91	3	37	0	0
21	29Aug89	28Oct89	2	40	0	0	71	13May89	10Dec92	43	25	0	1
22	29May89	27Nov89	6	34	1	1	72	9Aug90	8Sep90	1	38	0	1
23	16Nov90	14Nov95	60	25	0	0	73	18Dec91	17Jun92	6	32	0	1
24	9May90	8Apr91	11	32	0	1	74	23Aug90	21Jan95	53	34	0	1
25	10Sep91	9Nov91	2	42	1	0	75	19Jan91	19Mar92	14	29	0	1
26	26Dec91	26May92	5	47	0	1	76	26Aug91	25Dec91	4	36	1	1
27	29May91	27Sep91	4	30	0	0	77	16May91	13Nov95	54	21	0	1
28	1May90	31May90	1	47	1	1	78	20Mar89	19Apr89	1	26	1	1
29	24Mar91	22Apr92	13	41	0	1	79	5Oct91	4Nov91	1	32	1	1
30	18Jul89	17Oct89	3	40	1	1	80	21May91	19Jan92	8	42	0	1
31	16Sep90	15Nov90	2	43	0	1	81	10Jun91	9Nov91	5	40	1	1
32	22Jun89	22Jul89	1	41	0	1	82	31Aug89	30Sep89	1	37	1	1
33	27Apr90	25Oct92	30	30	0	1	83	28Dec91	27Jan92	1	47	0	1
34	16May90	14Dec90	7	37	0	1	84	29Sep90	28Nov90	2	32	1	1
35	19Feb89	20Jun89	4	42	1	1	85	20Nov91	19Jun92	7	41	1	0
36	17Feb90	18Oct90	8	31	1	1	86	2Jul89	1Aug89	1	46	1	0
37	6Aug91	5Jan92	5	39	1	1	87	11Oct91	10Aug92	10	26	1	1
38	10Aug89	10Jun90	10	32	0	1	88	11Oct90	10Oct92	24	30	0	0
39	27Dec90	25Feb91	2	51	0	1	89	5Dec90	5Jul91	7	32	1	1
40	26Apr89	24Jan90	9	36	0	1	90	8Sep89	8Sep90	12	31	1	0
41	4Dec90	3Dec93	36	43	0	1	91	10Apr90	9Aug90	4	35	0	1
42	28Apr91	28Jul91	3	39	0	1	92	11Dec90	9Sep95	57	36	0	1
43	9Jul91	7Apr92	9	33	0	1	93	15Dec90	14Jan91	1	41	1	1
44	31Dec89	1Apr90	3	45	1	1	94	13Jan89	13Jan90	12	36	1	0
45	20Dec89	18Nov92	35	33	0	1	95	22Aug91	21Mar92	7	35	1	1
46	22Jun91	20Feb92	8	28	0	1	96	2Aug91	1Sep91	1	34	1	1
47	11Apr90	11Mar91	11	31	0	1	97	22May91	21Oct91	5	28	0	1
48	22May90	19Jan95	56	20	1	0	98	2Apr90	1Apr95	60	29	0	0
49	11Nov91	10Jan92	2	44	0	0	99	1May91	30Jun91	2	35	1	0
50	18Jan91	19Apr91	3	39	1	1	100	11May89	10Jun89	1	34	1	1

for subject 1 is *exactly* 5 months while that of subject 2 is *at least* 6 months? We return to the univariate descriptive statistics problem shortly.

Suppose for the moment that we have performed the univariate analysis and wish to explore possibilities for an appropriate regression model. In linear regression modeling the first step is usually to examine a scatterplot of the outcome variable versus all continuous variables to see if the "cloud" of data points supports the use of a straight-line model. We also assess if there appears to be anything unusual in the scatter about a potential model. For example, is the linear model plausible except for one or two points? The fact that we have censored data presents a problem for the interpretation of a scatterplot with survival time data. If we were to ignore the censoring in survival time, then we would have an extension of the problem we noted with use of the arithmetic mean as an estimator of the "true" mean. The values obtained from any "line" fit to the cloud of points would not estimate the "mean" at that point. We would only know that the "mean" is *at least* as large as the point on the "line."

Regardless of this "at least" problem, a scatterplot is still a useful and informative descriptive tool with censored survival time data. However, to interpret the plot correctly we must keep track of the different types of observations by using different plotting symbols for the values assigned to the censoring variable. Figure 1.1 presents the scatterplot of TIME versus AGE for the data in Table 1.1, where different plotting symbols are used for the two levels of CENSOR. We formalize the statistical assumptions about the censoring later in Chapter 1, but for the moment we assume that it is independent of the values of survival time and all covariate variables.

Under the independence assumption the censored and non-censored points should be mixed in the plot with the mix dictated by the study design. Any trend in the plot is controlled by the nature and strength of the association between the covariate and survival time. For example, if age has a strong negative association with survival time, then observed survival times should be shorter for older subjects than for younger ones. If all subjects were followed for the *same fixed length of time*, then we would expect to find proportionally more censored observations among younger subjects than older ones. However, if subjects enter the study *uniformly over the study period* and independently of their age, then we would expect an equal proportion of censored observations at all ages. The example data are assumed to be from a study of