



## Partial Residuals for The Proportional Hazards Regression Model

David Schoenfeld

*Biometrika*, Vol. 69, No. 1. (Apr., 1982), pp. 239-241.

Stable URL:

<http://links.jstor.org/sici?sici=0006-3444%28198204%2969%3A1%3C239%3APRFTPH%3E2.0.CO%3B2-3>

*Biometrika* is currently published by Biometrika Trust.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/bio.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

## Partial residuals for the proportional hazards regression model

BY DAVID SCHOENFELD

*Harvard School of Public Health, Sidney Farber Cancer Institute, Boston,  
 Massachusetts, U.S.A.*

### SUMMARY

Residuals are defined for the proportional hazards regression model introduced by Cox (1972). These residuals can be plotted against time to test the proportional hazards assumption. Histograms of these residuals can be used to examine fit and detect outlying covariate values.

*Some key words:* Censoring; Failure time data; Proportional hazard; Residual.

### 1. INTRODUCTION

The proportional hazards regression model (Cox, 1972) provides a method of estimating the effect of covariates on failure time. Kay (1977) derived residuals for this model (Cox & Snell, 1968). The present paper defines residuals which do not depend on time so that the  $i$ th residual can be plotted against  $t_i$  to test the proportional hazards assumption. Furthermore, they do not involve an estimated hazard function and this simplifies their asymptotic distribution.

### 2. DEFINITION OF THE PARTIAL RESIDUALS

Suppose  $n$  individuals are indexed by  $i = 1, \dots, n$  and that each has a  $p$ -vector of covariates  $X_i = (X_{i1}, \dots, X_{ip})'$ . The proportional hazards regression model specifies that the hazard function of the  $i$ th individual is

$$h_i(t) = \lambda_0(t) \exp(\beta' X_i), \quad (1)$$

where  $\beta$  is a vector of  $p$  parameters and  $\lambda_0(t)$  is an arbitrary function.

Let  $D$  be the indices of the individuals who failed and let  $R_i$  be the indices of those under observation when the  $i$ th individual fails. Using partial (Cox, 1975) or marginal (Kalbfleisch & Prentice, 1980, p. 71) likelihood arguments one can estimate  $\beta$  by maximizing the likelihood function of the following model: for  $i \in D$ , an index  $m \in R_i$  is selected with probability

$$\exp(\beta' X_m) / \sum_{k \in R_i} \exp(\beta' X_k).$$

In this model  $X_i$  is a random variable with

$$E(X_{ij} | R_i) = \sum_{k \in R_i} X_{kj} \exp(\beta' X_k) / \sum_{k \in R_i} \exp(\beta' X_k),$$

and the maximum likelihood estimate of  $\beta$  is a solution to

$$\sum_{i \in D} \{X_{ij} - E(X_{ij} | R_i)\} = 0.$$

Denote this solution by  $\hat{\beta}$  and let  $\hat{E}(X_{ij} | R_i)$  be  $E(X_{ij} | R_i)$  with  $\hat{\beta}$  substituted for  $\beta$ .

Define the partial residual at  $t_i$  as the vector  $\hat{r}_i = (\hat{r}_{i1}, \dots, \hat{r}_{ip})'$ , where  $\hat{r}_{ik} = X_{ik} - \hat{E}(X_{ik} | R_i)$ . Thus the residual is the difference between the observed value of  $X_i$  and its conditional expectation given  $R_i$ . Since  $\hat{r}_i$  is a vector each  $\hat{r}_{ik}$  must be examined graphically. This is feasible wherever computer graphics are available.

### 3. THE ASYMPTOTIC DISTRIBUTION OF $r_i$

Define  $r_i$  to be  $\hat{r}_i$  with  $\beta$  replacing  $\hat{\beta}$ . The  $\{r_i\}$  will have discrete distributions determined by  $R_i$  and will be uncorrelated (Cox, 1975). Let  $U_i$  be the  $p \times p$  matrix with  $(k, s)$ th element  $\partial r_{ik} / \partial \beta_s$  evaluated at  $\hat{\beta}$ . Furthermore, let  $U = \sum_{i \in D} U_i$ . When  $\hat{r}_i$  is expanded about  $\beta$  in a Taylor series,

$$\hat{r}_i = r_i + U_i(\hat{\beta} - \beta) + o_p(n^{-\frac{1}{2}}).$$

When the score statistic is substituted for  $\hat{\beta} - \beta$  this yields

$$\hat{r}_i = r_i + U_i U^{-1} \sum_{k \in D} r_k + o_p(n^{-\frac{1}{2}}),$$

which expresses the  $\hat{r}_i$ , which depend on  $\hat{\beta}$ , in terms of the  $r_i$ .

Since the  $r_i$  are uncorrelated the variance covariance matrix of  $\hat{r}_i$  and  $\hat{r}_j$  is asymptotically  $\delta_{ij} U_i - U_i U^{-1} U_j'$  which can be used to find the variance of functions of the  $\{\hat{r}_j\}$ . Since  $U^{-1} \rightarrow 0$  the  $\hat{r}_i$  are asymptotically uncorrelated.

### 4. EXAMINING THE PROPORTIONAL HAZARDS ASSUMPTION

If proportional hazards holds  $E(\hat{r}_i) \simeq 0$  and a plot of  $\hat{r}_{ik}$  versus  $t_i$  will be centred about 0. However, suppose that

$$h_i(t) = \lambda_0(t) \exp \{ \beta' X_i + \theta g(t_i) X_{ik} \}$$

with  $g(t_i)$  varying about 0. Expanding  $E(X_{ik} | R_i)$  about  $g(t_i) = 0$ , we have

$$E(\hat{r}_{ik}) \simeq g(t_i) \{ E(X_{ik}^2 | R_i) - E(X_{ik} | R_i)^2 \}.$$

Since the term in brackets is positive the sign of  $E(\hat{r}_{ik})$  will depend on the sign of  $g(t_i)$ . Thus changes in  $g(t_i)$  will be reflected in a plot of  $\hat{r}_{ik}$  versus  $t_i$ .

### 5. EXAMINING GOODNESS OF FIT

In order to obtain residuals which have an approximately uniform distribution, let  $A_{ik}$  be the set of identifiers  $j \in R_i$  such that  $X_{jk} \leq X_{ik}$ . Define the uniform partial residual

$$\hat{s}_{ik} = \sum_{j \in A_{ik}} \exp(\hat{\beta}' X_j) / \sum_{j \in R_i} \exp(\hat{\beta}' X_j).$$

The residual  $\hat{s}_{ik}$  will have a discrete uniform distribution at each  $R_i$ . If there are many values of  $X_i$ , the histogram of  $\hat{s}_{ik}$  will appear uniform.

Figure 1 is a plot of the residuals of the data of Freireich (Cox, 1972). There is one covariate coded 0 or 1. Thus the residuals are  $1 - E(X | R_i)$  if  $X_i$  is 1 or  $-E(X | R_i)$  if  $X_i$  is 0. This gives rise to the two horizontal bands of residuals seen in Fig. 1. Ties were broken by the addition of a small random number to each failure time. For  $T > 16$  there are equal numbers of residuals at equal distances from 0. For  $5 < T \leq 15$  the positive residuals are closer to zero than the negative residuals but there are more positive residuals. Thus for  $T > 5$  there is no time trend. For  $T < 5$  there are no negative residuals indicating a failure of proportional hazards in this region. Using a chi-squared

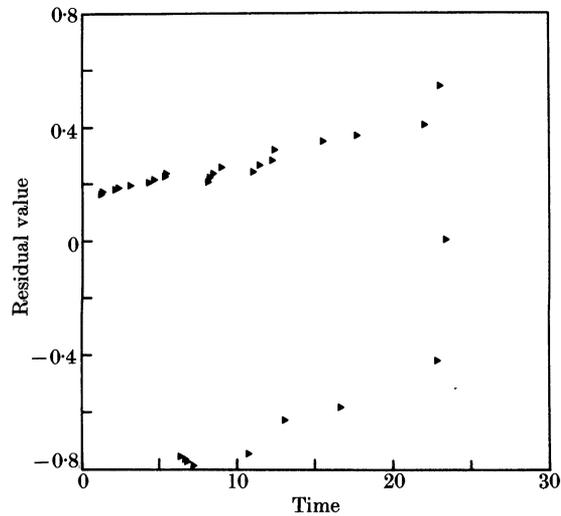


Fig. 1. Plot of  $\hat{r}_{i1}$  versus time.

goodness-of-fit test (Schoenfeld, 1980), dividing the time axis at  $T = 5$  yields a  $p$ -value of 0.08, ignoring the post hoc nature of the decision to divide the data at  $T = 5$ .

This investigation was supported by a U.S. Public Health Service Grant, awarded by the National Cancer Institute. I thank Stephen Lagakos for his review of the manuscript.

#### REFERENCES

- COX, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.
- COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.
- COX, D. R. & SNELL, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc. B* **30**, 248–75.
- KALBFLEISCH, J. D. & PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- KAY, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Appl. Statist.* **26**, 227–37.
- SCHOENFELD, D. (1980). Chi-squared goodness of fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–53.

[Received July 1980. Revised July 1981]