Applied Survival Analysis (171:242)

Section 1: Introduction

Brian J. Smith, Ph.D.

January 24, 2005

Table of Contents

1.1 Review of Basic Statistical Methods

1.1.1 Length of Hospital Stay Example

The Health Services Administration was interested in comparing the length of stay for patients at two different hospitals. Patients with the same diagnosis were enrolled in a two-day pilot study – 11 from the first hospital and 13 from the second. The length of time that each subject spent in the hospital during the study period was recorded. The data are given in Table 2.

Table 1. Variables in the Length of Hospital Stay Study

Variable	Description	Values
ID	Patient study identifier	integer
Hospital	Hospital identifier	1, 2
Hrs	Observed length of hospital stay in hours	0 – 48
Discharge	Patient was discharged	1 = yes 0 = no

Table 2. Data for the Length of Hospital Stay Study

ID	Hospital	Hrs	Discharge	ID	Hospital	Hrs	Discharge
1	1	21	1	13	2	27	1
2	1	10	1	14	2	10	1
3	1	32	1	15	2	48	0
4	1	48	0	16	2	48	0
5	1	8	1	17	2	48	0
6	1	44	1	18	2	48	0

ID	Hospital	Hrs	Discharge	ID	Hospital	Hrs	Discharge
7	1	29	1	19	2	48	0
8	1	5	1	20	2	35	1
9	1	13	1	21	2	48	0
10	1	26	1	22	2	48	0
11	1	33	1	23	2	44	1
12	2	48	0	24	2	48	0

An alternative summary of the data is given in Figure 1.

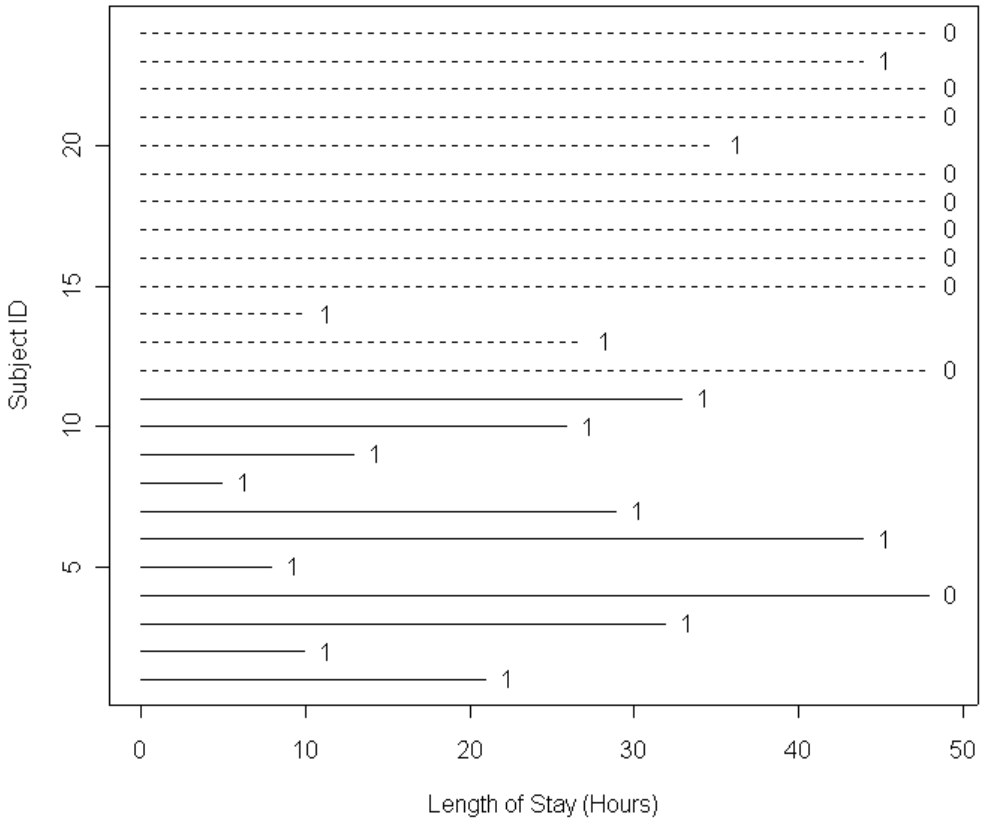


Figure 1. Length of hospital stays for the two-day pilot study.

Analysis Goal: *Test for differences in length of stay between the two hospitals.*

Outcome Variables

Note that outcome in this study consists of two random variables:

1. The observed length of stay (in hours) as a continuous variable.
2. A dichotomous variable that indicates whether the patient left the hospital within two days.

Basic statistical methods could be applied to each of the two random variables separately in order to test for differences between the hospitals.

1.1.2 Analysis 1: Length of Stay

One strategy might be to focus on the continuous length of stay data. These data are summarized below.

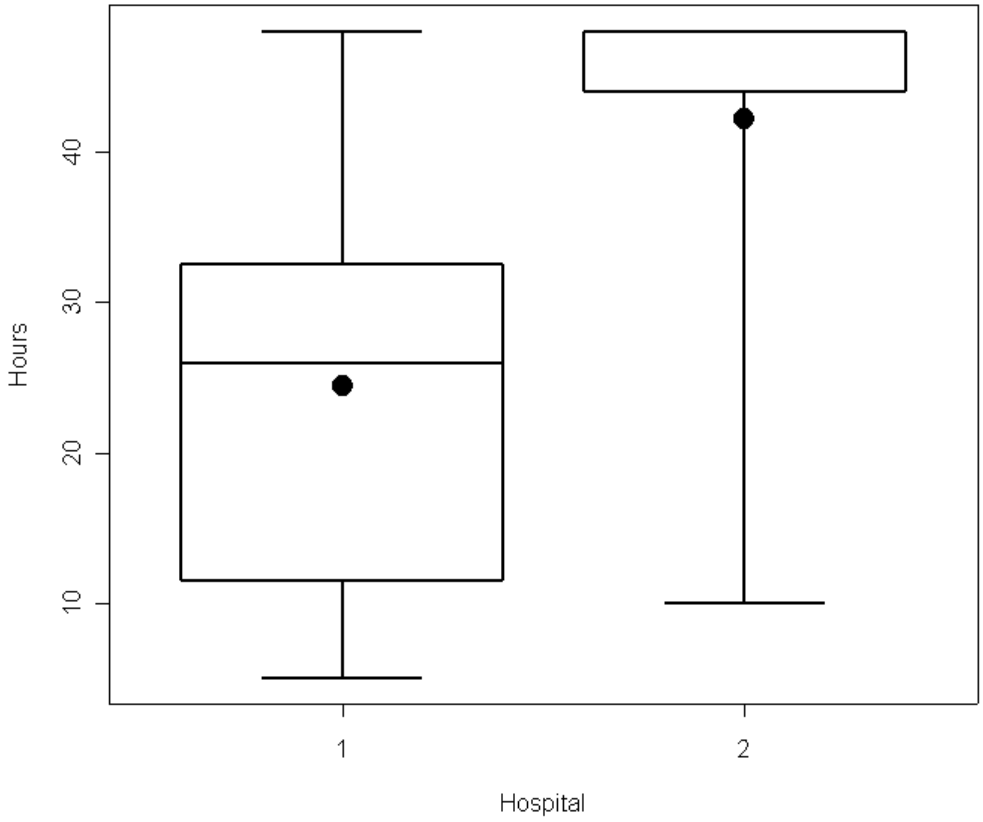


Figure 2. Box plot comparison of length of stay by hospital.

Table 3. Summary statistics for hospital stay study.

	Hospital 1	Hospital 2
N	11	13
Mean	24.5	42.2
Median	26.0	48.0
Standard Deviation	14.5	11.6

What characteristics of the length of stay data are important in choosing an appropriate statistical test?

What statistical tests could be used to compare a continuous variable across two groups?

Results

The Wilcoxon rank-sum test was used to compare the length of stay between the two hospitals. A two-sided p-value of 0.0029 was obtained. Therefore, at the 5% level of significance, the length of stay differs between the two hospitals. Length of stay at Hospital 2 is significantly greater than that at Hospital 1.

Pitfall

One possible strategy is to omit patients, from the analysis, who did not leave the hospital during the study period (see Figure 3). However, these patients provide valuable information, and their omission could severely bias the study results. All subjects should be included in the analysis.

- The fact that their length of stay is known to be greater than two days should be utilized.
- We are already familiar with one test that can be applied to the full data set.
- We will learn about other methods for analyzing such data.

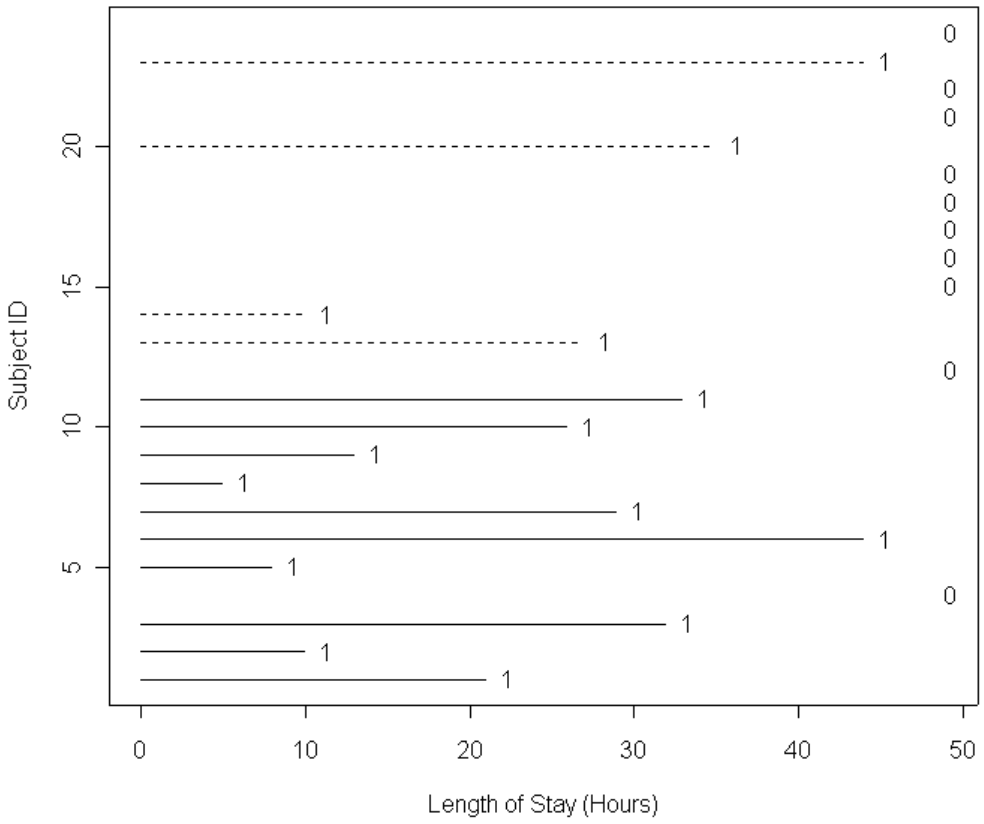


Figure 3. Length of hospital stays for patients who were discharged during the two-day study.

1.1.3 Analysis 2: Discharged within Two Days

A second strategy might be to focus on the simple Yes/No variable indicating whether the patient was discharged during the study period. Since everyone was followed for exactly two days, this dichotomous variable can be used to estimate the probability of being discharged within two days. Consequently, a two-sample test for binomial proportions can be performed to compare the hospitals.

Table 4. Number of subjects discharged within two days for each hospital.

Hospital	Time of Discharged		Totals
	< 48 Hours	≥ 48 Hours	
1	10	1	11
2	4	9	13

Results

Fisher's exact test was used to compare the hospitals with respect to the probability of being discharged within two days. A two-sided p-value of 0.0045 was obtained. Therefore, at the 5% level of significance, the probabilities differ between the two hospitals. Patients in Hospital 1 are more likely to be discharged within two days.

1.2 Basic Concepts

1.2.1 Survival Applications

The statistical techniques covered in this course are commonly referred to as “survival analysis” because many originated from studies of time to death data. Survival analysis, however, generally refers to statistical methods for the analysis of any time to some event outcome.

Potential time-to-event data:

Medical Field

- Death
- Relapse
- Occurrence of symptoms
- Disease onset

Sociology

- Divorce
- Career change
- Smoking cessation
- First marijuana use

Reliability

- Product failure
- Machine repair

Business/Economics

- Bankruptcy
- Unemployment assistance
- Divestiture of stocks
- Labor strike duration

Examples of public health applications:

- Therapeutic trials – Randomize patients to different treatments for bladder cancer. Follow patients to determine which treatment group has the longest disease-free survival.
- Intervention trials – One group of smokers participates in a support group; the other does not. Measure time to smoking cessation.
- Epidemiology – Enroll a cohort of uranium miners to study the effects of radon exposure on lung cancer risk. The event of interest is age at lung cancer diagnosis. What other risk factors should be considered in the analysis?

1.2.2 Censoring

Follow-up data result from subjects being observed or followed for a period of time. Subjects for whom disease is not observed at the end of their follow-up period are said to be **censored**. The **follow-up time** is defined to be the length of time from study entry until disease occurrence or censoring.

Reasons for censoring:

- Follow-up loss due to migration, lack of cooperation, withdrawal of consent, etc.
- Death from another cause.
- No longer at risk; e.g. hysterectomy.
- Termination of study prior to disease occurrence.

Different types of censoring:

- Right Censoring – events would necessarily take place after the follow-up period.
- Left Censoring - events that take place at some unknown time prior to the follow-up period.
- Interval Censoring – events that are known to occur only within a certain time interval.

- Type I Censoring – events are observed only if they occur prior to some pre-specified time.
- Type II Censoring – the study is terminated after a pre-specified number of events are observed.

The design of the hospital stay example was straightforward. All subjects were enrolled at the same point in time for an observation period of two days. In general, the entry times and follow-up periods can vary from subject-to-subject, as illustrated in Figure 4.

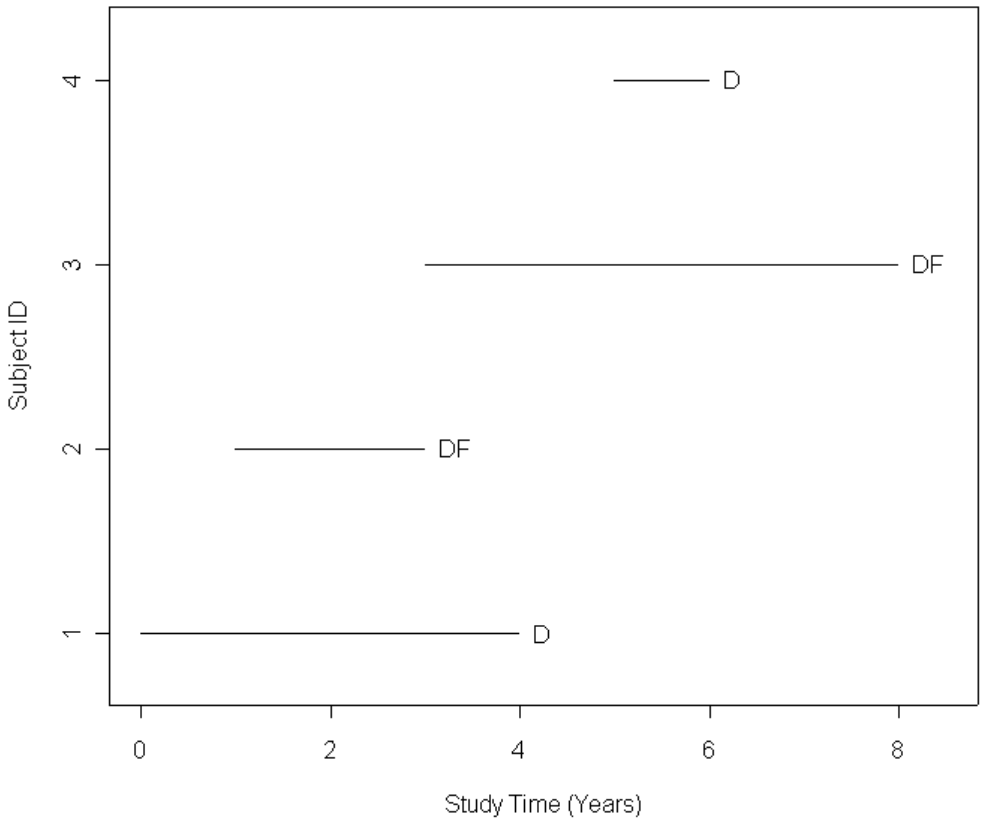


Figure 4. Subject follow-up versus study time.

Note that,

- D and DF denote subjects who were diseased and disease-free, respectively, at the end of their follow-up period.
- Subject 1 was enrolled at the start of the study and developed disease at year 4. Total follow-up time is 4 years.
- Subject 2 was enrolled 1 year into the study and withdrew, disease-free at year 3. Total follow-up time is 2 years.

- Subject 3 was enrolled 3 years into the study and was disease free at the end of the 8-year study. Total follow-up time is 5 years.
- Subject 4 was enrolled 5 years into the study and developed disease at year 6. Total follow-up time is 1 year.
- Subjects 2 and 3 are treated as censored observations since they did not develop disease during the time that they were enrolled in the study.

Table 5. Summary of Follow-up Data

ID	Start Year	Stop Year	Total Years	Status	Censored
1	0	4	4	Diseased	No
2	1	3	2	Withdrew	Yes
3	3	8	5	Disease-Free	Yes
4	5	6	1	Diseased	No

Alternatively, the data could be presented as total length of follow-up, as in Figure 5.

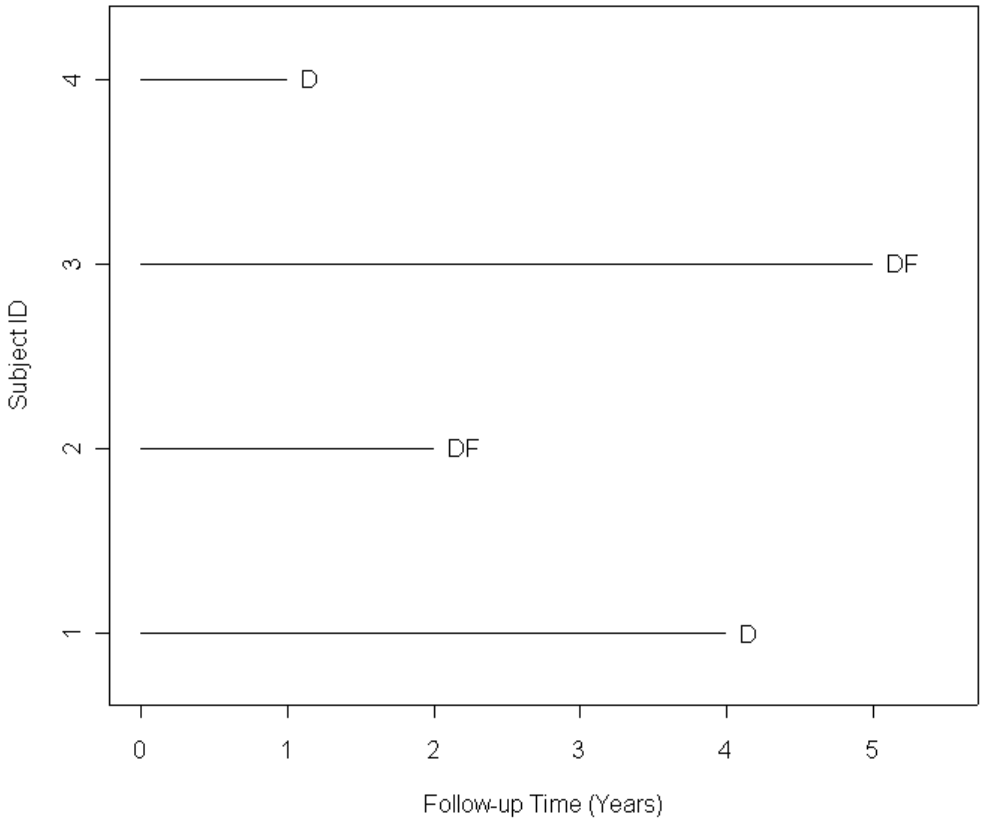


Figure 5. Subject participation in terms of the actual follow-up time.

The strategies discussed in Section 1.1 are not applicable to the more general problem of censoring.

1.2.3 Notation

T = random variable denoting the follow-up time, i.e. time to the event of interest or censoring.

δ = indicator for the event of interest (1 = event is observed, 0 = censoring)

Follow-up time for a **sample** of N individuals may be denoted as $\{(t_1,\delta_1),(t_2,\delta_2),\ldots,(t_N,\delta_N)\}$.

1.3 Summary

Survival data presents a new class of problems:

- Data may be censored
- Outcome is time to some event or censoring
- Outcome variable is not normally distributed

We need to learn new statistical methods to deal with survival data. The methods will provide the means to:

- Summarize and graphically display the data
- Quantify uncertainty in estimating survival
- Test hypotheses
- Model the effects of covariates on survival

Applied Survival Analysis (171:242)
Section 2: Kaplan-Meier Estimator

Brian J. Smith, Ph.D.

February 15, 2005

Table of Contents

2.1 Estimation of Survival Probabilities

2.1.1 Lymphoblastic Leukemia Clinical Trial Example

A clinical trial was conducted to study remission maintenance in children with acute lymphoblastic leukemia. Forty-two patients, who had achieved complete remission, were randomized to receive maintenance therapy with 6-mercaptopurine or placebo.

Table 1. Variables in the Leukemia Trial

Variable	Description	Values
id	Patient study identifier	integer
group	Treatment group	6-MP Placebo
weeks	Weeks until relapse or censoring	1-35
relapse	Relapse indicator variable	1 = yes 0 = no

The observed number of weeks until relapse or censoring (*) for the patients in the trial are given below.

6-MP (21 patients): 6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*

Placebo (21 patients): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Analysis Objective: *Provide a statistical summary of the time to relapse data.*

Leukemia Follow-up Data

A plot of the follow-up times is given in Figure 1. The plot provides a graphical display of the raw data. However, it does not summarize any particular feature of the data.

Note that the data in this trial exemplify the more general type of time-to-event data:

1. Patients in 6-MP group are censored
2. Outcome is time to relapse or censoring
3. Observation periods differ between patients

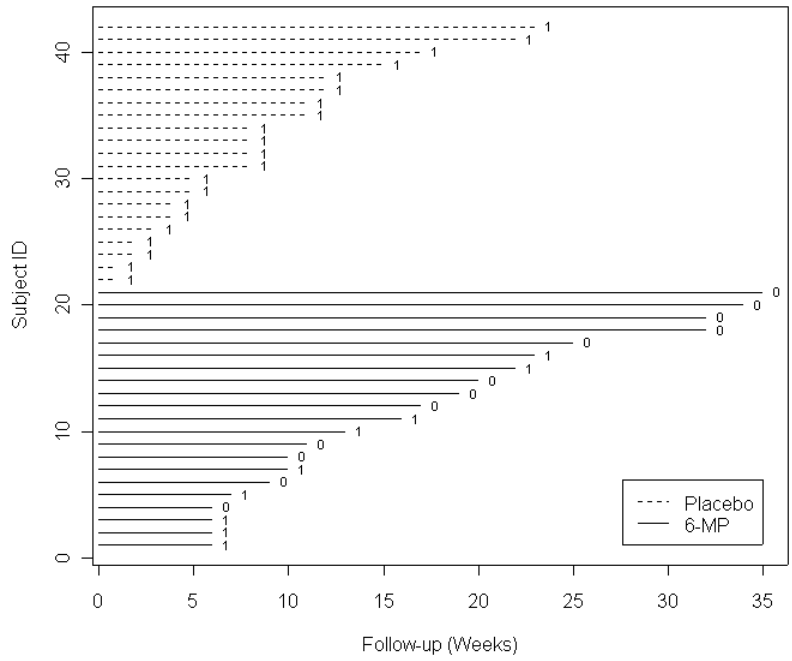


Figure 1. Follow-up times for the Leukemia Trial.

Analysis Strategy: We would like to compute summary statistics for the data. Classical approaches:

- Mean (Standard Deviation) time to relapse

- Median time to relapse

- Proportion of patients who relapse

2.1.2 Survival Function

Recall that the random variable T denotes the time to the event of interest or censoring. Traditionally, T is referred to as the survival time and observed events as failures.

Definition

The *survival function* $S(t) = \Pr[T > t]$ is the probability of surviving beyond time t. Note that the survival function $S(t)$ is a probability and therefore assumes values in the interval [0, 1].

Examples:

1. If the event of interest is death, then $S(t)$ represents the probability of living beyond time t.
2. If the event is relapse among leukemia patients, then $S(t)$ is the probability of being in remission beyond time t.
3. A theoretical survival function is plotted in Figure 2. The survival function gives the cumulative probability of survival as a function of time. The following information can be obtained from the survival curve in the figure:

    - For $t = 0$, $S(0) = 1$; all of the subjects (100%) survive beyond the initial follow-up point. By definition, the survival function must be equal to 1 at the start of follow-up.

- For t = 3.465, $S(3.465) = 0.500$; half of the subjects survive beyond this point in the study.
- For $t = 5$, $S(5) = 0.368$; 36.8% of the subjects survive beyond this point.

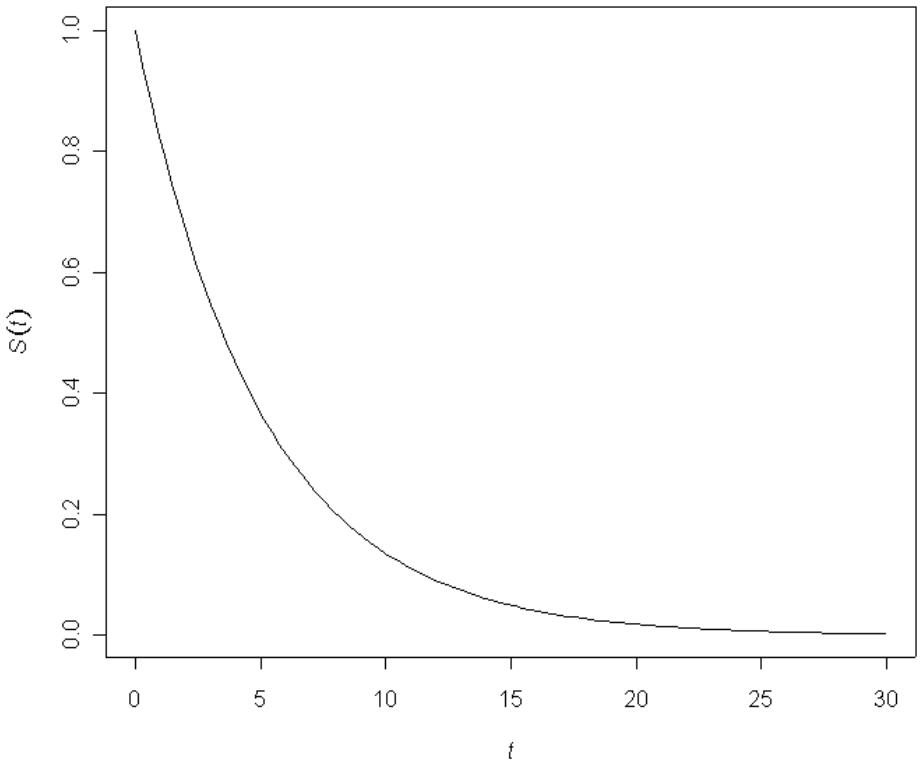


Figure 2. Plot of a theoretical survival function $S(t)$.

2.2 The Kaplan-Meier (Product Limit) Estimator

2.2.1 Overview

The methods of Kaplan-Meier provide an estimate of the survival function $S(t)$ using time-to-event data from a sample of subjects. The Kaplan-Meier estimator has the following properties:

- Nonparametric method for estimating survival.
- Yields an estimate of survival at any point in time during the follow-up period.
- Also referred to as the Product-Limit estimator.
- Allows for censoring and varying lengths of follow-up.

Leukemia Trial Example

We will use the times to relapse or censoring (*) in the 6-MP treatment group to illustrate the Kaplan-Meier estimator:

6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*

Step 1: Construct a table with a row for the starting time of follow-up and each subsequent time point at which an event (relapse) occurs.

Table 2. Kaplan-Meier Estimate of the Cumulative Survival Function in the Leukemia Trial

Week (t)	Number at Risk (n_t)	Number of Events (e_t)	p_t	s_t
0	21	0	1.000	1.000
6	21	3	0.857	0.857
7	17	1	0.941	0.806
10	15	1		
13	12	1		
16	11	1		
22	7	1		
23	6	1		

Step 2: Calculate the proportion p_t of subjects at risk at time t who do not experience an event

$$p_t = \frac{n_t - e_t}{n_t}.$$

This is referred to as the conditional probability of survival (remaining disease-free) at time t. For example,

$$p_0 = \frac{21-0}{21} = 1.00$$

$$p_6 = \frac{21-3}{21} = 0.857.$$

$$p_7 = \frac{17-1}{17} = 0.941$$

Step 3: Calculate the proportion of original subjects that are survivors at time t

$$s_t = \prod_{t_i \le t} p_{t_i} \ .$$

In our example,

$$s_0 = p_0 = 1.000$$
$$s_6 = p_0 p_6 = (1.000)(0.857) = 0.857 \quad .$$
$$s_7 = p_0 p_6 p_7 = (1.000)(0.857)(0.941) = 0.806$$

- This is called the Kaplan-Meier estimate of the cumulative survival function.
- s_t is the estimated probability of surviving beyond any time-point in the interval $[t, t')$, where t' is the next time point at which an event is observed.

2.2.2 Motivation

The Kaplan-Meier estimator provides an estimate of the probability of surviving beyond any given point in time during the follow-up period. Consider the 6-MP group in the Leukemia Trial. To remain in remission past time *t*, a patient has to have been in remission at every prior time point in the study.

Week 0

All patients are in remission at the start of the study. Thus, the corresponding survival probability is 1.

Week 6

The first set of relapses occurred at 6 weeks of follow-up. The patients who disease-free and at risk to relapse are referred to as the *risk set*.

- At week 6 there are _____ patients in the risk set.
- _____ patients in the risk set relapse.
- The probability of survival at week 6 is the number of patients who did not relapse divided by the number of patients in the risk set.

$$p_6 = \frac{\text{risk set at week 6 - relapses at week 6}}{\text{risk set at week 6}}$$
$$= \frac{21-3}{21} = \frac{18}{21} = 0.857$$

Week 7

The second set of relapses occurred 7 weeks.

- At week 7 there are _____ patients in the risk set.
- _____ patients in the risk set relapse.
- The probability of survival at week 7 is the number of patients who did not relapse divided by the number of patients in the risk set.

$$p_7 = \frac{\text{risk set at week 7 - relapses at week 7}}{\text{risk set at week 7}}$$
$$= \frac{17-1}{17} = \frac{16}{17} = 0.941$$

Notice that the 17 patients in the risk set at week 7 had necessarily survived past week 6. In fact, p_7 is a conditional probability. It is the probability of surviving beyond week 7, given that the patient survived beyond week 6. The survival function which we want to estimate is an unconditional probability; namely, the probability of surviving beyond week 7. The unconditional probability of interest can be written as

$$s_7 = p_0 p_6 p_7$$

This conditioning argument can be extended to estimate survival beyond any time point *t*. In general, p_t is the conditional probability of surviving beyond time *t*, and the unconditional probability can then be written as

$$s_t = \prod_{t_i \le t} p_{t_i}$$

which is also referred to as the *cumulative probability* of survival. To illustrate, the probability of surviving beyond week 10 is

$$s_{10} = p_0 p_6 p_7 p_{10}$$

and the probability of surviving beyond week 13 is

$$s_{13} = p_0 p_6 p_7 p_{10} p_{13}\,.$$

Q: Do we need to include conditional probabilities at time points for which a relapse did not occur, for instance, the estimated conditional probability at week 11?

A: The estimated probability of remission is equal to 1 at time points for which a relapse did not occur. Hence, only probabilities at time points for which an event occurred need to be explicitly included in the calculation. The cumulative probabilities of remission are summarized in Table 3.

Table 3. Estimated survival (remission) for the 6-MP group in the Leukemia Trial.

Time to Relapse	Censored	Relapse	Risk Set	Conditional Probability	Cumulative Probability
0	-	-	21	21/21=1.000	1.000
6	1	3	21	18/21=0.857	0.857
7	-	1	17	16/17=0.941	0.806
10	1	1	15	14/15=0.933	0.752
13	-	1	12	11/12=0.917	0.690
16	-	1	11	10/11=0.909	0.627
22	-	1	7	6/7=0.857	0.537
23	-	1	6	5/6=0.833	0.447

2.2.3 Kaplan-Meier Survival Plots

The estimated cumulative probability can be plotted against time to provide a graphical display of the probability of remission in the two treatment groups. The resulting *survival curves* start at 1 and changes values at each distinct relapse time.

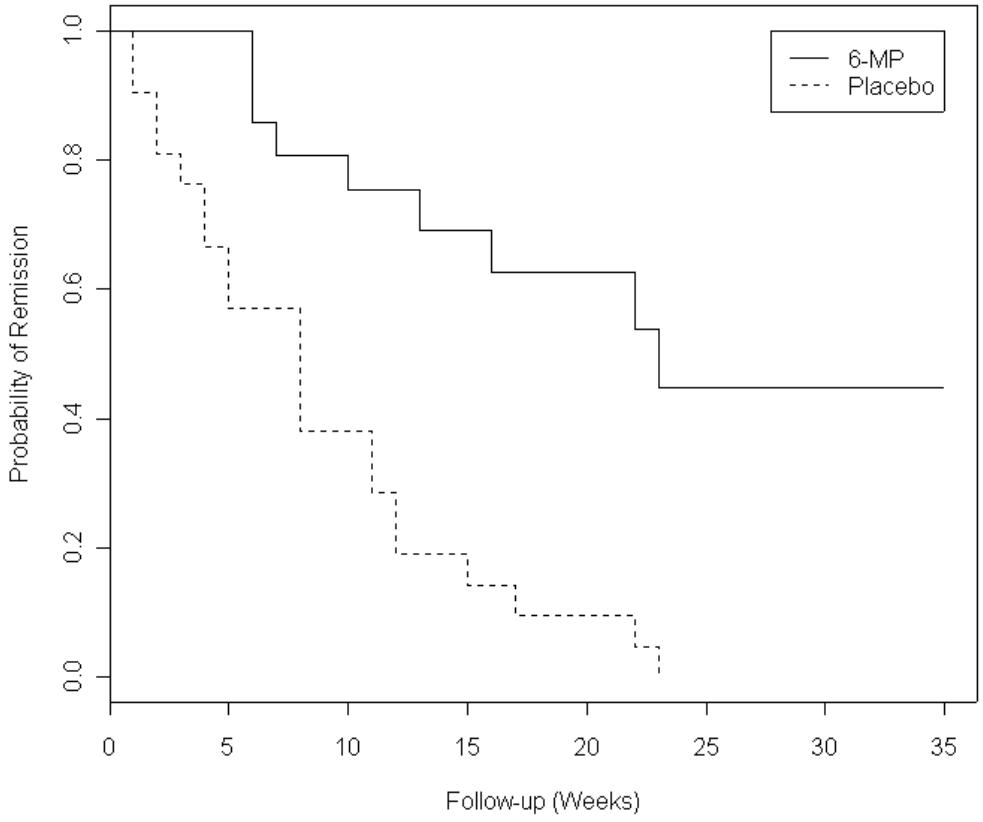


Figure 3. Kaplan-Meier plot of the cumulative probability of remission in the two treatment groups of the Leukemia Trial.

The plot indicates that patients in the 6-MP treatment group are likely to stay in remission longer than those in the placebo group. Indeed, after 23 weeks no patient would be expected to be in remission in the placebo group; whereas, 45% would be expected in the treatment group. Note that the survival curves in this plot provide no information about the uncertainty in our survival estimates. Thus, we cannot yet determine if there is a statistically significant difference between the two groups.

2.2.4 Notes

- The Kaplan-Meier method provides a non-parametric estimate of the survival function $S(t)$. No assumptions are made about the functional form of the distribution for survival times.
- Assumptions: 1) continuous survival times, and 2) the censoring mechanism is independent of the event of interest.
- The estimates from this method are commonly denoted $\hat{S}_{KM}(t)$.
- If censoring occurs at the greatest observed survival time, then the estimated survival function will not reach zero. Care should be taken when interpreting the tail of the survival curve since the corresponding number of subjects in the risk set may be small.

2.2.5 Variable Follow-up Periods

Breast Cancer Example

Suppose that the data below were collected in a prospective cohort study designed to estimate the risk of breast cancer as a function of age. Included in the table

are the ages at which subjects were enrolled as well as the ages at censoring or disease occurrence.

ID	Age at Enrollment	Age at Diagnosis or Censoring	Years	Breast Cancer
1	27	32	5	No
2	30	35	5	No
3	30	35	5	Yes
4	32	37	5	No
5	34	40	6	No
6	36	41	5	Yes
7	34	41	7	Yes
8	42	45	3	No
9	43	47	4	No
10	31	50	19	Yes

Use the methods of Kaplan-Meier to estimate and plot disease-free survival as a function of age.

Table 4. Kaplan-Meier Estimate of the disease-free survival in the Breast Cancer Example.

Age (t)	Number at Risk (n_t)	Number of Events (e_t)	p_t	s_t
27	1	0	1.000	1.000
35				
41				
50				

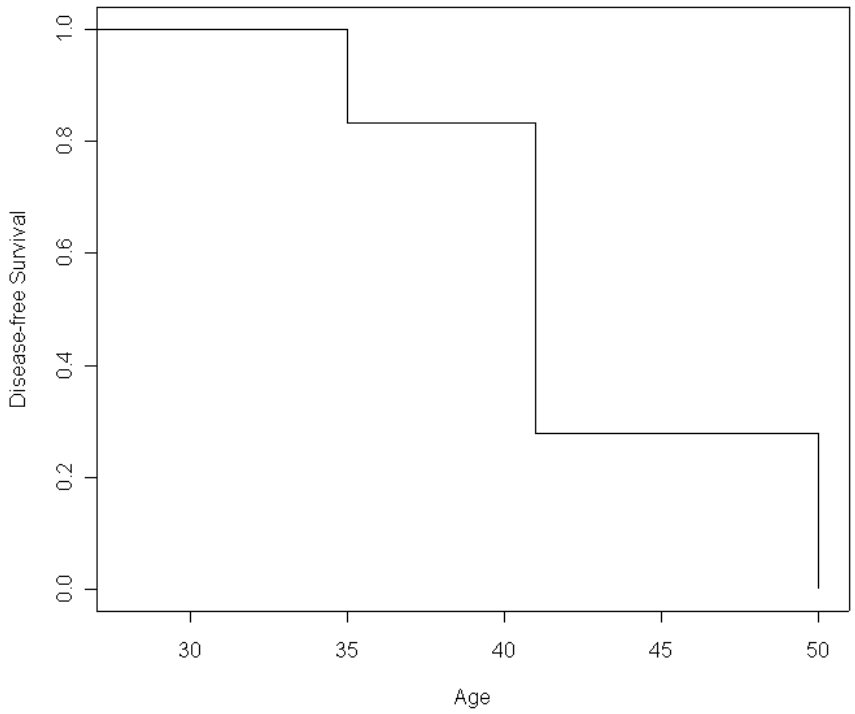


Figure 4. Kaplan-Meier plot of the disease-free survival in the Breast Cancer Example.

2.2.6 Precision of the Survival Estimates

The Kaplan-Meier estimator provides only point estimates of the survival function. It gives a statistical summary of the survival distribution and, like any other statistic, is subject to random variation. Methods for quantifying the variation in the survival estimates are presented in this section.

We will consider two confidence interval formulas for $S(t)$: the more commonly used formula of Greenwood and the formula of Kalbfleisch and Prentice.

Greenwood Formula

The Greenwood formula for the variance of $\hat{S}_{KM}(t)$ is

$$\text{var}_G(t_0) = 0$$

$$\text{var}_G(t_i) = \text{var}_G(t_{i-1}) \times \left(\frac{n_{t_i} - e_{t_i}}{n_{t_i}} \right)^2 + \hat{S}^2_{KM}(t_i) \times \frac{e_{t_i}}{n_{t_i}\left(n_{t_i} - e_{t_i}\right)}$$

The variances for the 6-MP group in the Leukemia Trial are calculated in Table 5. The variance formula is not a function of the censoring times. Thus, the table does not need to include the censored time points.

Table 5. Greenwood variance for the 6-MP group in the Leukemia Trial.

t	n_t	e_t	$\frac{n_t - e_t}{n_t}$	$\hat{S}_{KM}(t)$	$\frac{e_t}{n_t(n_t - e_t)}$	$\text{var}_G(t)$
6	21	3	0.857	0.857	0.00794	0.00583
7	17	1	0.941	0.807	0.00368	0.00756
10	15	1	0.933	0.753	0.00476	0.00928
13	12	1	0.917	0.690	0.00758	0.0114
16	11	1	0.909	0.628	0.00909	0.0130
22	7	1	0.857	0.538	0.0238	0.0164
23	6	1	0.833	0.448	0.0333	0.0181

In large samples, $\hat{S}_{KM}(t)$ is approximately normally distributed with mean $S(t)$ and variance given by Greenwood's formula. Hence, an approximate 95% confidence interval for $S(t)$ is

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\text{var}_G(t_i)}$$

for $t_i \le t < t_{i+1}$. Confidence intervals for the 6-MP group are given in Table 6. Unfortunately, Greenwood's formula can produce confidence limits outside of the range [0, 1]. Despite this limitation, Greenwood's formula is widely used in practice.

Table 6. Greenwood 95% confidence intervals for the 6-MP group in the Leukemia Trial.

$[t_i, t_{i+1})$	$\hat{S}_{KM}(t)$	$\text{var}_G(t_i)$	95% CI
[6, 7)	0.857	0.00583	(0.707, 1.01)
[7, 10)	0.807	0.00756	(0.637, 0.977)
[10, 13)	0.753	0.00928	(0.564, 0.942)
[13, 16)	0.690	0.0114	(0.481, 0.899)
[16, 22)	0.628	0.0130	(0.404, 0.852)
[22, 23)	0.538	0.0164	(0.287, 0.789)
[23, ∞)	0.448	0.0181	(0.184, 0.712)

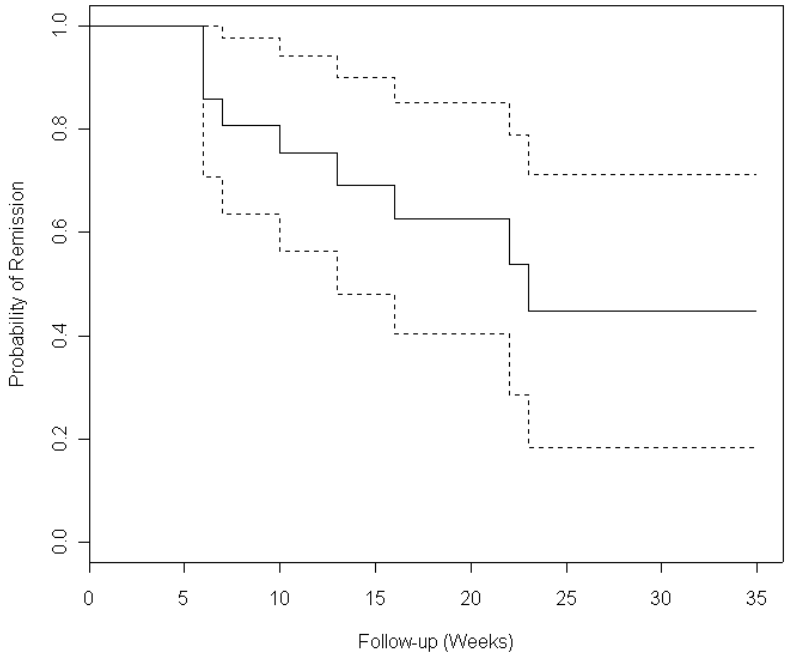


Figure 5. Kaplan-Meier survival curve for the 6-MP group with 95% confidence limits from Greenwood's formula.

Kalbfleisch and Prentice Formula

The Kalbfleisch and Prentice formula for the variance of $\log\left(-\log\left(\hat{S}_{KM}(t)\right)\right)$ is

$$\text{var}_{KP}(t_i) = \frac{\text{var}_G(t_i)}{\hat{S}^2_{KM}(t_i) \times \left[\ln\left(\hat{S}_{KM}(t_i)\right)\right]^2}.$$

The variances for the 6-MP group in the Leukemia Trial are calculated in Table 7.

Table 7. Kalbfleisch and Prentice variance for the 6-MP group in the Leukemia Trial.

t	$\hat{S}_{KM}(t)$	$\text{var}_G(t)$	$\text{var}_{KP}(t)$
6	0.857	0.00583	0.333
7	0.807	0.00756	0.252
10	0.753	0.00928	0.203
13	0.690	0.0114	0.174
16	0.628	0.0130	0.152
22	0.538	0.0164	0.148
23	0.448	0.0181	0.140

An approximate 95% confidence interval for the Kalbfleisch and Prentice approach is given by

$$\left(\hat{S}_{KM}(t)^{\exp\left(1.96\sqrt{\text{var}_{KP}(t_i)}\right)}, \hat{S}_{KM}(t)^{\exp\left(-1.96\sqrt{\text{var}_{KP}(t_i)}\right)}\right)$$

for $t_i \le t < t_{i+1}$. Confidence intervals for the 6-MP group are given in Table 8. Confidence limits based on this method will always fall within the range [0, 1]. For this reason, the Kalbfleisch and Prentice formula is sometimes preferred over the Greenwood method.

Table 8. Kalbfleisch and Prentice 95% confidence intervals for the 6-MP group in the Leukemia Trial.

$[t_i, t_{i+1})$	$\hat{S}_{KM}(t)$	$\text{var}_{KP}(t_i)$	95% CI
[6, 7)	0.857	0.333	(0.620, 0.954)
[7, 10)	0.807	0.252	(0.563, 0.923)
[10, 13)	0.753	0.203	(0.503, 0.889)
[13, 16)	0.690	0.174	(0.431, 0.849)
[16, 22)	0.628	0.152	(0.368, 0.805)
[22, 23)	0.538	0.148	(0.268, 0.747)
[23, ∞)	0.448	0.140	(0.188, 0.680)

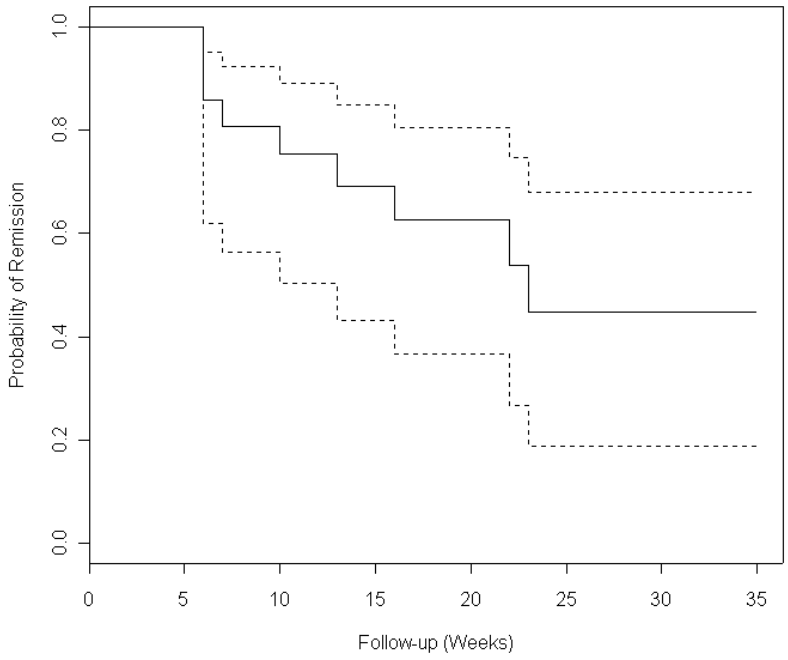


Figure 6. Kaplan-Meier survival curve for the 6-MP group with 95% confidence limits from the Kalbfleisch and Prentice formula.

2.2.7 Median Survival Time

The median survival time T_{50} is defined as the first time point beyond which 50% of the subjects are expected to survive. One can easily read the median off of the Kaplan-Meier curves. In general, these curves can be used to compute the p^{th} percentile T_p. However, the survival curve must drop below $p/100$ in order to estimate the corresponding percentile.

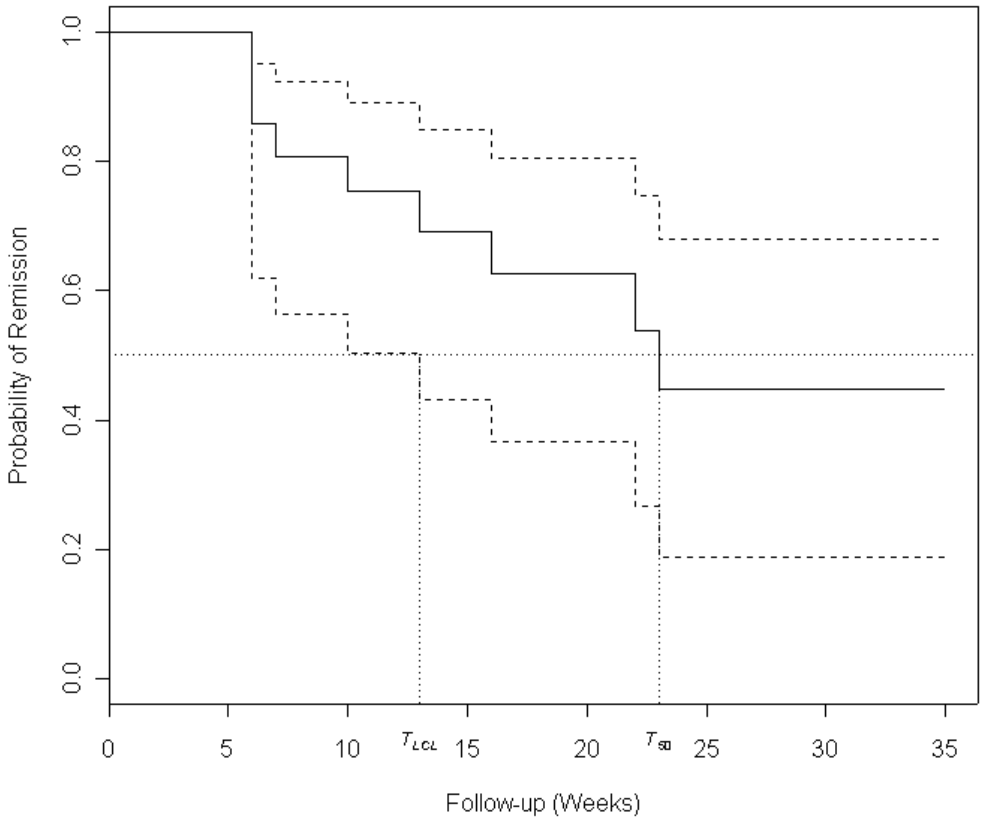


Figure 7. Kaplan-Meier estimate of the median survival time for the 6-MP group in the Leukemia Trial.

For the Leukemia Trial, the estimated median disease-free survival is 23 weeks for the 6-MP group. The confidence interval for the median is defined as the time points where the confidence intervals for the survival curve drop below 0.50. In this case the lower bound of the 95% confidence interval for the median is 13 weeks; the upper bound is not defined since the upper bound of the confidence interval for the survival curve does not drop below 0.50.

Likewise, certain percentiles, such as the 25th percentile T_{25} for the 6-MP group cannot be estimated because the survival curve does not drop below 0.44.

2.2.8 Mean Survival Time

The mean survival time is estimated as

$$\hat{\mu} = \sum_{i=1}^{k} \hat{S}_{KM}\left(t_{i-1}\right)\left(t_i - t_{i-1}\right)$$

with standard error given by

$$se\left(\hat{\mu}\right) = \sqrt{\frac{m}{m-1}\sum_{i=1}^{k-1}\frac{A_i^2}{n_i\left(n_i - e_i\right)}}$$

$$A_i = \sum_{j=i}^{k-1} \hat{S}_{KM}\left(t_j\right)\left(t_{j+1} - t_j\right) \ .$$

$$m = \sum_{j=1}^{k} e_j$$

where k is the number of unique time points at which events were observed. The mean survival for the 6-MP group in the Leukemia Trial is calculated in the table below.

t	$\hat{S}_{KM}(t)$	$t_i - t_{i-1}$	$\hat{S}_{KM}(t_{i-1})(t_i - t_{i-1})$
0	1.000	-	-
6	0.857	6	6
7	0.807	1	0.857
10	0.753	3	2.421
13	0.690	3	2.259
16	0.628	3	2.07
22	0.538	6	3.768
23	0.448	1	0.538
Total	-	-	17.91

The mean disease-free survival in the 6-MP group is 17.9 weeks. Note that when the largest observed time is censored, this estimator will underestimate the true mean.

2.3 Comments

2.3.1 Cumulative Probability of Failure

It may be helpful to plot the cumulative probability of failure when reporting results. The cumulative probability of failure can be estimated as $1-\hat{S}_{KM}(t)$. In other words, it can be estimated by subtracting the Kaplan-Meier estimates from 1. Figure 8 plots the estimated cumulative probability of relapse in the Leukemia Trial.

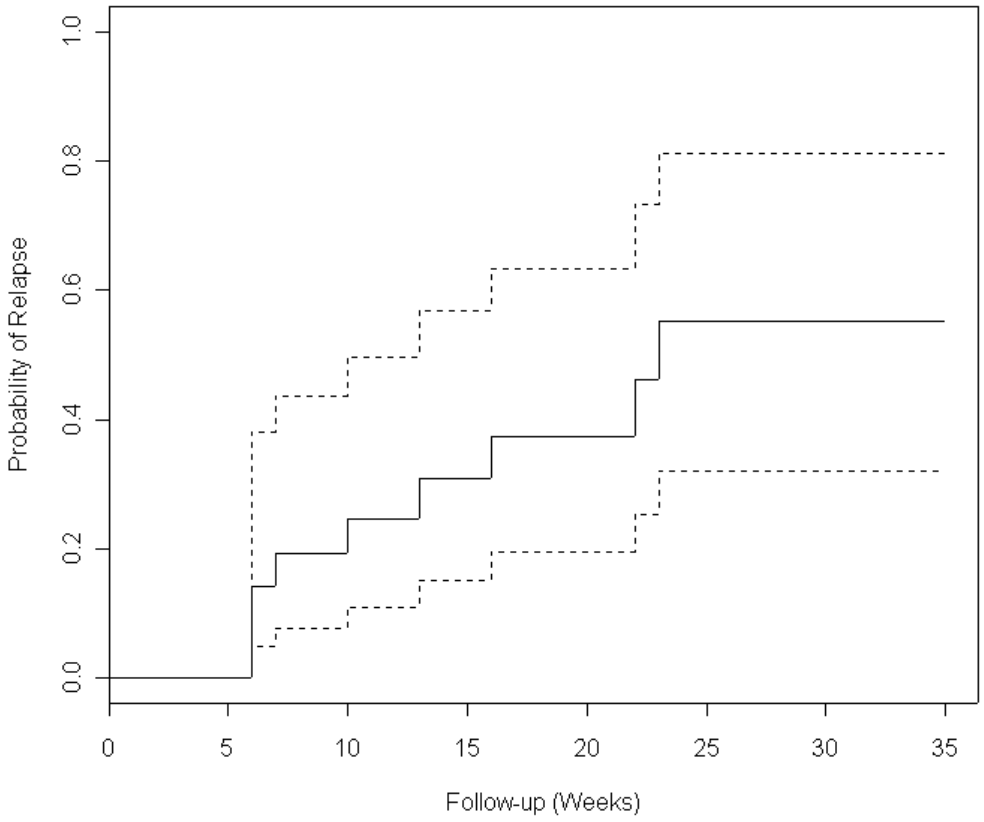


Figure 8. Kaplan-Meier plot of the cumulative probability of relapse for the 6-MP group in the Leukemia Trial.

2.3.2 Interpreting the Survival Curve

The estimated survival curves display patterns in the time-to-event data. Conclusions based on the fine detail of a curve (e.g. the size of a specific step or the length of a flat region) should be made with caution.

- Large steps in the survival curve do not necessarily imply a sudden increase in the risk of failure. For instance, one should not assume from Figure 8 that

the large steps around weeks 22 and 23 imply a sudden increase in the probability of relapse. Recall that there are a small number of subjects at risk during those time periods. Hence, there is more uncertainty in the corresponding estimates.

- Keep in mind that the Kaplan-Meier method only gives estimates of the survival function after the occurrence of the first failure. Consequently, the survival curve is reported to be equal to 1 until the first failure time. This does not imply that there is no risk of failure before that time.

- Flat regions of the curve do not necessarily imply that the risk of failure is negligible. Thus, one should not infer from the plot that there is minimal risk of relapse before week 6 and after week 23.

- Both the estimated survival curve and confidence intervals should be utilized when drawing conclusions about the survival process.

Applied Survival Analysis (171:242)

Section 3: Group Comparisons of Survival

Brian J. Smith, Ph.D.

April 9, 2005

Table of Contents

3.1 Hazard Function

3.1.1 Definition

Definition: The *hazard function* or *hazard rate* $\lambda(t)$ is the instantaneous rate of failure at time t for those at risk at time t. It is defined as

$$\lambda(t) = \lim_{h \to 0} \frac{\Pr[t \le T < t + h \mid T \ge t]}{h}.$$

The term $\Pr[t \le T < t + h \mid T \ge t]$ is the conditional probability that the event occurs during the interval $t \le T < t + h$ given that it has not occurred before time t. Dividing this probability by h, the length of the interval, gives us a rate. The limit of this rate as h approaches zero is the instantaneous event rate at time t.

Examples:

1. If the event of interest is death, then $\lambda(t)$ is the mortality rate as a function of time.
2. If the event is relapse among leukemia patients, then $\lambda(t)$ is the relapse rate as a function of time.
3. Theoretical hazard functions are displayed in Figure 1.

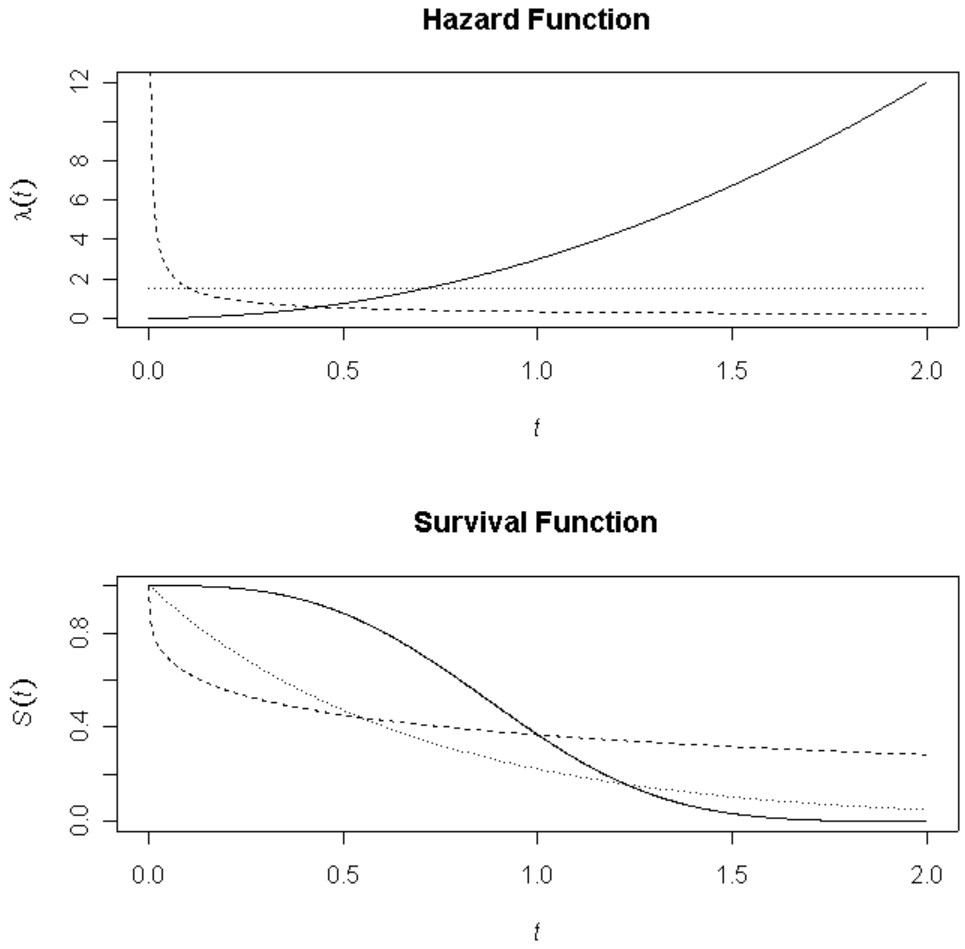


Figure 1. Plots of theoretical hazard functions with corresponding survival functions.

Can you think of an example where the hazard function changes from periods of increase to decrease over time?

Notes

1. The hazard function $\lambda(t)$ is a rate, not a probability, and assumes values in the interval $[0, \infty)$. The hazard rate is the rate of change in the cumulative probability of failure at time *t* relative to the corresponding survival probability. It can be thought of as the "force of mortality" at time *t*.

2. There is a one-to-one relationship between the hazard function and the survival function. The survival function is completely determined by the hazard function and vice versa (Figure 1).

3. Any non-negative function can serve as a hazard function. Over time, hazard functions may be increasing, decreasing, constant, or any combination thereof.

3.1.2 Estimation

Techniques for estimating the hazard function tend to be more involved than, say, the Kaplan-Meier estimator of the survival function. Sophisticated mathematical techniques are typically employed to produce a smooth estimate of the hazard function. Figure 2 displays smooth estimates of the hazard functions for the Leukemia Trial. Cubic splines were used for the data smoothing.

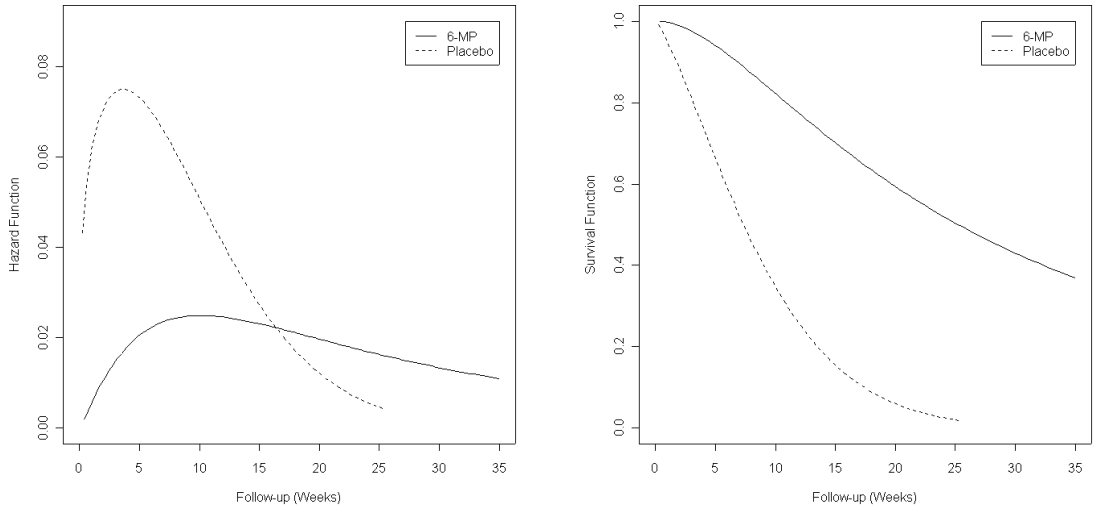


Figure 2. Smooth estimates of the hazard and survival functions for the Leukemia Trial.

The smoothed survival curves are displayed along with the Kaplan-Meier curves in Figure 3. Both are based on the cumulative probabilities discussed in Section 2 and give non-parametric estimates of the survival function.

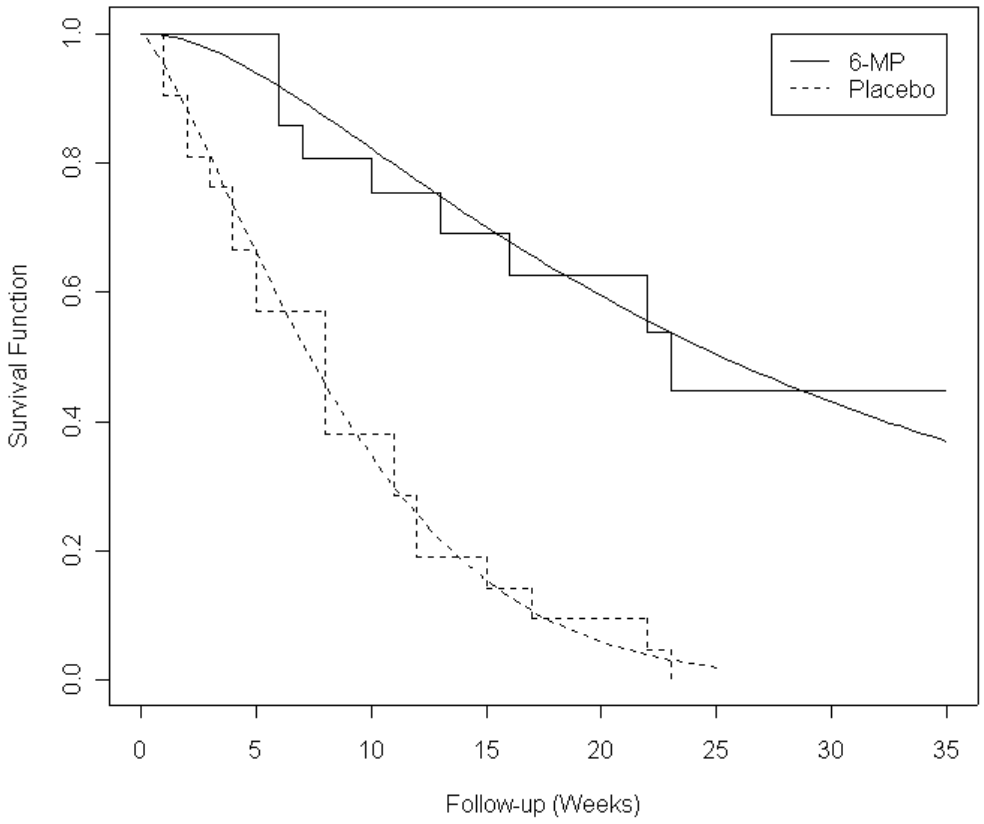


Figure 3. Kaplan-Meier and smoothed estimates of the survival curves for the Leukemia Trial.

Table 1. Comparison of the Kaplan-Meier and smoothed estimates of the survival function.

Kaplan-Meier Estimate	**Smoothed Estimate**
Step function	Smooth curve
Plots exact failure times	Displays survival as a continuous process
Uniquely determined	Subjective amount of smoothing

3.2 Log-Rank (Mantel-Haenzel) Test

3.2.1 Overview

Up to this point, the focused has been on descriptive methods for survival data. We now turn our attention to inferential methods; specifically, statistical tests for the comparison of survival between two or more groups.

Results from the Leukemia Trial

The purpose of this trial was to test the effectiveness of 6-mercaptopurine (6-MP) in delaying the time to relapse for lymphoblastic leukemia patients currently in remission.

The log-rank test was carried out to test the null hypothesis that the hazard rates between the two groups are equal, versus the two-sided alternative they differ; namely,

$$H_0 : \lambda_{6-MP}(t) = \lambda_{Placebo}(t)$$
$$H_A : \lambda_{6-MP}(t) \neq \lambda_{Placebo}(t).$$

A value of 16.8 was obtained for the χ_1^2 test statistic. The resulting p-value was 4.17e-5. Thus, at the 5% level of significance, it is concluded that the hazard rates differ between the two groups. In particular, the relapse rate is significantly lower in the 6-mercaptopurine group. The drug effectively delays the time to relapse in this patient population.

Questions:

1. What are the properties of the log-rank test?
2. When is the test appropriate?
3. How should the results be interpreted?

3.2.2 Methodology

Consider the entire collection of ordered time points at which a relapse was observed in the Leukemia Trial.

t	e_{1t}	n_{1t}	e_{2t}	n_{2t}
1	0	21	2	21
2	0	21	2	19
3	0	21	1	17
4	0	21	2	16
5	0	21	2	14
6	3	21	0	12
7	1	17	0	12
8	0	16	4	12
10	1	15	0	8
⋮	⋮	⋮	⋮	⋮
23	1	6	1	1

At time t there are n_{1t} subjects at risk in the 6-MP group, of which e_{1t} relapsed. n_{2t} and e_{2t} are similarly defined for the Placebo group. The events at time t can be summarized in a 2 x 2 table.

	Relapse	Disease-free	Risk Set
6-MP	e_{1t}	$n_{1t} - e_{1t}$	n_{1t}
Placebo	e_{2t}	$n_{2t} - e_{2t}$	n_{2t}
Totals	e_t	$n_t - e_t$	n_t

If we condition on knowing the table margins and assume a common rate of relapse, then e_{1t} is a hypergeometric random variable with mean and variance given by

$$E\left[e_{1t}\right] = n_{1t}\frac{e_t}{n_t}$$

$$\text{var}\left[e_{1t}\right] = \frac{n_{1t}n_{2t}e_t\left(n_t - e_t\right)}{n_t^2\left(n_t - 1\right)}.$$

The standard test for an association between the row and column factors for *independent* 2 x 2 tables is the Mantel-Haenszel statistic. This statistic is constructed by subtracting the expected number of events in Group 1 from the observed events, and then standardizing this difference by the square root of the variance:

$$X_{MH} = \frac{\sum_t \left(e_{1t} - E\left[e_{1t}\right]\right)}{\sqrt{\sum_t \text{var}\left[e_{1t}\right]}} \approx N\left(0,1\right).$$

The square of this statistic X_{MH}^2, which has an approximate chi-square distribution with one degree of freedom, is typically reported in practice.

- X_{MH}^2 is known as the log-rank statistic. It can be generalized to the comparison of more than two groups of subjects.

- P-value formulas for one and two-sided alternative hypotheses are as follows:

H_A	p-value
$\lambda_1(t) < \lambda_2(t)$	$p = \Pr\left[Z < X_{MH}\right]$
$\lambda_1(t) > \lambda_2(t)$	$p = \Pr\left[Z > X_{MH}\right]$
$\lambda_1(t) \neq \lambda_2(t)$	$p = 2\Pr\left[Z > \lvert X_{MH} \rvert\right]$ $= \Pr\left[\chi_1^2 > X_{MH}^2\right]$

Leukemia Example

The log-rank test for the leukemia data is based on 17 unique failure times, each of which can be summarized in a 2 x 2 table. The first seven tables are given below.

Week 1

	Relapse	Disease-free	Risk Set
6-MP	0	21	21
Placebo	2	19	21
	2	40	42
	E = 1.000	var = 0.488	

Week 2

	Relapse	Disease-free	Risk Set
6-MP	0	21	21
Placebo	2	17	19
	2	38	40
	E = 1.050	var = 0.486	

Week 3

	Relapse	Disease-free	Risk Set
6-MP	0	21	21
Placebo	1	16	17
	1	37	38
	E = 0.553	var = 0.247	

Week 4		Relapse	Disease-free	Risk Set
	6-MP	0	21	21
	Placebo	2	14	16
		2	35	37
		E = 1.135	var = 0.477	

Week 5		Relapse	Disease-free	Risk Set
	6-MP	0	21	21
	Placebo	2	12	14
		2	33	35
		E = 1.200	var = 0.466	

Week 6		Relapse	Disease-free	Risk Set
	6-MP	3	18	21
	Placebo	0	12	12
		3	30	33
		E = 1.909	var = 0.651	

Week 7		Relapse	Disease-free	Risk Set
	6-MP	1	16	17
	Placebo	0	12	12
		1	28	29
		E = 0.586	var = 0.243	

The calculations necessary for computing the log-rank statistic are summarized in the following worksheet.

t	e_{1t}	e_{2t}	e_t	n_{1t}	n_{2t}	n_t	$E[e_{1t}]$	$e_{1t} - E[e_{1t}]$	$var[e_{1t}]$
1	0	2	2	21	21	42	1.000	-1.000	0.488
2	0	2	2	21	19	40	1.050	-1.050	0.486
3	0	1	1	21	17	38	0.553	-0.553	0.247
4	0	2	2	21	16	37	1.135	-1.135	0.477
5	0	2	2	21	14	35	1.200	-1.200	0.466
6	3	0	3	21	12	33	1.909	1.091	0.651
7	1	0	1	17	12	29	0.586	0.414	0.243
8	0	4	4	16	12	28	2.286	-2.286	0.871
10	1	0	1	15	8	23	0.652	0.348	0.227
11	0	2	2	13	8	21	1.238	-1.238	0.448
12	0	2	2	12	6	18	1.333	-1.333	0.418
13	1	0	1	12	4	16	0.750	0.250	0.188
15	0	1	1	11	4	15	0.733	-0.733	0.196
16	1	0	1	11	3	14	0.786	0.214	0.168
17	0	1	1	10	3	13	0.769	-0.769	0.178
22	1	1	2	7	2	9	1.556	-0.556	0.302
23	1	1	2	6	1	7	1.714	-0.714	0.204
Total	9	21	30					-10.251	6.257

Summing the terms over all 17 failure times gives

$$X^2_{MH} = \frac{\left[(0-1.000)+(0-1.050)+\ldots+(1-1.714)\right]^2}{0.488+0.486+\ldots+0.204}$$
$$= \frac{(-10.251)^2}{6.257} = 16.79$$

Thus, the 2-sided p-value is

$$p = \Pr\left[\chi^2_1 > 16.79\right] = 4.17e-5$$

which agrees with the p-value given in the original statement of the analysis results.

Notes

1. The log-rank test is a non-parametric test. As in the Kaplan-Meier estimator, no assumptions are made about the functional form of the distribution for survival times.

2. Each of the 2 x 2 tables can be viewed as a comparison of the hazard rate at the corresponding point in time. Since each of the $e_{1t} - E\left[e_{1t}\right]$ differences receives equal weight in the test statistic, it can be shown that the log-rank test is most sensitive when the hazard rates are proportional, namely

$$\frac{\lambda_1(t)}{\lambda_2(t)} = c$$

for some constant *c*.

3. If the hazard rates cross or if they differ only over a subset of the follow-up times, then the log-rank test may have low power to detect a difference between the groups.

4. A non-significant result from the log-rank test does not imply that the hazard rates are equal; only that the test does not provide evidence to the contrary.

3.2.3 Weighted Log-Rank Test

Recall that the log-rank test weights the differences in the hazard curves equally over time. One might be interested in testing for earlier or later differences in the hazard functions. To this end, a generalization of the log-rank test was developed. The *weighted* log-rank test statistic is

$$X = \frac{\sum_t w_t \left(e_{1t} - E\left[e_{1t} \right] \right)}{\sqrt{\sum_t w_t^2 \operatorname{var}\left[e_{1t} \right]}} \approx N(0,1)$$

where w_t are weights defined at each failure time. Some commonly used weights are given in Table 2.

Table 2. Weights for the weighted log-rank statistic.

Weights	Test Name	Trend
$w_t = 1$	Log-rank	Constant
$w_t = n_t$	Gehan	Decreasing
$w_t \approx \hat{S}_{KM}(t)$	Peto-Prentice	Decreasing
$w_t = \left[\hat{S}_{KM}(t) \right]^{\rho}$	Harrington & Fleming G-rho	$\rho > 0$: Decreasing $\rho = 0$: Constant $\rho < 0$: Increasing

where $\hat{S}_{KM}(t)$ is the Kaplan-Meier estimate of the survival function computed over the entire set of subjects. A

comparison of the relative weights for select weighted log-rank statistics is given in Figure 4.

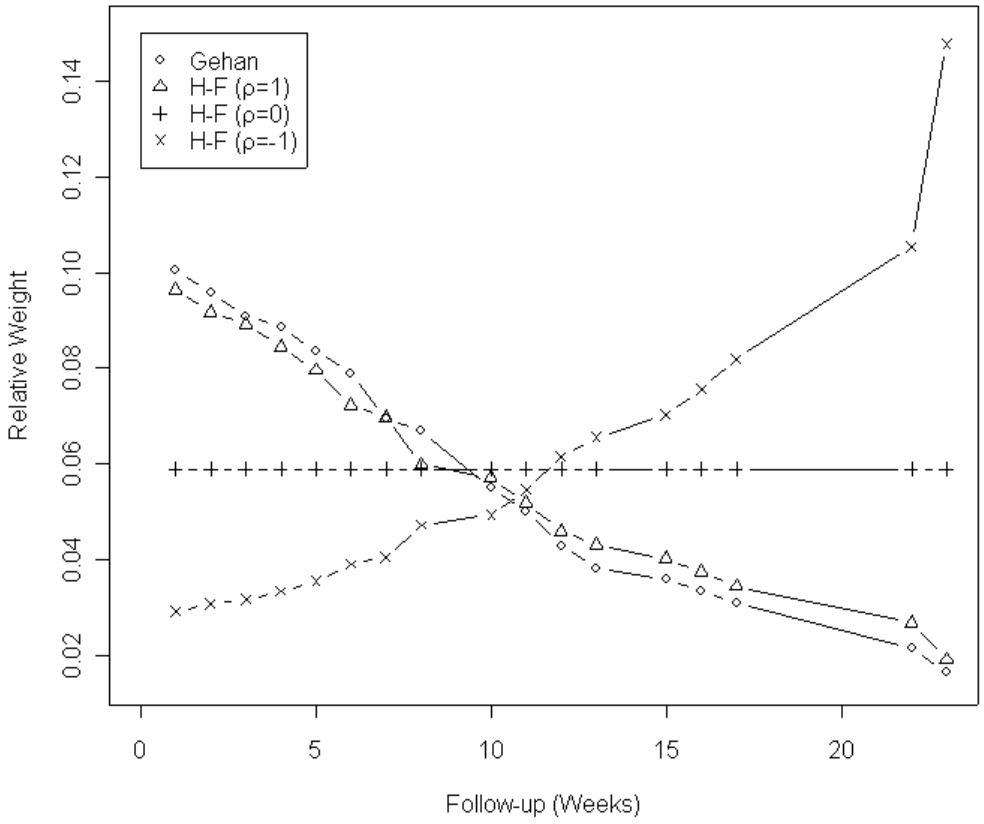


Figure 4. Relative weights for select weighted log-rank statistics applied to the failure times in the Leukemia Trial.

We will use the Harrington & Fleming (H-F) class of test statistics because of their flexibility. The ρ parameter should be chosen *a priori* based on the time period in the study for which differences are of most interest.

Leukemia Example: Various Harrington & Fleming test results for the Leukemia Trial are given in the table below.

Harrington & Fleming	Test Statistic (X)	p-value
$\rho = 1$	-3.802	1.43e-4
$\rho = 0$	-4.098	4.17e-5
$\rho = -1$	-4.087	4.38e-5

Notes

1. The log-rank test is the most powerful for detecting group differences when the hazard rates are proportional.

2. In the case of crossing hazard functions, the log-rank test statistic will tend toward zero.

3. The Harrington-Fleming tests can be used to detect either earlier or later differences in the hazard functions. This is useful in the case of crossing hazards or if the hazards differ only at the beginning or end of the study.

4. The Harrington-Fleming test is equal to the log-rank test for $\rho = 0$ and is equivalent to the Peto-Prentice test for $\rho = 1$.

5. A disadvantage of the Gehan test is that the weights $w_t = n_t$ are a function of censoring. In the presence of heavy censoring this test can be very misleading and should not be used.

3.2.4 Comparison of More than Two Samples

Carcinogenesis Experiment

A group of 29 laboratory mice were randomly assigned to receive one of three dose levels of a suspected tumor-causing agent. They were then followed until tumor development or censoring (*). The follow-up times are given below.

Group	Dose	Follow-up (Days)
1	2.0	41*, 41*, 47, 47*, 47*, 58, 58, 58, 100*, 117
2	1.5	43*, 44*, 45*, 67, 68*, 136, 136, 150, 150, 150
3	0.0	73*, 74*, 75*, 76, 76, 76*, 99, 166, 246*

Analysis Objective: Extend the log-rank statistic to allow for the comparison of more than two groups.

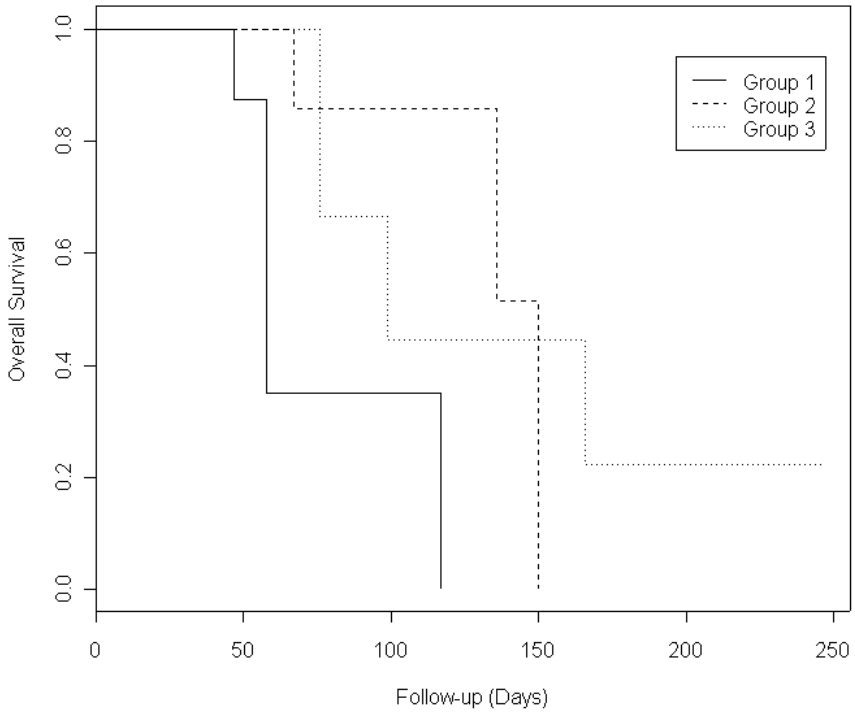


Figure 5. Kaplan-Meier plot of the survival curve for the Carcinogenesis Experiment.

Suppose that we have *G* groups of subjects and wish to test the null hypothesis that

$$H_0 : \lambda_1(t) = \ldots = \lambda_g(t) = \ldots = \lambda_G(t)$$
$$H_A : \lambda_g(t) \neq \lambda_{g'}(t) \text{ for some } g, g'$$

where the alternative hypothesis that at least two of the groups have different hazard rates.

Similar to the previous discussion of the log-rank test for two groups, the events at time *t* can be summarized in a *G* x 2 table.

	Event	Non-event	Risk Set
Group 1	e_{1t}	$n_{1t} - e_{1t}$	n_{1t}
⋮	⋮	⋮	⋮
Group g	e_{gt}	$n_{gt} - e_{gt}$	n_{gt}
⋮	⋮	⋮	⋮
Group G	e_{Gt}	$n_{Gt} - e_{Gt}$	n_{Gt}
Totals	e_t	$n_t - e_t$	n_t

If we condition on knowing the table margins and assume that the null hypothesis holds, then the e_{gt} have a multiple hypergeometric distribution with mean

$$E\left[e_{gt}\right] = n_{gt}\frac{e_t}{n_t}$$

Recall that the weighted log-rank test is constructed by summing the differences between the observed and expected number of failures. In group g, the sum of the differences is

$$U_g = \sum_t w_t\left(e_{gt} - E\left[e_{gt}\right]\right).$$

The presence of more than two groups makes for a more involved variance formula. The test statistic is computed as

$$X^2 = \mathbf{U}'\mathbf{V}^{-1}\mathbf{U} \approx \chi^2_{G-1}$$

where

$$\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_{G-1} \end{bmatrix}$$

and **V** is a (G - 1) x (G - 1) covariance matrix for the observed failure times in the first G - 1 groups.

Results from the Carcinogenesis Experiment

The log-rank statistic was used to test the null hypothesis that the hazard rates are equal across the three groups, versus the alternative that at least two differ, namely

$$H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$$
$$H_A : \lambda_g(t) \neq \lambda_{g'}(t) \text{ for some } g, g'.$$

A value of 8.0 was obtained for the χ^2_2 test statistic. The resulting p-value is 0.0179. Thus, at the 5% level of significance, the hazard rates are not equal across all three groups.

Notes

1. The log-rank test statistic can be used for the comparison of hazard rates across two or more groups of subjects.

2. The alternative hypothesis is that at least two of the groups differ. One cannot determine where the specific group differences occur with the overall test of equality for the hazard functions.

3. The formulation of the statistic given in this section yields an inherently two-sided test. It cannot be used to test for any particular ordering of the groups.

4. As in the 2-sample case, weighting functions, such as the one proposed by Harrington-Fleming, can be incorporated into the statistic to emphasize earlier or later differences in the hazard functions.

3.2.5 Pairwise Comparisons

A significant result from a log-rank test applied to multiple groups indicates that at least two of the groups differ. A natural question to ask then is where the group differences occur. To address this question, one can perform a series 2-sample tests comparing all $G(G - 1) / 2$ pairs of groups. In order to maintain an overall α level of significance, an adjusted significance level α' must be used for each

pairwise comparison. Two conservative methods for computing α' are

1. Bonferroni Method: $\alpha' = \dfrac{\alpha}{G(G-1)/2}$

2. Probability Method: $\alpha' = 1-(1-\alpha)^{1/(G(G-1)/2)}$

The second method is slightly less conservative (α' is closer to α) than the Bonferroni method.

Table 3. Adjusted significance level for use in multiple pairwise comparisons.

Exposure Levels G	Pairwise Comparisons $G(G-1)/2$	Overall Significance α	Individual Test Significance α'	
			Bonferroni	Probability
3	3	0.05	0.01667	0.01695
4	6	0.05	0.00833	0.00851
5	10	0.05	0.00500	0.00512

When there are a large number of groups, these two adjustments for multiple comparisons are of limited utility.

Carcinogenesis Experiment

The p-value from the log-rank test of overall equality was 0.0179. We need to determine where the groups differ. There are $G = 3$ groups with $G(G-1)/2 = 3$ possible pairwise comparisons. The p-values from the pairwise comparisons are given below.

Comparisons	p-value
1 vs. 2	0.00857
1 vs. 3	0.0801
2 vs. 3	0.531

If an overall 5% level of significance is desired, then the two methods yield the following adjustments:

1. $\alpha' = \dfrac{\alpha}{G(G-1)/2} = \dfrac{0.05}{3} = 0.0167$
2. $\alpha' = 1-(1-\alpha)^{1/(G(G-1)/2)} = 1-0.95^{1/3} = 0.0169$

Only the p-value for the pairwise comparison of groups 1 and 2 is less than the adjusted significance level, $\alpha' = 0.0169$. Therefore, we can conclude that groups 1 and 2 are significantly different. There is no significant difference between groups 1 and 3 or groups 2 and 3.

3.2.6 Tests for Trend across Groups

In the carcinogenesis experiment the subjects were grouped by the administered dose levels (2.0, 1.5, and 0.0). Thus, the investigators might be interested in testing for a trend in the hazard functions, say

$$H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$$
$$H_A : \lambda_1(t) > \lambda_2(t) > \lambda_3(t).$$

Recall that, in the log-rank statistic,

$$U_g = \sum_t w_t \left(e_{gt} - E\left[e_{gt} \right] \right)$$

is the sum of the observed minus expected number of events for the *g*th group. Our log-rank statistics have been based on the sum total of these differences. Thus, in order to test for trend, we might include an additional term z_g that weights each U_g by a group-specific variable. The resulting sum would be $\sum_g z_g U_g$, which can be written in matrix notation as $z'\mathbf{U}$ where

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_g \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_g \end{bmatrix}.$$

The trend test statistic is given by

$$X_{trend} = \frac{z'\mathbf{U}}{\sqrt{z'\mathbf{V}z}} \approx N(0,1),$$

where **V** is a *G* x *G* covariance matrix for the number of failures in each group. One and two-sided p-values may be computed according to the formulas in Table 4.

Table 4. p-value formulas for the log-rank trend statistic.

Alternative Hypothesis	Group Weights	p-value
$H_A : \lambda_1(t) < \ldots < \lambda_G(t)$	Increasing	$p = \Pr[Z > X_{trend}]$
	Decreasing	$p = \Pr[Z < X_{trend}]$
$H_A : \lambda_1(t) > \ldots > \lambda_G(t)$	Increasing	$p = \Pr[Z < X_{trend}]$
	Decreasing	$p = \Pr[Z > X_{trend}]$
$H_A : \lambda_1(t) < \ldots < \lambda_G(t)$ $\lambda_1(t) > \ldots > \lambda_G(t)$	Increasing or Decreasing	$p = 2\Pr[Z > \lvert X_{trend} \rvert]$

Carcinogenesis Experiment

A natural choice for group weights in the Carcinogenesis Experiment are the administered dose levels:

$$z = \begin{bmatrix} 2.0 \\ 1.5 \\ 0.0 \end{bmatrix}.$$

Use of equal weights in the log-rank statistic results in

$$\mathbf{U} = \begin{bmatrix} 3.209 \\ -0.803 \\ -2.405 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} 1.319 & -0.641 & -0.677 \\ -0.641 & 2.663 & -2.021 \\ -0.677 & -2.021 & 2.699 \end{bmatrix},$$

so that

$$\begin{aligned} z'\mathbf{U} &= \begin{bmatrix} 2.0 & 1.5 & 0.0 \end{bmatrix} \begin{bmatrix} 3.209 \\ -0.803 \\ -2.405 \end{bmatrix} \\ &= (2.0)(3.209) + (1.5)(-0.803) + (0.0)(-2.405) \\ &= 5.212 \end{aligned}$$

and

$$\begin{aligned} z'\mathbf{V}z &= \begin{bmatrix} 2.0 & 1.5 & 0.0 \end{bmatrix} \begin{bmatrix} 1.319 & -0.641 & -0.677 \\ -0.641 & 2.663 & -2.021 \\ -0.677 & -2.021 & 2.699 \end{bmatrix} \begin{bmatrix} 2.0 \\ 1.5 \\ 0.0 \end{bmatrix} \\ &= \begin{bmatrix} 1.676 & 2.711 & -4.387 \end{bmatrix} \begin{bmatrix} 2.0 \\ 1.5 \\ 0.0 \end{bmatrix} \\ &= 7.418 \end{aligned}$$

Therefore, the test statistic is equal to

$$X_{trend} = \frac{z'\mathbf{U}}{\sqrt{z'\mathbf{V}z}} = \frac{5.212}{\sqrt{7.418}} = 1.91$$

which gives a p-value of

$$p = \Pr\left[Z > 1.91\right] = 0.0278$$

Therefore, at the 5% level of significance, the hazard functions are decreasing relative to the administered dose levels.

Notes

1. Any set of weights could be assigned to the groups. Using the dose levels (2.0, 1.5, and 0.0) tests for a difference in hazards between groups 2 and 3 (1.5 – 0.0 = 1.5) that is 3 times as great as the difference between groups 1 and 2 (2.0 – 1.5 = 0.5).

2. If the integer ranks 1, 2, and 3 were assigned to the groups, then one would be testing for a constant incremental difference in the hazards.

3.2.7 Stratified Log-Rank Test

Bone Marrow Transplants for Non-Hodgkin's and Hodgkin's Lymphoma

A group of 43 patients with either non-Hodgkin's or Hodgkin's lymphoma were studied in order to asses disease-free survival after bone marrow transplant. The patients received either an allogeneic transplant from an HLA-matched sibling donor or an autologous transplant where their own marrow was cleansed and reinfused after a high dose of chemotherapy.

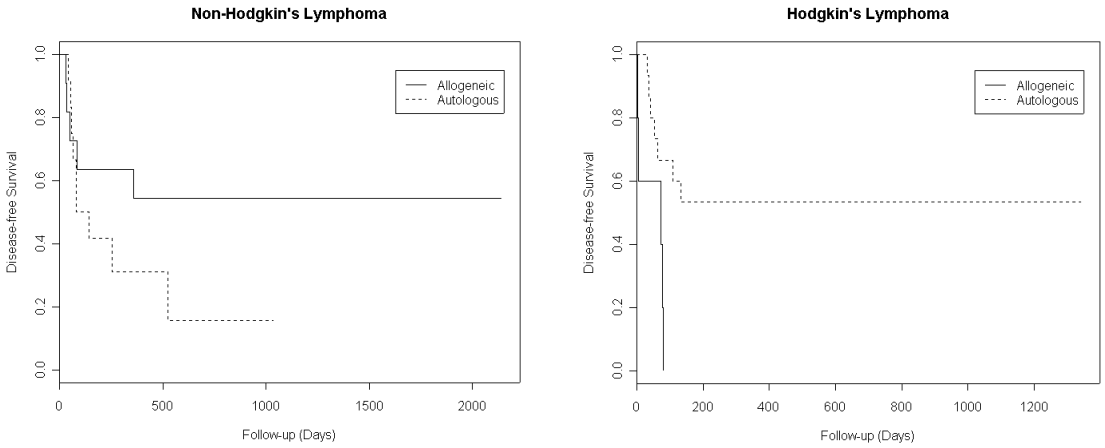


Figure 6. Kaplan-Meier plots of the survival functions in the Lymphoma Study.

Analysis Objective: Compare the disease-free survival rate between allogeneic and autologous transplant recipients, adjusting for the patient's disease type.

Stratification Variables

Oftentimes, there are several variables that affect the survival of subjects in the study population. In order to assess the effect of a specific variable, we need to consider the potential impact of other important risk factors. For example, we might have reason to believe that disease-free survival differs between non-Hodgkin's and Hodgkin's patients.

Confounding

A **confounder** is a covariate that is causally related to the outcome and is associated with the variable of interest. The potential effects of a confounder are illustrated in Figure 7. The figure shows hazard functions for two groups of subjects. When the subjects are divided into strata based on the levels of another variable, we see a consistent difference between the two groups. However, if the stratification variable is ignored and the groups pooled, the results may be unpredictable:

- There may be no association between the grouping and stratification variable. Consequently, in a given strata, one would expect a similar proportion of subjects from each group. In this case the stratification variable is not a confounder, and the data can be pooled to study the group differences.
- There is an association between the grouping and stratification variable. Groups contribute a different proportion of subjects within a given strata. When this is the case, the stratification variable is a confounder,

and pooling across strata to analyze group differences will lead to unpredictable results.

- In the presence of confounding, a pooled analysis of the data can lead to A) underestimation of the group differences, B) reordering of the hazard functions, or C) over-estimation of the group differences.

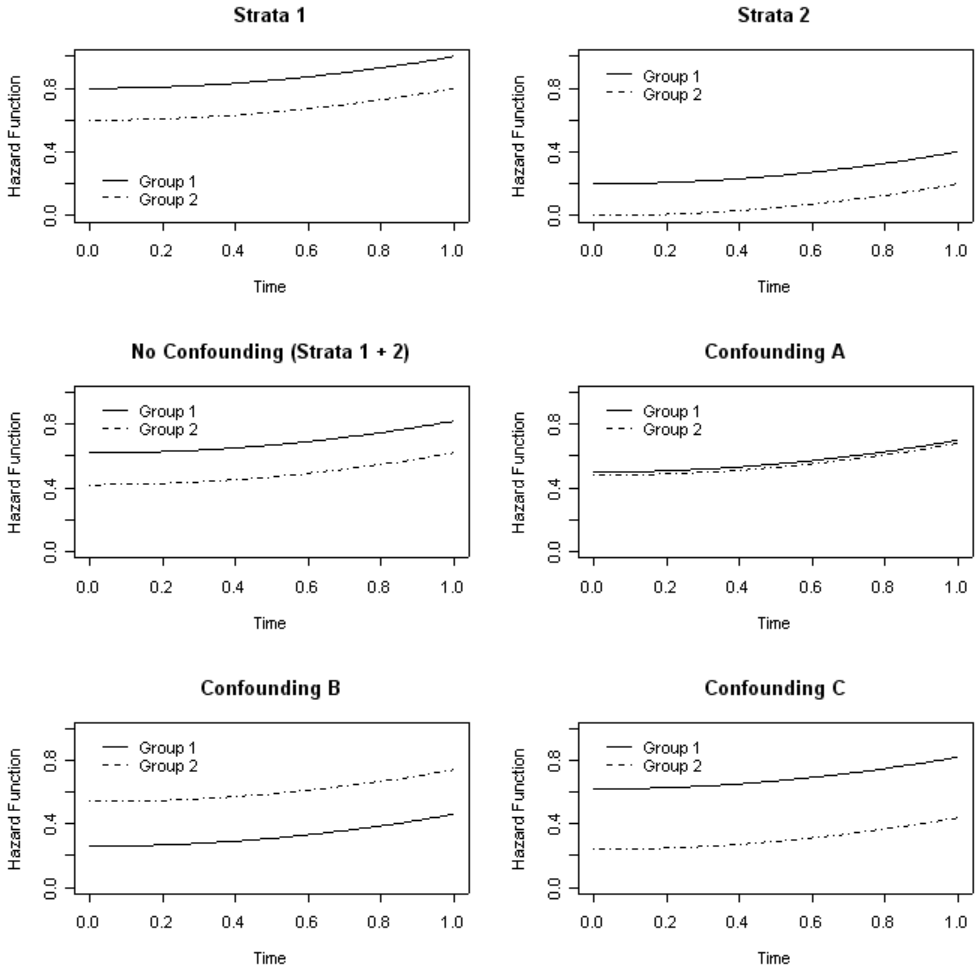


Figure 7. Comparison of hazard functions in the presence of confounding.

Q: Why not perform separate tests within each strata; e.g. perform two log-rank tests, one for strata 1 and another for strata 2?

A: If there is a persistent difference across the stratum, then an analysis of the complete data set would be more powerful than individual subset analyses of the data.

Methodology

Let $m = 1,...,M$ index the strata. The null and alternative hypotheses are

$$H_0 : {}_m\lambda_1 = \ldots = {}_m\lambda_G \text{ for all } m$$
$$H_A : \sum_m \left({}_m\lambda_g - {}_m\lambda_{g'} \right) \neq 0 \text{ for some } g, g'$$

The test statistic is formed by computing the difference between observed and expected failure times within each strata,

$${}_mU_g = \sum_t w_t \left({}_me_{gt} - E\left[{}_me_{gt} \right] \right).$$

The stratum-specific differences are summed together to yield

$$U_g = \sum_m {}_m U_g \,,$$

from which the log rank statistic is computed in the usual fashion; i.e.

$$X^2 = \mathbf{U}'\mathbf{V}^{-1}\mathbf{U} \approx \chi^2_{G-1}$$

where $\mathbf{V} = \sum_m {}_m \mathbf{V}$ is the covariance matrix.

Results from the Lymphoma Study

The log-rank statistic was used to test the null hypothesis of equal hazard rates across allogeneic and autologous transplant recipients, versus a 2-sided alternative. The statistic was stratified by type of disease (non-Hodgkin's and Hodgkin's lymphoma). A value of 0.10 was obtained for the χ^2_1 test statistic, giving a p-value of 0.729. Thus, at the 5% level of significance, the stratified log-rank test does not indicate a difference between the two groups.

However, log-rank tests applied to the strata individually suggested otherwise. P-values of 0.198 and 0.0117 were obtained from log-rank tests of the non-Hodgkin's and Hodgkin's patients, respectively. Why is the p-value from the stratified test so large in comparison?

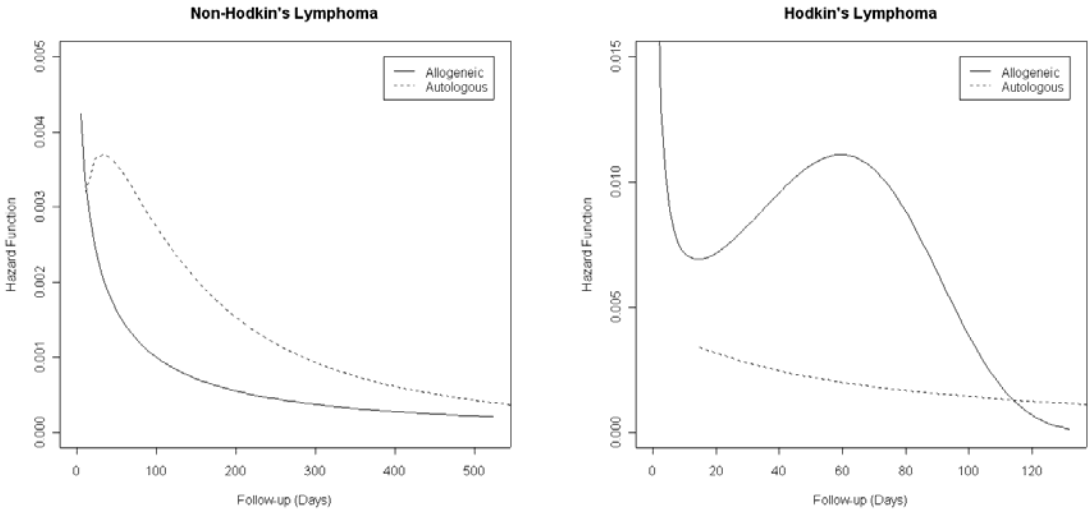


Figure 8. Estimated hazard functions for the lymphoma study.

Disease type is not a confounder, it is an **effect modifier**. The differences between the two hazard functions cancel each other in the stratified test because of their reversed ordering. Stratification is not the appropriate means to adjust for disease type.

Notes

1. The incorporation of a stratification variable in the test statistic affects only the computation of the expected number of failures and the variance. As before, the test statistic has a chi-square distribution with $G - 1$ degrees of freedom.

2. Weighting functions, such as the one proposed by Harrington-Fleming, can be incorporated to emphasize earlier or later differences in the hazard functions.

3. The alternative hypothesis is that there is a consistent difference across strata between at least two of the groups. Pairwise comparisons can be performed to identify where the group differences occur.

4. A test of trend in the hazard rates across groups can also be performed.

3.3 Sample Size

During the design and planning stages of a study, investigators are often faced with the task of justifying the number of subjects to be enrolled. Two key issues that must be considered in the determination of sample size are:

1. There are practical limits on the number of subjects that can be enrolled.
2. An adequate number of subjects should be enrolled to ensure sufficient powered.

Recall that *power* (1 - β) is defined as the probability of correctly rejecting the null hypothesis, and that the significance level (α) is the probability of incorrectly rejecting the null hypothesis.

The first step in estimating sample size is to specify the null and alternative hypotheses of interest, as well as the statistic to be used for the test. The sample size estimation is based on these specifications.

3.3.1 Freedman's Method

We will consider the issue of sample size estimation for tests of the hypotheses

$$H_0 : \lambda_1(t) = \lambda_2(t)$$
$$H_A : \lambda_1(t) \neq \lambda_2(t)$$

using the log-rank statistic. Freedman's sample size approach follows from an assumption that the analysis will occur at a fixed time *t** after the last patient has been enrolled. The sample size formula is

$$n_1 = \frac{d}{1 - S_1(t^*) + c\left[1 - S_2(t^*)\right]}$$
$$n_2 = cn_1$$

where *c* is the number of subjects in group 2 relative to group 1, as specified by the investigators, and

$$d = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{1+\theta}{1-\theta}\right)^2$$

$$\theta = \frac{\ln S_1(t^*)}{\ln S_2(t^*)}$$

In the case of a one-sided alternative hypothesis, $z_{1-\alpha/2}$ would be replaced by $z_{1-\alpha}$. The group-specific survival probabilities $S_1(t^*)$ and $S_2(t^*)$ must be supplied.

Adjuvant Chemotherapy Example

Investigators are interested in evaluating the effect of a chemotherapeutic agent as adjuvant post-surgery treatment compared to surgery alone. Based on the literature and personal experience, they suggest that

- The fiver year survival probability for surgically treated patients is 60%.
- A survival increment of 25% is clinically relevant.
- Two-sided tests are planned and will be carried out at the 5% level of significance.
- A power of 80% is desired.
- Twice as many patients will be enrolled to the surgery-only group.

$t^* = 5$ $\qquad$ $\alpha = 0.05$

$S_1(5) = 0.60$ $\qquad$ $1-\beta = 0.80$

$S_2(5) = 0.85$ $\qquad$ $c = n_2/n_1 = 1/2$

The sample size estimates are as follows:

$$\theta = \frac{\ln S_1(t^*)}{\ln S_2(t^*)} = \frac{\ln 0.60}{\ln 0.85} = 3.1431$$

$$d = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{1+\theta}{1-\theta}\right)^2$$

$$= \left(1.96 + 0.8416\right)^2 \left(\frac{1+3.1431}{1-3.1431}\right)^2 = 29.3335$$

and

$$n_1 = \frac{d}{1 - S_1(t^*) + c\left[1 - S_2(t^*)\right]} = \frac{29.3335}{1 - 0.6 + 0.5\left[1 - 0.85\right]} \cong 62.$$

$$n_2 = cn_1 = 0.5(62) \cong 31$$

Notes

1. To allow for the possibility that *x* percent of the subjects might withdraw from the study, the estimated sample sizes should be multiplied by 100/(100-*x*). For example, if 20% of subjects are expected to withdraw, then the previous sample size estimates become 62*100/80 = 78 and 31*100/80 = 39.

2. When the enrollment period is relatively long, the use of minimum follow-up time for t^* severely overestimates the number of subjects needed. In these situations, the *average* follow-up time should be used for t^*.

3. Freedman's method tends to overestimate the sample size when either of the survival probabilities is close to zero or one.

4. There are many techniques for estimating sample size. Indeed, there are complete software packages developed specifically for sample size estimation; e.g. PASS, Power and Precision, and nQuery.

5. Sample size determination is an inexact science. In order to estimate sample size, one must provide values for the parameters under study. However, if those values were known, there would be no need to do the study.

6. It may be helpful to give sample size estimates over a range of values for the parameters, the power, and the significance level.

7. Generally speaking, the more sophisticated the analysis, the more complicated the sample size estimation. Sample size techniques are often based on simplifying assumptions or a less sophisticated statistical test.

Applied Survival Analysis (171:242)

Section 4: Cox Proportional Hazards Regression

Brian J. Smith, Ph.D.

March 2, 2005

Table of Contents

4.1 Introduction

One of the advantages of regression modeling is the ability to examine the effect of multiple predictor variables on the outcome of interest.

Time to Breast-Feeding Cessation Study

The National Labor Survey of Youth is a random sample of youths, aged 14 to 21, who were interviewed yearly from 1978 through 1988. The survey data contain information on 927 mothers who had given birth to their first child and chose to breast-feed. There is interest in identifying factors that are associated with breast-feeding cessation. The following variables were collected in the study:

Variable	Description	Values
weaned	Indicator for breast-feeding cessation	1 = yes 0 = no
weeks	Length of follow-up in weeks	continuous
age	Subject age	continuous
alcohol	Alcohol use at time of birth	1 = yes 0 = no
care3	Use of prenatal care after first trimester	1 = yes 0 = no
education	Years of education	continuous
poverty	Below poverty level	1 = yes 0 = no
race	Subject race	1 = white 2 = black 3 = other
smoke	Smoking at time of birth	1 = yes 0 = no

The data are summarized in Table 1 and Table 2. Note that 35 mothers were still breast-feeding at the end of the study.

Table 1. Descriptive statistics for the categorical variables in the Breast-Feeding Study.

Variable	Levels	N	Percents
weaned	0	35	3.8%
	1	892	96.2%
alcohol	0	848	91.5%
	1	79	8.5%
care3	0	763	82.3%
	1	164	17.7%
poverty	0	756	81.6%
	1	171	18.4%
race	1	662	71.4%
	2	117	12.6%
	3	148	16.0%
smoke	0	657	70.9%
	1	270	29.1%

Table 2. Descriptive statistics for the continuous variables in the Breast-Feeding Study.

Variable	Mean	SD	Min	Max
weeks	16.18	17.92	1	192
age	21.54	2.67	15	28
education	12.21	1.93	3	19

Analysis Goal: *Perform a multivariate regression analysis of the data.*

Multivariate regression objectives:

- Identify the variables associated with breast-feeding cessation.
- Determine if race is a significant predictor after controlling for age and socio-economic status (alcohol use, education, prenatal care, poverty, and smoking).
- Assess whether the covariates interact in their effect on the breast feeding cessation.
- Estimate the effect of age, education, race, etc.

4.2 Cox Proportional Hazards Model

In Section 3 log-rank statistics were used to test for group survival (hazard rate) differences. Likewise, log-rank statistics could be used for the breast-feeding data to test for differences across the levels of any one of the variables; although, categorical variables would have to be created from the continuous variables age and years of education. Stratification could be employed in an attempt to control for confounding.

Log-rank tests are particularly useful when interest centers on individual variables, such as treatment or exposure to a potential carcinogen. However, as is the case here, we are often interested in the relationship of the outcome to several (continuous and categorical) covariates simultaneously. Moreover, we need to quantify and test these relationships. Thus, a multivariate regression approach is needed.

You are already familiar with two regression techniques: 1) linear regression and 2) logistic regression. Recall that regression methods model the outcome as a function of the covariates. For example, the linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

whereas, the logistic regression model is

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

or

$$\begin{aligned}\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} &= \exp\left\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\} \\ &= \exp\left\{\beta_0\right\}\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}\end{aligned} .$$

The Cox proportional hazards model is the most widely used regression technique for censored time-to-event data. In general, the multivariate Cox proportional hazards regression model is of the form

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

where there are *p* predictor variables x_i. We will use the notational convention that $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. $\lambda_0(t)$ is called the *baseline hazard rate*. It is the hazard rate when all of the covariates are equal to zero, since

$$\begin{aligned}\lambda(t;\mathbf{x}=\mathbf{0}) &= \lambda_0(t)\exp\{\beta_1 0 + \beta_2 0 + \ldots + \beta_p 0\} \\ &= \lambda_0(t)\end{aligned}.$$

The Cox model is non-parametric because it does not specify a functional form for the baseline hazard. $\lambda_0(t)$ is a completely arbitrary function except for the constraint that it must be ≥ 0. The model is given the name *proportional hazards* because the ratio of the hazards for two groups

$$\begin{aligned}\frac{\lambda(t,\mathbf{x}')}{\lambda(t,\mathbf{x}'')} &= \frac{\lambda_0(t)\exp\{\beta_1 x_1' + \beta_2 x_2' + \ldots + \beta_p x_p'\}}{\lambda_0(t)\exp\{\beta_1 x_1'' + \beta_2 x_2'' + \ldots + \beta_p x_p''\}} \\ &= \exp\{\beta_1(x_1' - x_1'') + \beta_2(x_2' - x_2'') + \ldots + \beta_p(x_p' - x_p'')\}\end{aligned}$$

does not depend on *t*. In other words, the hazard ratio is constant across time.

Breast-Feeding Study:

The following Cox regression model was fit to the data from the Breast-feeding Study:

$$\lambda(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}$$

where the terms for race are indicator variables defined as,

$$race1 = \begin{cases}1 & race = 1 \text{ (white)} \\ 0 & otherwise\end{cases}$$

$$race2 = \begin{cases}1 & race = 2 \text{ (black)} \\ 0 & otherwise\end{cases}$$

$$race3 = \begin{cases}1 & race = 3 \text{ (other)} \\ 0 & otherwise\end{cases}.$$

SAS Program and Output

```
proc import datafile="H:\breast_fed2005.txt"
   out=temp
   dbms="TAB"
   replace;

data breastfed;
   set temp;
   race1 = (race = 1);
   race2 = (race = 2);
   race3 = (race = 3);

proc phreg data=breastfed;
   model weeks*weaned(0) = age alcohol care3 education poverty
         race2 race3 smoke;
run;
```

Syntax

- PROC IMPORT reads the data from the tab-delimited file **breast_fed.txt** into the SAS dataset **temp**.
- A new dataset **breastfed** is created. It contains the original data plus new indicator variables for race.
- PROC PHREG performs multivariate Cox regression.
- The time-to-event variable **weeks** is specified in the model statement, followed by the event indicator **weaned**. The value for censored observations may be specified in parentheses. By default, subjects with a value of 1 for their event indicator are assumed to be censored.

Summary of Results

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
age	$\hat{\beta}_1$	0.0197	0.0165	1.44	0.231
alcohol	$\hat{\beta}_2$	0.1583	0.1225	1.67	0.1962
care3	$\hat{\beta}_3$	-0.0224	0.0898	0.06	0.8035
education	$\hat{\beta}_4$	-0.0516	0.0229	5.09	0.024
poverty	$\hat{\beta}_5$	-0.1898	0.0932	4.15	0.0418
race2	$\hat{\beta}_6$	0.1736	0.1052	2.72	0.0988
race3	$\hat{\beta}_7$	0.2894	0.0972	8.86	0.0029
smoke	$\hat{\beta}_8$	0.2395	0.0793	9.13	0.0025

These are the maximum likelihood estimates for the eight predictor variables in the model.

- The estimate for a given beta is interpreted as the effect of the associated predictor in the Cox model after controlling for the remaining covariates.
- Note that there are two predictor (indicator) variables for the effect of race.
- Each predictor has an estimate, standard error, Wald chi-square statistic, and p-value.
- As in multivariate linear regression, the individual p-values are used to determine if the associated parameter is significant, given that the remaining predictors are in the model.

4.3 Inference

4.3.1 Hazard Ratio Estimation

The hazard ratio is the hazard rate in one group relative to the rate in another. In the multivariate regression setting, it is often of interest to estimate the hazard ratio for subjects with covariates $\mathbf{x}'$ relative to those with covariates $\mathbf{x}''$. The general steps for computing hazard ratios from the results of a Cox regression are:

1. Write out the ratio of hazards using the model specified for the Cox regression,

$$HR = \frac{\lambda(t, \mathbf{x}')}{\lambda(t, \mathbf{x}'')} = \frac{\lambda_0(t)\exp\left\{\beta_1 x_1' + \beta_2 x_2' + \ldots + \beta_p x_p'\right\}}{\lambda_0(t)\exp\left\{\beta_1 x_1'' + \beta_2 x_2'' + \ldots + \beta_p x_p''\right\}}.$$

2. Reduce this equation to a form that is the exponential of the estimated regression parameters.

$$HR = \exp\left\{\beta_1\left(x_1' - x_1''\right) + \beta_2\left(x_2' - x_2''\right) + \ldots + \beta_p\left(x_p' - x_p''\right)\right\}$$

3. Insert the regression estimates for the parameters in order to calculate the hazard ratio.

If the value of a predictor variable is the same in the numerator and denominator hazards, then that predictor does not factor into the calculation of the hazard ratio. For instance, if $x_p' = x_p''$ then

$$\beta_p\left(x_p' - x_p''\right) = 0$$

and so the term for the p^{th} predictor drops out of the equation.

In the Breast-feeding example, the hazard is modeled as

$$\lambda(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}$$

for which the estimates from the Cox regression analysis are

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
0.0197	0.1583	-0.0224	-0.0516	-0.1898	0.1736	0.2894	0.2395

Breast-Feeding Example 1: Age

Goal: Estimate the hazard ratio for individuals aged 25, relative to those aged 20.

Q: In the multivariate setting, we have variables other than age to consider in computing the hazard ratio. What values should be use for them?

A: Our goal is to estimate the hazard ratio associated with age, while controlling for the effects of the other covariates in the model. This is done by comparing the hazards for individuals who differ only with respect to their ages (25 vs. 20). The individuals are assumed to share the same values for the remaining the covariates (alcohol, pre-na, smoking, education).

Specifically, the model estimates of the numerator and denominator hazards are

$$1.\ \hat{\lambda}(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\hat{\beta}_1 25 + \hat{\beta}_2 alcohol + \hat{\beta}_3 care3 \\ +\hat{\beta}_4 education + \hat{\beta}_5 poverty \\ +\hat{\beta}_6 race2 + \hat{\beta}_7 race3 + \hat{\beta}_8 smoke\end{array}\right\}$$

$$2.\ \hat{\lambda}(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\hat{\beta}_1 20 + \hat{\beta}_2 alcohol + \hat{\beta}_3 care3 \\ +\hat{\beta}_4 education + \hat{\beta}_5 poverty \\ +\hat{\beta}_6 race2 + \hat{\beta}_7 race3 + \hat{\beta}_8 smoke\end{array}\right\}$$

so that

$$\begin{aligned}\widehat{HR} &= \frac{\hat{\lambda}(t,\mathbf{x}')}{\hat{\lambda}(t,\mathbf{x}'')} = \frac{\hat{\lambda}(t;age=25)}{\hat{\lambda}(t;age=20)} \\ &= \exp\left\{\hat{\beta}_1 \times (25-20)\right\} = \exp\left\{0.0197 \times 5\right\}. \\ &= 1.104\end{aligned}$$

The other terms do not contribute because the values for those predictor variables are held constant. The estimated hazard ratio indicates that the rate of breast-feeding cessation for 25 year-olds is 1.104 times the rate for 20 year-olds, after controlling for alcohol use, prenatal care, education, poverty, race, and smoking.

Breast-Feeding Example 2: Race

Goal: Estimate the hazard ratio for each of black and "other" races, relative to whites.

The indicator variables that were created for race, so that it could be included as a categorical variable in the model, are summarized below.

Race	race1	race2	race3
White*	1	0	0
Black	0	1	0
Other	0	0	1

* Reference group; omitted from the regression model.

Blacks versus Whites

The hazard ratio comparing blacks to whites is

$$\widehat{HR} = \frac{\hat{\lambda}(t,\mathbf{x}')}{\hat{\lambda}(t,\mathbf{x}'')} = \frac{\hat{\lambda}(t;race2=1)}{\hat{\lambda}(t;race2=0)}$$
$$= \exp\left\{\hat{\beta}_6 \times (1-0)\right\} = \exp\{0.1736 \times 1\}.$$
$$= 1.341$$

The rate of breast-feeding cessation for blacks is 1.341 times the rate for whites, after controlling for age, alcohol use, prenatal care, education, poverty, and smoking.

"Other" versus Whites

The hazard ratio comparing race category "other" to whites is

$$\widehat{HR} = \frac{\hat{\lambda}(t, \mathbf{x}')}{\hat{\lambda}(t, \mathbf{x}'')} = \frac{\hat{\lambda}(t; race3 = 1)}{\hat{\lambda}(t; race3 = 0)}$$
$$= \exp\left\{\hat{\beta}_7 \times (1-0)\right\} = \exp\{0.2894 \times 1\}$$
$$= 1.498$$

The rate of breast-feeding cessation for those of race "other" is 1.498 times the rate for whites, after controlling for age, alcohol use, prenatal care, education, poverty, and smoking.

Breast-Feeding Example 3: Age and Race

Goal: Estimate the hazard ratio for blacks aged 25, relative to those aged 20 and of race "other".

The hazard ratio estimate is

$$\widehat{HR} = \frac{\hat{\lambda}(t; age = 25, race2 = 1)}{\hat{\lambda}(t; age = 20, race3 = 1)}$$
$$= \exp\left\{\hat{\beta}_1 \times (25-20) + \hat{\beta}_6 \times (1-0) + \hat{\beta}_7 \times (0-1)\right\}$$
$$= \exp\left\{\hat{\beta}_1 \times 5 + \hat{\beta}_6 - \hat{\beta}_7\right\}$$
$$= \exp\{0.0197(5) + 0.1736 - 0.2894\}$$
$$= 0.989$$

The rate of breast-feeding cessation for blacks aged 25 is 0.989 times the rate for those of race "other" aged 20, after controlling for alcohol use, prenatal care, education, poverty, and smoking.

4.3.2 Wald Confidence Intervals

Estimates of hazard ratios are often accompanied by confidence intervals and p-values in order to provide measures of statistical significance. Suppose that a Cox regression model of the form

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

is fit to a dataset, and interest lies in making inference about hazard ratios

$$\begin{aligned} HR &= \exp\left\{c_1\beta_1 + c_2\beta_2 + \ldots + c_p\beta_p\right\} \\ &= \exp\left\{\sum c_i\beta_i\right\} \end{aligned}$$

where c_i are arbitrary constants. The **Wald 100(1 - α)% confidence interval** for this hazard ratio is

$$\begin{aligned} CI &= \exp\left\{\sum c_i\beta_i \pm z_{1-\alpha/2}se\left(\sum c_i\beta_i\right)\right\} \\ &= HR\exp\left\{\pm z_{1-\alpha/2}se\left(\sum c_i\beta_i\right)\right\} \end{aligned}$$

Breast-Feeding Example 1: Age

The estimated hazard ratio for subjects aged 25 relative 20 is

$$\widehat{HR} = \exp\left\{5\hat{\beta}_1\right\} = 1.104.$$

Thus, the Wald 95% confidence interval is

$$
\begin{aligned}
CI &= \exp\left\{5\beta_1 \pm 1.96 se\left(5\beta_1\right)\right\} \\
&= \exp\left\{5\beta_1 \pm 1.96(5) se\left(\beta_1\right)\right\} \\
&= \exp\left\{5(0.0197) \pm 1.96(5)(0.0165)\right\} \\
&= (1.07, 1.14)
\end{aligned}
$$

where the estimates and standard errors are obtained from the regression output.

Breast-Feeding Example 2: Race

The estimated hazard ratio comparing blacks to those of race "other" is

$$\widehat{HR} = \exp\left\{\hat{\beta}_6 - \hat{\beta}_7\right\} = 0.89.$$

for which the resulting Wald 95% confidence interval is

$$
\begin{aligned}
CI &= \exp\left\{\hat{\beta}_6 - \hat{\beta}_7 \pm 1.96 se\left(\hat{\beta}_6 - \hat{\beta}_7\right)\right\} \\
&= \exp\left\{0.1736 - 0.2894 \pm 1.96(0.1287)\right\} \\
&= (0.69, 1.15)
\end{aligned}
$$

Note that the standard error for a linear combination of two or more parameters cannot be obtained directly from the regression output, i.e.

$$se\left(\hat{\beta}_6 - \hat{\beta}_7\right) \neq se\left(\hat{\beta}_6\right) - se\left(\hat{\beta}_7\right).$$

In general, the standard error is a function of the coefficients and estimated covariance matrix for the parameters involved in the calculation of the hazard ratio.

Breast-Feeding Example 3: Age and Race

The estimated hazard ratio comparing blacks aged 25 to those aged 20 of race "other" is

$$\widehat{HR} = \exp\left\{5\hat{\beta}_1 + \hat{\beta}_6 - \hat{\beta}_7\right\} = 0.99.$$

for which the resulting Wald 95% confidence interval is

$$\begin{aligned} CI &= \exp\left\{5\hat{\beta}_1 + \hat{\beta}_6 - \hat{\beta}_7 \pm 1.96 se\left(5\hat{\beta}_1 + \hat{\beta}_6 - \hat{\beta}_7\right)\right\} \\ &= \exp\left\{-0.0172 \pm 1.96(0.1606)\right\} \\ &= (0.72, 1.35) \end{aligned}$$

Although PROC PHREG in SAS does not provide an option to generate these confidence intervals, the variance necessary to compute them manually can be obtained.

SAS Program and Output

```
proc phreg data=breastfed;
   model weeks*weaned(0) = age alcohol care3 education poverty
         race2 race3 smoke / risklimits;
   Age25v20: test 5*age / print;
   Race2v3: test race2 - race3 / print;
   Age25Race2vAge20Race3: test 5*age + race2 - race3 / print;
run;
```

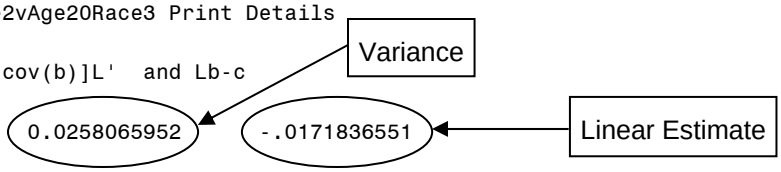
Syntax

- The **risklimits** option displays confidence intervals for the hazard ratios computed for each individual term in the model.
- The **test** statement can be used to obtain variance estimates for any linear combination of the model parameters.
- A label to appear in the SAS output. It should be given at the beginning of the test statement, followed by a colon. The label for the first statement in this example is “Age25v20”.
- Any linear combination of the parameters may be specified using numbers and arithmetic operators. Numbers must appear before the associated variable name.
- The **print** option displays the estimate and variance for the specified linear combination.

```
 Test Age25v20 Print Details

            L[cov(b)]L'  and Lb-c

Equation 1      0.0067719319      0.0985778750

Ginv(L[cov(b)]L')  and Ginv(L[cov(b)]L')(Lb-c)

Equation 1       147.6683488        14.5568320

      Test Age25v20 Results

       Wald
 Chi-Square      DF    Pr > ChiSq

     1.4350       1        0.2310
```

```
Test Race2v3 Print Details

            L[cov(b)]L'  and Lb-c

Equation 1      0.0165649030      -.1157615301

Ginv(L[cov(b)]L')  and Ginv(L[cov(b)]L')(Lb-c)

Equation 1       60.36859966       -6.98836147


      Test Race2v3 Results

       Wald
 Chi-Square      DF    Pr > ChiSq

     0.8090       1        0.3684
```

```
Test Age25Race2vAge20Race3 Print Details

            L[cov(b)]L'  and Lb-c

Equation 1      0.0258065952      -.0171836551

Ginv(L[cov(b)]L')  and Ginv(L[cov(b)]L')(Lb-c)

Equation 1       38.74978436       -0.66586293


Test Age25Race2vAge20Race3 Results

       Wald
 Chi-Square      DF    Pr > ChiSq
     0.0114       1        0.9148
```



4.3.3 Wald Test Statistic

Hazard ratios that differ from unity indicate a difference between the hazard rates for the two groups being compared. Thus, the null hypotheses in testing for a significant hazard ratio is

$$H_0 : HR = 1$$

where, in general,

$$\begin{aligned} HR &= \exp\left\{c_1\beta_1 + c_2\beta_2 + \ldots + c_p\beta_p\right\} \\ &= \exp\left\{\sum c_i\beta_i\right\} \end{aligned}.$$

An equivalent way to write the hypothesis is

$$H_0 : \sum c_i\beta_i = 0\,.$$

The **Wald test statistic** for assessing the significance of the hazard ratio is

$$X_{Wald} = \frac{\sum c_i\beta_i}{se\left(\sum c_i\beta_i\right)} \sim N(0,1)$$

or

$$X^2_{Wald} = \left(\frac{\sum c_i\beta_i}{se\left(\sum c_i\beta_i\right)}\right)^2 \sim \chi^2_1\,.$$

The p-value formulas for one and two-sided alternative hypotheses are:

$H_A : HR < 1$	$p = \Pr\left[Z < X_{Wald}\right]$
$H_A : HR > 1$	$p = \Pr\left[Z > X_{Wald}\right]$
$H_A : HR \neq 1$	$p = 2\Pr\left[Z > \left\|X_{Wald}\right\|\right]$ $= \Pr\left[\chi_1^2 > X_{Wald}^2\right]$

Wald test statistics and Wald confidence intervals give consistent results. In other words, for a given alpha-level, the Wald two-sided confidence interval will include unity if and only if the two-sided p-value is non-significant. For this reason, in summaries that include both confidence intervals and hypothesis testing, Wald methods are used for both.

Breast-Feeding Example 1: Age

Suppose that we are interested in testing that the estimated hazard ratio for subjects aged 25 relative 20 is significantly different from unity.

$$\widehat{HR} = \exp\left\{5\hat{\beta}_1\right\} = 1.104.$$

The null and alternative hypotheses are

$$H_0 : HR = 1$$
$$H_A : HR \neq 1$$

for which the Wald test statistics is

$$X_{Wald} = \frac{5\hat{\beta}_1}{se\left(5\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{0.0197}{0.0165} = 1.19.$$

The resulting p-value is $2\Pr\left[Z > \left|1.19\right|\right] = 0.2130$.

Therefore, at the 5% level of significance, the hazard ratio

for 25 year-olds versus 20 year-olds is not significant. Note that for this model, the Wald statistic will be the same regardless of the ages being compared. In other words, age is not significantly associated with the rate of breast-feeding cessation, after controlling for alcohol use, prenatal care, education, poverty, race, and smoking.

4.3.4 Proportional Hazards Assumption

The proportional hazards assumption of the Cox model implies that the hazard rates are constant multiples of the baseline hazard $\lambda_0(t)$. For example, consider the Cox regression model and parameter estimates for the Breast-Feeding Study given below.

$$\lambda(t;\mathbf{x}) = \lambda_o(t)\exp\{\beta_{black}race2 + \beta_{other}race3\}$$

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
race2	$\hat{\beta}_{black}$	0.0940	0.1024	0.84	0.3586
race3	$\hat{\beta}_{other}$	0.2406	0.0924	6.78	0.0092

A plot of hypothetical hazard rates for this model is given in Figure 1.

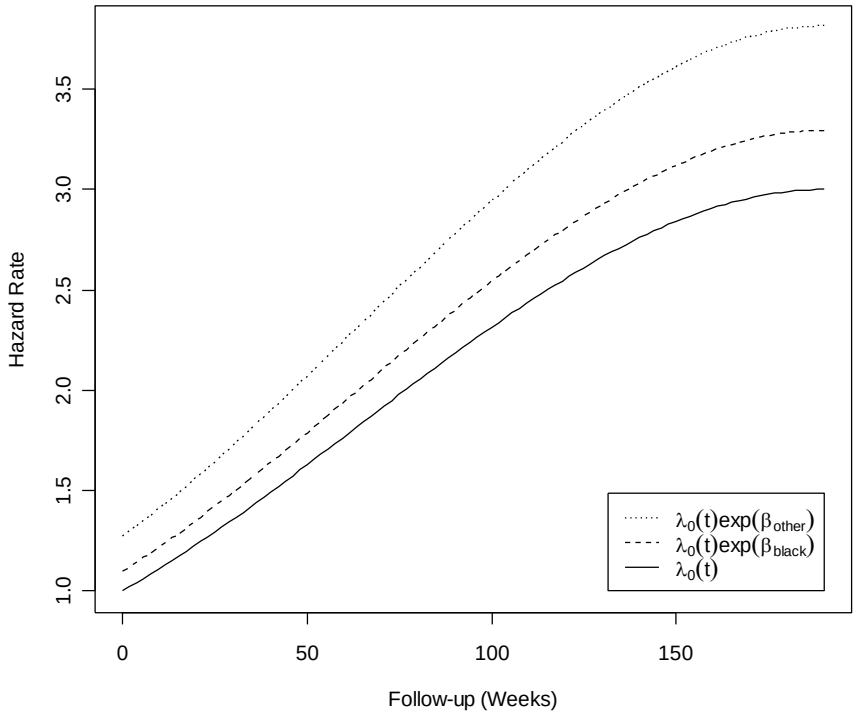


Figure 1. Multiplicative effect on the hazard rate.

The proportional hazards assumption can be further understood by considering the natural logarithm of the hazard rate

$$\ln \lambda(t, \mathbf{x}) = \ln \lambda_0(t) + \beta_{black} race2 + \beta_{other} race3 .$$

Thus, we see that the log of the hazard rate for the regression model is an additive function of the covariates (Figure 2). Likewise, the effects of continuous variables, such as age, would also be additive on the log-scale.

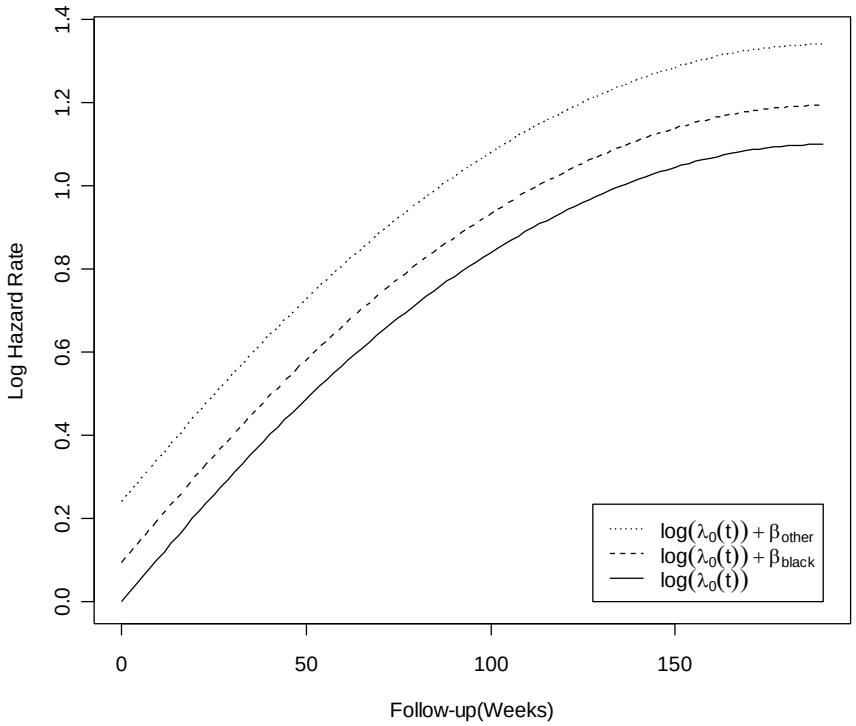


Figure 2. Additive effect on the log-hazard rate.

The previous are hypothetical plots of the hazard function, meant to illustrate the additive effects of covariates on the log-scale. In practice, the baseline hazard is not estimated in Cox regression. Instead, we use a proportional hazards assumption to obtain a hazard ratio that is independent of the baseline hazard. Logistic regression is used similarly to model the odds ratio without need to estimate the probability of the outcome of interest.

4.4 Likelihood Estimation

The method of maximum likelihood is used to estimate the parameters in Cox regression. Maximum likelihood is also used to estimate parameters in logistic regression. The maximum likelihood estimates are the values of the model parameter that maximize a likelihood function comparable to

$$L(\mathbf{t};\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left\{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\right\}}{\sum_{j \in r(t_i)} \exp\left\{\beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj}\right\}}$$

where $i = 1,\ldots,k$ indexes the events, and j indexes the subjects at risk at the time event i occurs. Thus, the data contribute to the estimation of the regression parameters only at the failure times. Computational algorithms are employed to find the values of $\boldsymbol{\beta}$ that maximize this function.

Comments

Besides yielding estimates of the parameters, the method of maximum likelihood yields several results that are useful in comparing models, testing hypotheses, and constructing confidence intervals. Maximum likelihood provides:

- The maximized value of the likelihood $L\left(\mathbf{t};\hat{\boldsymbol{\beta}}\right)$. Larger values of the likelihood indicate a better fit to the data.
- Estimates of the variances (standard errors) and covariances (correlations) for the parameter estimates.

Furthermore, note that the likelihood function naturally allows for

1. Varying follow-up period, i.e. the start times for the subjects can differ since it is only important to keep track of who is in the risk set at any given failure time. This is comparable to the allowance for varying follow-up periods in computing the Kaplan-Meier survival estimates.
2. The covariate values for a subject may vary over the follow-up period. Note that at each failure time t_i, the covariate values for the subject that failed and the subjects at risk are compared. Subjects may contribute to more than one risk set, and their covariate values need not remain constant from one risk set to another.

4.5 Summary

1. The proportional hazards model is a relative rate model. The hazard ratio gives the factor by which the rate is increase over the comparison group. A hazard ratio > 1 indicates an increased relative rate; a ratio < 1 indicates a decreased relative rate; a ratio = 1 indicates no difference in the rates.

2. Given two sets of covariates x' and x'', the hazard ratio is a multiplicative function of the covariates, namely

$$\frac{\lambda(t,\mathbf{x}')}{\lambda(t,\mathbf{x}'')} = \frac{\lambda_0(t)\exp\{\beta_1 x_1' + \ldots + \beta_p x_p'\}}{\lambda_0(t)\exp\{\beta_1 x_1'' + \ldots + \beta_p x_p''\}}$$
$$= \exp\{\beta_1(x_1' - x_1'') + \ldots + \beta_p(x_p' - x_p'')\}$$
$$= \exp\{\beta_1(x_1' - x_1'')\} \times \ldots \times \exp\{\beta_1(x_p' - x_p'')\}$$

Alternatively, the log-hazard ratio is an additive function of the covariates

$$\ln\frac{\lambda(t,\mathbf{x}')}{\lambda(t,\mathbf{x}'')} = \ln\lambda(t,\mathbf{x}') - \ln\lambda(t,\mathbf{x}'')$$
$$= \beta_1(x_1' - x_1'') + \ldots + \beta_p(x_p' - x_p'')$$

.

This implies a constant difference between the log-hazard rates; similar to the log-odds ratio in the logistic regression model.

Applied Survival Analysis (171:242)

Section 5: Cox Regression with Time-Dependent Covariates

Brian J. Smith, Ph.D.

February 21, 2005

Table of Contents

5.1 Introduction

Up to this point, our analyses have focused on the effect, on the hazard, of a set of covariates measured at baseline. For example, in the breast-feeding study we considered the effects of the baseline characteristics age, alcohol consumption, prenatal care, education, race, and smoking. Remember, though, that the subjects are being followed forward in time. Thus, there is an opportunity to measure covariates as they change during the follow-up period. Covariates that change over time are referred to as *time-dependent* or *time-varying covariates*.

Colorado Plateau Uranium Miners Study

The Colorado Plateau uranium miners study was one of the earliest of the modern epidemiologic studies to document increased lung cancer risk with exposure to radon. Listed below are a few details of the study.

- 3,347 miners were enrolled in the four States of Arizona, Colorado, New Mexico, and Utah, who had completed at least one month of underground uranium mining, volunteered for at least one medical examination between 1950 and 1960, and provided personal and occupational data of sufficient detail for exposure estimation.
- 258 lung cancer deaths were observed; 2,661 subjects reported smoking at some point in their lives.
- Follow-up ended in the 1980s.

In this section, we will focus on the lung cancer effect of smoking among the uranium miners.

Table 1. Description of variables in the Uranium Miners Study dataset.

Variable	Description	Values
id	Subject identifier	numerical
deathlung	Death from lung cancer	1 = yes 0 = no
age1	Age at entry to study	continuous
age2	Age at exit from study	continuous
smoke1	Age started smoking	continuous
smoke2	Age last known to smoke	continuous
smoketot	Total cigarette smoking (100s of packs)	continuous
sexp5	Cigarette smoking during ages 1-5	continuous
sexp10	Cigarette during ages 6-10	continuous
⋮	⋮	⋮
sexp90	Cigarette smoking during ages 86-90	continuous

Table 2. Follow-up statistics for the 3,347 subjects.

Variable	Mean	SD	Min	Median	Max
age1	35.4	11.6	15.8	34.0	80.0
age2	57.4	11.0	19.2	56.5	98.5
Years of follow-up	22.0	7.3	0.1	23.9	32.5

Table 3. Smoking statistics for the 2,661 smokers.

Variable	Mean	SD	Min	Median	Max
smoke1	15.3	5.4	1.0	16.0	56.0
smoke2	55.3	12.1	16.0	55.2	89.5
smoketot	140.3	77.7	0.0	130.9	676.3
Years of smoking	40.0	12.2	5.0	40.1	77.9

5.2 Choice of Time-Scale

In follow-up studies there may be more than one choice of time-scale over which to define the baseline hazard in the Cox regression model. Two possible choices for the Uranium Miners Study are time-on-study and age. Recall the general form of the Cox model,

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}.$$

The baseline hazard function offers the greatest flexibility in controlling for the effects of a time-varying covariate on the hazard rate because the baseline hazard is modeled non-parametrically. Thus, the recommended choice of time-scale is the one over which the hazard rate is most variable.

In the Miners study we would expect the force of mortality to differ more as a function of age then of time-on-study. Hence, age will be used as the time-scale for the baseline hazard in the Cox regression analyses.

Guidelines

- Time-on-study is more appropriate for studies in which enrollment coincides with some intervention; e.g. clinical trials.
- Age is more appropriate for prospective observational studies of health populations; e.g. NCI Iowa and North Carolina Agricultural Health Study. When age is used as the time-scale, the age intervals during which subjects are enrolled in the study must be specified in the Cox regression analysis.

SAS Program and Output

```
data radonmod;
   set radon;
   smoker = (smoke1 > 0);

proc phreg data=radonmod;
   model (age1,age2)*deathlung(0) = smoker smoketot;

run;
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smoker	1	1.36620	0.26502	26.5749	<.0001	3.920
smoketot	1	-0.00122	0.0007860	2.3955	0.1217	0.999

Syntax

- Risk intervals may be specified in the model statement for the time-scale variable. The syntax is *(t1, t2) <*censor(value)>*, where *t1* and *t2* define the beginning and end of the risk interval, respectively, and censoring/events are assumed to occur at the end of the interval.

5.3 Time-Dependent Covariates

A baseline measure may not be sufficient to capture the effect of a covariate on the outcome of interest. Hence, covariates may be measured repeatedly during the follow-up period. For instance, measures of smoking behavior are available at 5-year age intervals in the Miners Study. Consider the following data for one of the subjects:

id	deathlung	age1	age2	smoke1	smoke2
1979	1	36.1	51.0	21	41
sexp20	sexp25	sexp30	sexp35	sexp40	sexp45
0	22.3	27.3	27.3	21.4	0

This subject's smoking status as well as the amount smoked is changing during follow-up, as illustrated in Figure 1.

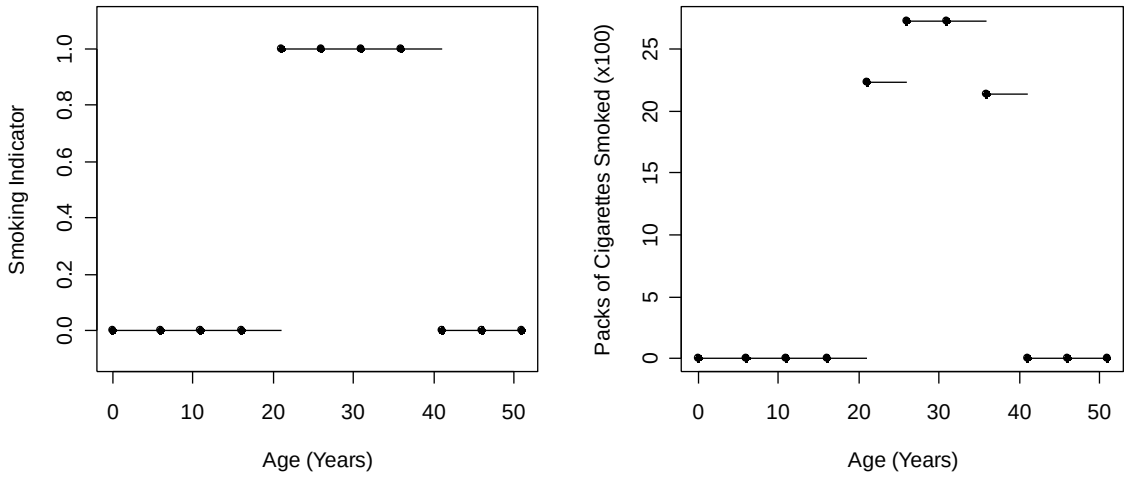


Figure 1. Change in smoking status over time.

Furthermore, we might be interested in using the cumulative packs smoked as a function of time. Depicted in Figure 2 are two possible ways to quantify cumulative smoking over time. The first approach assumes that the smoking variable remain constant until the next follow-up time point. The second uses linearly interpolate to estimate how the variable might be changing from one follow-up time point to another.

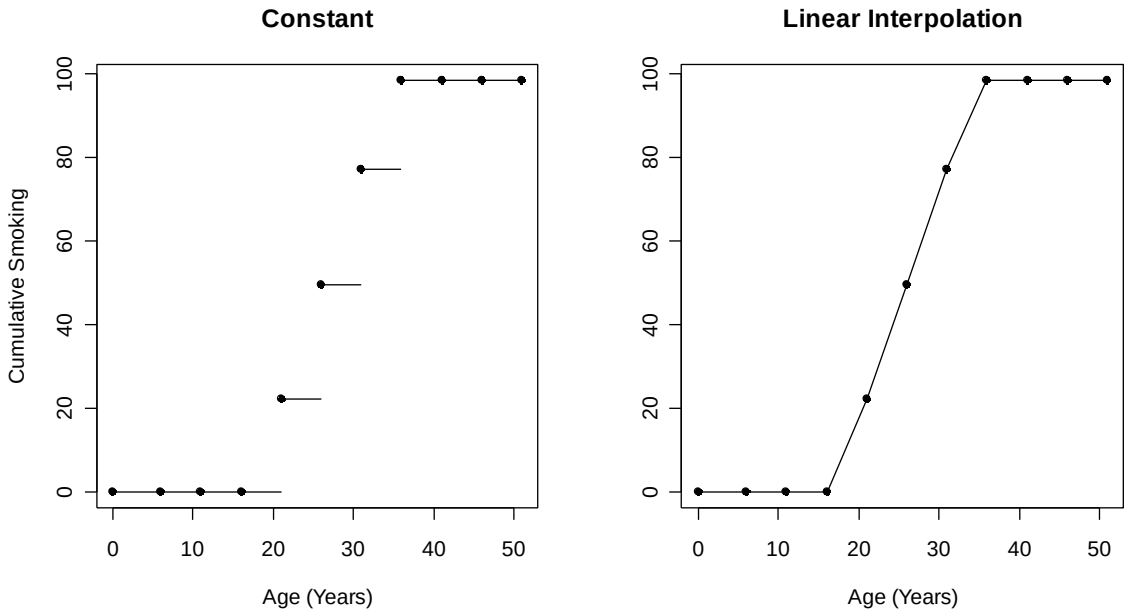


Figure 2. Cumulative packs of cigarettes smoked over time.

Similarly, time-dependent covariates could be created for years of cigarette smoking or time since smoking cessation.

Note that measurements for time-dependent covariates are rarely available at every point in time. Rather, the covariates are measured at a finite number of times during the study; e.g. every 5 years.

5.3.1 Notation

To account for time-dependent covariates, the Cox regression model may be written as

$$\lambda(t,\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1(t) + \beta_2 x_2(t) + \ldots + \beta_p x_p(t)\right\}$$

where each covariate could potentially be changing over time. The covariates must be defined for every time point *t*. Baseline covariates are simply defined as $x_j(t) = x_j$.

5.3.2 Example 1: Smoking Indicator

Suppose that our time-dependent covariate $x(t)$ is an indicator variable for smoking in the Miners Study, defined as

$$x(t) = \begin{cases} I(sexp5 > 0) & 0 < t < 6 \\ I(sexp10 >) & 6 \le t < 11 \\ \vdots & \vdots \\ I(sexp90 > 0) & 86 \le t < 91 \end{cases}.$$

and we are interested in fitting the Cox model,

$$\lambda(t,\mathbf{x}) = \lambda_0(t)\exp\{\beta x(t)\}.$$

SAS Program and Output

```
proc phreg data=radon;
   model (age1,age2)*deathlung(0) = smoker;
   if age2 < 6 then smoker = (sexp5 > 0);
   else if age2 < 11 then smoker = (sexp10 > 0);
   else if age2 < 16 then smoker = (sexp15 > 0);
   else if age2 < 21 then smoker = (sexp20 > 0);
   else if age2 < 26 then smoker = (sexp25 > 0);
   else if age2 < 31 then smoker = (sexp30 > 0);
   else if age2 < 36 then smoker = (sexp35 > 0);
   else if age2 < 41 then smoker = (sexp40 > 0);
   else if age2 < 46 then smoker = (sexp45 > 0);
   else if age2 < 51 then smoker = (sexp50 > 0);
   else if age2 < 56 then smoker = (sexp55 > 0);
   else if age2 < 61 then smoker = (sexp60 > 0);
   else if age2 < 66 then smoker = (sexp65 > 0);
   else if age2 < 71 then smoker = (sexp70 > 0);
   else if age2 < 76 then smoker = (sexp75 > 0);
   else if age2 < 81 then smoker = (sexp80 > 0);
   else if age2 < 86 then smoker = (sexp85 > 0);
   else if age2 < 91 then smoker = (sexp90 > 0);
run;
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smoker	1	0.62198	0.16010	15.0926	0.0001	1.863

5.3.3 Example 2: Cumulative Smoking

Suppose that our time-dependent covariate $x(t)$ is cumulative smoking in the Miners Study, defined as

$$x(t) = \begin{cases} sexp5 & 0 < t < 6 \\ sexp5 + sexp10 & 6 \le t < 11 \\ \vdots & \vdots \\ sexp5 + \ldots + sexp90 & 86 \le t < 91 \end{cases}.$$

and we are interested in fitting the Cox model,

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp\{\beta x(t)\}.$$

Note that this definition for cumulative smoking assumes that the value is constant within each of the 5-year age intervals, which is consistent with the first plot in Figure 2. Alternatively, linear interpolation could be used to allow cumulative smoking to increase linearly across the intervals.

SAS Program and Output

```
data radonmod;
   set radon;
   csexp10 = sexp5 + sexp10;
   csexp15 = csexp10 + sexp15;
   csexp20 = csexp15 + sexp20;
   csexp25 = csexp20 + sexp25;
   csexp30 = csexp25 + sexp30;
   csexp35 = csexp30 + sexp35;
   csexp40 = csexp35 + sexp40;
   csexp45 = csexp40 + sexp45;
   csexp50 = csexp45 + sexp50;
   csexp55 = csexp50 + sexp55;
   csexp60 = csexp55 + sexp60;
   csexp65 = csexp60 + sexp65;
   csexp70 = csexp65 + sexp70;
   csexp75 = csexp70 + sexp75;
   csexp80 = csexp75 + sexp80;
   csexp85 = csexp80 + sexp85;
   csexp90 = csexp85 + sexp90;

proc phreg data=radonmod;
   model (age1,age2)*deathlung(0) = smokecum;
   if age2 < 6 then smokecum = sexp5;
   else if age2 < 11 then smokecum = csexp10;
   else if age2 < 16 then smokecum = csexp15;
   else if age2 < 21 then smokecum = csexp20;
   else if age2 < 26 then smokecum = csexp25;
   else if age2 < 31 then smokecum = csexp30;
   else if age2 < 36 then smokecum = csexp35;
   else if age2 < 41 then smokecum = csexp40;
   else if age2 < 46 then smokecum = csexp45;
   else if age2 < 51 then smokecum = csexp50;
   else if age2 < 56 then smokecum = csexp55;
   else if age2 < 61 then smokecum = csexp60;
   else if age2 < 66 then smokecum = csexp65;
   else if age2 < 71 then smokecum = csexp70;
   else if age2 < 76 then smokecum = csexp75;
   else if age2 < 81 then smokecum = csexp80;
   else if age2 < 86 then smokecum = csexp85;
   else if age2 < 91 then smokecum = csexp90;
run;
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smokecum	1	0.00358	0.0006083	34.6295	<.0001	1.004

5.3.4 Multiple Records Format

In the previous two examples, the time-dependent covariates were created with programming statements in the body of PROC PHREG. This approach is most useful when there is a fixed set of time points at which the values of the covariates change for all subjects; e.g. 5-year age intervals.

When the time points vary from subject-to-subject, another approach can be used. Specifically, a new dataset can be created with a separate row for each time interval in which the subject has a different set of covariate values.

Consider, again, the subject with covariate values

id	deathlung	age1	age2	smoke1	smoke2
1979	1	36.1	51.0	21	41
sexp20	sexp25	sexp30	sexp35	sexp40	sexp45
0	22.3	27.3	27.3	21.4	0

During the follow-up period for this subject (ages 36.1 – 51.0) his smoking status changes ones. Thus, the dataset could be restructured so that there are separate rows to

store the covariate values corresponding to the period when he was and was not a smoker.

id	deathlung	age1	age2	smoker	smokecum
1979	0	36.1	39.9	1	98.3
1979	1	39.9	51.0	0	98.3

The process of creating this new dataset can be summarized in the following steps:

1. Determine the time points at which the covariate values change for a given subject.
2. Partition the on-study time into intervals defined by the time points in Step 1.
3. For each time interval, create a new row in the dataset that contains the start and stop times for the interval and all (baseline and time-dependent) covariates.
4. The event indicator variable should be included in each row, but only take on a value of one for the interval in which the event occurs; otherwise it should be coded as a zero.

Applied Survival Analysis (171:242)

Section 6: Cox Proportional Hazards Model Diagnostics

Brian J. Smith, Ph.D.

March 30, 2007

Table of Contents

6.1 Introduction

6.1.1 Linear Regression

Residuals are most easily understood in the context of linear regression, where the response variable y_i for each subjects is modeled as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i ,$$

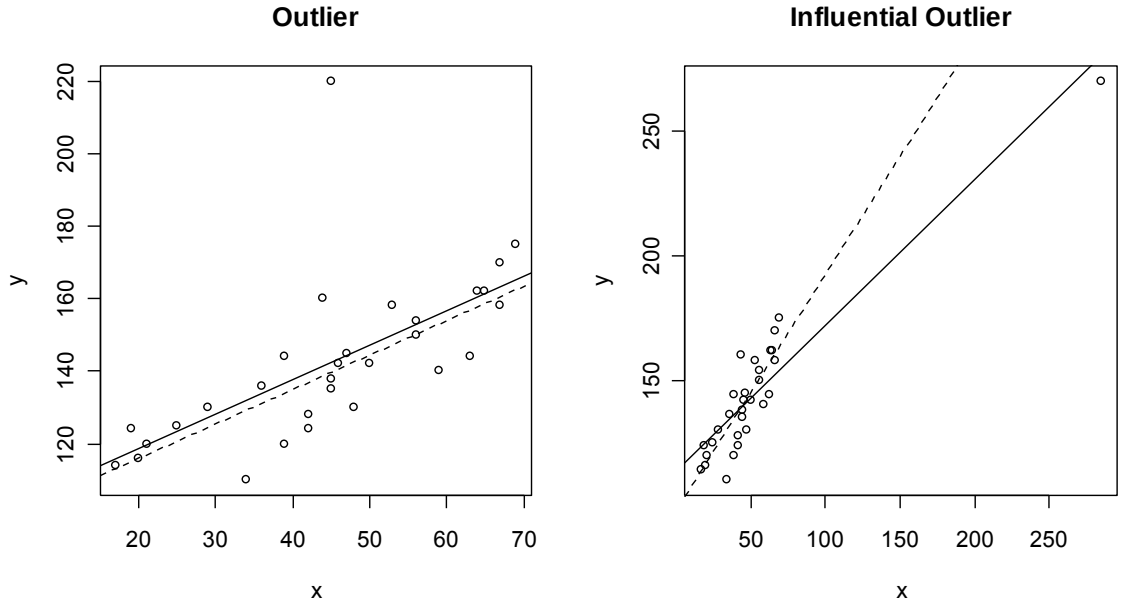
Recall that the *residuals* r_i are defined as the difference between the observed y_i and the predicted $\hat{y}_i$. These differences can then be used to examine the fit of the linear regression model. Notationally,

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_p x_{pi} .$$

The residuals provide estimates of the error terms ε_i in the model. Hence, they may be used to check the assumption that $\varepsilon_i \sim N\left(0, \sigma^2\right)$; that the error terms are normally distributed with constant variance.

6.1.2 Outliers

Regression diagnostics should be performed to identify subjects whose outcome and/or predictor variables are different from the majority of the sample. Such subjects are referred to as outliers.



Notes

- The solid line represents the regression fit with the outlier in the analysis; the dashed line represents the fit without the outlier.
- The first plot depicts an outlier whose response value is not explained well by the predictor in the model. In other words, there is a relatively large difference between the observed and predicted response (the residual value).

- The second, influential outlier, has a substantial impact on the estimated effect of the predictor, but does not have a large residual value.

6.1.3 Goodness-of-fit

In least squares linear regression, the R^2 statistic is often reported as a measure of the amount of variability in the data that is explained by the model. Recall that the sum-of-squared errors

$$SSE = \sum (y_i - \hat{y}_i)^2$$

measures the aggregate deviation of the predicted values from the observed. We would like for *SSE* to be small. The R^2 statistic

$$R^2 = \frac{SST - SSE}{SST},$$

where

$$SST = \sum (y_i - \bar{y})^2 \text{ and } SSR = \sum (\hat{y}_i - \bar{y})^2,$$

provides a measure of the overall fit of the model to the data. Specifically, it measures the amount of variability in the response variable explained by the predictors.

6.2 Residuals

In linear regression, residuals are simply computed as the observed response variables minus the predicted values. We can then plot the residuals to check that they are normally distributed. They can also be plotted against covariates not included in the model to explore possible relationships that are not accounted for. We would like to perform comparable residual analyses in the Cox regression setting. However, here we are modeling the hazard rate

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \ldots + \beta_p x_p\right\}$$

which is not directly observable. As a result, the construction of residuals is more involved. In fact, there are many different proposed methods for computing residuals in the Cox regression setting. We will discuss two:

1. Martingale Residuals
2. Deviance Residuals

6.2.1 Martingale Residuals

Martingale residual can be explained as the difference between the number of events (0 or 1) occurring for the i^{th} individual during follow-up and the number expected under the model. These residuals are used primarily to identify patterns in the data that are not explained by the model.

Lymphoma Example

In the Kaplan-Meier analysis of the Lymphoma Study we found an interaction between the method of bone marrow transplant and disease type. Thus, we might propose the following model:

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\{\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl\}$$

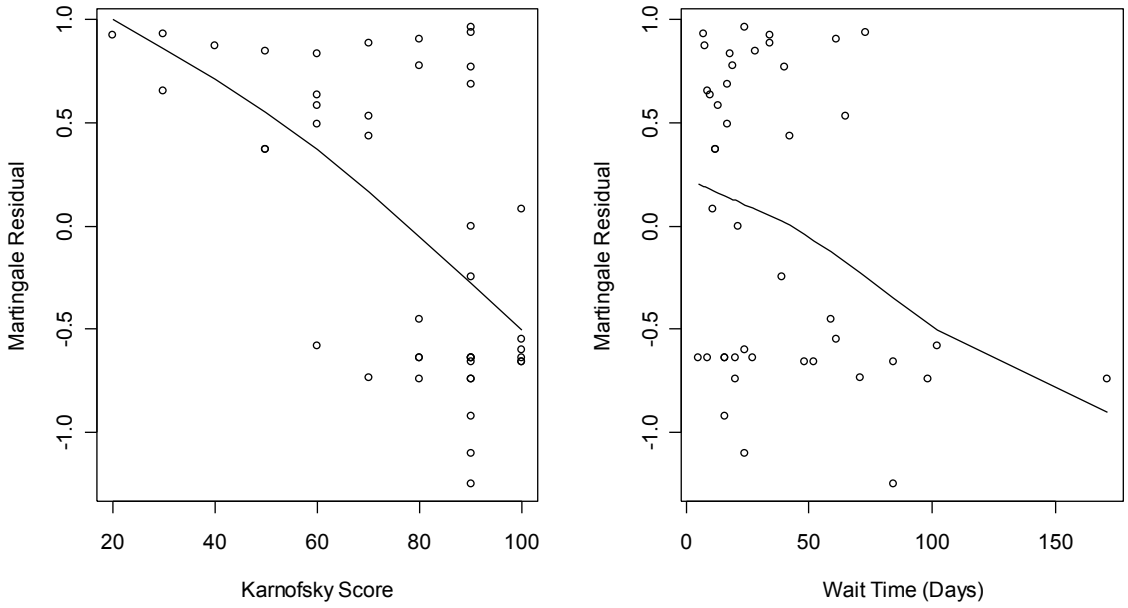
where

$$auto = \begin{cases} 1 & \text{if } graft = 2 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad nhl = \begin{cases} 1 & \text{if } disease = 1 \\ 0 & \text{otherwise} \end{cases}.$$

The regression estimates for this model are given in the table below.

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
auto	$\hat{\beta}_1$	-1.6762	0.6200	7.3101	0.0069
nhl	$\hat{\beta}_2$	-1.8298	0.6753	7.3424	0.0067
auto*nhl	$\hat{\beta}_3$	2.3400	0.8517	7.5489	0.0060

We may be interested in adding wait time and Karnofsky scores to the model. Martingale residuals are useful for examining the relationship between the disease rate and covariates not included in the model.



The plots suggest a linear effect for Karnofsky scores and a nonlinear effect for wait time. Notice that the residuals in the latter case are predominantly less than zero after about 70 days; otherwise, the residuals are more evenly scattered about zero. Consequently, we might create the indicator variable

Variable	Levels	N	Percents
wait70	0 = wait < 70 1 = wait $\geq$ 70	36 7	84% 16%

and fit the Cox model

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{matrix}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70\end{matrix}\right\}$$

The resulting parameter estimates are given below.

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
auto	$\hat{\beta}_1$	-1.8600	0.7335	6.4312	0.0112
nhl	$\hat{\beta}_2$	-2.7276	0.8273	10.8691	0.0010
auto*nhl	$\hat{\beta}_3$	2.4845	0.9849	6.3641	0.0116
karnofsky	$\hat{\beta}_4$	-0.0539	0.0123	19.3641	<.0001
wait70	$\hat{\beta}_5$	-1.5140	0.7449	4.1305	0.0421

6.2.2 Deviance Residuals

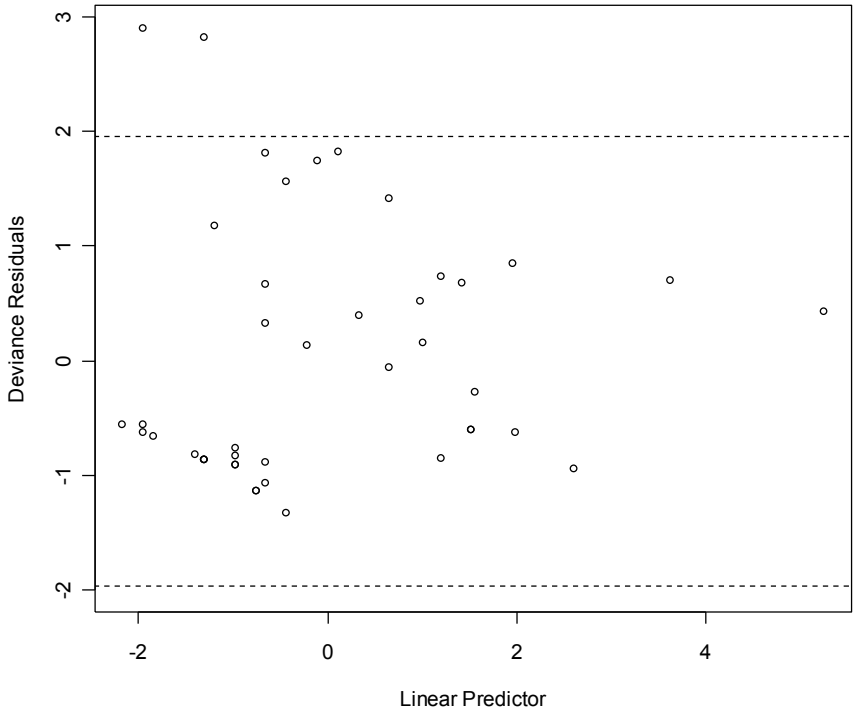
Deviance residuals are a transformed version of the Martingale residuals and defined so as to generate results that tend toward the standard normal distribution. These serve as the semi-parametric analog to the residuals utilized in linear regression. The deviance residuals are often plotted against the values of the linear predictor in the Cox regression model. Observations that deviate from the specified model will result in relatively large residuals. Thus, deviance residuals are useful in detecting outliers and points in the data that are not adequately described by the model.

Lymphoma Study

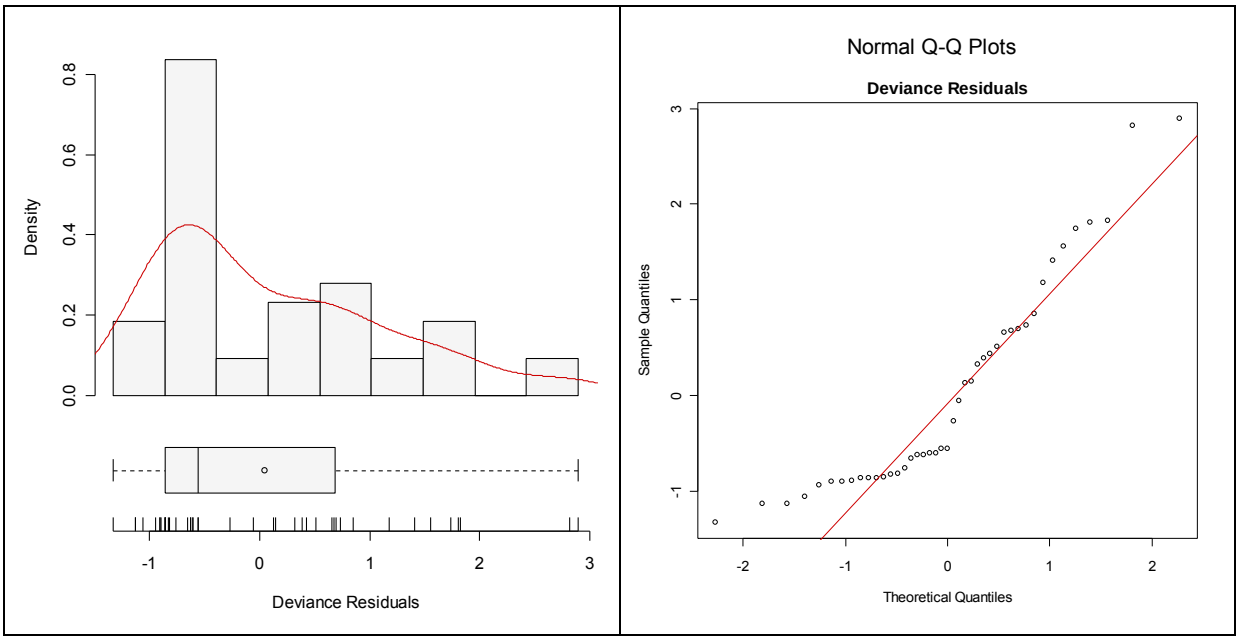
Suppose that we are evaluating the following model for predicting recurrence in the Lymphoma Study:

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{matrix}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70\end{matrix}\right\}.$$

The deviance residuals that result in fitting this model are displayed in the plot below.



Approximately 95% of the deviance residuals would be expected to fall within the interval (–1.96, 1.96). In this example, 95.3% of the residuals are within this range.



Notes

- If the model provides an adequate fit to the data, the Deviance residuals will have an approximate mean of 0 and variance of 1.
- Few extreme positive or negative Deviance residuals would be expected.
- About 95% of the residuals fall between -1.96 and +1.96.
- About 99% of values fall between -2.32 and +2.32. Values substantially outside of this range should be investigated as potential outliers.

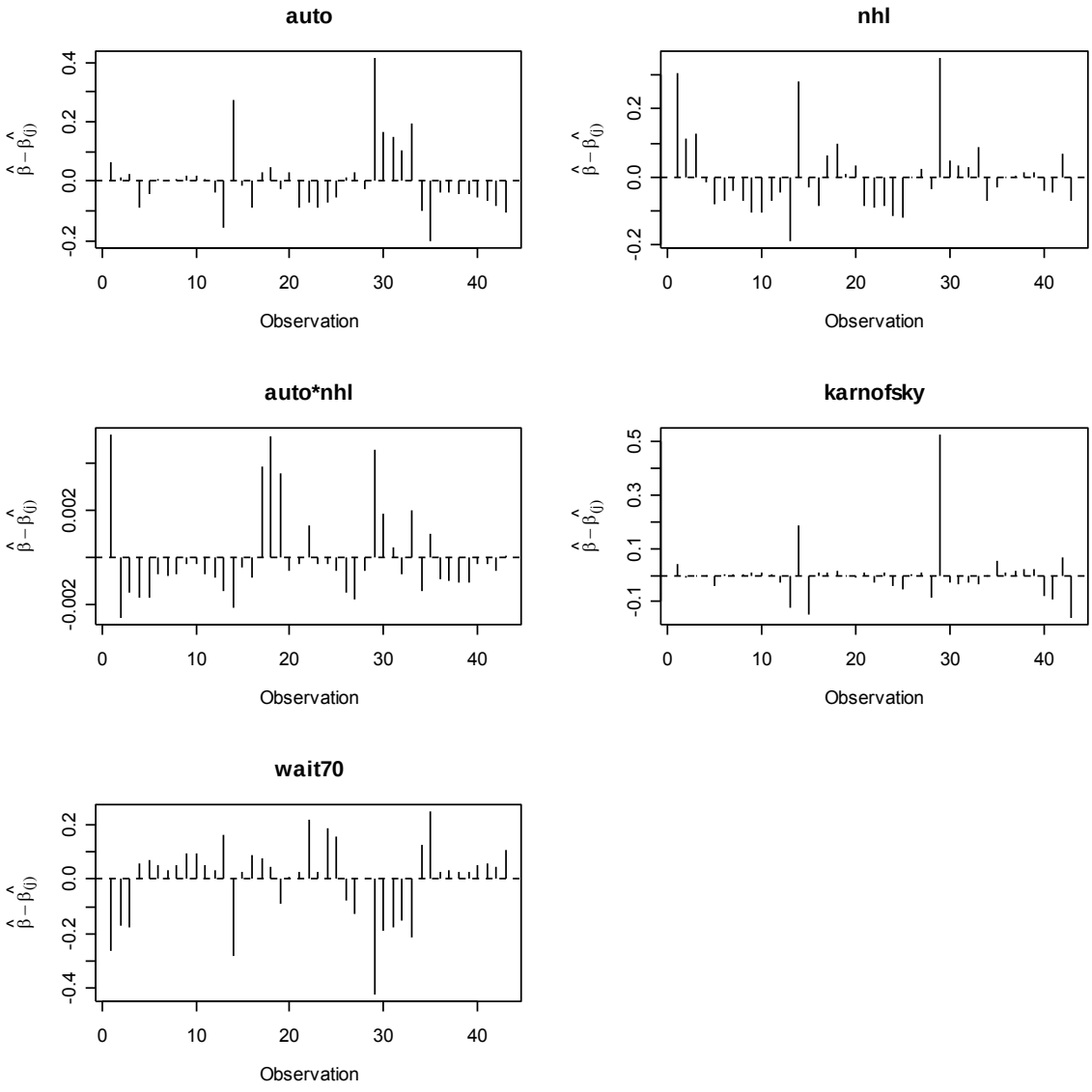
6.2.3 Delta-Beta Plots

Delta-Beta plots are one method of checking the influence of each observation on the estimated model parameters.

- The idea is to compare the estimate $\hat{\beta}$ for a given parameter with all observations in the analysis, to the estimates $\hat{\beta}_{(j)}$ after excluding the j^{th} observation.
- This is done for every observation in the data set and the changes $\Delta_j = \hat{\beta} - \hat{\beta}_{(j)}$ are reported as the delta-beta values.
- Observations that exert undue influence on the parameter estimates have large delta-betas.
- A delta-beta plot may be constructed for each term in the regression model.

<u>Lymphoma Example</u>

Delta-beta plots were constructed by excluding observations one-at-a-time and "refitting" the final model to obtain the associated changes in the parameters. This was done for each of the 43 subjects in the data set. The changes are plotted below against the observation numbers.



One of the subjects appears to have a relatively large delta-beta in the plot of the Karnofsky scores. This is subject 29, whose covariate values are

days	event	auto	nhl	karnofsky	wait70
90	1	0	0	90	1

The mean number of follow-up days for subjects with Karnofsky scores greater than 50 is over 500 days. Thus, with a Karnofsky score of 90, this subject has an unusually short time to disease recurrence.

SAS Code

```
data lymphomamod;
   set lymphoma;
   auto = (graft = 2);
   nhl = (disease = 1);
   auto_nhl = auto * nhl;
   wait70 = (wait >= 70);

proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl;
   output out=residuals1 resmart=martingale;

proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl karnofsky wait70;
   output out=residuals2 resdev=deviance
          dfbeta=db_auto db_nhl db_auto_nhl db_karnofsky db_wait70;
run;
```

Syntax

- The **output** statement saves the Martingale residuals (**resmart)** and the Deviance residuals (**resdev**) in the SAS data sets **residuals1** and **residuals2** with variable names **martingale** and **deviance**, respectively. The saved residuals may then be plotted with appropriate graphing software.
- Delta-betas are also saved in the **residual2** dataset. Note that variable names must be given to the delta-betas for the terms listed to the right of the equal sign in the **model** statement.

6.2.4 Handling Outliers

If a subject appears to be an outlier, there are several steps that should be taken.

1. Verify that the data were collected and entered correctly for the subject in question.
2. Examine the covariate values for the subject. If the covariate pattern falls within the population to which the results will be generalized, then the subject is often included in the analysis. On the other hand, if there is no interest in generalizing the results to individuals with similar covariate patterns, then the subject is often excluded.
3. Assess the influence of this subject on the parameter estimates. If an influential outlier is to be retained in the analysis, modifications to the model may be needed.

6.3 Test for Proportional Hazards

A key assumption in the proposed Cox model is that of proportional hazards. The assumption can be seen from the equation for the hazard ratio

$$\frac{\lambda(t;\mathbf{x}')}{\lambda(t;\mathbf{x}'')} = \exp\left\{\beta_1(x_1' - x_2'') + \ldots + \beta_p(x_p' - x_p'')\right\}.$$

For instance, a unit increase in the first covariate is associated with a hazard ratio of

$$\frac{\lambda(t;x_1+1)}{\lambda(t;x_1)} = \exp\{\beta_1(x_1+1-x_1)\} = \exp\{\beta_1\}$$

which is constant across time. We may want to test this assumption.

A formal test of proportionality, with respect to one of the covariates, can be performed using the Cox regression model. Consider the following model

$$\lambda(t;x) = \lambda_0(t)\exp\{\beta_1 x_1 + \ldots + \beta_p x_p + \gamma x_i g(t)\}$$

where x_i is one of the *p* covariates to test for non-proportional hazards, $g(t)$ is a specified function of time, and γ is the estimated effect of the interaction between the covariate and time. Common choices of *g* are

$$g(t) = \log(t)$$

and

$$g(t) = t.$$

Under the proposed model, the hazard ratio for a unit increase in x_i is

$$\frac{\lambda(t;x_i+1)}{\lambda(t;x_i)} = \exp\{\beta_i + \gamma g(t)\}$$

which is constant across time if $\gamma = 0$. Thus, a formal test of proportionality can be carried out by fitting this Cox model and testing if the γ parameter is statistically significant.

Lymphoma Example

The Lymphoma Study has several covariates that could be tested for proportional hazards. We will test the proportional hazards assumption for the Karnofsky score variable. The following models can be used to illustrate two different approaches for testing the proportional hazards assumption:

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{l}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70 \\ +\gamma karnofsky * \log(t)\end{array}\right\}$$

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{l}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70 \\ +\gamma_1 karnofsky * t2 + \gamma_2 karnofsky * t3\end{array}\right\}$$

where

$$t2 = \begin{cases} 1 & 72 < t \le 80 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad t3 = \begin{cases} 1 & t > 80 \\ 0 & \text{otherwise} \end{cases}.$$

The resulting parameter estimates for the Karnofsky-time interaction variables are

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
*log(t)	$\hat{\gamma}$	-0.0052	0.0104	0.2467	0.6194
*t2	$\hat{\gamma}_1$	-0.0203	0.0463	0.1925	0.6608
*t3	$\hat{\gamma}_2$	-0.0848	0.0381	4.9614	0.0259

The interaction with the continuous time variable is non-significant (p=0.6194). However, the categorical time-interaction variables suggest that the hazard rate for Karnofsky scores may not be constant across all time. This can be seen by calculating the hazard ratio for a unit increase in Karnofsky scores within the different time-intervals defined by the categorical variables *t2* and *t3*.

Time Interval	Hazard Ratio
$0 < t \le 72$	$\dfrac{\lambda(t; karnofsky + 1, t2 = 0, t3 = 0)}{\lambda(t; karnofsky, t2 = 0, t3 = 0)}$ $= \exp\{\beta_4\} = 0.963$
$72 < t \le 80$	$\dfrac{\lambda(t; karnofsky + 1, t2 = 1, t3 = 0)}{\lambda(t; karnofsky, t2 = 1, t3 = 0)}$ $= \exp\{\beta_4 + \gamma_1\} = 0.944$

Time Interval	Hazard Ratio
$t > 80$	$\frac{\lambda(t; karnofsky + 1, t2 = 0, t3 = 1)}{\lambda(t; karnofsky, t2 = 0, t3 = 1)}$ $= \exp\{\beta_4 + \gamma_2\} = 0.885$

The model estimates indicate that there is no difference between the hazard ratios for the first two time intervals (p = 0.6608). However, the hazard ratios do differ significantly between the first and third time intervals (p = 0.0259); thus, providing evidence of non-proportional hazards for the Karnofsky scores.

SAS Code

```
proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl karnofsky wait70
         tkarnofsky;
   tkarnofsky = karnofsky * log(days);

proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl karnofsky wait70
         knfsky2 knfsky3;
   knfsky2 = karnofsky * (72 < days <= 80);
   knfsky3 = karnofsky * (days > 80);
run;
```

6.4 Model Fit

Several authors have proposed methods for computing an R^2 statistic for Cox regression. One method due to Nagelkerke (1991) defines the R^2 statistic as

$$R^2 = 1 - \exp\left\{-\frac{2}{n}\left(\ln L\left(\hat{\beta}\right) - \ln L(0)\right)\right\}$$

where $\ln L\left(\hat{\beta}\right)$ and $\ln L(0)$ denote the likelihoods for the Cox regression models with and without the covariates, respectively. The R^2 given by this definition has the following properties:

1. It has the same interpretation as the R^2 in linear regression. Specifically, it measures the proportion of variation explained by the model, or rather, 1 - R^2 is the proportion of unexplained variation.
2. For a given model, it achieves the largest value at the maximum likelihood estimates.
3. It is independent of the sample size *n*.
4. It is independent of the units used for the response and predictor variables.

Lymphoma Example

For the regression model

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{matrix}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70\end{matrix}\right\}$$

the likelihood estimates from SAS are

$$-2\ln L\left(\hat{\beta}\right) = 141.197$$
$$-2\ln L\left(0\right) = 174.595 \quad .$$

Based on these estimates and the sample size of n = 43, the value of the R^2 statistics is

$$\begin{aligned} R^2 &= 1-\exp\left\{-\frac{2}{n}\left(\ln L\left(\hat{\beta}\right)-\ln L\left(0\right)\right)\right\} \\ &= 1-\exp\left\{\frac{1}{n}\left(-2\ln L\left(\hat{\beta}\right)-\left(-2\ln L\left(\hat{\beta}\right)\right)\right)\right\}. \\ &= 1-\exp\left\{\frac{1}{43}\left(141.197-174.595\right)\right\} \\ &= 0.540 \end{aligned}$$

Thus, 54% of the variation in the time-to-recurrence variable is explained by the covariates for disease type, bone marrow transplant method, Karnofsky score, and waiting time.

SAS Code and Output

```
proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl karnofsky wait70;
run;
```

```
         Model Fit Statistics

                  Without          With
Criterion      Covariates     Covariates

-2 LOG L          174.595        141.197
AIC               174.595        151.197
SBC               174.595        157.487
```

Applied Survival Analysis (171:242)

Section 7: Comparing Cox Regression Models

Brian J. Smith, Ph.D.

March 2, 2005

Table of Contents

7.1 Introduction

There are often several potential models that may be constructed from the predictor variables in a given dataset. In this section three methods for comparing such models are introduced; namely,

1. Likelihood Ratio Test
2. Wald Test
3. Akiake Information Criterion (AIC)

Lymphoma Example

Suppose that we are interested in determining the "best" way to characterize the effect of Karnosfsky scores on the time to cancer recurrence. Specifically, we will use Cox regression to choose among the following variables:

Variable	Description	Values
karnofsky	Karnofsky scores	20 – 100
karn	Categorical variable for Karnofsky scores	1 = (karnofsky < 35) 2 = (35 $\leq$ karnofsky < 65) 3 = (karnofsky $\geq$ 65)
karn1 karn2 karn3	Indicator variables for the three categories	I(karn = 1) I(karn = 2) I(karn = 3)

Four different Cox regression models will be considered. The linear predictors for the models are summarized below.

Model	Linear Predictor	Effect
1	$\beta_1 karnofsky$	Continuous (Linear)
2	$\beta_1 karnofsky + \beta_2 karnofsky^2$	Continuous (Quadratic)
3	$\beta_1 karn$	Categorical (Integer)
4	$\beta_1 karn2 + \beta_2 karn3$	Categorical (Nominal)

7.2 Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is the recommended method for comparing the fit of two regression models when one is "nested" within the other. A model whose predictor variables are a subset of another model is said to be "nested". For instance, the second model below is nested within the first.

1. $\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \ldots + \beta_p x_p + \ldots + \beta_{p+k} x_{p+k}\right\}$
2. $\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\beta_1 x_1 + \ldots + \beta_p x_p\right\}$

The first model is referred to as the "full" model and the second as the "reduced" model. The LRT is carried out as follows:

1. Fit the "full" model with $p + k$ predictor variables.
2. Fit the "reduced" model with p predictors.
3. Calculate the difference in the log-likelihood functions
$$X^2 = -2\left(\ln L_{\text{reduced}} - \ln L_{\text{full}}\right) \sim \chi^2_k .$$
4. If the difference, as measured by the p-value

$$p = \Pr\left[\chi_k^2 \geq X^2\right],$$

is significant then conclude that full model provides a better fit to the data than the reduced model. In other words, the *k* predictor variables are significant in the model.

Lymphoma Example

The values of the log-likelihood functions for the four models, plus the model no covariates (Model 0), can be obtained from PROC PHREG.

Model	Parameters	Log-Likelihood
0	null	-87.30
1	$\beta_1 karnofsky$	-75.81
2	$\beta_1 karnofsky + \beta_2 karnofsky^2$	-75.78
3	$\beta_1 karn$	-79.31
4	$\beta_1 karn2 + \beta_2 karn3$	-79.26

Note that Model 1 is nested within Model 2, since the latter simply adds a quadratic effect. In other words, Model 2 contains all of the terms found in Model 1; i.e. the linear effect for Karnofsky scores. The LRT is equivalent to a test of

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

and yields a test statistic value of

$$X^2 = -2\left(-75.81-\left(-75.78\right)\right)$$
$$= 0.06 \sim \chi_1^2$$

for which $p = \Pr\left[\chi_1^2 \geq 0.06\right] = 0.8065$. Therefore, at the 5% level of significance, the full model does not provide a better fit than the reduced model; the quadratic effect is not significant.

It may not be obvious that Model 3 is nested within Model 4, but that is the case.

- When indicator variables are used to model the effect of a categorical variable, no assumption is made about the function form of the relationship (linear, quadratic, etc.) between the categorical levels and the outcome.
- Thus, Model 4 allows for the most general relationship between the three categorical Karnofsky score levels and time to recurrence.
- Model 3 is nested because it can be expressed as a special case of Model 4,

$$\begin{aligned}\text{Model 3} &= \lambda_0(t)\exp\{\beta karn\} \\ &= \lambda_0(t)\exp\{\beta(1+karn2+2*karn3)\} \\ &= \lambda_0(t)\exp\{\beta\}\exp\{\beta karn2+2\beta karn3\} \\ &= \lambda_0'(t)\exp\{\beta karn2+2\beta karn3\}\end{aligned}.$$

  In other words, for any β, values of β_1 and β_2 can be found so that Model 4 equals Model 3; namely, $\beta_1 = \beta$ and $\beta_2 = 2\beta$. However, it is not true that a value of β can be found so that Model 3 equals Model 4 for any

β_1 and β_2. Thus, Model 3 is nested within Model 4, but the opposite is not true.

The LRT statistic comparing Model 3 to Model 4 is

$$X^2 = -2\left(-79.31-\left(-79.26\right)\right)$$
$$= 0.10 \sim \chi_1^2$$

for which $p = \Pr\left[\chi_1^2 \geq 0.10\right] = 0.7518$. Therefore, at the 5% level of significance, a linear effect for the categorical Karnofsky variable provides an adequate fit to the data.

Model 0 does not include an effect for Karnofsky scores. It is nested within the other four and may be used to test the significance of the Karnofsky variables in each model.

Model	Log-Likelihood	LRT		
		Statistic	df	p-value
0	-87.30	0	-	-
1	-75.81	23.0	1	p < 0.0001
2	-75.78	22.2	2	p < 0.0001
3	-79.31	16.0	1	p < 0.0001
4	-79.26	16.1	2	p = 0.0003

We see that the Karnofsky score variables are significant in all of the models.

- Each null hypothesis is a global test of the Karnofsky variables in the model.

Model	H_0	H_A
1	$\beta_1 = 0$	$\beta_1 \neq 0$
2	$\beta_1 = 0, \beta_2 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$
3	$\beta_1 = 0$	$\beta_1 \neq 0$
4	$\beta_1 = 0, \beta_2 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$

Comments

- The LRT is the most appropriate method for comparing nested models.
- This method requires fitting both the full and reduced model.
- The LRT cannot be used to compare Models 1 and 2 to Models 3 and 4, since they are not nested. Neither the categorical nor the continuous variable can be written as a linear combination of the other.

7.3 Wald Test

The Wald test can also be used to compare nested models. Specifically, the test may be used to assess the significance of terms in a given model. We have already used the Wald test for the hypotheses

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

where β is a parameter in the regression model. The test statistic is

$$X^2 = \left(\frac{\hat{\beta}}{\text{se}\left(\hat{\beta}\right)} \right)^2 \sim \chi_1^2$$

for which $p = \Pr\left[\chi_1^2 \geq X^2\right]$. So far, we have only used the Wald test for a single parameter. The test has a general form that allows for the simultaneous testing of multiple parameters.

Lymphoma Example

The following results were obtained for Models 1 and 2:

Model	Term	Parameter Estimate	SE	Wald	
				Chi-Square	p-value
1	karnofsky	-0.0524	0.0110	22.7301	<.0001
2	karnofsky	-0.0386	0.0612	0.3974	0.5284
	karnofsky^2	-0.0001	0.0005	0.0529	0.8181

The comparison of Models 1 and 2 is equivalent to a test of the hypotheses

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0.$$

The Wald statistic for this test is

$$X^2 = \left(\frac{\hat{\beta}_2}{\text{se}\left(\hat{\beta}_2\right)}\right)^2 = \left(\frac{-0.0001}{0.0005}\right)^2 = 0.04$$

for which $p = \Pr\left[\chi_1^2 \geq 0.04\right] = 0.8415$. The quadratic term is not significantly different from zero. Therefore, Model 2 is not significantly different from Model 1.

The Wald test could also be used to test the joint significance of the linear and quadratic terms in Model 2, for which the hypotheses would be

$$H_0 : \beta_1 = 0, \beta_2 = 0$$
$$H_A : \beta_1 \neq 0 \text{ or } \beta \neq 0.$$

The test statistic and p-value are

$$X^2 = 22.1$$
$$p = \Pr\left[\chi^2_2 > 22.1\right] < 0.0001.$$

We will rely on SAS to jointly test the significance of parameters with the Wald statistic.

SAS Code and Output

```
proc phreg data=lymphomamod;
   model days*event(0) = karnofsky karnofsky2;
   karnofsky2 = karnofsky**2;
   Test1: test karnofsky, karnofsky2;
run;
```

```
The PHREG Procedure

    Linear Hypotheses Testing Results

                Wald
 Label     Chi-Square       DF     Pr > ChiSq

 Test1        22.1021        2         <.0001
```

Comments

- The Likelihood Ratio and Wald test are alternative methods of comparing nested models.
- The LRT is preferred because the test statistic has better distributional properties.
- The Wald test statistic is often easier to compute since a second, reduced model need not be fit. Plus, Wald p-values and confidence intervals give equivalent inferential results. Thus, the Wald test is more commonly used in practice

7.4 Akiake Information Criterion (AIC)

Neither the Likelihood Ratio test nor the Wald test can be used to compare models that are not nested. There are several methods to handle this problem. We will discuss one, the Akaike Information Criterion (AIC).

Akaike (1972) proposed a method of comparison based on both the log-likelihood and the number of parameters in the model. The AIC is defined as

$$AIC = -2\ln L + 2p$$

where p is the number of parameters in the model. Based on this criterion the preferred model is the one with the lowest AIC.

Lymphoma Example

Suppose that we want to compare the model with a linear effect for Karnofsky scores (Model 1) to the model with a categorical effect (Model 4).

Model	Parameters	Log-Likelihood
1	$\beta_1 karnofsky$	-75.81
4	$\beta_1 karn2 + \beta_2 karn2$	-79.26

The AIC for each model is

Model	$-2\ln L$	$2p$	AIC
1	151.62	2	153.62
4	158.52	4	162.52

Model 1 has the smaller AIC and would be preferred based on this criterion.

Comments

- The AIC is a method for choosing among competing models. It is does not provide a test for detecting statistically significant differences.
- The Likelihood Ratio or Wald tests should be used to compare nested models.
- The AIC may be used to compare models that are not nested. It is often referred to as a goodness-of-fit statistic.

Applied Survival Analysis (171:242)

Section 8: Cox Regression Variable Selection

Brian J. Smith, Ph.D.

March 9, 2005

Table of Contents

8.1 Introduction

There are three main categories of variables to consider for inclusion in a regression model:

1. Predictors – variables for which risk estimates are desired.
2. Confounders – variables that are confounded with the predictors.
3. Effect Modifiers – variables that interact or modify the effect of the predictors.

Goal: Select the set of covariates that results in the "best" model within the scientific context of the problem.

In our approach, we will try to strike a balance between the following two objectives:

1. Traditional – Seek the most parsimonious model that "explains" the data.
    - Smaller models are more likely to be numerically stable. The standard errors for the parameter estimates tend to increase as additional variables are added to the model.
    - The dependence of the model on the data set increases with the number of variables. Consequently, large models are less generalizable.
    - Parsimonious models are easier to interpret.

2. Biological – Include all scientifically relevant variables in the model.
    - We want to ensure that confounding and interaction are accounted for in the model; e.g. covariates may not show confounding individually, but do so when analyzed together.

Advise: Beware of overfitting, especially when there are a large number of covariates relative to the number of cases and controls. Also, think about the interpretation of the variables in the models that you are fitting.

Breast-Feeding Study

Suppose that we would like to select among the variables in the Breast-Feeding Study to build a final Cox model for breast-feeding cessation. The variables are summarized below.

Variable	Description	Values
weaned	Indicator for breast-feeding cessation	1 = yes 0 = no
weeks	Length of follow-up in weeks	continuous
age	Subject age	continuous
alcohol	Alcohol use at time of birth	1 = yes 0 = no
care3	Use of prenatal care after first trimester	1 = yes 0 = no
education	Years of education	continuous
poverty	Below poverty level	1 = yes 0 = no
race	Subject race	1 = white 2 = black 3 = other

Variable	Description	Values
smoke	Smoking at time of birth	1 = yes 0 = no

8.2 Model Building

We will take the following steps to build a final regression model:

Step 1. Descriptive summaries of the data

Step 2. Univariate analyses

Step 3. Variable selection

Step 4. Model Diagnostics

If problems with the model fit are identified in Step 4, then return to the variable selection in Step 3 and repeat until the model diagnostics are satisfactory.

8.2.1 Descriptive Statistics

Tables, such as those given below for the Breast-Feeding Study, should be given to summarize the variables available for the analyses, even if they are not all included in the final regression model.

Table 1. Summary of the categorical variables in the Breast-Feeding Study.

Variable	Levels	N	Percents
weaned	0 1	35 892	3.8% 96.2%
alcohol	0 1	848 79	91.5% 8.5%

Variable	Levels	N	Percents
care3	0	763	82.3%
	1	164	17.7%
poverty	0	756	81.6%
	1	171	18.4%
race	1	662	71.4%
	2	117	12.6%
	3	148	16.0%
smoke	0	657	70.9%
	1	270	29.1%

Table 2. Summary of the continuous variables in the Breast-Feeding Study.

Variable	Mean	SD	Min	Max
weeks	16.18	17.92	1	192
age	21.54	2.67	15	28
education	12.21	1.93	3	19

In addition, a write-up of the analyses should include a description of the variables, how they were measured, and the range of possible values for each.

8.2.2 Univariate Analyses

The goal in model building is to identify a set of variables that offers a satisfactory explanation of the outcome in the study population.

- Our final model should be scientifically valid; that is, there should be a biologically plausible explanation for the effect of our chosen variables on the outcome.

- We begin with a pool of variables that will be considered for inclusion in the final model. Any of the variables in this pool could end up in the model.
- Therefore, it is at the beginning, before any statistical tests are performed, that we should narrow our pool to only those variables for which an association with the outcome makes sense.
- Another way to frame this problem is to ask, “How will the effect of variable *X* be explained if it is included in the model?”

Once a pool of scientifically relevant variables has been identified, it is often helpful to further narrow the pool by examining the effect of each variable individually in a univariate Cox regression model.

Categorical Covariates

- Construct Kaplan-Meier plots to graphically display the survival functions across levels of the covariate. Use Kalbfleish and Prentice method for confidence intervals.
- Report the estimated median survival times and confidence intervals
- Create indicator variables and fit univariate Cox models to assess significance and estimate hazard ratios.

Continuous Covariates

- Assess significance with univariate Cox models; estimate hazard ratios and confidence intervals.

- Test for nonlinearity in the effect of the covariate on survival:
    - Divide the covariate into three or more categories.
    - Inspect plots of the Martingale residuals versus the covariate.
    - Try adding a quadratic or other non-linear terms to the Cox regression model.

Present a table with the Cox results, including estimates and p-values. If there is evidence of a nonlinear effect for any continuous covariate, report that as well.

Table 3. Tests of the main effect for each variable; based on separate univariate Cox regression models.

Variable	df	HR	p-value
age	1	0.99	0.632
age1 age2 age3	2	1.00 0.94 0.95	0.723
ns(age)	3	*	0.146
alcohol	1	1.18	0.178
care3	1	1.04	0.691
education	1	0.96	0.009
ns(education)	2	*	0.025
poverty	1	0.93	0.357
race1 race2 race3	2	1.00 1.12 1.29	0.022
smoke	1	1.25	0.002

* Estimates are a non-linear function of the continuous variable.

Table 4. Tests of interaction for select variables; based on separate Cox regression models.

Variable	df	p-value
alcohol*smoke	1	0.900
education*poverty	1	0.041
education*race	2	0.415
* Each model included main effects for the terms in the interaction.		

Univariate Analysis of Age

Three models for age were considered in the univariate analysis:

$$\text{Model 1: } \lambda(t;x) = \lambda(t)\exp\{\beta_1 age\}$$

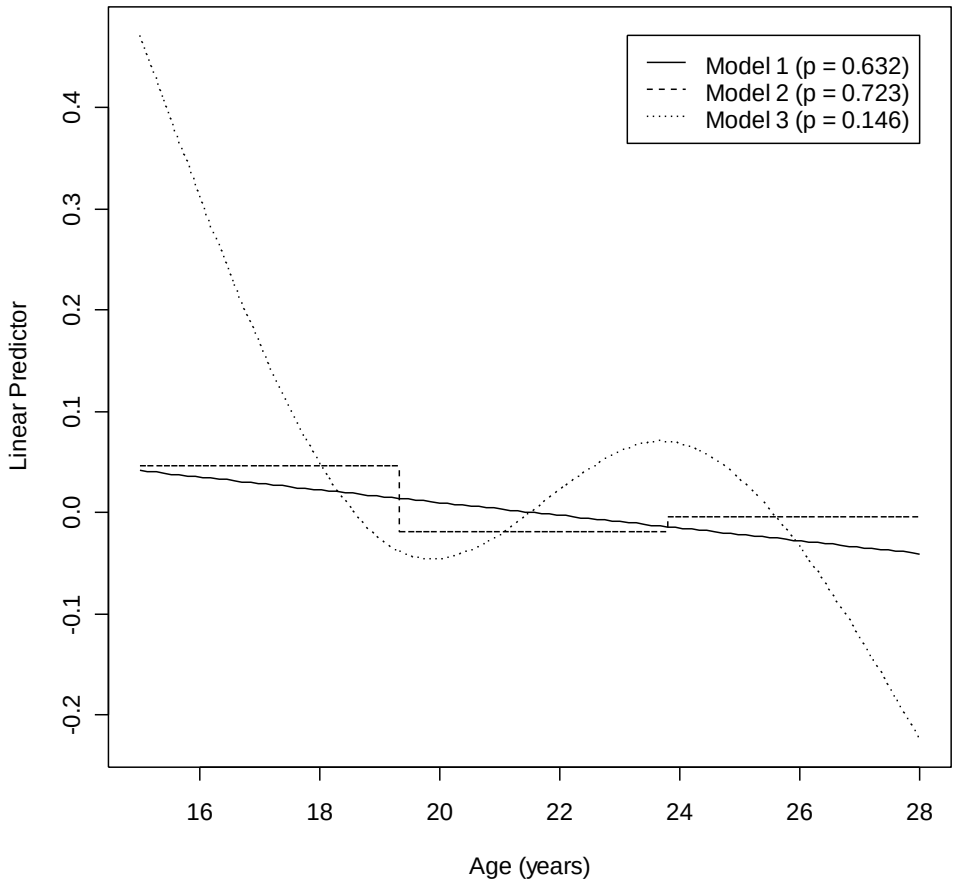
$$\text{Model 2: } \lambda(t;x) = \lambda(t)\exp\{\beta_1 age2 + \beta_2 age3\}$$

$$\text{Model 3: } \lambda(t;x) = \lambda(t)\exp\left\{\begin{array}{c}\beta_1 ns_1(age) + \beta_2 ns_2(age) \\ +\beta_3 ns_3(age)\end{array}\right\}$$

where *age1* (referent), *age2*, and *age3* are indicator variables for three equally spaced age intervals; namely,

$$age1 = I(age \le 19.3)$$
$$age2 = I(19.3 < age \le 23.7)$$
$$age3 = I(age > 23.7)$$

and $ns_i(age)$ is a non-linear transformation of age. The estimated linear predictors for these three models are displayed in the figure below.



The three univariate models for age are summarized below.

Model 1. A linear effect for age is included in the model. This assumes the same hazard ratio for any one-year

increase in age. The linear effect of age on the log-hazard rate is not significant (p = 0.632).

Model 2. Age is modeled as a nominal categorical variable. The hazard rate is allowed to vary between the three predefined age intervals. However, the hazard rate is assumed constant within each interval. In other words, the hazard ratio is independent of age for individuals within the same interval. Categorical variables are sometimes useful for uncovering non-linear effects of a continuous variable. Non-significant results are obtained in this example (p = 0.723).

Model 3. A non-linear function of age is included in the model using natural smoothing splines. The approach used assumes no particular function form for the effect of age; the shape of the effect is instead estimated from the data. The estimates suggest that the hazard rate decreases for the youngest and oldest subjects. Overall the non-linear effect is more significant than the linear or categorical effects (p = 0.146).

8.2.3 Variable Selection

A few possible variable selection strategies are:

1. <u>All possible regressions</u> Only possible when there are very few covariates. Given *p* covariates, there are 2^p possible models. A very good method when possible.

2. Best subset regressions Derive best one-covariate model, best two-covariate model, best 3-covariate model, etc. until it is too much work to continue. p models are fit to find the best one-covariate model; $\binom{p}{2}$ to get the best two-covariate model; $\binom{p}{3}$ to get the best three-covariate model; etc. Can be very time consuming. Which model is "best"?

3. Forward variable selection

   - Variables are *added* to the model one-at-a-time, provided that their p-value is *smaller* than some prespecified cutoff.
   - The variable with the smallest univariate p-value is the first to be added.
   - At each step, the remaining variable with the smallest p-value is added to the model.
   - This process iterates until all of the p-values for the remaining variables are greater than the prespecified cutoff.

4. Backward variable selection

   - Variables are *removed* from the model one-at-a-time, provided that their p-value is *larger* than some prespecified cutoff.
   - An initial model is fit with all of the variables.
   - At each step, the variable in the model with the largest p-value is removed.

- This process iterates until all of the p-values for the variables in the model are less than the prespecified cutoff.

5. Stepwise variable selection
   - Starts like forward selection.
   - At each subsequent step, variables may either enter or leave the model.
   - p-value cutoffs for variable entry into the model and variable removal from the model must be specified.
   - Common choices of p-value cutoffs are 0.20, 0.15, 0.10, and 0.05. A larger value for the cutoff to enter or the cutoff to remove will result in more variables in the model. The same cutoff is typically used for both.

6. Important Variables method Upon completion of the univariate analyses, consider as a candidate for the multivariate model any covariate whose univariate p-value is sufficiently small (e.g. < 0.25) or which is believed to be biologically important. If the number of candidates is small, one might try the "all possible regressions" approach or the "best subset" approach. If these are not feasible, one might try a stepwise selection method.

7. Hybrid method This method is really just a combination of several of the above methods, including forcing in a few important covariates and doing stepwise selection on the remaining covariates;

performing stepwise selection among those identified in the "important variables method"; and "all possible regressions" on a very small subset of important variables.

Breast-Feeding Example

We will use a hybrid of variable selection techniques to build a final model.

Goal: Estimate the effect of race on time to breast-feeding cessation while controlling for other important covariates.

Strategy: We will start with the model

$$\lambda(t;x) = \lambda(t)\exp\{\beta_1 race2 + \beta_2 race3\}$$

and use stepwise variable selection among the remaining variables with the following criteria:

- Inclusion or exclusion of variables will be based on the AIC, as opposed to p-values.
- *race2* and *race3* will be included in all models.
- The interaction between education and poverty will be considered. If the interaction term is included in the model then so to must the associated main effects.

Step 1

The first step starts with a model containing only the indicator variables for race. The other variables are added individually to inspect their impact on the AIC. Addition of the smoking variable leads to the smallest AIC. Thus, smoking is added to the model in this step.

Model	df	AIC
race2 + race3	-	*10378.6*
+ ns(age)	3	10379.2
+ alcohol	1	10378.4
+ care3	1	10380.6
+ education	1	10375.9
+ poverty	1	10379.0
+ smoke	**1**	**10367.4**
+ education + poverty + education*poverty	3	10371.4

Step 2

In the second step we again look at the effect on the AIC of adding each remaining variable individually to the model. In addition, the AIC is calculated for the removal of terms already in the model. Race is forced into the model, so we do not consider the effect of removing this covariate. Addition of the education-poverty interaction and main effects leads to the smallest AIC.

Model	df	AIC
race2 + race3 + smoke	-	*10367.4*
+ ns(age)	3	10368.1
+ alcohol	1	10368.4
+ care3	1	10369.4
+ education	1	10367.7
+ poverty	1	10366.2
- smoke	1	10378.6
+ education + poverty + education*poverty	**3**	**10362.7**

Step 3

The AIC is again calculated for the addition or removal of variables. Note that consideration is given to the removal of the interaction term, but not the associated main effects. Addition of age results in the smallest AIC.

Model	df	AIC
*race2 + race3 + education + poverty + smoke + education*poverty*	-	*10362.7*
+ ns(age)	**3**	**10360.9**
+ alcohol	1	10362.7
+ care3	1	10364.6
- smoke	1	10371.4
- education*poverty	1	10364.4

Step 4

The addition of alcohol reduces the AIC and is selected for inclusion to the model.

Model	df	AIC
*race2 + race3 + ns(age) + education + poverty + smoke + education*poverty*	-	*10360.9*
- ns(age)	3	10362.7
+ alcohol	**1**	**10360.7**
+ care3	1	10362.8
- smoke	1	10369.3
- education*poverty	1	10364.4

Step 5

No further changes to the model will reduce the AIC.

Model	df	AIC
*race2 + race3 + ns(age) + alcohol + education + poverty + smoke + education*poverty*	-	*10360.7*
- ns(age)	3	10362.7
- alcohol	1	10360.9
+ care3	1	10362.5
- smoke	1	10367.6
- education*poverty	1	10364.5

Therefore, we stop the stepwise selection process and propose a final model of

$$\lambda(t;x) = \lambda(t)\exp\left\{\begin{array}{l}\beta_1 race2 + \beta_2 race3 + \beta_3 ns_1(age) \\ \quad + \beta_4 ns_2(age) + \beta_5 ns_3(age) + \beta_6 alcohol \\ \quad + \beta_7 education + \beta_8 poverty + \beta_9 smoke \\ \quad + \beta_{10} education * poverty\end{array}\right\}$$

The parameter estimates for this model are

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
race2	$\hat{\beta}_1$	0.161	0.105	2.332	0.130
race3	$\hat{\beta}_2$	0.315	0.097	10.563	0.001

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
ns_1(age)	$\hat{\beta}_3$	0.092	0.176	0.274	0.600
ns_2(age)	$\hat{\beta}_4$	-0.925	0.534	3.000	0.083
ns_3(age)	$\hat{\beta}_5$	-0.040	0.249	0.025	0.870
alcohol	$\hat{\beta}_6$	0.185	0.123	2.253	0.130
education	$\hat{\beta}_7$	0.240	0.079	9.175	0.003
poverty	$\hat{\beta}_8$	-0.074	0.025	8.791	0.003
smoke	$\hat{\beta}_9$	-1.439	0.517	7.767	0.005
education *poverty	$\hat{\beta}_{10}$	0.109	0.045	5.910	0.015

Confounding

Confounding is the bias in a risk estimate that can result when the predictor-response relationship of interest is partially or wholly explained by the effects of an extraneous variable.

Suppose that Cox regression is used to estimate the effect of a predictor variable *X*.

- A confounder is any variable associated with *X* as well as with the outcome of interest.
- A variable is a confounder if and only if its inclusion in the model changes the estimated effect of *X*. The result could be to increase or decrease the estimate for *X*.
- Any confounding variable that has an appreciable impact on the effect of *X* should be considered for

inclusion, even if the confounder itself is not statistically significant in the model.

- The confounder should be properly controlled for in the regression model. This involves:
  1. The identification of potential confounders at the study design phase.
  2. The measurement of detailed and complete information on the confounders during the data collection phase.
  3. The inclusion of important confounding variables in the regression model during the data analysis phase.

One way to monitor for confounding during variable selection is to look for important changes in the parameter estimates for the predictors of interest.

Step	race2		race3	
	Estimate	SE	Estimate	SE
1	0.111	0.102	0.254	0.093
2	0.153	0.103	0.320	0.094
3	0.158	0.015	0.307	0.097
4	0.165	0.106	0.315	0.176
5	0.161	0.105	0.315	0.097

This approach is more effective when backward variable selection is used because changes can be assessed relative to the starting model, which contains all covariates.

Notes

1. In any variable selection approach, one should also consider interaction terms of interest and nonlinear terms (e.g. Age^2). If there is a treatment variable of interest, interactions between treatment and the other variables should be checked.

2. A very subtle problem can arise in model building. Suppose that you have a covariate in the model with a non-significant p-value. You decide to drop it from the model, but in doing so the coefficients for the remaining covariates change substantially. It is possible that the non-significant covariate makes some necessary adjustment for another covariate in the model. So, be sure to check changes in the coefficients as well as the test statistic when dropping (or adding) a covariate.

3. Never use a variable in the disease pathway as a covariate. For example, suppose you have smoking status and cough as covariates for time to lung cancer. The disease pathway is

   $$\text{Smoking} \rightarrow \text{Cough} \rightarrow \text{Lung Cancer}$$

   In other words, smoking causes cough and cough is an early sign of lung disease. Adjusting for the effects of cough attenuates the effect of smoking in the

model, leading to an underestimate of smokers' risk of lung cancer.

4. Stepwise procedures can lead to a biologically implausible model and can select biologically irrelevant variables. The data analyst must carefully review each step of the selection process. Knowledge of the biology is invaluable in this review. Discussions with the investigator about the modeling process are important. The ease and availability of the stepwise procedures has reduced some analysts to the role of "assisting the computer in model selection, rather than the more appropriate reverse relationship." The analyst, not the computer, is responsible for the final model.

8.2.4 Model Diagnostics

There are other valid model building approaches that could lead to different final models. Note that we are not done. The next step is to perform model diagnostics in order to answer the questions:

- Are there outliers in the data set that need to be excluded from the analysis (Deviance Residual and Delta-Beta Plots)?
- Does the model fit the data (R^2)?

If subjects are excluded or problems are identified with the fit of the model at the diagnostic stage, the variable selection should be revisited.

Applied Survival Analysis (171:242)

Section 9: Cox Analysis of the Stanford Heart Transplant Study

Brian J. Smith, Ph.D.

March 30, 2005

Table of Contents

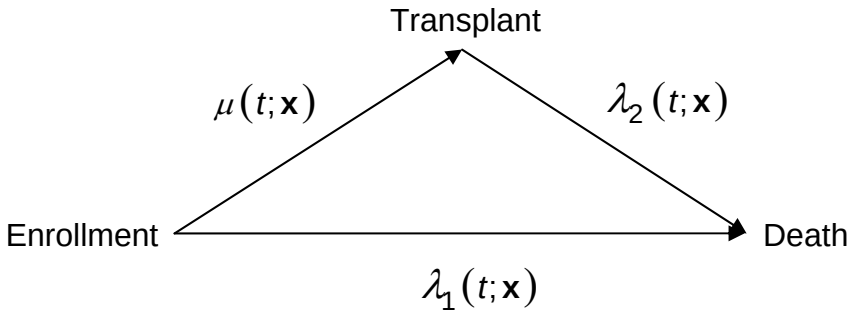
9.1 Introduction

The Stanford Heart Transplant Study is quite famous in the statistical literature and has been analyzed by many people. We will consider the approach of Crowley and Hu (JASA 1977).

9.1.1 Study Overview

Between 1967 and 1974 the study enrolled 103 patients that had been accepted for heart transplants. These patients then waited for a heart to become available.

- The availability of a suitable transplant heart depended on factors not under the control of the study investigators, such as donor-recipient tissue matching, and some patients died before a transplant was performed.
- The outcome of interest is time from enrollment to death. Patients are treated as censored observations if they 1) improved and no longer needed a transplant, 2) were lost to follow-up, or 3) were alive at the end of the study.
- All patients were followed until April 1, 1974.
- Cox regression was used to model the death rate. Note that one might expect the death rate to differ between patients who had and had not received a transplant. This potential difference in hazards is depicted graphically below.



where $\mathbf{x}$ is a set of covariates, $\mu(t;\mathbf{x})$ is the rate at which transplantation takes place, $\lambda_1(t;\mathbf{x})$ is the death rate for an untransplanted patient, and $\lambda_2(t;\mathbf{x})$ is the death rate for a transplanted patient.

9.1.2 Study Variables

The variables for the Stanford Study are summarized in the following table:

Variable	Description	Values
event	Death indicator variable	1 = yes 0 = no
start	Follow-up time (days) at start of risk interval	continuous
stop	Follow-up time (days) at end of risk interval	continuous
age	Age at enrollment in years (minus 48)	continuous
year	Year of enrollment (minus November 1, 1967)	continuous
prior	Bypass surgery prior to enrollment	1 = yes 0 = no

Variable	Description	Values
transplant	Received transplant	1 = yes 0 = no
id	Patient identifier	integer

The dataset is structured so that separate rows are included for the time periods that patients were pre and post-transplant. For example, a subset of the data for the first three patients is

id	start	stop	event	transplant
1	0	50	1	0
2	0	6	1	0
3	0	1	0	0
3	1	16	1	1
⋮	⋮	⋮	⋮	⋮

The data show that

1. The first subject did not receive a transplant and died at 50 days of follow-up.
2. Likewise, the second subject did not receive a transplant and died at six days of follow-up.
3. The third subject received a transplant after 1 day of follow-up and, subsequently, died at 16 days. Thus, there are two rows for this subject, the first contains data relevant to the pre-transplant follow-up period and the second contains data relevant to the post-transplant period.

9.2 Cox Regression Model

In the analysis, the effect of the covariates should be allowed to differ between patients who have and have not received a transplant. If we are interested in the effects of age at enrollment, year of enrollment, prior bypass surgery, and transplant; the two models below might be considered.

Pre-Transplant Death Rate

$$\lambda_1(t;\mathbf{x}) = \lambda_0(t)\exp\{\beta_{11}age + \beta_{12}year + \beta_{13}prior\}$$

Post-Transplant Death Rate

$$\lambda_2(t;\mathbf{x}) = \lambda_0(t)\exp\{\beta_{21}age + \beta_{22}year + \beta_{23}prior + \beta_4\}$$

where β_4 is the multiplicative effect of transplantation on the common baseline hazard. These two models can be written as a single model through the use of a time-dependent covariate. Let W denote the time from enrollment to transplant, and define the time-dependent covariate

$$x(t) = \begin{cases} 1 & \text{if } t > W \\ 0 & otherwise \end{cases}$$

indicating whether transplantation has occurred. Thus, the two models can be represented by

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 year + \beta_3 prior + \beta_4 x(t) \\ \quad + \beta_5 x(t) age + \beta_6 x(t) year \\ \quad + \beta_7 x(t) prior\end{array}\right\}.$$

Here, $\beta_4,\ldots,\beta_7$ measure the effect of transplantation on the risk of death.

9.3 Results

There are other covariates measuring tissue matching that we won't consider here. The following tables summarize a considerable number of regression models fit in an attempt to find a model that best describes the data. Every model that contains a covariate-transplant interaction term also contains both the main effect for the covariate as well as the transplant status.

	Main Effect Terms				***Interaction Terms***		
Model	***age***	***year***	***prior***	***x*(*t*)**	***x*(*t*)*age***	***x*(*t*)*year***	***x*(*t*)*prior***
Term	β_1	β_2	β_3	β_4	β_5	β_6	β_7
1	.0307* (.0143)						
2	.0119 (.0183)			.075 (.321)	.0413 (.0283)		
3		-.191* (.070)					
4		-.265* (.105)		-0.282 (0.514)		0.136 (0.141)	
5			-.739* (.359)				

	Main Effect Terms				***Interaction Terms***		
Model	***age***	***year***	***prior***	***x*(*t*)**	***x*(*t*)*age***	***x*(*t*)*year***	***x*(*t*)*prior***
Term	β_1	β_2	β_3	β_4	β_5	β_6	β_7
6			-.518 (.610)	.187 (.305)			-.337 (.757)
7	.027† (.014)	-.178* (.070)					
8	.0155 (.0173)	-.274* (.106)		-0.588 (0.543)	.0339 (.0279)	.201† (.143)	
9	.0139 (.0176)	-.166* (.071)		0.029 (.325)	.0303 (.0279)		
10	.0294* (.0141)	-.278* (.106)		-0.606 (0.540)		.186 (.142)	
11	.0307* (.0136)		-.771* (.360)				
12	.0138 (.0181)		-.546 (.611)	.118 (.328)	.0348 (.0273)		-.292 (.758)
13	.0141 (.0180)		-.743* (.361)	.089 (.318)	.0348 (.0273)		
14	.0304* (.0139)		-.576 (.610)	-.043 (.318)			-.293 (.757)
15		-.162* (.070)	-.597 (.366)				
16		-.254* (.108)	-.236 (.628)	-.292 (0.506)		.164 (.142)	-.550 (.776)
17		-.240* (.104)	-.621† (.367)	-.284 (.505)		.145 (.138)	
18		-.162* (.070)	-.358 (.615)	.179 (.308)			-.361 (.757)
19	.0270* (.0134)	-.146* (.070)	-.636† (.367)				
20	.0167 (.0173)	-.262* (.108)	-.258 (.629)	-.605 (.535)	.0304 (.0270)	.230 (.143)	-.557 (.777)

	Main Effect Terms				*Interaction Terms*		
Model	***age***	***year***	***prior***	***x*(*t*)**	***x*(*t*)*age***	***x*(*t*)*year***	***x*(*t*)*prior***
Term	β_1	β_2	β_3	β_4	β_5	β_6	β_7
21	.0169 (.0172)	-.248* (.105)	-.647† (.369)	-.598 (.534)	.0305 (.0271)	.210 (.140)	
22	.0150 (.0176)	-.136† (.071)	-.419 (.616)	.077 (.332)	.0270 (.0271)		-.298 (.758)
23	.0299* (.0137)	-.266* (.108)	-.274 (.629)	-.632 (.532)		.218 (.143)	-.556 (.777)
24	.0152 (.0175)	-.136† (.071)	-.621† (.368)	.048 (.322)	.0271 (.0271)		
25	.0299* (.0137)	-.252* (.105)	-.663† (.368)	-.622 (.531)		.197 (.139)	
26	.0280* (.0137)	-.146* (.071)	-.428 (.615)	-.018 (.323)			-.305 (.757)
27				.126 (.301)			
28	.0307* (.0145)			-.006 (.312)			
29		-.191* (.070)		.122 (.303)			
30			-.747* (.360)	.156 (.297)			

* $p < 0.05$
† $0.05 \leq p < 0.10$

9.3.1 Age at Enrollment

Age at enrollment was often significant as a main effect term but never as an interaction term. Apparently, the older the patient, the greater the risk of death; it was about the same regardless of transplant status.

9.3.2 Year of Enrollment

In every model that was tried, year of enrollment was significant or close to it. The sign of its coefficient is always negative whereas the sign of its time-dependent interaction coefficient is always positive. it appears that its effect on overall survival prior to transplantation is beneficial (negative coefficient, smaller hazard) whereas its effect on post-transplant survival is not that wonderful. For example, in Model 4:

$$\beta_2 = -0.265$$
$$\beta_2 + \beta_6 = -0.265 + 0.136 = -0.129$$

which suggests that the overall health of patients being accepted into the study was improving over calendar time but that the survival time of these patients was not increasing at the same rate. This brings up the issue of selection bias.

1. Did the general health of newly enrolled patients improve over time?
2. Did patient selection for transplant change as a function of general health over time?

9.3.3 Prior Surgery

Prior surgery was a significant factor in improving overall survival (negative coefficient). It had no interaction with

transplant status. Note that other interaction terms, such as age-prior, should still be considered before arriving at a final model.

9.3.4 Model Interpretation

Consider Model 25

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{c} 0.0299age - 0.252year - 0.663prior \\ -0.622x(t) + 0.197x(t)year \end{array}\right\}.$$

Age and Prior Bypass Surgery

To better interpret this model, suppose we wish to compare two subjects entering the study during the same year and having the same transplant status at any given time. Suppose, however, that they differ on prior surgery and age at enrollment. The hazard ratios are

age	*prior*	Hazard Ratio
x years	0	1.00
	1	0.51
x+10 years	0	1.35
	1	0.69

where x years and no prior surgery is the reference.

Year of Enrollment

Since year interacts with transplant status in the model, the estimated hazard ratio for year must be estimated separately. In the following calculations prior surgery and age at enrollment are held constant.

Pre-Transplant: The estimated hazard ratio for a 5 year difference in enrollment for patients that had not received a transplant is

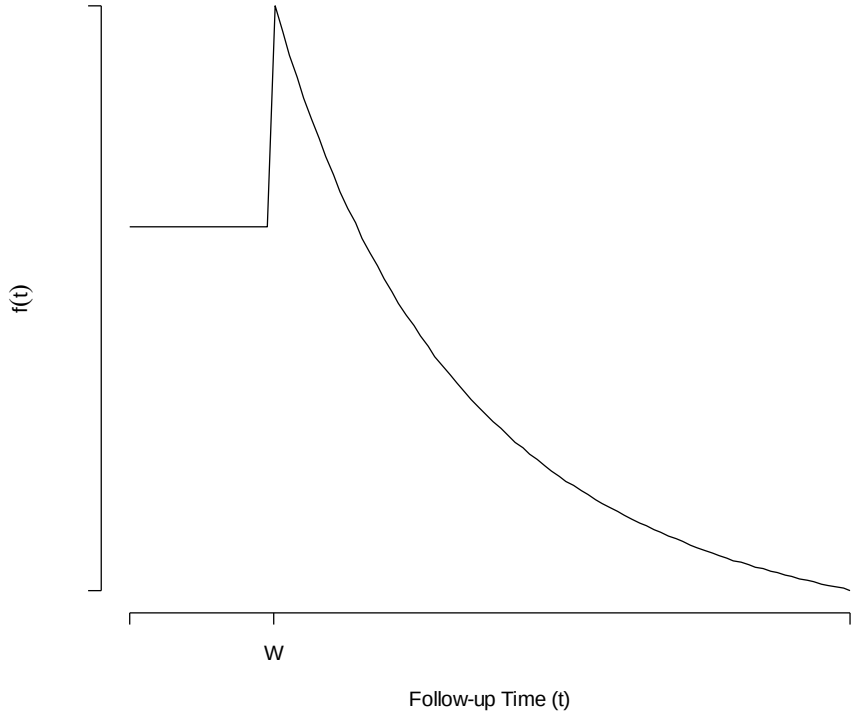
$$\frac{\lambda\left(t;year+5,x\left(t\right)=0\right)}{\lambda\left(t;year,x\left(t\right)=0\right)}=\exp\left\{-0.252\left(5\right)\right\}=0.28.$$

Post-Transplant: The estimated hazard ratio for a 5 year difference in enrollment for patients that had received a transplant is

$$\frac{\lambda\left(t;year+5,x\left(t\right)=1\right)}{\lambda\left(t;year,x\left(t\right)=1\right)}=\exp\left\{-0.252\left(5\right)+0.197\left(5\right)\right\}=0.76.$$

9.4 Alternative Effect of Transplantation

Since risk usually increases right after invasive procedures, like transplant surgery, before settling down to a long term level, one might prefer to model the hazard like the function shown in the figure below.



Algebraically, this hazard function behavior could be modeled in our example with the following regression equation:

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{c}\beta_1 age + \beta_2 year + \beta_3 prior + \beta_4 x(t) \\ +\beta_5 x(t)\exp\{-\beta_6(t-W)\}\end{array}\right\}.$$

Applied Survival Analysis (171:242)

Section 10: Parametric Regression for Survival Data

Brian J. Smith, Ph.D.

March 30, 2005

Table of Contents

10.1 Introduction

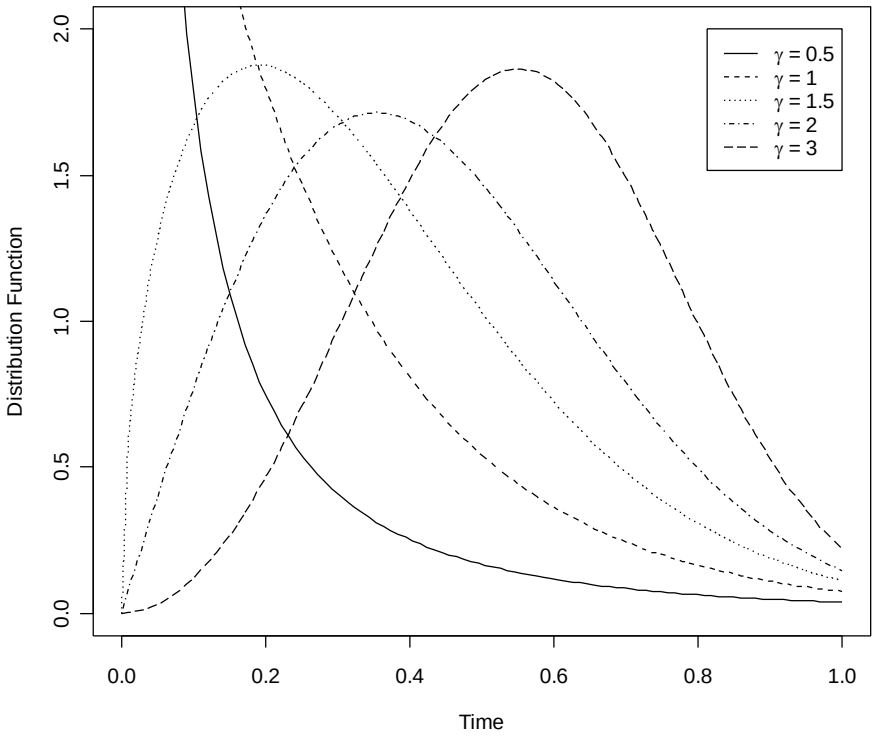
The Cox regression model is popular because it does not make any distributional assumptions about the survival times. Rather, nonparametric methods are used to estimate the baseline hazard function. This approach has the advantage of being robust to different distributions, but does not allow for direct estimation of the hazard rate and may be less powerful when it is appropriate to make assumptions about the distribution of survival times.

10.1.1 Weibull Distribution

The Weibull distribution is commonly used to model time-to-event data. This distribution takes on non-negative values and is defined by two parameters – a **scale** parameter α and a **shape** parameter γ. The functional form of the Weibull distribution is

$$f(t) = \alpha\gamma t^{\gamma-1} \exp\left\{-\alpha t^{\gamma}\right\}$$

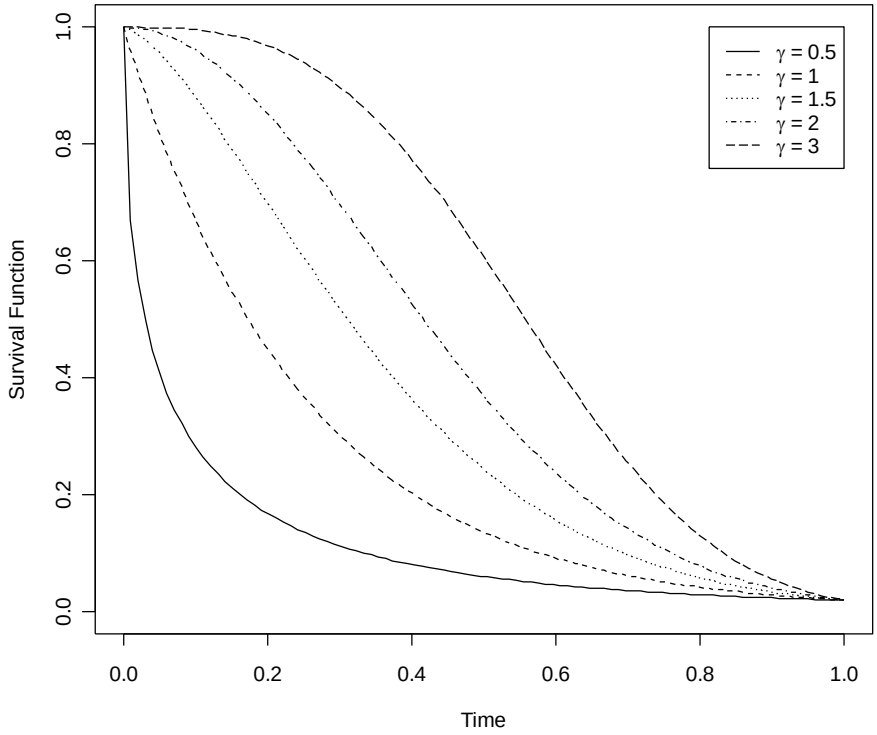
for $t \geq 0$, and constants α and γ which we will estimate from the data.



There is a one-to-one relationship between the probability distribution, hazard, and survival functions. The survival function corresponding to the specified Weibull distribution is

$$S(t) = \exp\left\{-\alpha t^{\gamma}\right\}.$$

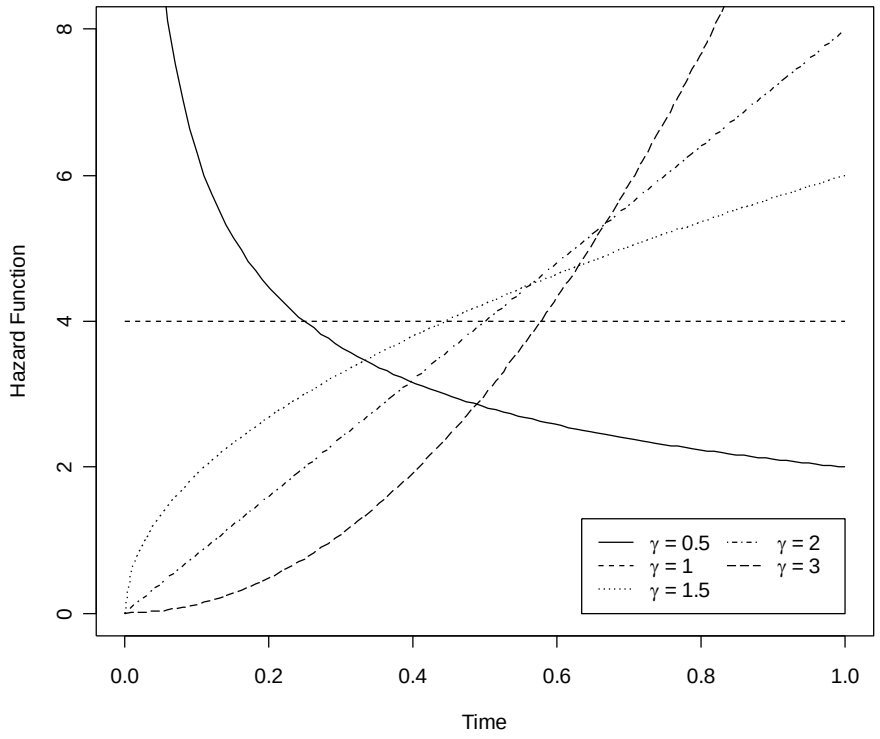
A plot of the survival functions for different choices of gamma is given below.



The corresponding hazard function is

$$\lambda(t) = \alpha\gamma t^{\gamma-1}$$

and is similarly plotted in the following figure.



The Weibull is a popular distribution for modeling survival time because of the wide range of shapes that it can take on, including decreasing ($\gamma < 1$), increasing ($\gamma > 1$), and constant ($\gamma = 1$) hazards.

10.2 Weibull Regression Model

The Weibull regression model has the form

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1} \exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

where $x_1,\ldots,x_p$ are covariates, and α, γ, and $\beta_1,\ldots,\beta_p$ are parameters to be estimated from the data. Unlike Cox

regression, the baseline hazard $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$ is estimated directly in Weibull regression analyses.

The Weibull model is a *proportional hazards* model because the hazard ratio comparing individuals with different sets of covariates

$$HR = \frac{\lambda(t,\mathbf{x}')}{\lambda(t,\mathbf{x}'')} = \frac{\lambda_0(t)\exp\left\{\beta_1 x_1' + \beta_2 x_2' + \ldots + \beta_p x_p'\right\}}{\lambda_0(t)\exp\left\{\beta_1 x_1'' + \beta_2 x_2'' + \ldots + \beta_p x_p''\right\}}$$
$$= \exp\left\{\beta_1(x_1' - x_1'') + \beta_2(x_2' - x_2'') + \ldots + \beta_p(x_p' - x_p'')\right\}$$

is a multiplicative function of the covariates that does not depend on *t*. In other words, the hazard ratio is constant as a function of time.

Breast-Feeding Study

Recall the regression model that was fit to the data from the Breast-feeding Study,

$$\lambda(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ \quad + \beta_4 education + \beta_5 poverty \\ \quad + \beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}$$

where the following indicator variables were defined to model the effect of race:

$$race1 = I(whites)$$
$$race2 = I(blacks).$$
$$race3 = I(other)$$

Previously, Cox regression was used to fit this model. Here we use Weibull regression to obtain the following estimates for this model:

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
age	$\hat{\beta}_1$	0.0215	0.0166	1.68	0.1946
alcohol	$\hat{\beta}_2$	0.1654	0.1225	1.82	0.1770
care3	$\hat{\beta}_3$	-0.0194	0.0897	0.05	0.8288
education	$\hat{\beta}_4$	-0.0543	0.0229	5.63	0.0176
poverty	$\hat{\beta}_5$	-0.2084	0.0930	5.02	0.0250
race2	$\hat{\beta}_6$	0.2048	0.1049	3.81	0.0508
race3	$\hat{\beta}_7$	0.3351	0.0969	11.96	0.0005
smoke	$\hat{\beta}_8$	0.2710	0.0792	11.70	0.0006
scale	$\hat{\alpha}$	0.0675	0.0229	-	-
shape	$\hat{\gamma}$	0.9882	0.0249	-	-

Comments

The scale α and shape γ parameters control the behavior of the baseline hazard function. When $\gamma = 1$, the Weibull distribution reduces to an exponential distribution, whose hazard rate is constant over time. Thus, one could test

whether the exponential model provides as good a fit to the data by testing the hypotheses

$$H_0 : \gamma = 1$$
$$H_A : \gamma \neq 1.$$

The Wald or likelihood ratio statistic can be used to carry out this test; although the later is preferred in practice.

10.3 Inference

10.3.1 Hazard Ratio Estimation

Hazard ratios for the Weibull model are computed the same way that they are for a Cox model.

Breast-Feeding Study

The estimated hazard ratio for individuals aged 25, relative to those aged 20 is

$$\begin{aligned}\widehat{HR} &= \frac{\hat{\lambda}(t,\mathbf{x}')}{\hat{\lambda}(t,\mathbf{x}'')} = \frac{\hat{\lambda}(t;age=25)}{\hat{\lambda}(t;age=20)} \\ &= \exp\left\{(25-20)\hat{\beta}_1\right\} = \exp\left\{5(0.0215)\right\} \\ &= 1.11\end{aligned}$$

with a 95% confidence interval of

$$
\begin{aligned}
CI &= \exp\left\{5\hat{\beta}_1 \pm 1.96 se\left(5\hat{\beta}_1\right)\right\} \\
&= \exp\left\{5\hat{\beta}_1 \pm 1.96(5) se\left(\hat{\beta}_1\right)\right\} \\
&= \exp\left\{5(0.0215) \pm 1.96(5)(0.0166)\right\} \\
&= (0.95, 1.31)
\end{aligned}
.
$$

10.3.2 Hazard Rate Estimation

One of the advantages of parametric survival regression is the ability to estimate the hazard rate as a function of time and the covariates. Recall that the baseline hazard function in the Weibull model is $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$. Estimates for the parameters are available from the regression analysis; e.g.

$$
\begin{aligned}
\lambda_0(t) &= \hat{\alpha}\hat{\gamma}t^{\hat{\gamma}-1} = (0.0675)(0.9882)t^{0.9882-1} \\
&= 0.0667t^{-0.0118}
\end{aligned}
.
$$

In general, the hazard rate in the Weibull model is a function of the covariates, i.e.

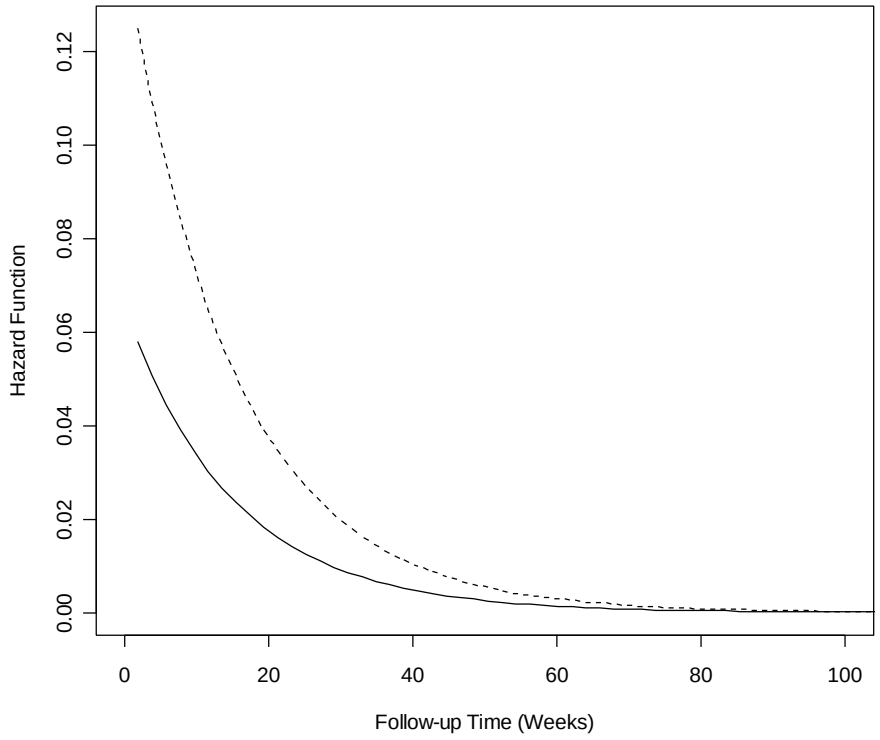
$$
\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1}\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}.
$$

Thus, the hazard rate could be plotted for any set of covariate values. For example, the estimated hazard

function for 20 years old of race “other”, with 8 years of education, and a history alcohol and cigarette usage is

$$\lambda(t;\mathbf{x}) = 0.0667t^{-0.0118}\exp\left\{\begin{array}{l} 20(0.0215)+0.1654 \\ \quad +8(-0.0543)+0.3351 \\ \quad +0.2710 \end{array}\right\}$$
$$= 0.0667t^{-0.0118}\exp\{0.7671\}$$

This (dashed line) and the baseline hazard (solid line) function are displayed in the plot below.



10.3.3 Survival Function Estimation

The survival function for the Weibull regression model is

$$S(t) = \exp\left\{-\alpha t^{\gamma}\right\}^{\exp\left\{\beta_1 x_1 + \beta_2 x_x + \ldots + \beta_p x_p\right\}}.$$

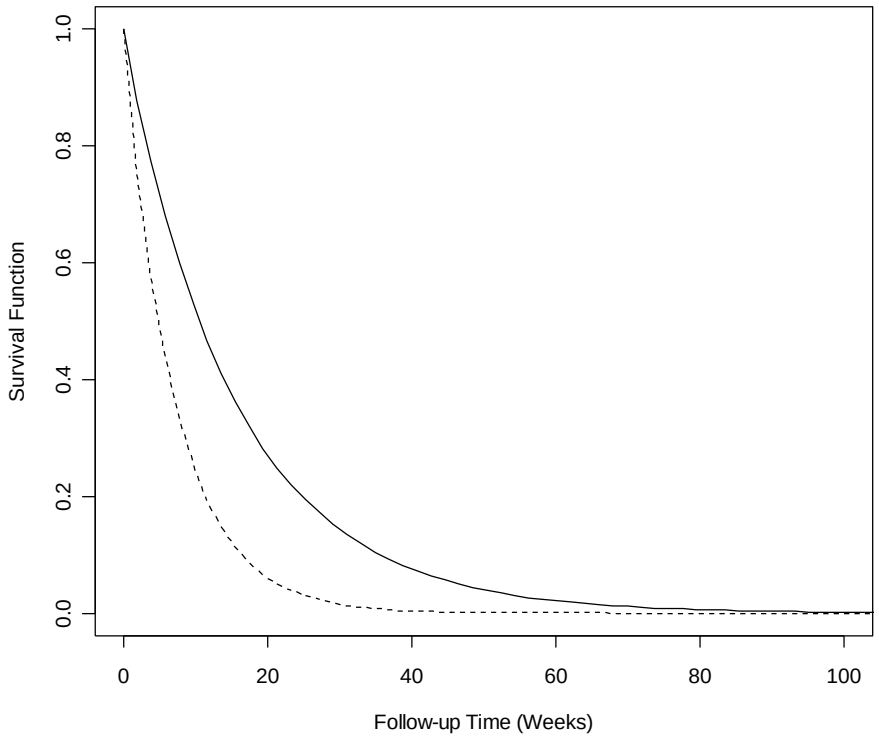
The survival functions that correspond to the hazard functions for the Breast-Feeding example in the previous section are

$$S_0(t) = \exp\left\{-0.0675 t^{0.9882}\right\}$$

and

$$S(t) = \exp\left\{-0.0675 t^{0.9882}\right\}^{\exp\{0.7671\}}$$

These functions are displayed in the figure below.


Survival Function
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Follow-up Time (Weeks)

10.4 Software Considerations

The SAS procedure PROC LIFEREG fits parametric regression models to survival data that can be censored. Similar routines are available in R and other statistical software packages

SAS Program and Output

```
proc lifereg data=breastfed;
   model weeks*weaned(0) = age alcohol care3 education poverty
         race2 race3 smoke / dist=weibull;
run;
```

Syntax

- Specification of the model statement is similar to that in PROC PHREG.
- Since LIFEREG fits several different types of parametric regression models, the distribution can be specified with the **dist** option (default: weibull). Specification of “exponential” will fit an exponential regression model, which is a special case of the Weibull for which the shape parameter is equal to 1.

The LIFEREG Procedure

Model Information

Data Set	WORK.BREASTFED
Dependent Variable	Log(weeks)
Censoring Variable	weaned
Censoring Value(s)	0
Number of Observations	927
Noncensored Values	892
Right Censored Values	35
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Weibull
Log Likelihood	-1418.106283

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	1.6864	0.1941
alcohol	1	1.8241	0.1768
care3	1	0.0468	0.8288
education	1	5.6349	0.0176
poverty	1	5.0442	0.0247
race2	1	3.8257	0.0505
race3	1	12.0155	0.0005
smoke	1	11.8031	0.0006

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.7275	0.3302	2.0804	3.3746	68.24	<.0001
age	1	-0.0217	0.0167	-0.0545	0.0111	1.69	0.1941
alcohol	1	-0.1674	0.1239	-0.4103	0.0755	1.82	0.1768
care3	1	0.0196	0.0908	-0.1583	0.1976	0.05	0.8288
education	1	0.0550	0.0232	0.0096	0.1004	5.63	0.0176
poverty	1	0.2109	0.0939	0.0269	0.3950	5.04	0.0247
race2	1	-0.2073	0.1060	-0.4150	0.0004	3.83	0.0505
race3	1	-0.3392	0.0978	-0.5309	-0.1474	12.02	0.0005
smoke	1	-0.2742	0.0798	-0.4306	-0.1178	11.80	0.0006
Scale	1	1.0120	0.0257	0.9627	1.0637		
Weibull Shape	1	0.9882	0.0251	0.9401	1.0387		

Scaled Parameter Estimates

In order to estimate the model parameters, SAS (and R) use a different, but equivalent, formulation of the Weibull distribution. Essentially, a log-linear model of the form

$$\ln t = \mu + \zeta_1 x_1 + \zeta_2 x_2 + \ldots \zeta_p x_p + \sigma w$$

is fit, for which $\zeta_1, \zeta_2, \ldots, \zeta_p$ are the estimated effects of the covariates on the log-transformed survival times and w is a random variable that has the standard extreme value distribution. The SAS output provides the following:

Variable	Parameter	Estimate	SE	Wald	
				Chi-Square	p-value
Intercept	$\hat{\mu}$	2.728	0.330	68.24	<.0001
age	$\hat{\zeta}_1$	-0.022	0.017	1.69	0.1941
alcohol	$\hat{\zeta}_2$	-0.167	0.124	1.82	0.1768
care3	$\hat{\zeta}_3$	0.020	0.091	0.05	0.8288
education	$\hat{\zeta}_4$	0.055	0.023	5.63	0.0176
poverty	$\hat{\zeta}_5$	0.211	0.094	5.04	0.0247
race2	$\hat{\zeta}_6$	-0.207	0.106	3.83	0.0505
race3	$\hat{\zeta}_7$	-0.339	0.098	12.02	0.0005
smoke	$\hat{\zeta}_8$	-0.274	0.080	11.8	0.0006
Scale	$\hat{\sigma}$	1.012	0.026	-	-

These results must be transformed in order to obtain the parameter estimates for our proportional hazards model

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1} \exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

according to the following relationships:

$$\begin{aligned} \alpha &= \exp\{-\mu/\sigma\} \\ \gamma &= 1/\sigma \\ \beta_j &= -\zeta_j/\sigma \end{aligned} \quad .$$

Unfortunately, the methods needed to compute the variances for these parameters are beyond the scope of this class.

Applied Survival Analysis (171:242)

Section 11: Parametric Regression Model Diagnostics

Brian J. Smith, Ph.D.

March 28, 2005

Table of Contents

11.1 Residuals

In linear regression, residuals are simply computed as the observed response variables minus the predicted values. We can then plot the residuals to check that they are normally distributed or plot them against covariates not included in the model to explore possible relationships that are not accounted for. Similar residual analyses can be performed for parametric survival regression models. In this section, we present diagnostic results for the regression analysis of the Breast-Feeding study using the Weibull model

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1}\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}$$

which is not directly observable. We will discuss two types of residuals:

1. Standardized Residuals
2. Deviance Residuals

11.1.1 Standardized Residuals

Examination of the model fit can be done with **standardized residuals** based on the linear model representation

$$\ln t = \mu + \zeta_1 x_1 + \zeta_2 x_2 + \ldots \zeta_p x_p + \sigma w\,.$$

The standardized residuals are defined by analogy to those used in normal theory regression as

$$r_i = \frac{\ln t_i - \left(\hat{\mu} + \hat{\zeta}_1 x_{1i} + \hat{\zeta}_2 x_{2i} + \ldots + \hat{\zeta}_p x_{pi}\right)}{\hat{\sigma}}.$$

These could be used to assess the distributional assumption for the *w* term in the regression model. For Weibull regression, *w* is assumed to have the extreme value distribution with survival function

$$S_W(w) = \exp\left\{-\exp\{w\}\right\}$$

which implies

$$\ln\left(-\ln S_W(w)\right) = w.$$

Thus, a plot of the log-log transformed survival estimates $\ln\left(-\ln \hat{S}_{KM}(r)\right)$ should fall on a 45-degree line if the extreme value distribution assumption is satisfied.

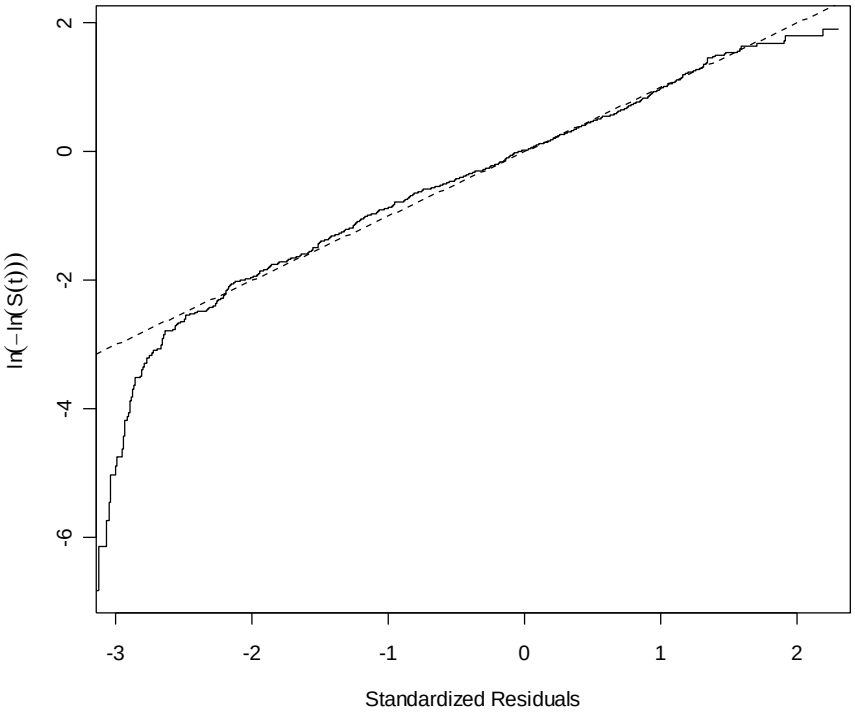
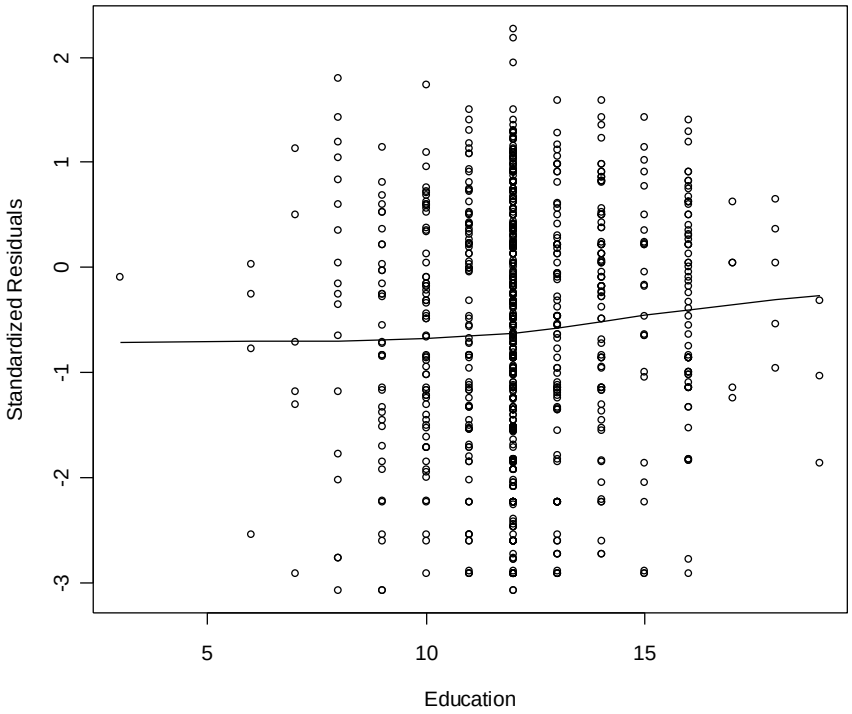


Figure 1. Residuals for the Weibull regression model fit to the Breast-Feeding Study data.

We see from the plot that the Weibull model gives a reasonable fit to the data.

One could also use residual plots to examine the effect of covariates not included in the model, in much the same way that Martingale residuals are used in Cox regression. For example, suppose we exclude education from the model and examine the relationship between the resulting standardized residuals.



Based on the plot, we might decide to use a dichotomous effect of education in the model instead of a linear effect. The question, then, is what cutpoint to use to dichotomize education. The plot suggests a cutpoint around 12 years. A more object strategy would be to fit the regression model using different choices of cutpoints to find the one that maximized the likelihood function.

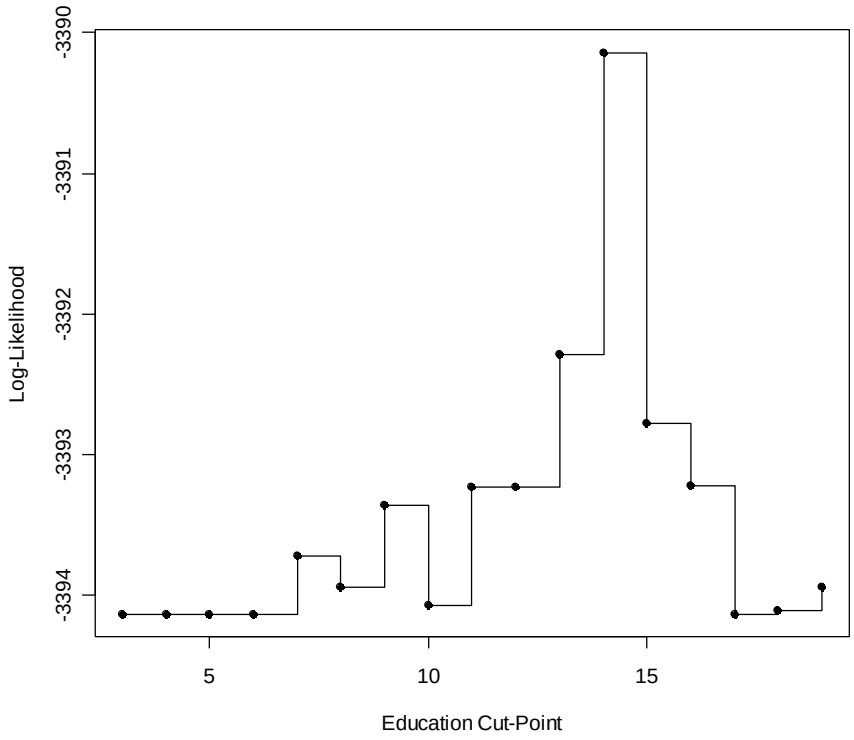
A cut point of 14 years (<14, ≥14) results in the largest value of the likelihood function; hence, we might group subjects based on years of education above or below 14 with the dichotomous variable

Variable	Levels	N	Percents
education14	0 = education < 14	746	80.5%
	1 = education ≥ 14	181	19.5%

and fit the Weibull model

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1}\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education14 + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}.$$

The resulting log-likelihood values for all possible cutpoint values are displayed in the following figure.

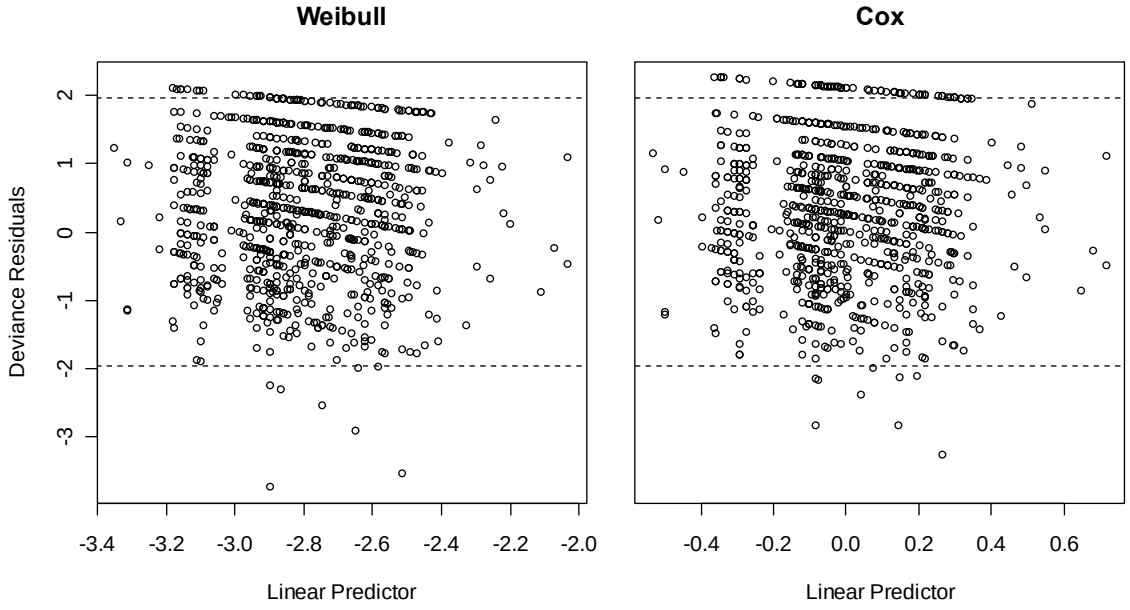


Interestingly, this model has a larger log-likelihood value than the original model with a linear effect of education.

Thus, the categorical variable for education provides a better fit to the data.

11.1.2 Deviance Residuals

Deviance residuals were used to identify outliers in Cox regression and will be used similarly here for the Weibull model. The deviance residuals from the regression analyses of the Breast-Feeding data are given below.

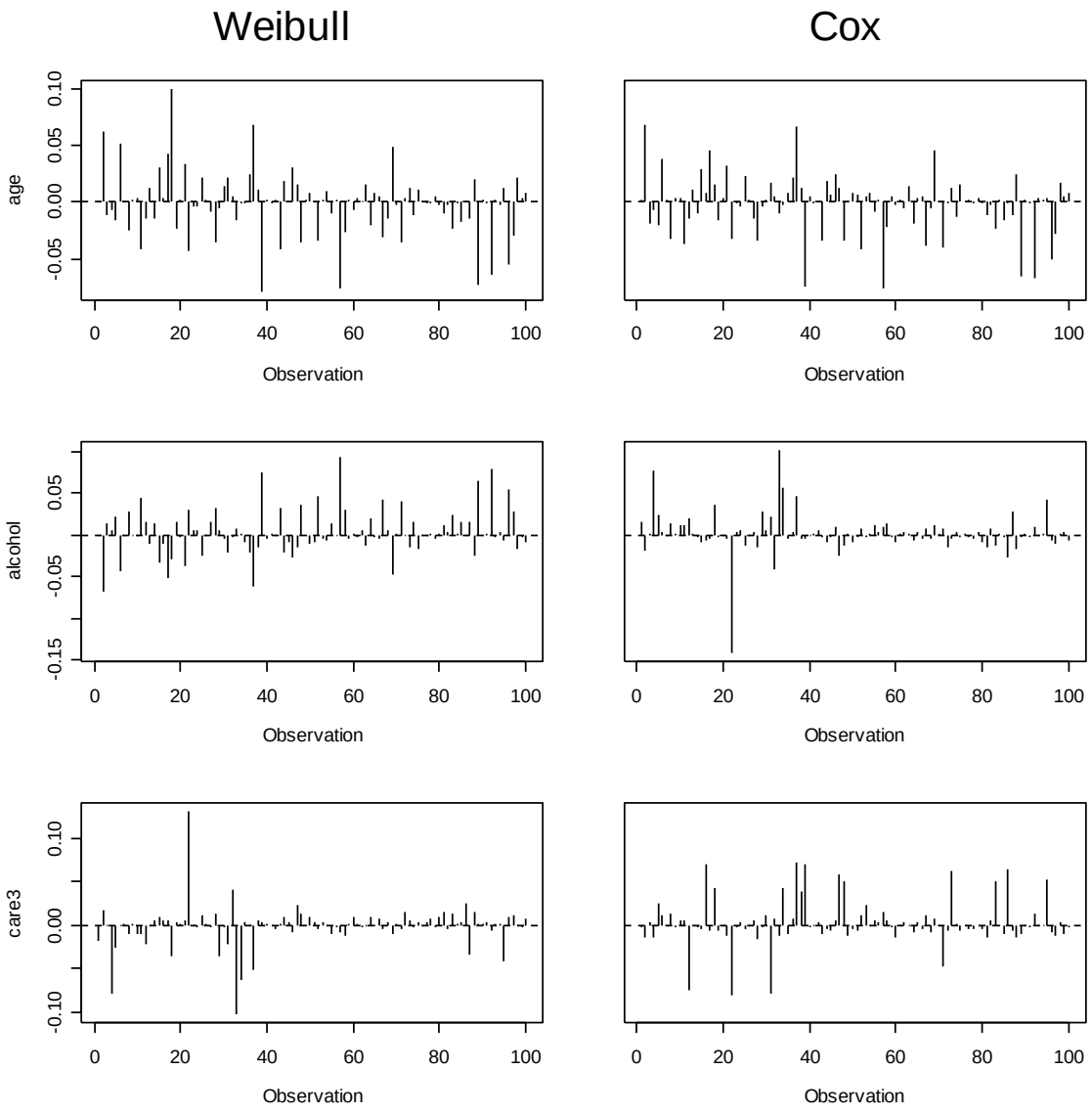


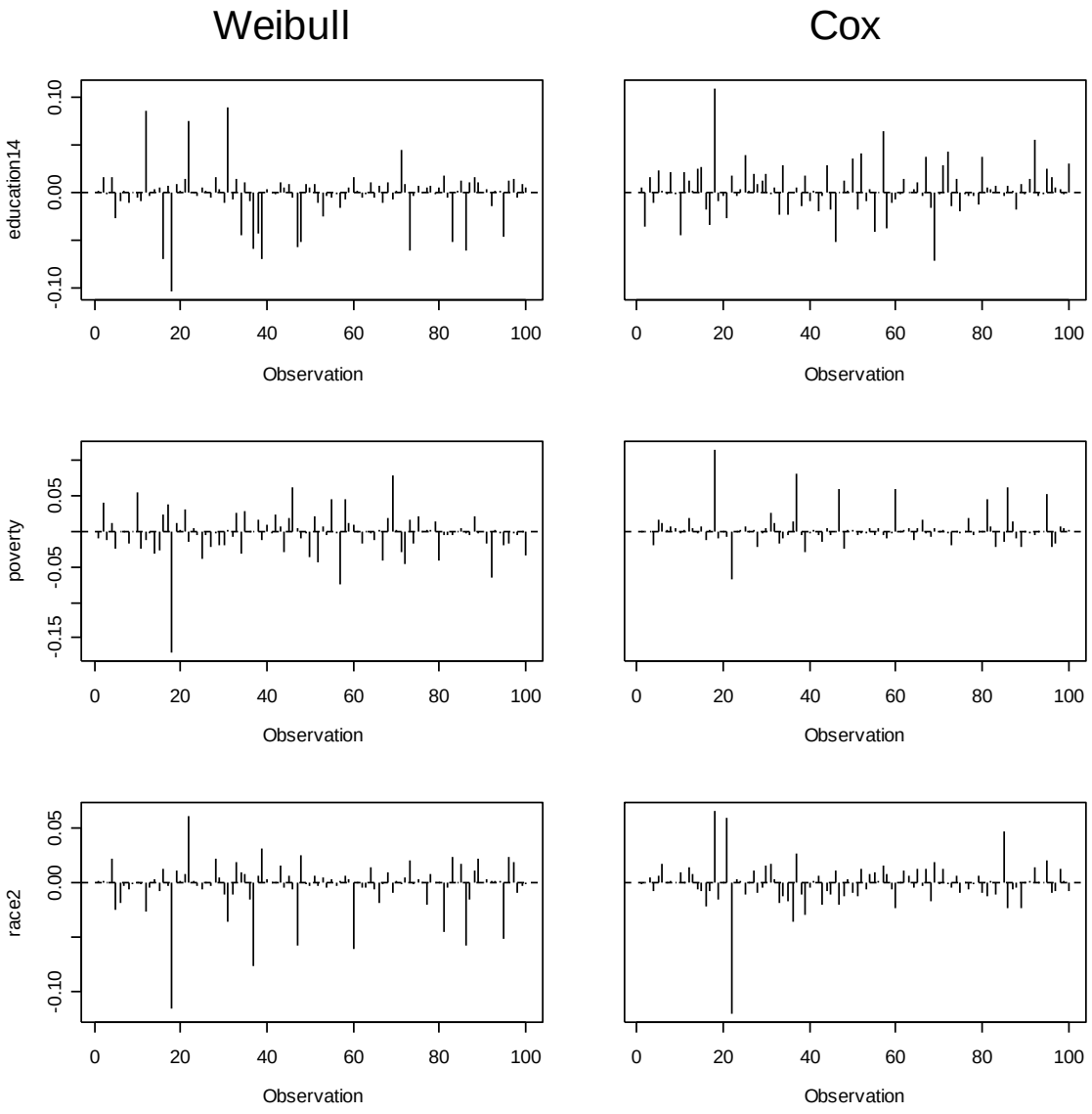
	Mean	Variance	±1.96
Weibull	-0.31	1.01	97.4%
Cox	-0.30	1.01	91.8%

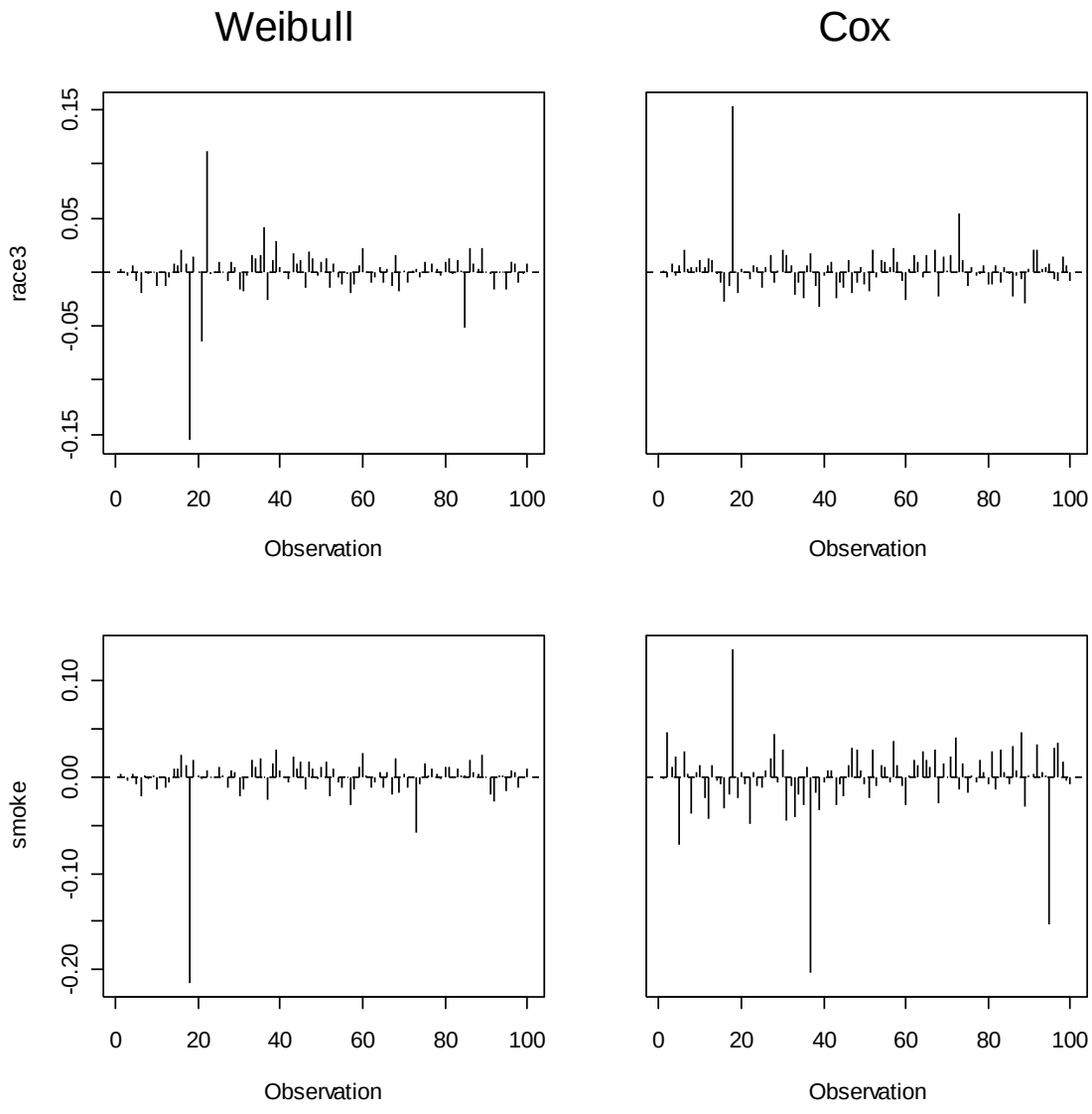
The residuals for the Weibull and Cox models are similar and, in this example, would not indicate that one model provides a better fit than the other.

11.1.3 Delta-Beta Plots

Delta-Beta plots are one method of checking the influence of each observation on the estimated model parameters.




Weibull
Cox
education14
poverty
race2
0.10
0.00
-0.10
0.05
-0.05
-0.15
0
20
40
60
80
100
Observation



11.2 Test for Proportional Hazards

11.2.1 Likelihood Ratio Test

An inferential approach to testing the proportional hazards assumption, similar to that of time-dependent covariates in Cox regression, can be used in the parametric setting.

Suppose that proportional hazards assumptions is to be tested for the model

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1}\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education14 + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}.$$

For example, say we want to test the proportional hazards assumption with respect to education. The test proceeds as follows:

1. Stratify subjects based on the levels of the covariate, *X*, to be test for proportional hazards. Fit the parametric regression model separately to each stratum. Sum the values of the log-likelihood functions across the stratum-specific model fits.

Education	Model	$\ln L(\hat{\boldsymbol{\beta}})$
< 14	$\alpha_1\gamma_1 t^{\gamma_1-1}\exp\left\{\begin{array}{l}\beta_{11} age + \beta_{12} alcohol + \beta_{13} care3 \\ +\beta_{15} poverty + \beta_{16} race2 \\ +\beta_{17} race3 + \beta_{18} smoke\end{array}\right\}$	-2725.7
≥ 14	$\alpha_2\gamma_2 t^{\gamma_2-1}\exp\left\{\begin{array}{l}\beta_{21} age + \beta_{22} alcohol + \beta_{23} care3 \\ +\beta_{25} poverty + \beta_{26} race2 \\ +\beta_{27} race3 + \beta_{28} smoke\end{array}\right\}$	-657.6
		-3383.3

2. Fit the parametric regression model to the full dataset, allowing the effects of the covariates to differ across the levels of *X*. The model

$$\alpha\gamma t^{\gamma-1}\exp\left\{\begin{array}{l}\beta_1 age+\beta_2 alcohol+\beta_3 care3+\beta_4 education14\\ +\beta_5 poverty+\beta_6 race2+\beta_7 race3+\beta_8 smoke\\ +\beta_9 age*education14+\beta_{10} alcohol*education14\\ +\beta_{11} care3*education14+\beta_{12} poverty*education14\\ +\beta_{13} race2*education14+\beta_{14} race3*education14\\ +\beta_{15} smoke*education14\end{array}\right\}$$

fit to the entire Breast-Feeding dataset has a log-likelihood value of -3389.2.

3. Calculate the difference in the log-likelihoods from the two analyses

$$X^2_{LR}=-2\left(\ln L_{\text{reduced}}-\ln L_{\text{full}}\right)\sim\chi^2_k.$$

The stratum-specific analyses allow the distributional parameters α and γ to differ across the levels of *X*. In other words, the baseline hazard is allowed to vary across the strata. Thus, the resulting composite log-likelihood value can be treated as coming from a "full" model. The regression model fit to the entire dataset is a "reduced" model because α and γ are the same across all strata; i.e. all subjects have the same baseline hazard function. The degrees of freedom *k* is the difference in the number of parameters between the two analyses.

4. If the difference, as measured by the p-value

$$p = \Pr\left[\chi_k^2 \geq X_{LR}^2\right],$$

is significant, then conclude that proportional hazards does not hold for the covariate *X*. In our example,

$$\begin{aligned} X_{LR}^2 &= -2(-3389.2 + 3383.3) \\ &= 11.8 \\ p &= \Pr\left[\chi_1^2 \geq 11.8\right] = 0.0006 \end{aligned}$$

and so the proportional hazards assumption does not hold for education.

11.2.2 Graphical Check

The survival function of

$$S(t;\mathbf{x}) = \exp\left\{-\alpha t^{\gamma}\right\}^{\exp\{x_1\beta_1+\ldots+x_p\beta_p\}}$$

for the Weibull regression model may be re-written as

$$\ln\left(-\ln S(t;\mathbf{x})\right) = \beta_1 x_1 + \ldots \beta_p x_p + \ln\alpha + \gamma \ln t\,.$$

which is a linear function of the log-transformed survival times. If the proportional hazards assumptions holds, then plots of the non-parametric estimates $\ln\left(-\ln\hat{S}_{KM}(t)\right)$

calculated separately for subjects with covariate patterns $\mathbf{x}'$ and $\mathbf{x}''$ should be approximately parallel.

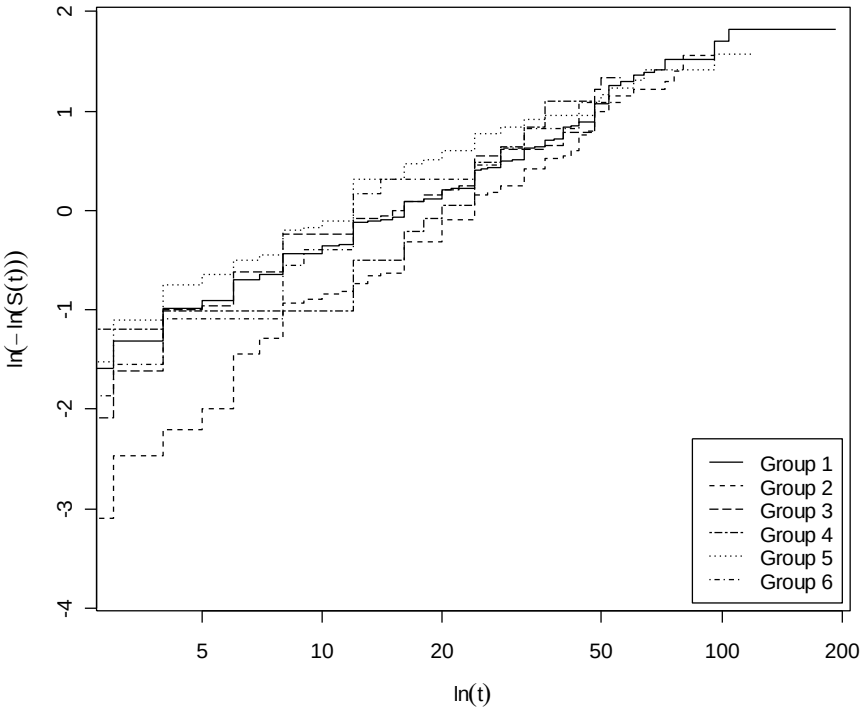
This method is most useful when there are a small number of categorical variables in the regression model. Consider the model

$$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1}\exp\{\beta_1 race2 + \beta_2 race3 + \beta_3 education14\}.$$

The proportional hazards assumption could be checked by plotting $\ln\left(-\ln\hat{S}_{KM}(t)\right)$ for each of the six groups defined by the covariates in the regression model

Group	Race	Education (years)
1	whites	< 14
2	whites	≥ 14
3	blacks	< 14
4	blacks	≥ 14
5	others	< 14
6	others	≥ 14

The results are shown below.



11.3 Model Fit

The method of Nagelkerke (1991) can be used to compute the R^2 statistic

$$R^2 = 1 - \exp\left\{-\frac{2}{n}\left(\ln L\left(\hat{\boldsymbol{\beta}}\right) - \ln L\left(0\right)\right)\right\}$$

where $\ln L\left(\hat{\boldsymbol{\beta}}\right)$ and $\ln L\left(\mathbf{0}\right)$ denote the likelihoods for the regression models with and without the covariates, respectively. The resulting values for the Weibull and Cox regression fits for the proposed regression model are

	n	$\ln L(\mathbf{0})$	$\ln L(\hat{\boldsymbol{\beta}})$	R^2
Weibull	927	-3408.6	-3390.1	3.9%
Cox	927	-5191.1	-5174.5	3.5%

11.4 Parametric Models

Our consideration of parametric regression models for survival data has been limited to the Weibull model. There are many other parametric models available. A few other popular choices are given in the table below.

Model	Hazard Rate
Exponential	$\lambda(t;\mathbf{x}) = \alpha \exp\{x_1\beta_1 + \ldots + x_p\beta_p\}$
Weibull	$\lambda(t;\mathbf{x}) = \alpha\gamma t^{\gamma-1} \exp\{x_1\beta_1 + \ldots + x_p\beta_p\}$
Log-logistic	$\lambda(t;\mathbf{x}) = \dfrac{\alpha\gamma t^{\gamma-1} \exp\{x_1\beta_1 + \ldots + x_p\beta_p\}}{1+\alpha t^{\gamma} \exp\{x_1\beta_1 + \ldots + x_p\beta_p\}}$
Log-normal	*

* Complex function of the covariates

The AIC could be used to select among different parametric models.

Model	Log-Likelihood	p	AIC
Exponential	-3390.2	9	6798.5
Weibull	-3390.1	10	6800.3
Log-logistic	-3407.2	10	6834.4
Log-normal	-3383.8	10	6787.6

11.5 Model Specification

The specification of a survival model requires consideration of (1) which variables to include and (2) the functional form of the relationship between the variables and the survival times. In practice, there is often a lack of evidence to support an *a priori* choice of one survival distribution over another.

Furthermore, when the primary concern is in evaluating the effects of covariates, there may be little to be gained in choosing a parametric model over a nonparametric model, like the Cox model. Estimated covariate effects tend to be similar, as is the case for the estimates from the model

$$\lambda(t;\mathbf{x}) = \lambda_o(t)\exp\left\{\begin{array}{l}\beta_1 age + \beta_2 alcohol + \beta_3 care3 \\ +\beta_4 education14 + \beta_5 poverty \\ +\beta_6 race2 + \beta_7 race3 + \beta_8 smoke\end{array}\right\}.$$

Variable	Parameter	Weibull		Cox	
		Estimate	SE	Estimate	SE
age	$\hat{\beta}_1$	0.0196	0.0156	0.0173	0.0155
alcohol	$\hat{\beta}_2$	0.1642	0.1225	0.1668	0.1227
care3	$\hat{\beta}_3$	-0.0114	0.0895	-0.0179	0.0896
education14	$\hat{\beta}_4$	-0.2765	0.0992	-0.2764	0.0994
poverty	$\hat{\beta}_5$	-0.1707	0.0908	-0.1725	0.0913
race2	$\hat{\beta}_6$	0.1903	0.1046	0.1792	0.1050
race3	$\hat{\beta}_7$	0.3605	0.0954	0.3312	0.0957

Variable	Parameter	Weibull		Cox	
		Estimate	SE	Estimate	SE
smoke	$\hat{\beta}_8$	0.2717	0.0786	0.2513	0.0786

Thus, if one is only interested in hazard ratio estimation for this study, there is little difference between the two models. However, other factors may make one method more attractive than the other. Below is a comparison of the Weibull and Cox models that can be fit with standard SAS or R survival regression routines.

	Weibull	Cox
Hazard rate and survival estimation	yes	no
Interval censoring	yes	no
Multiplicative effect of covariates	yes	yes
Proportional hazards	yes	yes
Time-dependent covariates	no	yes
Varying follow-up intervals	no	yes

Regression diagnostics should always be performed before selecting a final regression model. R^2 values may be used to compare fits between a parametric and non-parametric model. Likewise, analyses of residuals and model assumptions may provide evidence that one model formulation provides a better fit than others that are under consideration.

Applied Survival Analysis (171:242)

Appendix A: Comments on the Log-Rank Test and Alternative Methods

Brian J. Smith, Ph.D.

February 9, 2005

Table of Contents

A.1 Comments on the Log-Rank Test and Alternative Methods

A.1.1 Proportional hazards assumption

The log-rank test is most powerful in the presence of proportional hazards. One method for determining if the hazards are proportional is to plot smooth estimates of the hazard functions. Another graphical check is to plot $\log\left(-\log\hat{S}(t)\right)$, the log-log transformed Kaplan-Meier estimates (see Figure 1).

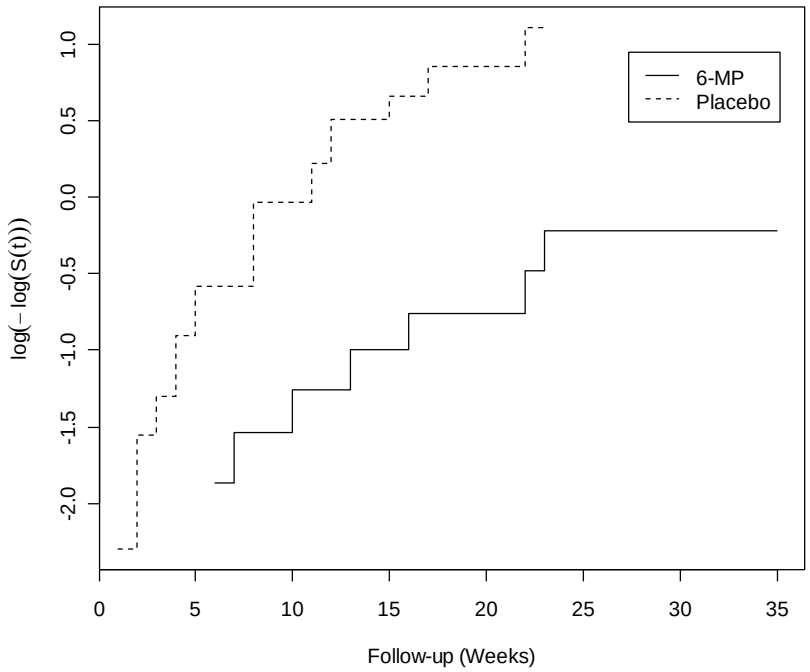


Figure 1. Log-log transformed Kaplan Meier curves for the Leukemia Trial.

Notes

1. If the hazards are proportional the log-log survival curves will be parallel.

2. The advantage of this method over the use of smoothed hazard plots is that the Kaplan-Meier estimates are deterministic and do not depend on subjective choices of smoothing functions.

A.1.2 The effect of outliers

In forming the weighted log-rank statistic, a 2x2 table is formed at each failure time. Results of those tables are summed and used to construct the test statistic. The actual failure times are not used in the calculation of the statistic. This has both advantages and disadvantages. The advantage is that the log-rank statistics are robust to outliers among the failure times.

<u>Leukemia Example</u>

Recall the data from the leukemia clinical trial of children treated with 6-mercaptopurine versus placebo:

6-MP (21 patients): 6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*

Placebo (21 patients): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, **23**

Suppose that the time of the last failure in the Placebo group was changed from 23 to 50. A before and after comparison of the survival curves is given in Figure 2. Although the curves decrease at different rates at the beginning of the study, the outlier leads to curves that seem to behave more similar at the end of the study. What impact does this have on the results from the log-rank test?

The test statistic for the original data was 16.8 (p = 4.17e-5). With the outlier, the resulting statistic is 14.3 (p = 1.57e-4). The outlier has very little effect on the test results.

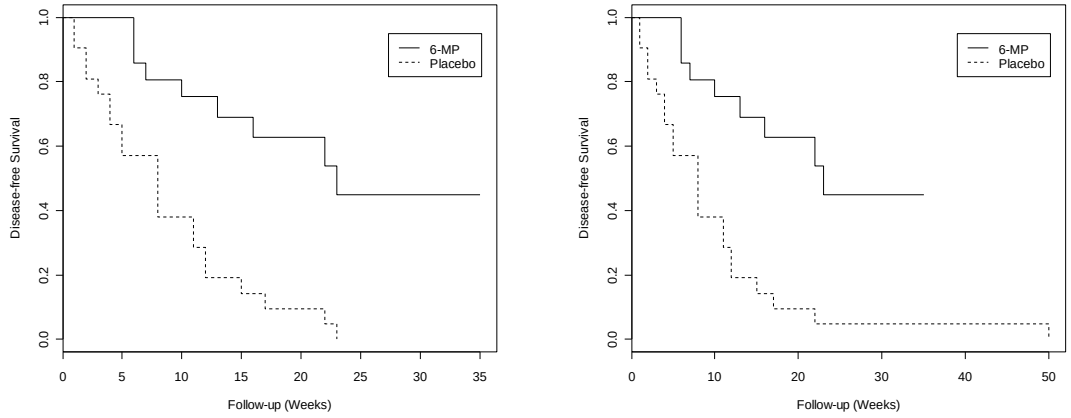
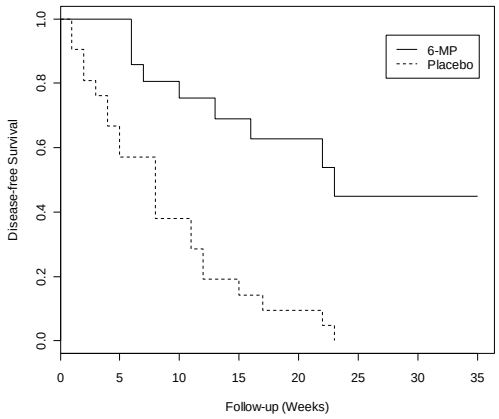
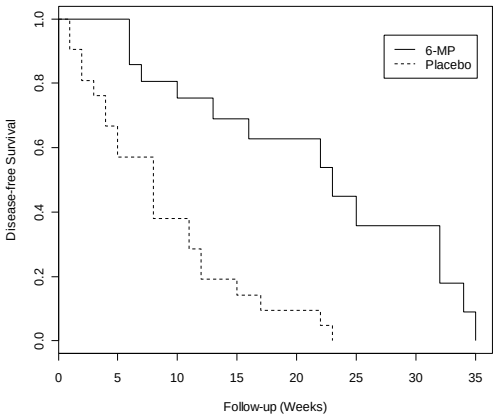


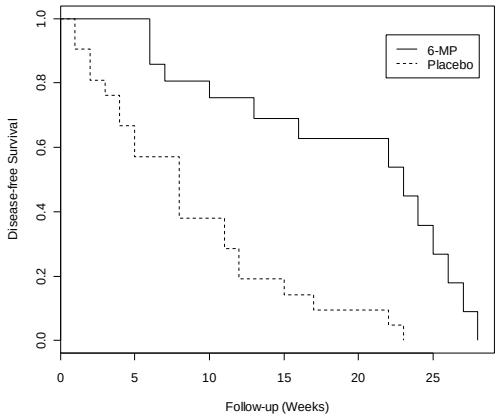
Figure 2. Effects of an outlier on Kaplan-Meier curves using the Leukemia Trial data.

A.1.3 Limitations of the log-rank tests

The actual failure times are not used in the calculation of the statistics. The disadvantage is that these times may contain valuable information about the survival experience. For example, the following three configurations

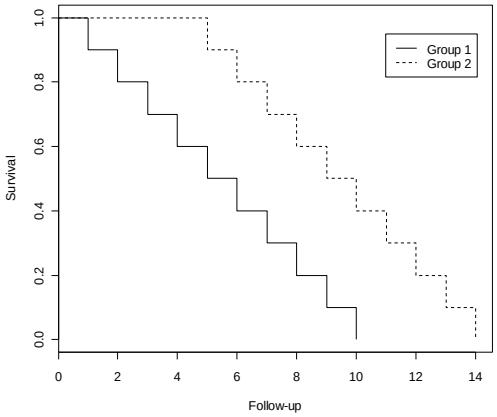


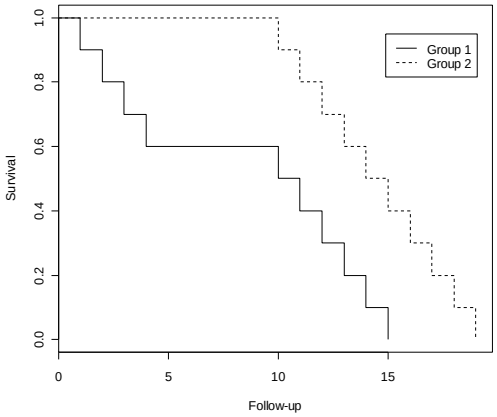




all yield the same log-rank statistic, even thought the survival experience in the 6-MP group gets progressively worse in the plots. The test statistics do not reflect the changing situations.

Furthermore, the log-rank tests do not distinguish between the next two configurations because the actual failure times are not used.





Notes:

1. At each failure time, the log-rank test only compares those groups which still contain subjects at risk.

2. Curves cannot be compared beyond the follow-up period for the risk set.

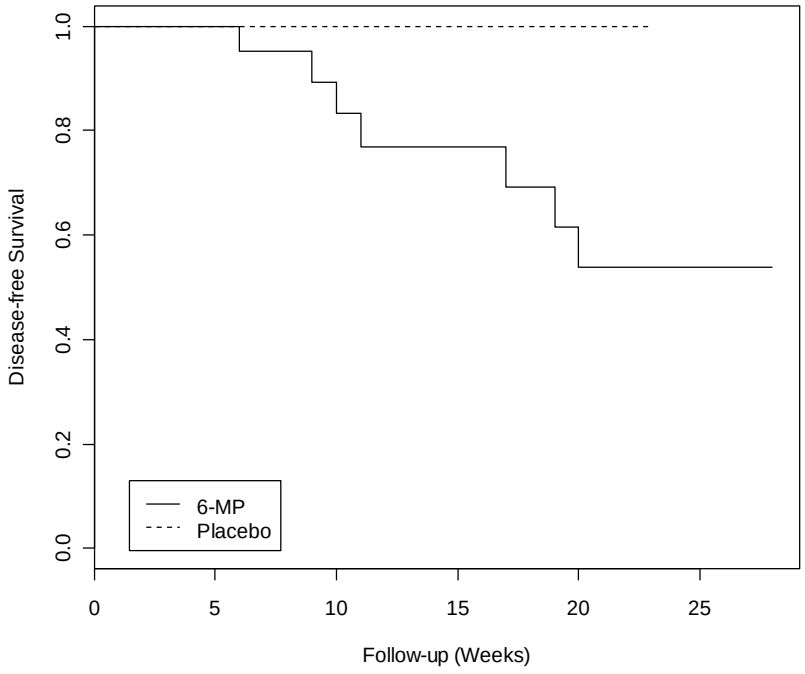
A.1.4 An alternative two-sample testing procedure

One natural way to compare the survival curves would be to accumulate the area between them over the length of the study period, i.e.

$$\int_t \left(\hat{S}_1(t) - \hat{S}_2(t) \right) dt .$$

where $\hat{S}_i(t)$ is the Kaplan-Meier estimator. Under H_0: $S_1 = S_2$ the area between the survival curves ought to be equal to zero.

Pepe and Fleming proposed a weighted version of the above statistic. Let $\hat{C}_i(t)$ be the estimated probability that censoring does not occur before time t. $\hat{C}_i(t)$ is computed the sample way that $\hat{S}_i(t)$ is except that failure and censoring are interchanged. For example, the following plot displays the $\hat{C}_i(t)$ in the leukemia trial.



The Pepe-Fleming statistic, called the weighted Kaplan-Meier statistic, is

$$WKM = \sqrt{\frac{n_1 n_2}{n}} \sum_i \hat{w}(t_i) \left[\hat{S}_1(t_i) - \hat{S}_2(t_i) \right] \{ t_i - t_{i-1} \}$$

where

$$\hat{w}(t_i) = \frac{\hat{C}_1(t_i)\hat{C}_2(t_i)}{\frac{n_1}{n}\hat{C}_1(t_i) + \frac{n_2}{n}\hat{C}_2(t_i)}.$$

Notes

1. *WKM* is a weighted sum of the area between the survival functions, making use of the actual failure times.

2. The test statistic $WKM / \sqrt{V_{WKM}}$, where V_{WKM} is the variance estimate, has an approximate $N(0,1)$ distribution under the null hypothesis.

Applied Survival Analysis (171:242)

Appendix B: Stratified Cox Regression Models

Brian J. Smith, Ph.D.

March 2, 2005

Table of Contents

B.1 Stratified Cox Regression Models

B.1.1 Introduction

Lymphoma Study

In Section 6 the proportional hazards assumption was tested for the Karnofsky variable in the model

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{matrix}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70\end{matrix}\right\}.$$

by considering for significant interaction between this variables and time. The following results were obtained:

Interaction	Wald	
	Chi-Square	p-value
$*\log(t)$	0.2467	0.6194
$*I(72 < t \leq 80)$	0.1925	0.6608
$*I(t > 80)$	4.9614	0.0259

Hence, there is evidence that the hazard ratio for Karnofsky scores may not be proportional across time. One solution is to keep the important time-dependent covariate in the model; i.e.

$$\lambda(t;\mathbf{x}) = \lambda_0(t)\exp\left\{\begin{array}{l}\beta_1 auto + \beta_2 nhl + \beta_3 auto * nhl \\ +\beta_4 karnofsky + \beta_5 wait70 \\ +\gamma karnofsky * I(t > 80)\end{array}\right\}.$$

There may be a better approach because there are an infinite number of Karnofsky-time interaction terms that could be used in the model, and we only considered two. An alternative approach to dealing with the non-proportional hazards problem is to allow the baseline hazard to vary across levels of the covariate. This can be accomplished with a *stratified* Cox regression model.

B.1.2 Stratification Variables

Suppose that we wish to allow the baseline hazard function to vary across *M* levels of a covariate, such that

$$\lambda_m(t;\mathbf{x}) = \lambda_{0m}(t)\exp\left\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

where $m = 1,\ldots,M$ indexes the strata. This model assumes that the hazards are proportional within each m^{th} stratum:

$$\begin{aligned}\frac{\lambda_m(t;\mathbf{x}')}{\lambda_m(t;\mathbf{x}'')} &= \frac{\lambda_{0m}(t)\exp\left\{\beta_1 x_1' + \ldots + \beta_p x_p'\right\}}{\lambda_{0m}(t)\exp\left\{\beta_1 x_1'' + \ldots + \beta_p x_p''\right\}} \\ &= \exp\left\{\beta_1(x_1' - x_1'') + \ldots + \beta_p(x_p' - x_p'')\right\}\end{aligned}.$$

Note that the hazards are not necessarily proportional between strata because of the allowance for different baseline hazards. However, the regression parameters β_i are the same across strata. In other words, the covariates in the model have the same multiplicative effect regardless of the baseline hazard or stratum.

Lymphoma Study

The Karnofsky scores are summarized in the table below.

Event	Karnofsky Scores								
	20	30	40	50	60	70	80	90	100
0	0	0	0	0	0	0	3	9	5
1	1	2	1	3	5	4	3	6	1

The results of fitting a Cox regression model stratified by the 9 possible Karnofsky scores are

Covariate	Unstratified		Stratified	
	Coefficient	p	Coefficient	p
auto	-0.6245	0.2919	-1.14189	0.1351
nhl	0.2431	0.6618	-0.31412	0.6442
auto*nhl	2.4845	0.0116	2.2256	0.0515
karnofsky	-0.0539	<.0001	-	-
wait70	-1.5140	0.0421	-0.87183	0.264

SAS Code

```
proc phreg data=lymphomamod;
   model days*event(0) = auto nhl auto_nhl wait70;
   strata karnofsky;
run;
```

Syntax

- The **strata** statement allows for the specification of variables over which to stratify the baseline hazards in the Cox model. Multiple variables can appear in the statement.

B.1.3 Likelihood Function

Recall that the likelihood function in Cox regression has the general form

$$L(\mathbf{t};\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left\{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\right\}}{\sum_{j \in r(t_i)} \exp\left\{\beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj}\right\}}.$$

In the stratified Cox model, the $\boldsymbol{\beta}$ parameters are estimated by maximizing

$$L(\mathbf{t};\boldsymbol{\beta}) = \prod_{m=1}^{M} L_m(\boldsymbol{\beta})$$

where the $L_m(\mathbf{t};\boldsymbol{\beta})$ are the likelihoods within each stratum. The individual likelihoods are thus constructed from the observed events and subjects at risk within the stratum.

Q: What happens if there are no events within a given stratum?

B.1.4 Notes

1. Stratification is attractive because the effect of the covariate need not be modeled as a specified function of time.

2. Stratification is a way of controlling for the main effects of a covariate.

3. A drawback is that one cannot estimate the effect of the stratification variable on survival.

4. When stratification is employed, the tests of hypotheses for the regression coefficients will have good power only if the deviations from the null are the same in all strata.

5. The tests of the regression coefficients are appropriate when either the number of failures within each stratum is large or the number of strata is large.

6. If there are strata in which no events are observed, then a loss of power will result. Consequently, continuous variables should be categorized if they are to be used as stratification variables in a Cox model.