MARTINGALE BASED RESIDUALS FOR SURVIVAL MODELS

by

Terry Therneau, Patricia Grambsch, and Thomas Fleming

Technical Report Series #40

April 1988

# Martingale Based Residuals for Survival Models

Terry Therneau Patricia Grambsch Thomas Fleming

April 11, 1988

## 1 Introduction

### 1.1 Model

Consider a set of n subjects such that for the ith subject in this set, the counting process $N_i \equiv \{N_i(t), t \geq 0\}$ indicates the number of observed events experienced over the passage of time. The sample paths of the $N_i$ are assumed to be right continuous step functions with jumps of size +1 and with value zero at time zero The intensity function for $N_i$ at time t is given by

$$Y_i(t)\ d\Lambda(t, Z_i(t)) \tag{1}$$

where $Y_i(t)$ is a left continuous 0-1 process indicating whether the ith subject is in the risk set at time $t$, and $Z_i(t)$ is a p dimensional vector of left continuous covariate processes having right hand limits Unless specified otherwise, we will assume that

$$d\Lambda(t, Z_i(t)) = \exp(\beta' Z_i(t))\ d\Lambda_0(t) \tag{2}$$

for cumulative hazard function $\Lambda_0$ and vector of regression coefficients $\beta$ We assume that $\Lambda_0$ is an absolutely continuous function and that no two processes jump at the same time, so that $(N_1, N_2, \ldots, N_n)$ is a multivariate counting process.

Several familiar survival models fit into this framework. The Anderson-Gill (1982) generalization of the Cox(1972) model arises when $\Lambda_0(t)$ is completely unspecified The further restriction that $N_i(t) = 1 \Longrightarrow Y_i(s) = 0$ for all $s > t$ yields the Cox model The parametric form $\Lambda_0(t) = \lambda t$ yields a Poisson (if there are multiple events) or an exponential (if there is only a single event) model, and $\Lambda_0(t) = (\lambda t)^p$ a Weibull model. Our attention will focus primarily on the Anderson-Gill and Cox models, however, the methods to be developed will largely apply to both the parametric and semi-parametric case.

### 1.2 Martingales

For measure theoretic reasons, assume our model is endowed with a right continuous non-decreasing family $(\mathcal{F}_t, t \in [0, \infty))$ of $\sigma$ algebras, where $\mathcal{F}_t$ can be thought of as containing all of the information through time $t$ In particular, $N_i(t), Y_i(t)$, and $Z_i(t)$ are all measurable with respect to $\mathcal{F}$ It follows that

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} d\Lambda_0(s) \tag{3}$$

is a local square integrable martingale. The term *local* can be dropped if $E(N_i(t)) < \infty$ for all t and if, for j=1,2,. ,p, $\sup_t (Z_i(t))_j$ is bounded Hereafter, assume this to hold.

## 1.3 Martingale Residuals

In the parametric or semi-parametric models above, the vector of regression parameters $\beta$ and the baseline hazard $\Lambda_0$ are commonly estimated by maximum likelihood or partial likelihood methods Well known techniques are then employed to develop the relevant tests of hypotheses and confidence intervals

Of importance in such regression analyses are diagnostic tools for assessing model adequacy We will discuss certain types of residuals which are useful diagnostic tools, focusing in particular on graphical applications We will consider the use of residuals to assess·

1. the functional form for the influence of a covariate, in a model already accounting for other covariates,
2. model adequacy, particularly with respect to proportional hazards assumptions,
3. the leverage exerted by each subject in parameter estimation,
4. the accuracy of the model in predicting the outcome for a particular subject

The martingales defined in (3) form the basis for these residuals In particular, let $\beta$ and $\Lambda_0$ be estimated by maximum likelihood for the parametric models, and for the non-parametric models let $\beta$ be estimated by the maximum partial likelihood estimator and the cumulative hazard by the Breslow(1974) estimate

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\left(\sum_{j=1}^n Y_j(s) e^{\hat{\beta}' Z_j(s)}\right)}$$

(Other estimators of $\Lambda_0$ are available for the semi-parametric case, our reasons for preferring the Breslow estimate will become clear ) Then the martingale residual is defined to be

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\hat{\beta}' Z_i(s)} d\hat{\Lambda}_0(s) \tag{4}$$

with $\widehat{M}_i$ as a shorthand for $\widehat{M}_i(\infty)$. The residual can be interpreted, at each t, as the difference over [0,t] in the observed number of events minus the expected number given the model, or as excess deaths Note that for a Cox model with constant (non time-dependent) covariates this residual reduces to the simple form

$$\widehat{M}_i = \delta_i - \hat{\Lambda}_0(t)e^{\hat{\beta}' Z_i}$$

a residual that has been proposed from a different perspective by Kay(1977).

# 2 Properties

## 2.1 Sum

The next lemma will be useful in establishing properties of the residuals in parametric models.

**Lemma 2.1** Consider the model given by (1), where $\Lambda$ is differentiable and specified parametrically. If $\hat{\Lambda}$ is the MLE estimate for $\Lambda$ and the solution space is scalable, i.e , for any potential solution $\hat{\Lambda}$ then $k\hat{\Lambda}$ is also in the solution space for all $k > 0$, then

$$\sum_{i=1}^{n} \int_0^{\infty} dN_i(s) = \sum_{i=1}^{n} \int_0^{\infty} Y_i(s)\, d\hat{\Lambda}_i(s)$$

**Proof:** For parametric $\Lambda$ we can write the likelihood as

$$L = \prod_{i=1}^{n} \prod_{s>0} (1 - \lambda_i(s))^{Y_i(s)(1-dN_i(s))} \, (\lambda_i(s))^{Y_i(s)dN_i(s)} \tag{5}$$

(where $\lambda = d\Lambda$) so that

$$\log L = \sum_{i=1}^{n} \int_0^{\infty} \ln(\lambda_i(s))dN_i(s) - Y_i(s)d\Lambda_i(s)$$

Note that $\int Y_i\, dN_i \equiv \int dN_i$, since a process cannot be observed to jump when not under observation. The maximized value of the log likelihood can be written as

$$H(k) = \sum \int \left( \ln(k\hat{\lambda}(s))dN_i(s) - Y_i(s)kd\hat{\Lambda}_i(s) \right),$$

where the nuisance parameter $k$ has been added The maximum with respect to $k$ occurs when

$$0 \equiv \frac{\partial H}{\partial k} = \sum \int \frac{1}{k} dN_i(s) - \sum \int Y_i(s)d\hat{\Lambda}_i(s).$$

By hypothesis this occurs when $k = 1$ □

Using lemma 2 1, we must have $\sum \widehat{M}_i(\infty) = 0$ for any parametric model that satisfies (1) and is scalable A sufficient condition for a scalable solution space is a $\beta_0$ term is the exponent, similar to the condition that guarantees the residuals will sum to zero in a linear model.

The lemma does not directly apply to the semi-parametric model, which arises in (2) when $\Lambda_0$ is unspecified. However, it is easy to verify that when the Breslow estimate is used the even stronger condition

$$\sum \widehat{M}_i(t) = 0, \forall t \tag{6}$$

holds independent of the estimate $b$ of $\beta$.

$$\begin{aligned}
\sum \widehat{M}_i(t) &= \sum \left( \int dN_i(s) - \int Y_i(s)\, e^{b'Z_i(s)} d\hat{\Lambda}_0(s) \right) \\
&= \sum \left( \int dN_i(s) - \int Y_i(s)\, e^{b'Z_i(s)} \left[ \frac{\sum dN_j(s)}{\sum Y_j(s)\, e^{b'Z_j(s)}} \right] \right) \\
&= 0.
\end{aligned}$$

The converse is also true: equation (6) uniquely defines the Breslow estimate.

## 2.2 Expectation

Let $b$ be some estimate, not necessarily the MLE, of $\beta$ If perchance $b = \beta$, then $E(\widehat{M}_i(t)) = 0$ for either the parametric or semi-parametric models For the semi-parametric model, we also have mean zero when $b = 0$

$$\begin{aligned}
E\widehat{M}_i(t) &= E \int_0^t \left\{ dN_i(s) - \frac{Y_i(s) \sum dN_j(s)}{\sum Y_j(s)} \right\} \\
&= E \left[ \int_0^t \{ dN_i(s) - Y_i(s) d\Lambda_0(s) \} \right. \\
&\quad - \left. \int_0^t \frac{Y_i(s)}{\sum Y_j(s)} \sum \{ dN_j(s) - Y_j(s) d\Lambda_0(s) \} \right] ,
\end{aligned}$$

which is the expectation of the sum of a zero mean martingale and zero mean martingale transform

For $b = \hat{\beta}$, $E(\widehat{M}_i(t))$ converges to zero by standard martingale convergence theorems The asymptotic covariance of $\widehat{M}_i$ and $\widehat{M}_j$ goes to zero, while

$$\mathrm{var}(\widehat{M}_i(t)) \to \int_0^t Y_i(s)\, e^{\beta' Z_i(s)}\, d\Lambda_0(s) \,.$$

### 2.3 Score Vector and Score Residuals

For the semi-parametric model arising in (2) when $\Lambda_0$ is unspecified, we can write the partial likelihood as

$$L_p = \prod_{i=1}^{n} \prod_{s>0} \left\{ \frac{Y_i(s)\, e^{\beta' Z_i(s)}}{\sum_j Y_j(s)\, e^{\beta' Z_j(s)}} \right\}^{dN_i(s)} ,$$

so that, for k=1, . .,p,

$$\frac{\partial \log L_p}{\partial \beta_k} = \sum_{i=1}^{n} \int_0^\infty \{Z_{ik}(s) - \bar{Z}_k(\beta, s)\} dN_i(s) ,$$

where

$$\bar{Z}_k(b, s) \equiv \frac{\sum Y_i(s) e^{b' Z_i(s)} Z_{ik}(s)}{\sum Y_i(s) e^{b' Z_i(s)}} \tag{7}$$

is a weighted mean of the covariates over the risk set at time $s$. If $\hat{\beta}$ denotes the maximum partial likelihood estimate of $\beta$,

$$\begin{aligned}
0 &= \left. \frac{\partial \log L_p}{\partial \beta_k} \right|_{\beta=\hat{\beta}} \\
&= \sum_{i=1}^{n} \int_0^\infty (Z_{ik}(s) - \bar{Z}_{ik}(\hat{\beta}, s))\, dN_i(s) \\
&= \sum \int_0^\infty (Z_{ik}(s) - \bar{Z}_{ik}(\hat{\beta}, s))\, d\widehat{M}_i(s) \\
&\equiv \sum L_{ik}(\hat{\beta}, \infty).
\end{aligned}$$

In parallel, consider the parametric model. The derivative of log L in (5) with respect to the betas is

$$\frac{\partial \log L}{\partial \beta_k} = \sum_{i=1}^{n} \int_0^\infty Z_{ik}(s)\, dM_i(s)$$

Evaluating at the maximum likelihood estimate,

$$\begin{aligned}
0 &= \left. \frac{\partial \log L}{\partial \beta_k} \right|_{\beta=\hat{\beta}} \\
&= \sum_{i=1}^{n} \int_0^\infty Z_{ik}(s)\, d\widehat{M}_i(s) \\
&\equiv \sum L_{ik}(\hat{\beta}, \infty)
\end{aligned} \tag{8}$$

Define $L_{ik}(\cdot)$ as the *score process*, and $L_{ik}(\infty)$ as the *score residual* of the ith subject and the kth variable (Our use of the same symbol for both the parametric and semi-parametric models is an abuse of notation, but the proper definition will always be clear from the context) In both the parametric and semi-parametric cases, the score vector's terms appear in the form $f(\text{data}_i) *$ $\text{residual}_i$, a form reminiscent of that found in the generalized linear models literature

The score residuals are an example from the broader class of *martingale transform residuals.* In particular, let the process $W_i = \{W_i(s), s \geq 0\}$ be bounded, predictable, and adapted with respect to our family $(\mathcal{F}_t \; t \in [0, \infty))$ of $\sigma$ algebras (e.g., it suffices for $W_i$ to be an adapted left continuous bounded process with right hand limits) Then $\int W_i(s) dM_i(s)$ is a martingale transform and hence also is a martingale In turn, $\int_0^\infty W_i(s) d\widehat{M_i}(s)$ is a martingale-transform residual. If each component of the random variable $Z_{ij}(t)$ is bounded, it follows that the score residual is a martingale–transform residual These residuals will be found quite useful in diagnosis of each subject's leverage on parameter estimates and in assessing model assumptions such as proportional hazards

## 2.4 Deviance residuals

One deficiency of the martingale residual $\widehat{M}_i$, particularly in the one event setting of the Cox model, is its skewness. In a one event setting, its maximum value is 1 while its minimum is $-\infty$. As a visual aid in certain plots, particularly when assessing the accuracy of the model in predicting the failure rate of a given subject, it may be helpful to transform the residual to achieve a more normal shaped distribution. One such transformation is motivated by the deviance residuals found in the general linear models literature (McCullagh and Nelder(1983)) Define the deviance as

$$D = 2\left\{\text{loglik(saturated)} - \text{loglik}(\hat{\beta})\right\},$$

where a saturated model is one in which $\beta$ is completely free, i.e , each observation is allowed its own private $\hat{\beta}$ vector. There may be other nuisance parameters $\theta$ which are held constant across the two models, such as $\sigma^2$ in a normal errors linear model. In our models the nuisance parameter is the actual baseline hazard $\Lambda_0$ Let $h_i$ be the individual per-subject estimates of $\beta$, then the deviance for non time-dependent covariates is

$$\begin{aligned} D = 2 \sup_h \sum \Bigg\{ \int (\ln e^{h_i' Z_i} &- \ln e^{\hat{\beta}' Z_i}) dN_i(s) \\ &- \left\{ \int Y_i(s) \left( e^{h_i' Z_i} - e^{\hat{\beta}' Z_i} \right) d\Lambda_0(s) \right\} . \end{aligned}$$

Because terms separate, we may optimize $h_i$ for each subject separately Applying a similar argument to that of lemma 2 1 to these samples of size 1 (with

$\hat{\Lambda} \equiv \exp(h_i' Z_i)\,\Lambda_0),$

$$\int_0^\infty Y_i(s) e^{h_i' Z_i} d\Lambda_0(s) = \int_0^\infty dN_i(s)$$

Let $\widetilde{M}_i(t) \equiv N_i(t) - \int_0^t \exp(\hat{\beta}' Z_i) d\Lambda_0(s)$, i.e , the martingale residual with $\beta$ estimated and $\Lambda$ known. Substituting gives

$$\begin{aligned} D &= -2 \sum \left\{ \widetilde{M}_i + \ln\left( \frac{e^{\hat{\beta}' Z_i}}{e^{h_i' Z_i}} \right) \int dN_i(s) \right\} \\ &= -2 \sum \left\{ \widetilde{M}_i + N_i(\infty) \ln\left( \frac{N_i(\infty) - \widetilde{M}_i}{N_i(\infty)} \right) \right\}. \end{aligned} \tag{9}$$

The last step above requires a factorization

$$\int Y_i(s) e^{\hat{\beta}' Z_i} d\Lambda_0(s) = e^{\hat{\beta}' Z_i} \int Y_i(s) d\Lambda_0(s)$$

which is not valid for time dependent $Z$.

For the Gaussian density the nuisance parameter $\sigma$ cancels out completely, not so here. We need to estimate $\Lambda_0$, which results in the replacement of $\widetilde{M}_i$ by $\widehat{M}_i$ in the formula. Equation (9) is equivalent to the deviance formula for a Poisson model found in McCullagh and Nelder (1983) with $y_i$ replaced by $N_i(\infty)$ and $\mu = \lambda t_i$ replaced by the observed cumulative hazard $\int Y_i(s) \exp(\hat{\beta}' Z_i) d\hat{\Lambda}_0$. The deviance residual is defined as the signed square root of this expression. Note that the deviance residual is zero if and only if $\widehat{M}_i = 0$ Also note that for the Cox model the deviance residuals are

$$d_i = \operatorname{sign}(\widehat{M}_i) \sqrt{-2(\widehat{M}_i + \delta_i \ln\left(\delta_i - \widehat{M}_i\right))}\,.$$

The log function "expands" residuals close to one, while the square root contracts the large negative values.

# 3 Functional Form

A key aspect of the model (2) is the functional form $\exp(\beta' Z)$ specified for the covariates Perhaps one of the variables $Z_j$ should be replaced by $\sqrt{Z}$, or by $I_{\{Z>c\}}$, or by some other transform? To investigate this for the semi-parametric model, consider a model with a single non time-dependent covariate and

$$\Lambda(t, Z) = h(Z)\Lambda_0(t)$$

for some unknown positive function h. We can think of the outcomes $(Y(\,).N(\cdot))$ as coming from a mixture distribution in Z with a crude hazard function

$$\begin{aligned}
\Lambda(t) &= \Lambda_0(t)\, E(h(Z) \text{ at time t}) \\
&= \Lambda_0(t)\, \frac{\int P(Y(t,z)=1)\, h(z) dF_Z(z)}{\int P(Y(t,z)=1)\, dF_Z(z)} \\
&\equiv \Lambda_0(t)\, \bar{h}(t)
\end{aligned}$$

where $F_Z(z)$ is the distribution function of $Z$ Then after fitting a model will all the variables except the $Z$ of interest (a null model in this case),

$$\begin{aligned}
E(\widehat{M}(t)|Z) &= E(N(t|Z)) + E\int_0^t -Y(s|Z) d\Lambda(s) \\
&\quad + E\int_0^t Y(s|Z)\,(d\Lambda(s) - d\hat{\Lambda}_0(s)) \\
&= term1 + term2 + term3
\end{aligned}$$

Consider these terms individually.

Term3· If there are no other covariates in the model, then

$$term3 = E\sum\int \frac{Y(s|Z)}{\sum Y_j(s)}(Y_i(s)d\Lambda(s) - dN_i(s))$$

Since $Y(s|Z)/\sum Y_j(s)$ is a predictable process, the entire term is the expectation of a mean zero martingale, so term3=0

Term2: Using the expression for $\Lambda$ above and then centering about $\bar{h}(t_o)$ for some fixed time $t_o$,

$$\begin{aligned}
term2 &= -E\int_0^t Y(s|Z)\,\frac{\bar{h}(t_o)}{h(Z)}\, h(Z)\, d\Lambda_0(s) \\
&\quad - E\int_0^t Y(s|Z)\,(\bar{h}(s) - \bar{h}(t_o))\, d\Lambda_0(s) \\
&= -\frac{\bar{h}(t_o)}{h(Z)} E(N(t)|Z) \;+\; \text{remainder.}
\end{aligned}$$

Thus

$$E(\widehat{M}(t)|Z) = \left(1 - \frac{\bar{h}(t_o)}{h(Z)}\right) E(N(t)|Z) \;+\; \text{remainder} \tag{10}$$

Equation 10 has a natural interpretation.

E(# excess deaths) ≈ (1 - hazard ratio) E(# events per subject)

Figures 1, 2, and 3 show the results of three calculations In each Z is uniformly spaced over (0,1), and $h(z)$ is $\exp(z)$, $\exp(5z)$, and $\exp(5z)$ respectively.

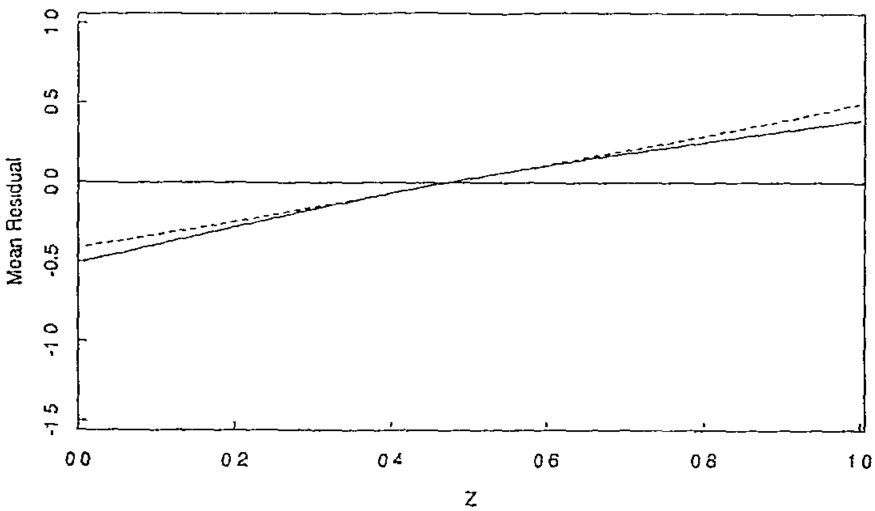


Figure 1· Martingale Residuals with $\beta = 1$ and no censoring

There is no censoring in figures 1 and 2, and a censoring of 50% in figure 3 (censoring is an independent uniform variable) The value of $t$ is $\infty$. Plotted are the actual value of $E(\widehat{M}|Z)$ (solid line) for a Cox model, calculated by a straightforward simulation with 1000 replications and a sample size per replication of 100, and the function $-\ln[1 - E(\widehat{M}|Z)/E(N|Z)]$ (dashed line) This latter is obtained by solving equation (10) for $\ln(h(Z)) - \ln(\bar{h}(t_o))$ This is the functional that should be placed in the exponent of a proportional hazards model (The term '$\ln(\bar{h}(t_o))$' is just a multiplicative constant, though, and would be absorbed into $\Lambda_0$ when a model is run) Note that $E(N|Z) \equiv 1$ for figures 1 and 2, and that $E(E(N(t)|Z)) = .5$ for figure 3.

In figures 1 and 3 the function $E(\widehat{M}|Z)$ is nearly straight, suggesting that when this is rewritten in the form of equation (2), with $f(z) \equiv \ln(h(z))$, that the further approximation is tolerable:

$$\begin{aligned}\left(1 - \frac{\bar{h}}{h(z)}\right) E(N(t)|Z) &= \left(1 - e^{\bar{f}-f}\right) E(N(t)|Z) \\ &\approx (f - \bar{f})\, E(E(N(t)|Z)) \\ &= (f - \bar{f}) \cdot \text{constant}\end{aligned}$$

That is, if the fit is not "pushed down" by the +1 boundary, a smoothed plot of the $\widehat{M}_i$ versus a covariate will give *approximately* the correct functional form to place in the exponent of a Cox model. A major advantage to plotting the "raw"

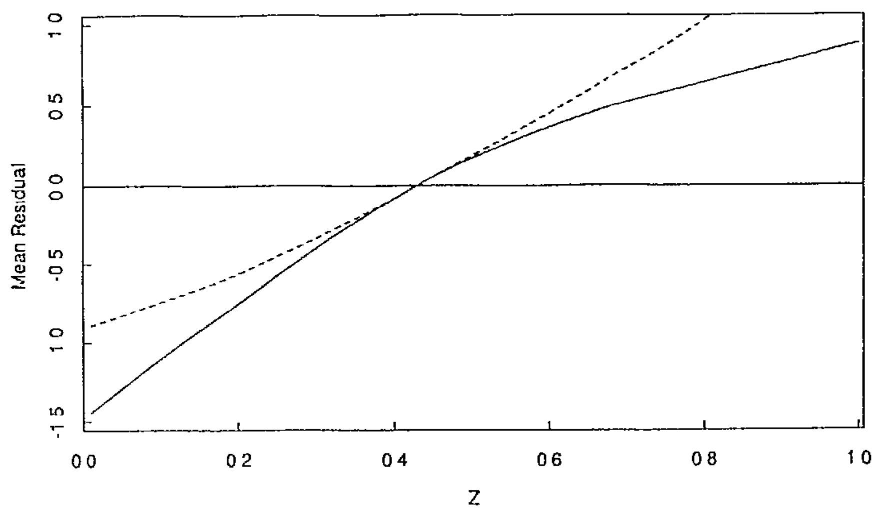


Figure 2: Martingale Residuals with $\beta = 5$ and no censoring

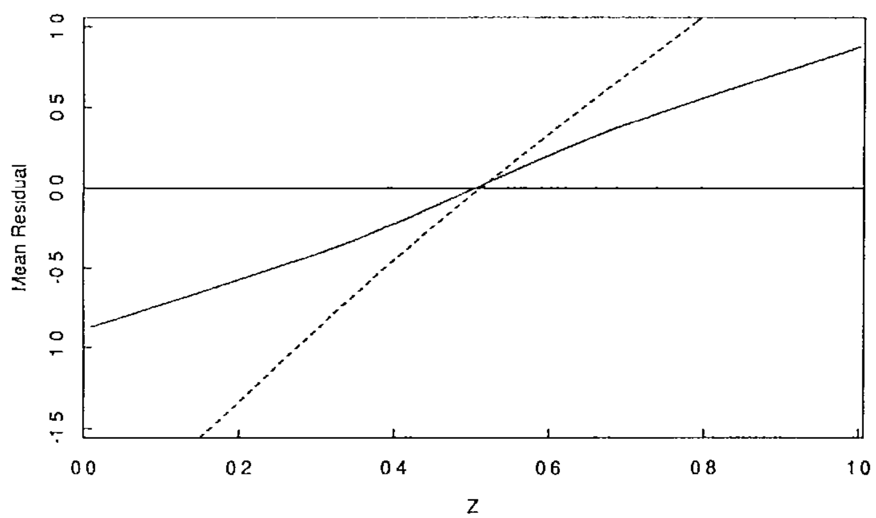


Figure 3. Martingale Residuals with $\beta = 5$ and 50% censoring

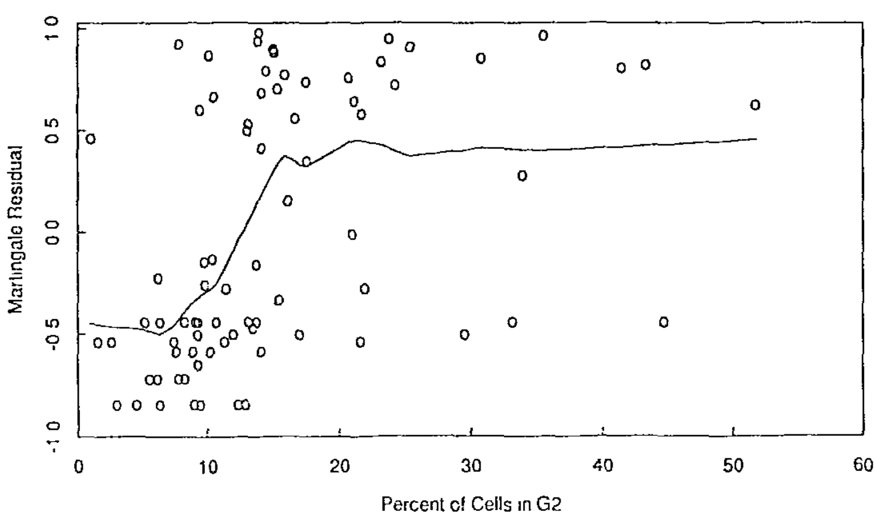


Figure 4: Martingale Residuals, Stage D1 Prostate Cancer

martingale residuals rather than the transformed function is interpretability, the y axis is in a direct scale of excess deaths.

Figure 4 shows the result of such a fit for a data set of patients with surgically treated stage D1 prostate cancer and is taken from Winkler et al (1988). The $x$ variable, percent of cells in g2 phase, is a measure of the proliferation rate of the resected tissue, and the y variable is the martingale residual from a null Cox model exploring time to recurrence of disease. Before the study began, a provisional cutoff of 13% had been set as "mean + 3 s.d " of the g2% values from 60 non-cancerous tissue specimens. A smooth fit to the $\widehat{M}_i$, using the *lowess* function of S (Becker and Chambers, 1984), bears out the initial guess.

One of the desirable features of assessing functional form through the martingale residuals is illustrated in figure 4 The display of the smooth fit in relation to the individual residuals provides insight into both the variability of and the influence of specific individuals on the estimate of the functional form.

The remainder term in (10) was small in our simulations, and one might expect it to always be so since it is based on a difference in means, a "1/n" effect compared to the lead term This is difficult to make precise without further restrictions in Y, however. Two special cases are

1. $Y(\cdot)$ independent of Z. Such would be the case in data from a Poisson process situation, where observation time for each object is not affected by the number of events produced. In this case $\bar{h}(t)$ is constant, and the remainder is zero.

2 A Cox model with uncensored data Then $Y(t)$ has exactly one jump from 1 to 0 for each subject, and $E(Y(t|Z)) = P[Y(t,Z) = 1] = \exp(-\Lambda_0(t)h(Z))$ Some further manipulation shows the remainder to be of the form

$$E_{A|Z}(\ln E_B e^{-AB})$$

where $A$ is the cumulative hazard at the time of failure, $B$ has mean zero, and $E(A) = 1$ A Taylor expansion of this has a leading term of O(1/n)

The argument given above extends to multiple covariates in a straightforward way. It does not apply to the parametric models, since $term3 \not\to 0$, nor to time dependent variates These latter areas need exploration.

# 4 Model Adequacy

An important use of residuals is in the graphical or analytical assessment of the validity of model assumptions. One such, functional form, has been discussed above. Three others, the limiting value of $\Lambda_0$, proportional hazards, and overall lack of fit are discussed below.

## 4.1 Crude and net hazards

One subtle model assumption relates to the interpretability of the function $\lambda(t; Z_i(t))$ By definition, this function satisfies the relationship

$$E\{N_i(t+dt) - N_i(t)|\mathcal{F}_t\} = Y_i(t)\,\lambda(t; Z_i(t))\,dt. \tag{11}$$

Thus, the interpretation of $\lambda$ is intrinsically tied to the censoring process $Y(\,)$. To understand the impact of this more clearly, consider the classical no covariate setting in which $T_i$ and $U_i$ represent an absolutely continuously distributed true survival time and censoring time for the ith subject. If $X_i = \min(T_i, U_i)$ is the observation time, suppose $Y_i(t) \equiv I_{\{X_i \geq t\}}$ and $N_i(t) \equiv I_{\{X_i \leq t, X_i = T_i\}}$ In this setting, it can be shown that the $\lambda$ in (11) is the "crude hazard",

$$\lambda_c(t) = \frac{-\frac{\partial}{\partial u}P(T \geq t; U \geq u)|_{u=t}}{P(T \geq t; U \geq u)},$$

and that the Breslow estimate in section 1.2, (which reduces to the Nelson (1969) estimate since there are no covariates), is a consistent estimator for $\int \lambda_c(s)ds$ The problem of interpretability arises in that one often interprets the parameter $\lambda$ in (11) in this classical setting to be the "net hazard"

$$\lambda_n(t) = \frac{-\frac{d}{dt}P(T \geq t)}{P(T \geq t)},$$

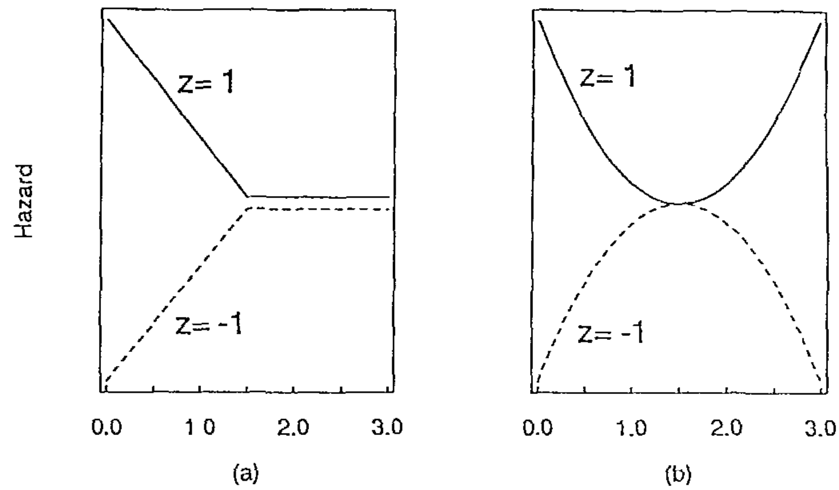


Figure 5: Two possible *non-proportional* hazards.

which is independent of U If one does wish to interpret $\lambda$ in (11) to be $\lambda_n$, then this would represent an additional assumption to the structure already imposed by (1) and (2) Unfortunately, this assumption is untestable (see Tsiatis 1975) using martingale residuals or any other approach.

## 4.2 Proportional Hazards

In this section, we will focus on the use of martingale and score residuals in the evaluation of the proportional hazards assumption, in the model where $Z(t)$ is independent of $t$.

For motivation, begin by considering the special case in which our model has a *single dichotomous* covariate, i.e., $Z = \pm 1$. In this setting, we wish to determine whether the hazard ratio $\lambda(t; Z = 1)/\lambda(t; Z = -1)$, estimated in the model to be $\exp(2\hat{\beta})$, is indeed independent of t. Consider the two non-proportional hazards situations illustrated in Figure 5 Because the martingale residuals sum to zero, it follows for $t_o$=0, 1, or 2 that

$$\begin{aligned}
&\sum_{i=1}^{n} I_{\{Z_i=1\}}\{\widehat{M}_i(t_o+1) - \widehat{M}_i(t_o)\} \\
&\qquad = -\sum_{i=1}^{n} I_{\{Z_i=-1\}}\{\widehat{M}_i(t_o+1) - \widehat{M}_i(t_o)\} \\
&\qquad \equiv A(t_o)
\end{aligned}$$

In either illustration (a) or (b), it is clear that $|A(t_o)|$ will be stochastically much larger than would be the case if proportional hazards were valid. Setting $\bar{Z}(b, s)$ as in (7), one might reject the proportional hazards assumption if a "large" value is obtained for $\sup_t \sum L_i(t)$ where

$$L_i(t) = \int_0^t \{Z_i - \bar{Z}(\hat{\beta}, s)\}\, d\widehat{M}_i(s)$$

If Z is any discrete or continuous covariate, this proportional hazards test statistic should be quite sensitive to alternatives for which

$$\frac{\lambda(t; Z = \beta)}{\lambda(t; Z = \alpha)}\ \forall \alpha < \beta \text{ is monotonically strictly decreasing (increasing) in } t. \tag{12}$$

To derive the distribution of the statistic $\sup_t \sum L_i(t)$ , consider $U(\hat{\beta}, t)$, the partial likelihood score statistic (using information over [0,t]). Then,

$$\sum_{i=1}^{n} L_i(t) = U(\hat{\beta}, t).$$

In turn, $\sum_{i=1}^{n} L_i(t) = 0$ for $t = 0$ and $t = \infty$, by the definition of $\hat{\beta}$ The next lemma establishes that a standardized version of this process converges asymptotically to a tied down Brownian Bridge process The lemma also addresses the $p > 1$ covariate vector situation and generalizes an earlier result obtained by Wei(1984)

**Lemma 4.1** As in section 2.1 let $U(\beta, t)$ denote the score vector process and $\hat{\beta}$ denote the maximum partial likelihood estimate of $\beta$ Define the information matrix

$$\mathcal{I}(\beta, \cdot) \equiv \sum_{i=1}^{n} \int_0 \left\{ \frac{\sum Y_j(s) Z_j Z_j' w_j}{\sum Y_j(s) w_j} - \left( \frac{\sum Y_j(s) Z_j w_j}{\sum Y_j(s) w_j} \right)^{\otimes 2} \right\} dN_i(s),$$

where $w_i \equiv \exp(\beta' Z_i)$, and $a^{\otimes 2} = aa'$ for any column vector $a$. Denote the in probability limit of $n^{-1}\mathcal{I}(\beta, \cdot)$ by $\Sigma(\cdot)$ (defined in Anderson and Gill(1984)). Then

a. Let $B(\cdot)$ be a mean zero vector of Gaussian processes having independent increments and covariance matrix $\Sigma(\ )$ Then under regularity conditions specified in Anderson and Gill(1984),

$$\frac{1}{\sqrt{n}} U(\hat{\beta}, \cdot) \Rightarrow B(\cdot) - \Sigma(\cdot)\{\Sigma(\infty)\}^{-1} B(\infty) \tag{13}$$

where $\Rightarrow$ denotes weak convergence over the relevant interval

b. If $(\Sigma(t))_{jk} = 0$ for all $k \neq j$ and for any $t$, then for $j = 1, 2, \ldots, p$,

$$\sqrt{\mathcal{I}^{-1}(\hat{\beta}, \infty)_{jj}}\, U(\hat{\beta},\ )_j \Rightarrow W^{\circ} \left( \frac{\sigma_{jj}(t)}{\sigma_{jj}(\infty)} \right) \tag{14}$$

where $\sigma_{jj}(t) \equiv \Sigma(t)_{jj}$ and where $\{W^{\circ}(t) : 0 \leq t \leq 1\}$ is distributed as a Brownian Bridge

**Proof:**

**a.** By Taylor's series expansion,

$$\frac{1}{\sqrt{n}}U(\hat{\beta},\cdot) = \frac{1}{\sqrt{n}}U(\beta,\cdot) - \frac{1}{n}\mathcal{I}(\beta^*,\cdot)\sqrt{n}(\hat{\beta}-\beta), \tag{15}$$

for some $\beta^*$ on a line segment connecting $\beta$ and $\hat{\beta}$. In turn,

$$\sqrt{n}(\hat{\beta}-\beta) = \left\{\frac{1}{n}\mathcal{I}(\beta^*,\infty)\right\}^{-1} \frac{1}{\sqrt{n}}U(\beta,\infty).$$

Inserting this into (15), we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}U(\hat{\beta},\cdot) &= \frac{1}{\sqrt{n}}U(\beta,\ ) \\ &\quad - \frac{1}{n}\mathcal{I}(\beta^*,\cdot)\left\{\frac{1}{n}\mathcal{I}(\beta^*,\infty)\right\}^{-1} \frac{1}{\sqrt{n}}U(\beta,\infty).\end{aligned}$$

Now (13) follows from Anderson and Gill's results, based upon the martingale structure of $U(\beta,\ )$, which establishes that $\frac{1}{\sqrt{n}}U(\beta,\ ) \Rightarrow B(\cdot)$ and $\frac{1}{n}\mathcal{I}(\beta^*,\ ) \Rightarrow \Sigma(\cdot)$ in probability.

**b.** To obtain the large sample covariance structure of $\frac{1}{\sqrt{n}}U(\hat{\beta},\cdot)$, observe for any set $s \le t$

$$\begin{aligned}&E([B(s) - \Sigma(s)\{\Sigma(\infty)\}^{-1}B(\infty)][B(t) - \Sigma(t)\{\Sigma(\infty)\}^{-1}B(\infty)]') \\ &\qquad = \Sigma(s) - \Sigma(s)\{\Sigma(\infty)\}^{-1}\Sigma(t)\end{aligned} \tag{16}$$

When $(\Sigma(t))_{jk} = 0$ for any $k \neq j$ and for any $t$, (14) follows from (16) and Andersen and Gill's result that $\frac{1}{n}\mathcal{I}(\hat{\beta},\infty) \Rightarrow \Sigma(\infty)$ in probability. □

When the jth component of the covariate vector satisfies the proportional hazards assumption, (14) indicates that the proportional hazards test statistic $\sqrt{\mathcal{I}^{-1}(\hat{\beta},\infty)_{jj}}\ \sup_t \sum_i L_{ij}(t)$ asymptotically has the well known distribution of $\sup_{0\le t\le 1} W_0(t)$, as long as $(\Sigma(t))_{jk} = 0$ for any $t$ In essence, this condition requires $Z_j$ to be orthogonal to the other covariates. In fact, by its definition in Lemma 4 1, the consistent estimator $\frac{1}{n}\mathcal{I}(\hat{\beta},\infty)$ of $\Sigma(\infty)$ can be interpreted to be the sum over death times of the covariance of $Z$ at each death time. Thus, for example, $(\Sigma(t))_{jk} \approx 0$ in intervention studies in which the jth covariate represents randomly assigned treatment, as long as strong treatment by factor interactions do not exist. Further efforts are necessary to address the situation in which the assumption $(\Sigma(t))_{jk} = 0$ fails to hold

For the parametric model, analogous results hold. A proportional hazards test statistic based on the standardized supremium of the score process $\sum_i L_{ik}(\ )$ also is distributed asymptotically as a time transformed Brownian bridge

When one has adequate data, it is often desirable to have flexible graphical and analytical methods to detect more general proportional hazards departures not characterized by (12), such as the alternative in figure 5(b). By choosing band widths $\Delta$ and $\delta$, we can make graphical assessments by plotting

$$f_{\Delta,\delta}(x,t) \equiv \sum_{i=1}^{n} I_{\{x-\Delta \le Z_i \le x+\Delta\}} \int_{t-\delta}^{t+\delta} d\widehat{M}_i(s)$$

as a function of $t$, for selected values of $x$. For discrete covariates, one can set $\Delta = 0$. Trends in the plots of $f_{\Delta,\delta}(x,\cdot)$ signal the nature of the departure from proportional hazards. To enable analytical inferences, one can obtain an expression for the conditional distribution of any term

$$T_A(s,t) \equiv \sum_{i\in A} \int_s^t d\widehat{M}_i(u) ,$$

where A is any subset of $\{1, 2, \ldots, n\}$ and $s \le t$ Specifically, $T_A(s,t)$ can be thought of as a sum over the L distinct failure times occurring over the interval $(s,t]$. At the lth of these L failure times, $t_{(l)}$, $\sum_{i\in A} \Delta N_i(t_{(l)})$ is the number of failures occurring in the set $A$. In turn, $\sum_{i\in A} \Delta N_i(t_{(l)})$ has the distribution arising from sampling $\sum_{i\in A} Y_i(t_{(l)})$ items without replacement from a set of $\sum_{i=1}^{n} Y_i(t_{(l)})$ items, which includes $\sum_{i=1}^{n} \Delta N_i(t_{(l)})$ total failures, and where each item has a relative probability $Y_i(t_{(l)})e^{\hat{\beta}' Z_i} / \sum_k Y_k(t_{(l)})e^{\hat{\beta}' Z_k}$ of being sampled In particular then, $\sum_{i\in A} \Delta N_i(t_{(l)})$ has expectation

$$\begin{aligned}\sum_{i\in A} \left[ \left\{ \frac{Y_i(t_{(l)})e^{\hat{\beta}' Z_i}}{\sum_k Y_k(t_{(l)})e^{\hat{\beta}' Z_k}} \right\} \sum_k \Delta N_k(t_{(l)}) \right] \\ = \sum_{i\in A} Y_i(t_{(l)})\, e^{\hat{\beta}' Z_i}\, \Delta\hat{\Lambda}_0(t_{(l)}) ,\end{aligned}$$

so indeed $T_A(s,t)$ has zero expectation in this sampling framework. Finally, the distribution of $T_A(s,t)$ is obtained by taking $\{\sum_{i\in A} \Delta N_i(t_{(l)}) : l = 1, 2, \ldots, L\}$ to be a collection of independent random variables

Many other methods for testing proportional hazards have been proposed, notably by Schoenfeld(1980), Andersen (1982), and Aranda-Osdaz (1983). One advantage of the the statistic in (14) is the lack of the need for an arbitrary discretization of the continuous time axis.

As an illustration of these ideas, we will use a data set which has been collected to model survival in patients suffering from primary biliary cirrhosis, a chronic and eventually fatal liver disease (Dickson, et al, 1988). A population of 418 patients was followed from the date of their referral to a tertiary care center until death or censoring at study closure. There were 161 deaths. An extensive database of medical variables measured at the time of referral is available A Cox

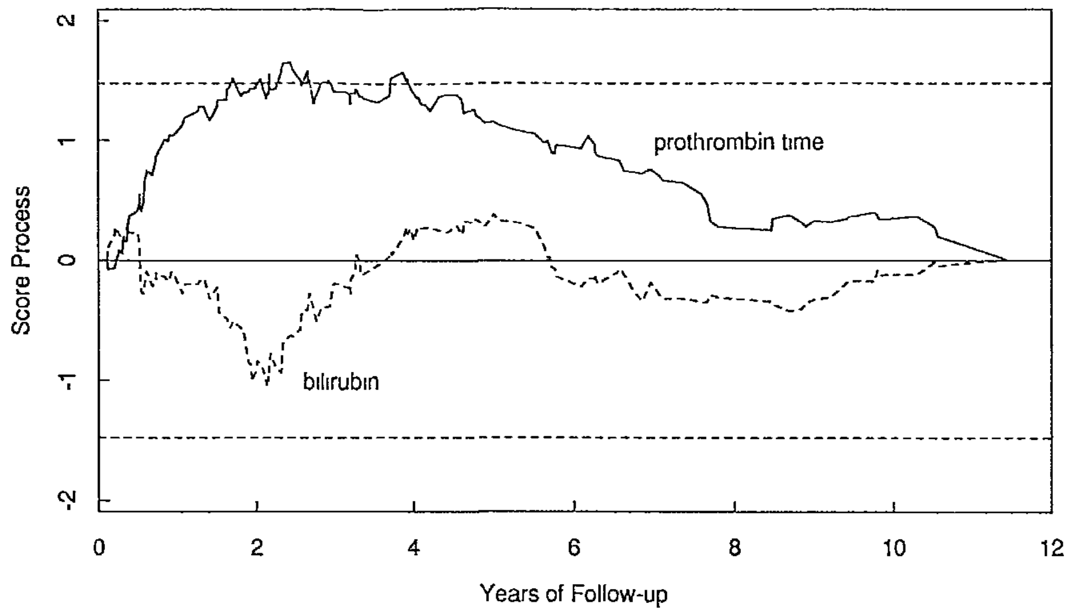


Figure 6. $\sum L_i(t)$ for two predictors of liver disease

regression model using five of the covariates — total serum bilirubin, albumin, prothrombin time, age, and edema — was found to fit the survial experience rather well. Figure 6 shows plots of the standardized score process,

$$\sqrt{\mathcal{I}^{-1}(\hat{\beta},\infty)_{jj}}\ \sum_i L_{ij}(t),$$

as a function of $t$ for two of the predictors. If the proportional hazards assumption is correct, we would expect each of these plots to be a tied down random walk, this may be true for bilirubin, but the pattern in the process for prothrombin time is obvious. One possible explanation is that in this disease prothrombin time can be readily modified by drug therapy, but bilirubin can not. The critical values for the supremum of a Brownian Bridge are also indicated on the plots (see Koziol and Byar (1975)). Because the predictor variables in this data set are mildly correlated, the critical values may need some adjustment

The increments in the (unstandarized) process are the ***partial residuals*** introduced by Schoenfeld (1982). Another test that may be applied, therefore, is one proposed by Harrell (1986). This is based on the Pearson correlation between the partial residuals and the rank order of the failure times, along with the standard z-transform of Fisher. When applied to this data set, the z-value for prothrombin time was -4 64 (p<.0001) and for bilirubin the value was 0.78 (p=.44).

### 4.3 Overall Measure of Fit

In more standard parametric models, the overall "size" of the residuals gives a clue to the overall fit of the model, and this holds for the parametric proportional hazards models also For a series of models with the same $\theta$ in each, the sum of the squared deviance residuals can be used as a surrogate for the log likelihood; the difference in this sum for two models will be a chi-square statistic on the appropriate degrees of freedom For a series of Cox models A, B, C, ..., however, the estimation of $\hat{\beta}_A, \hat{\beta}_B, \ldots$ by partial likelihood implies a reestimation of $\Lambda_0$ for each. The sums $D = \sum d_i^2$ cannot be used as a surrogate for the log likelihoods as is done in GLIMs, because $\Lambda_0$ does not cancel out in the derivation above (section 2.4). In fact, we have found in examples that the change $\Delta D$ between two nested models does not necessarily correlate with the change in partial likelihood.

The "lack of size" condition is stronger than this experience. As pointed out by Crowley and Hu (1977), when there is no censoring the values of the Breslow estimate at each event time are exactly the order statistics from an exponential distribution. The martingale residuals at $\hat{\beta} = 0$ have a distribution of (1 - exponential order statistics), while the martingale residuals evaluated at $\hat{\beta} = \beta$ have distribution of (1- exponential data sample) Thus, for uncensored data at least, the global distribution of the martingale residuals is the same under null and perfect models

# 5 Influential Observations

The influence of an observation on model fit depends on both the residual from the fit and on the extremity of its covariate value, roughly $(Z_i - \bar{Z}) *$ residual. In the Anderson-Gill model specified by (2), $\bar{Z}$ is a function of time: the mean over the risk set at time t (see (7) above) This suggests using a "time average" value of $Z_{ij} - \bar{Z}_j$, which leads to the score residual

$$L_{ij} \equiv \int_0^\infty (Z_{ij}(s) - \bar{Z}_j(\hat{\beta}, s))\, d\widehat{M}_i(s)$$

as an influence measure

To formalize this, we may use the approach of Cain and Lange (1984) and define a weighted score vector

$$U(\hat{\beta}, w) \equiv \sum_{i=1}^{n} w_i \int_0^\infty Y_i(t)\,(Z_{ij}(t) - \tilde{Z}_j(t))\, dN_i(t)$$

where $\tilde{Z}$ is the reweighted mean at $\hat{\beta}$

$$\tilde{Z}_j(t) = \frac{\sum Y_l(t) w_l e^{\hat{\beta}' Z_l(t)} Z_{lj}(t)}{\sum Y_l(t) w_l e^{\hat{\beta}' Z_l(t)}}.$$

Then

$$\frac{\partial\hat\beta}{\partial w_i} = \left(\frac{\partial\hat\beta}{\partial U}\right)\left(\frac{\partial U}{\partial w_i}\right) = -\mathcal{I}(\hat\beta)^{-1}\frac{\partial U}{\partial w_i};$$

evaluation of this quantity at $w = 1$ is the infinitesimal jackknife estimate of influence. In our case

$$\begin{aligned}\frac{\partial U_j}{\partial w_i} &= \int_0^\infty Y_i(s)\,(Z_{ij}(s) - \bar Z_j(s))\,dN_i(s)\\ &\quad - \sum_l w_l \int_0^\infty Y_l(s)Y_i(s)\,(Z_{ij}(s) - \bar Z_j(s))\left(\frac{e^{\beta' Z_i(s)}}{\sum_k Y_k(s) w_k e^{\beta' Z_k(s)}}\right) dN_l(s)\end{aligned}$$

The last term in the second integrand is just the component of the Breslow estimate of $\Lambda_0(s)$, so that

$$\begin{aligned}\left.\frac{\partial U_j}{\partial w_i}\right|_{\substack{w=1\\ \beta=\hat\beta}} &= \int Y_i(s)\,(Z_{ij}(s) - \bar Z_j(\hat\beta, s))\,(dN_i(s) - e^{\hat\beta' Z_i} d\hat\Lambda_0(s)) \qquad (17)\\ &= L_{ij}\end{aligned}$$

In the special case of a Cox model (17) reduces to equation 4 of Cain and Lange and so generalizes their work. The influence of the ith subject on the estimation of $\beta$ is then approximately the Newton-Raphson step $-\mathcal{I}^{-1}(\hat\beta)(L_{i1}, L_{i2}, \ldots, L_{ip})'$ A similar, though simpler, derivation holds for the parametric models and yields the score residual $L_{ij}$ defined in (8)

This method may underestimate the true jackknife, especially for extreme values of z, because $\mathcal{I}$ also changes when the observation is removed. Another method is to compute the 1-step update in $\hat\beta$ when a single covariate $Z_{p+1}$ is added, with $Z_{p+1}$ equal to 1 for subject $i$ and equal to 0 for all others This is explored for the Cox model by Storer and Crowley (1985). For the Anderson-Gill model at $(\hat\beta^{(1\mathrm{xp})}, 0)$,

$$\begin{aligned}\bar Z_{p+1}(\hat\beta, t) &= \frac{Y_i(t) e^{\hat\beta' Z_i}}{\sum_k Y_k(t) e^{\hat\beta' Z_k(t)}}\\ U_j &= 0 \text{ for } j = 1, 2, \ldots, p \quad (\text{since we are at } \hat\beta)\\ U_{p+1} &= \sum_l \int_0^\infty Y_l(s)\,(Z_{l,p+1}(s) - \bar Z_{p+1}(\hat\beta, s))\,dN_l(s)\\ &= \int_0^\infty dN_i(t) - \int_0^\infty Y_i(s)\, e^{\hat\beta' Z_i(s)}\, d\hat\Lambda_0(s)\\ &= \widehat{M}_i\end{aligned}$$

The same process for the new information matrix yields

$$\mathcal{I}_{new} = \begin{pmatrix} \mathcal{I}(\hat\beta) & \gamma_i \\ \gamma_i' & \eta_i \end{pmatrix}$$

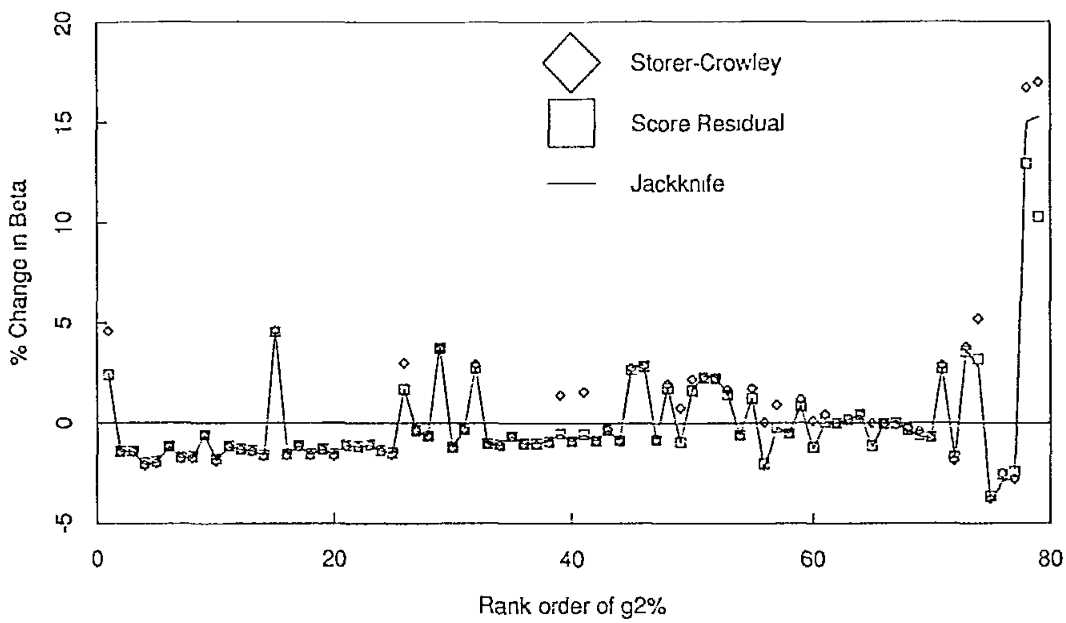


Figure 7· Comparison of two approximate measures with the jackknife

where

$$\begin{aligned}\gamma_{i,j} &= \int_0^\infty Y_i(s)\,(Z_{ij}(s) - \bar{Z}_j(\hat{\beta}, s))\, e^{\hat{\beta}' Z_i(s)}\, d\hat{\Lambda}_0(s) \\ \eta_i &= \int_0^\infty Y_i(s)\,(1 - \bar{Z}_{p+1}(\hat{\beta}, s))\, e^{\hat{\beta}' Z_i(s)}\, d\hat{\Lambda}_0(s)\,.\end{aligned}$$

Then the change in $\beta$ is $(-\mathcal{I}_{new})^{-1}U$, and using a standard formula for the inverse of a partitioned matrix:

$$\hat{\beta}_{(i)} = \frac{-\mathcal{I}(\hat{\beta})^{-1}\gamma_i}{\eta_i - \gamma_i'\mathcal{I}(\hat{\beta})^{-1}\gamma_i}\widehat{M_i}\,,$$

which extends the results of Storer and Crowley to the Anderson-Gill model.

In practice, the two forms are not very far apart; high leverage points are highlighted by both For the prostate cancer data presented earlier, figure 7 presents the results of the actual jackknife, score residual, and one-step influence measures after fitting the variable *%g2* as a linear covariate. Interestingly, there are a few subjects for which the Storer-Crowley approach gives the wrong sign, but they are all of small leverage (the value of $\beta_{p+1}$ is grossly overestimated at the first step for each of these subjects, compared to its value if iteration is allowed to continue).

Figure 8 illustrates an influential point in the liver disease data set. The score residuals for the variable age are plotted against age, and show that the oldest

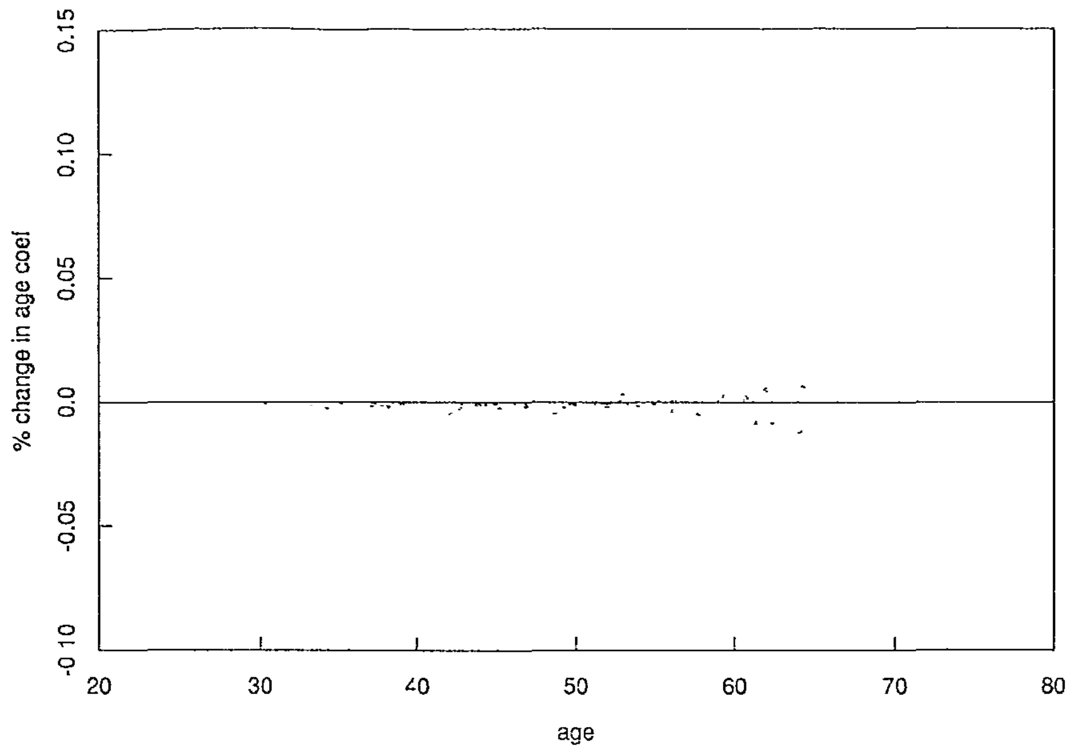


Figure 8· Score residuals from the cirrhosis data set

individual has a disparate amount of influence on the coefficient Interestingly, this observation led to the identification of a data error: the true age of the patient was 54, not 78.

Though the score and Storer-Crowley residuals are similar in numeric magnitude, the score residuals $L_{ij}$ have several technical advantages:

a) There is a simplicity of interpretation as components of the score statistic.

b) They are available for all values of $\beta$, not just the solution point $\hat{\beta}$. For instance, at $\beta = 0$ they are components of the log-rank statistic.

c) As a martingale transform, powerful theoretical tools are available. Computation of variance, for instance, is a simple exercise.

# 6 Model Accuracy for Individual Subjects

An important use of residuals is in graphical assessment of poor prediction by a model for individual subjects. The size of the individual's residual $\widehat{M_i}$ indicates model accuracy with a large positive value for a subject who has more events than predicted by the model (dies "too soon") and a large negative residual for any with fewer events than predicted by the model (lives "too long"). In the one event models such as the Cox model, the martingale residuals are heavily skewed and this skewness distorts the appearance of a standard residual plot.

It is nearly impossible to detect outliers of the "died too early" type because so many points are crowded up close to the value +1 A point with value .99999 does not appear any different than one with value of .9 The long right hand tail of the martingale residuals may also produce spurious outliers among those who "live too long" The deviance transform symmetrizes the martingale residuals and helps to alleviate this problem When censoring is minimal, <25% or so, the distribution of the deviance residuals is very close to a normal distribution. For censorings greater than 40%, a large bolus of points with residuals near 0 distorts the normal approximation, but the transform is still helpful in symmetrizing the set of residuals.

Figure 9 compares the martingale and deviance residuals for the liver disease data set presented earlier For each individual in the data set we have computed both the residuals and the risk score $\hat{\beta}' z_i$. Panel A shows the martingale residuals plotted against the risk score and panel B the deviance residuals. The deviance transform suggests that the 3 individuals (with risk score $\approx$ 8) who look like outliers in the martingale plot are, in fact, not outliers at all The heavy censoring in this data (62%) makes the normality of the deviance residuals' tails somewhat suspect; one might wish to further check the patients with the 2 largest and 2 smallest residuals as a precaution The latter two patient's values are not even distinguishable in the first plot.

Simulation results have shown that constructed outliers in the form of subjects who "live too long" are readily detected by the either the deviance or martingale residuals, though the scaling is visually more interpretable in the former Outlier subjects who "died too early", however, can be seen only in the deviance transform, and even then not always reliably This seems to be because in a proportional hazards framework even subjects with a very low risk have an appreciable probability of dying early. In a semi-parametric model, the automatic scaling afforded by the Breslow estimate virtually guarantees that a singleton small outlier will go unnoticed.

# 7 Discussion

We have defined a residual applicable to both parametric and semi-parametric proportional hazards models which is effective for exploration of functional form, model validity, leverage, and fit of individual subjects. The martingale formulation gives these residuals a strong theoretical underpinning and allows rigorous investigation of their properties. Computation of the residuals and their transforms is straightforward, and can easily be added to existing computer routines for the Cox or other proportional hazards models.

For any single one of the uses outlined above, it might be argued that a better method exists, e.g., actual jackknife values for assessing leverage, or estimating functional form by directly maximizing the likelihood over a spline or other flexible curve. A readily available residual can have unforseen benefits, however.

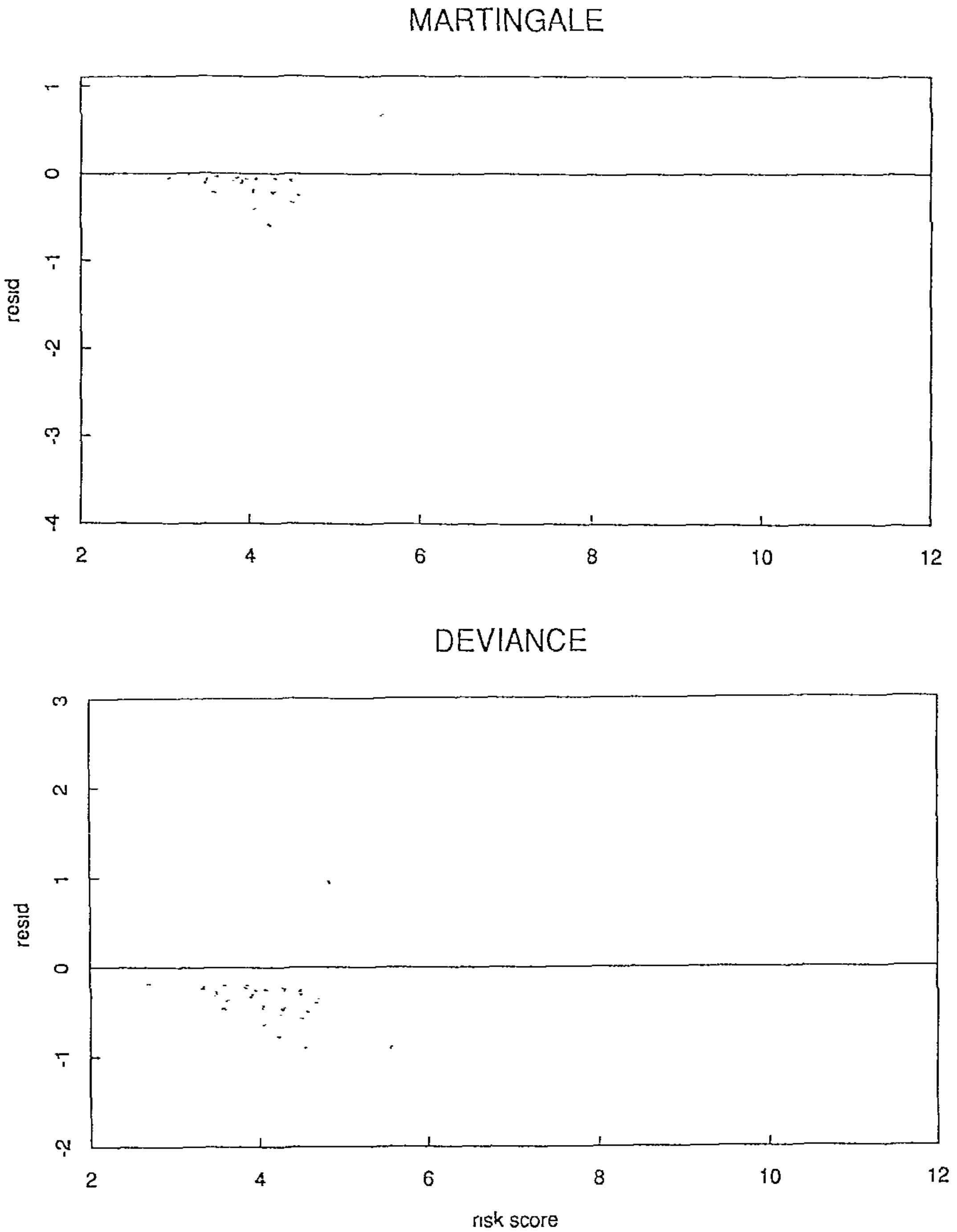


Figure 9: Martingale and deviance residuals for the cirrhosis data

An example from our own experience was the discovery that martingle residuals from a null Cox model could be used as input to the CART (Classification and Regression Trees) model of Breiman, et. al (1984), and that the marriage seems to work quite well This has allowed the direct use for survival data of a methodology designed for a continuous y variate, without a major overhaul of the algorithm or its computer code. In one particular data set, the first splits produced by CART appeared to be mimicking a linear age effect. This was verified using the plots of section 3 above, and CART re-run using residuals from a model that included age. Interactions such as this may be useful for other analysis methods as well

While this paper was in draft, we became aware of some related work by Barlow and Prentice (1988), which includes a more thorough discussion of the material in our §2.3 and 2.4 for the semi-parametric case, and also has some overlap with our §5.

# References


[1] Andersen, P K (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics* **38**:67-77

[2] Andersen, P K. and Gill, R.D. (1982) Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**:1100-1120

[3] Aranda-Ordaz, F J (1983) An extension of the proportional hazards model for grouped data *Biometrics* 39·109-117.

[4] Barlow, W.E. and Prentice R L. (1988) Residuals for relative risk regression. *Biometrika* (to appear)

[5] Becker, R A and Chambers, J.M (1984) *S. An Interactive Environment for Data Analysis and Graphics*, Wadsworth, Belmont.

[6] Breiman, L ; Friedman, J.; Olshen, R ; and Stone C (1984) *Classification and Regression Trees*, Wadsworth, Belmont.

[7] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**:80-99

[8] Cain, K.C. and Lange, N T. (1984) Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* **40**:493-499.

[9] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant data *J. American Stat. Assoc.* **72**:27-36.

[10] Cox, D R. (1972) Regression models and life tables (with discussion) *J. R. Statistical Soc. B* **34**:187-220

[11] Dickson, E.R.; Grambsch, P., Fleming, T R ; Fisher, L.D., and Langworthy, A (1988). Prognosis in primary biliary cirrhosis· Model for decision making. In submission

[12] Harrell, F. (1986). The PHGLM procedure. *SAS Supplimental Library User's Guide*, Version 5. SAS Institute Inc., Cary, N.C

[13] Kay, R. (1977). Proportional hazards regression models and the analysis of censored survival data. *Applied Statistics* **26**:227-237.

[14] Koziol, J A. and Byar, D.P. (1975). Percentage points of the asymptotic distributions of one and sample K-S statistics for truncated or censored data. *Technometrics* **17**:507-510.

[15] McCullagh, P. and Nelder, J A. (1983). *Generalized Linear Models* Chapman and Hall, London.

[16] Nelson, W (1969) Hazard plotting for incomplete failure data *J. Quality Tech.* **1**:27-52.

[17] Schoenfeld, D. (1980) Chi-squared goodness of fit tests for the proportional hazards regression model *Biometrika* **67**:145-153.

[18] Schoenfeld, D (1982) Partial residuals for the proportional hazards regression model *Biometrika* **69**:239-241

[19] Storer, B.E. and Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods *J. American Stat. Assoc.* **80**:139-147

[20] Tsiatis, A.A. (1975). A non-identifiability aspect of the problem of competing risks. *Proc. National Academy Sci.* **72**, No 1.

[21] Wei, L.J (1984) Testing goodness of fit for the proportional hazards model with censored observations *J. American Stat Assoc* **79** 649-652

[22] Winkler, H.Z , Rainwater, L M.; Meyers, R P.; Farrow, G M ; Therneau, T M., Zinke, H., and Lieber, M.M. (1988). Stage D1 Prostatic Adenocarcinoma: Significance of Nuclear DNA Ploidy Patterns Studied by Flow Cytometry. *Mayo Clinic Proceedings* **63**.103-112.