A NOTE ON LINEAR RANK TESTS AND
GILL AND SCHUMACHER'S TESTS OF
PROPORTIONALITY

Rick Chappell

Department of Human Oncology
and
Department of Statistics

UNIVERSITY OF WISCONSIN-MADISON
A NOTE ON LINEAR RANKED TESTS AND GILL AND SCHUMACHER’S TESTS OF PROPORTIONALITY

Rick Chappell

University of Wisconsin-Madison
Biostatistics Center

Biostatistics Technical Report No. 63

September 1990
Summary

Gill and Schumacher [1] (abbreviated GS) assess proportionality of hazards in the two-sample case by comparison of two generalized linear rank tests with different weights. They consider the difference of the two linear rank statistics; under proportionality, the expected value of this difference is zero and its variance may be estimated. The purpose of this note is to show that one can perform an similar procedure within the traditional regression setting. The procedure has the advantage of generalizability to a multi-sample or continuous data and is computable using existing regression programs. This note may also serve as a useful overview of linear rank tests.

Keywords: linear rank tests, proportional hazards, diagnostics.
1 Introduction

A linear rank test statistic of the equality of hazard functions for two groups,

$$H_0 : \lambda_A(t) = \lambda_B(t) \, \forall t,$$

which is generalized to accommodate right censored observations with rank invariant weight $K(t)$ and no tied failure times may be expressed as

$$Q_K = \sum K(t)[d_A(t) - n_A(t)/(n_B(t) + n_A(t))],$$

where the sum is over failure times and

$$\lambda_A(t) = \text{the hazard in group A at time } t,$$
$$\lambda_B(t) = \text{the hazard in group B at time } t,$$
$$d_A(t) = \text{number of failures in group A at time } t,$$
$$n_A(t) = \text{number in risk set of group A at time } t, \text{ and}$$
$$n_B(t) = \text{number in risk set of group B at time } t.$$

This form is similar to that used by GS except that their weight differs from the one given above:

$$K_{GS}(t) = K(t) \times n_A(t)n_B(t)/[n_A(t) + n_B(t)].$$

A test of equality of hazards for two groups is obtained by comparing $Q_K$ divided by its standard error to an appropriate normal quantile [2]. Note that

$$K(t) = 1 \text{ for the log-rank test [3]}$$
= \ n_B(t) + n_A(t) \text{ for Gehan's generalization of the Wilcoxon test} [4]
= \hat{S}(t) \text{ for Peto and Peto's generalization of the Wilcoxon test} [5]
= [\hat{S}(t)]^\rho \text{ for Harrington and Fleming's class of tests} [6], \text{ and}
= [\hat{S}(t)]^{\rho}[1 - \hat{S}(t)]^\gamma \text{ for an even wider class.}

\hat{S}(t) \text{ can be taken to be any reasonable estimate of the survival function } S(t).
Q_K \text{ can be considered a weighted difference in cumulative hazard estimates}
where \ K_{GS}(t) = 1 \text{ yields Peto and Peto's generalization of the Wilcoxon test}
using Nelson's estimate [7] for } S(t). \text{ Each test (and thus each weight function)}
corresponds to optimal sensitivity for a certain pattern of hazard ratios:

- The log rank test is most sensitive to constant hazard ratios (proportionality).

- The Wilcoxon generalizations are most sensitive to monotone hazard ratios
  with approximate unity at large survival times.

- Harrington and Fleming's generalization is most sensitive to a broader
  range of such monotone hazard ratios, given known } \rho.

- [\hat{S}(t)]^{\rho}[1 - \hat{S}(t)]^\gamma \text{ is most sensitive to a very broad range of patterns of the}
  hazard ratio, given known } \rho \text{ and } \gamma, \text{ including situations in which the ratio}
  is not monotone but approaches unity at large and small survival times.
2 The GS test using step-function weights

GS point out that generalized rank statistics should roughly agree when divided by their standard errors regardless of the weight functions chosen if the hazard ratio between groups A and B is constant in time. This is true, although one must pick the weight functions to correspond to particular alternatives to proportionality. This is not necessarily the most sensible course, however. Consider the simplest choices of weight functions:

\[ K_1(t) = 1 \]
\[ K_2(t) = 1, \quad t < \tau \]
\[ = 0, \quad t \geq \tau. \]

These are weights for the log-rank tests over the whole sample and for smaller times only, respectively, and comparison of the tests is equivalent to comparing the estimated relative risk for the first part of the data with that of the whole set. The pair of weight functions could also be

\[ K_1(t) = 1 \]
\[ K_3(t) = 0, \quad t < \tau \]
\[ = 1, \quad t \geq \tau, \]

with similar results. The generalized linear rank test using \( K_1 \), the log-rank test, is equivalent to the score test of

\[ H_0 : \beta = 0 \]
from the simple proportional hazards model

$$
\lambda_A(t) = \lambda(t; I_A) \propto e^{\beta \times I_A},
$$

with $I_A$ an indicator variable for enrollment of an observation in group $A$. One may also combine the chi-square statistics from the two tests with weights $K_2$ and $K_3$, thus forming a stratified log-rank test equivalent to the score test of

$$
H_0 : \beta_1 = \beta_2 = 0
$$

from the proportional hazards model

$$
\lambda(t; I_A) \propto e^{(\beta_1 + \beta_2 I_r(t)) \times I_A},
$$

where $I_r(t)$ is an indicator variable for $t \geq \tau$ (see Kalbfleisch and Prentice [8]). This test might be 'compared' to the ordinary log-rank test with weight $K_1(t)$ (score test of equality of hazards without stratification), but there is no point: the obvious and long-practiced recourse is to test $\beta_2 = 0$ in a regression setting while allowing $\beta_1$ to be arbitrary. This type of diagnostic was recommended in Cox's seminal paper [9] and is quite simple to compute. Additionally, it is obviously extendable to the multi-sample case.

3 The general case

Other pairs of weight functions have their analogs in more complicated time dependent covariates although, unlike Cox's suggestion of including $t$ or $\log(t)$
as a covariate for diagnostic purposes, the induced covariates for reasonable weight functions are usually rank invariant.

The score test statistic in the proportional hazards model \( \lambda(t; I_A) \propto e^{\beta f(t)I_A} \) for a defined rank-invariant time-dependent covariate \( f(t) \) is

\[
\frac{\partial \log \text{(partial likelihood)}}{\partial f} = \sum_{A,B} [f(t)I_A - C(t)/(n_A(t) + n_B(t))]
\]

\[
= \sum_A f(t) - \sum_{A,B} f(t) \times n_A(t)/(n_A(t) + n_B(t)),
\]

where the range of summation gives the sets over whose failure times the summation should be taken, and \( C(t) \) is the sum of \( f(t) \) over those observations in group \( A \) which are at risk at time \( t \). The second representation is identical to the form of the generalized rank test statistic given above with weight function \( K(t) = f(t) \). One can conclude that a two-sample generalized rank test with weight \( K(t) \) is equivalent to testing the significance of the coefficient of \( K(t) \times I_A \) in the Cox regression model, where \( I_A \) is an indicator variable for the observation which fails at time \( t \) belonging to group \( A \); this fact was pointed out for the multi-sample case by Andersen [10]. A comparison of two tests with weight functions \( K_1(t) \) and \( K_2(t) \) is thus analogous to the comparison of two regression models with single covariates: \( \lambda(t; I_A) \propto e^{\beta_1 K_1(t)I_A} \) and \( \lambda(t; I_A) \propto e^{\beta_2 K_2(t)I_A} \), in which \( K_1(t) \) is typically unity. Interpreted in this light, GS's comparison does not make as much sense as examining a model \( \lambda(t; I_A) \propto e^{\beta_1 I_A + \beta_2 K(t)} \) for significance of \( \beta_2 \), with \( K(t) \) an appropriate rank invariant function of time.

Embedding proportionality in a more general model is easily extendable to more
than two groups, or even to continuous covariates, and to multiple covariates. A separate construct seems redundant.

4 Computations

The GS diagnostic is calculated using quantities ordinarily computed in the course of conducting the two particular linear rank tests. Thus, rank test programs can be fairly easily modified to provide the diagnostic. On the other hand, linear rank test programs in many common statistical packages (e.g., SAS SURVDIFF [11] and LIFETEST [12] procedures, BMDP1 [13]) do not output these quantities. A researcher using these packages must find other methods.

The general regression diagnostic of Section 3 can be fit with any Cox regression program that allows time-dependent covariates. By stratifying on the time line, even the simplest programs can be made to accommodate a step-function weight.

References


Figure 1: Examination of Gill & Schumacher's diagnostic for \( K(t) \) a step function

In the example used in the text, we pick \( K_1(t) \) to be unity and \( K_2(t) \) an indicator function for a stratum on the time line. \( K_3(t) \) can be defined similarly. The tests induced by \( K_1(t) \), \( K_2(t) \), and \( K_3(t) \) are depicted schematically in parts a), b), and c). The tests of parts b) and c) may be combined to form that of part d). G & S would compare the statistic from a) with that of b), c), or d). A more direct way would be the test depicted in e), and described in the text.

Figure 2: A general schematic of G & S's diagnostic

The example described in Figure 1 may be generalized. G & S would compare the test of parts a) and b). The test of part c) is more direct.
a) $K_1(t)$ - log rank test

b) $K_2(t)$ - log rank test on 1st stratum

c) $K_3(t)$ - log rank test on 2nd stratum

d) Stratified log rank test

e) Test of equality across strata

---

estimated hazard ratio

predicted ratio under $H_0$
a) Log-rank test (constant $K(t)$)

b) Rank test with decreasing $K(t)$

c) Test of time-dependency