

RESEARCH ARTICLE

Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study

W.W.M. Abeysekera and M.R. Sooriyarachchi*

Department of Statistics, Faculty of Science, University of Colombo, Colombo 03.

Submitted: 13 July 2008; Accepted: 17 October 2008

Abstract: Cox proportional hazards (PH) model is one of the finest techniques in identifying combined effects of several covariates on the relative risk (hazard). This model assumes that the hazards of the different strata formed by the levels of the covariates are proportional. The primary objective of this paper is to illustrate the usefulness of a global goodness-of-fit test proposed by Schoenfeld for testing the PH assumption. Though several classical methods have been discussed in previous studies there is no one research paper that compares Schoenfeld's method with these. Moreover, programmes are developed in SAS for constructing this global goodness-of-fit test. In this paper the proposed test is applied to a real, large scale data set that involves several covariates, whereas Schoenfeld has used only a small data set with only one covariate to illustrate this new test.

Using Kaplan-Meier curves, a preliminary analysis was conducted on the survival data. Then, a Cox PH model was fitted to the data. All the methods and residual analysis including the global goodness-of-fit test indicated that for the data set used the assumption of PH is violated. However, other than for the global goodness-of-fit test all other techniques are based on graphical methods and are thus subjective. Hence, for cases where the violation of the PH assumption is marginal these graphical methods may be inadequate to detect this departure. However, as the global goodness-of-fit test is an objective test it is recommended as the best among the methods compared.

Keywords: Cox proportional hazards model, Cox-Snell residuals, goodness-of-fit, residual analysis, Schoenfeld residuals.

INTRODUCTION

In the comparison of two survival functions in a clinical trial, it is useful to have a means to measure the difference between the two survival curves. If the corresponding hazard functions are proportional, then the interpretation of relative risk (hazard) can be done using the maximum

partial likelihood estimator proposed by Cox¹. The descriptive method of identifying proportionality of hazards for levels of a single covariate involves the plotting of Kaplan-Meier estimates. However, this simple method is cumbersome when there are many covariates. Thus, the estimation of relative risk for a group of subjects depending on several explanatory variables (covariates which can be categorical or continuous) is assessed by a parametric fitting of a proportional hazard (PH) model¹.

After fitting an appropriate PH model for the given data, it is vital to check the goodness-of-fit and the residuals of the fitted model. The goodness-of-fit of a PH model mainly focuses on checking the validity of the assumption of the proportional hazards (i.e. whether the effects of covariates on risk remain constant over time). For a more general Cox's regression model, the PH property is one of the restrictions to using the model with time-fixed covariates. It assumes that the hazard ratio between two sets of covariates is constant over time, because the common baseline hazard function cancels out in the ratio of the two hazards. However, the impact of time-varying covariates leads the hazard to vary over time, thus violating the assumption of PH in the model. To test this PH assumption of Cox's regression model, Schoenfeld² and Moreau *et al.*³ introduced a dummy time-dependent covariate. There, the observed numbers of events in the cells arising from a partition of the Cartesian events product of the range of covariates (or ranges defined by the predicted partial likelihood estimates of the model) and the time axis, (with expected numbers of events predicted by the model) in the cells are computed, and a chi-squared statistic for the fit is obtained. The number of partitions of the time-axis is arbitrary but the defined portions (time intervals) should be non-overlapping.

*Corresponding author (roshini@mail.cmb.ac.lk)

A variety of residuals have been developed for a fitted PH model such as Cox-Snell residuals⁴, Schoenfeld residuals⁵ etc. Plots of these residuals are useful in detecting non-proportionality of predicted hazards of the fitted model over the covariate space for each covariate.

The objective in the present paper is to illustrate the usefulness of the global goodness-of-fit test proposed by Schoenfeld² and to discuss other classical methods of testing validity of Cox PH models. This is achieved by examining each method and applying all the mentioned model validation techniques including the new global goodness-of-fit test proposed by Schoenfeld² to a large, real life data set that includes several covariates. The data is about the times that heroin addicts remain in a clinic for methadone maintenance treatment⁶. Several possible explanatory variables are recorded along with the termination times for 238 individuals. Simply, the approach discussed in this paper extends the ideas developed for goodness-of-fit testing and residual analysis for a censored set of real data.

METHODS AND MATERIALS

Assumption of proportional hazards: Suppose two groups, namely group 1 and 2 (for example say, group 1 is receiving the new treatment and group 2 is receiving the standard treatment), are compared with respect to the hazard of each group. Let $\lambda_1(t | \text{group 1})$ and $\lambda_2(t | \text{group 2})$ be the hazard functions of group 1 and group 2 respectively, where $t > 0$. Then the two groups are said to have proportional hazard, when the hazard ratio Ψ is constant over time. That is,

$$\frac{\lambda_1(t | \text{group 1})}{\lambda_2(t | \text{group 2})} = \psi, \text{ for all } t$$

Kaplan-Meier curves: The Kaplan-Meier method⁷ estimates the survival function that summarizes the survival data. The curves of Kaplan-Meier survival functions work best for time fixed covariates with few levels. If the predictor satisfies the proportional hazard assumption then the graph of the log [-log(survival)] versus log of survival time should result in parallel lines.

The Cox regression model¹: Let T_i be the failure time for subject i , $i = 1, \dots, n$. If T_i follows the Cox proportional hazards regression model, then the hazard function for T_i at time $t > 0$, conditional on the $p \times 1$ covariate vector Z_i , is

$$\lambda(t | Z_i) = \lambda_0(t) \exp(\beta'Z_i) \dots(1)$$

where $\lambda_0(t)$ is the baseline hazard function (i.e. the hazard function when all covariates take value zero) and β is a

$p \times 1$ vector of regression coefficients. Statistics are designed to check whether interaction terms between elements of Z_i or higher order terms in the elements of Z_i need to be added to $\beta'Z_i$.

Using counting process notation, the information in the data can be represented by

$$\{N_i(t), Y_i(t), Z_i : 0 < t < \infty\}$$

where $N_i(t)$ takes value one if subject i has been observed to fail prior to time t and takes value zero otherwise and $Y_i(t)$ takes value one if subject i is at risk at time t and takes value zero otherwise. Then the Cox partial likelihood score vector equals

$$u(\beta) = \sum_{j=1}^n \int_0^{\infty} \{Z_i - \bar{Z}(s, \beta)\} dN_i(s)$$

where $\bar{Z}(s, \beta) = \frac{\sum_{j=1}^n Z_j Y_j(s) e^{\beta'Z_j}}{\sum_{j=1}^n Y_j(s) e^{\beta'Z_j}}$ is a weighted average of the

Z_i 's and $dN_i(s) = N_i(s) - N_i(\bar{s})$ is a binary random variable that equals one if subject i fails at time s and equals zero otherwise. The maximum partial likelihood estimate $\hat{\beta}$ is the solution to $u(\hat{\beta}) = 0$.

Model validation for Cox PH model:

Cox-Snell residuals⁴ and the log-cumulative hazard plot of the Cox-Snell residuals: The Cox-Snell residuals for the Cox PH model are given by

$$(r_{\text{cox-snell}})_i = \exp(\hat{\beta}'Z_i) \hat{\Lambda}_0(t) \dots(2)$$

where $\hat{\Lambda}_0(t) = \sum_{j=1}^n \int_0^t \frac{dN_j(s)}{\sum_{j=1}^n Y_j(s) \exp(\hat{\beta}'Z_j(s))}$ is the estimated

cumulative baseline hazard function at time t .

When the fitted model is correct, the Cox-Snell residuals $(r_{\text{cox-snell}})_i$, are a plausible sample of observations from a unit exponential distribution. Thus, a plot of Cox-Snell residuals versus observations (or time) will not lead to a symmetric display.

The log-cumulative hazard plot of residuals is given by plotting $\log\{-\log \hat{S}(r_{\text{cox-snell}})_i\}$ values against $\log(r_{\text{cox-snell}})_i$. A straight line plot with unit slope and zero intercept will then indicate that the fitted survival model is adequate in satisfying the proportional hazard assumption.

Schoenfeld residuals⁵: The Schoenfeld residual vector is calculated on a per event time basis as

$$U_i(t) = Z_i(t) - \bar{Z}(t) \dots(3)$$

where $\bar{Z}(t)$ is a weighted average of the covariates over the risk set at time t and is

$$\text{given by, } \bar{Z}(t) = \frac{\sum_{j=1}^n Z_j(t) Y_j(t) \exp(\hat{\beta}' Z_j(t))}{\sum_{j=1}^n Y_j(t) \exp(\hat{\beta}' Z_j(t))}$$

Under the proportional hazards assumption, the Schoenfeld residuals have the sample path of a random walk; therefore, they are useful in assessing time trend or lack of proportionality. Due to time dependent covariates the generalized linear regression of the Schoenfeld residuals on functions of time gives a non-zero slope. Thus, a non-zero slope is an indication of a violation of the proportional hazard assumption. As with any regression it is recommended to look at the graph of the regression in addition to performing the tests of non-zero slopes. There are certain types of non-proportionality that will not be detected by the tests of non-zero slopes alone but that might become obvious when looking at the graphs of the residuals such as nonlinear relationship (i.e., a quadratic fit) between the residuals.

Goodness-of-fit statistics δ : In the Cox regression model, the hazard ratio for subject i versus subject j at time t is

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{\lambda_0(t) \exp(\beta' Z_i)}{\lambda_0(t) \exp(\beta' Z_j)} = \exp(\beta' Z_i - \beta' Z_j)$$

When comparing two individuals, the individual with the larger value of $\exp(\beta' Z_i)$ has greater risk of death at time t . To form the proposed goodness-of-fit statistics, the partial likelihood estimate of $\phi_i = \exp(\beta' Z_i)$ is obtained, first, which is $\hat{\phi}_i = \exp(\hat{\beta}' Z_i)$. Then the subjects are grouped or partitioned into regions based on the percentiles of $\hat{\phi}_i$, which we call percentiles of risk. Following Hosmer and Lemeshow's⁹ approach with binary data, it is suggested to form G regions of approximately equal size so that the first group contains the n/G subjects with the smallest $\hat{\phi}_i$'s, and the last group contains the n/G subjects with the largest $\hat{\phi}_i$'s. In general, this classification leads to grouping subjects that are considered similar in that they have similar risks of death at any given time i .

Given the partition of the data, the goodness-of-fit statistic is formulated by defining the $(G - 1)$ group indicators

$$I_{ig} = \begin{cases} 1 & \text{if } \hat{\phi}_i \text{ is in region } g \\ 0 & \text{if otherwise,} \end{cases}$$

$g = 1, \dots, G - 1$. Then, in order to assess the goodness-of-fit of the model (1), we consider the alternative Cox model,

$$\lambda(t|Z_i) = \lambda_0(t) \exp\left(\beta' Z_i + \sum_{g=1}^{G-1} \gamma_g I_{ig}\right) \quad \dots(4)$$

If model (1) is correctly specified, then $\gamma_1 = \gamma_2 = \dots = \gamma_{G-1} = 0$ in (4). Although I_{ig} is based on the

random quantities $\hat{\phi}_i$'s, Moore and Spruill¹⁰ showed that, asymptotically, one can treat the partition as if it was based on the true ϕ_i 's (and thus, regard I_{ig} as a fixed covariate). To test the goodness-of-fit of model (1) versus alternative (4), the likelihood ratio, Wald, or score statistic can be used to test $H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_{G-1} = 0$. If model (1) has been correctly specified, each of these statistics has an approximate chi-squared distribution with $(G-1)$ degrees of freedom (d.f.) when the sample size is large.

Although the score, Wald, and likelihood ratio statistics are asymptotically equivalent, the score statistic for $H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_{G-1} = 0$ in (4), is proposed over others since it has a nice intuitive interpretation. For model (4), the score vector is

$$u(\beta, \gamma) = \begin{bmatrix} u_1(\beta, \gamma) \\ u_2(\beta, \gamma) \end{bmatrix} = \sum_{i=1}^n \begin{bmatrix} \int_0^\infty \{Z_i - \bar{Z}(s, \beta, \gamma)\} dN_i(s) \\ \int_0^\infty \{I_i - \bar{I}(s, \beta, \gamma)\} dN_i(s) \end{bmatrix}$$

where $\gamma = [\gamma_1, \dots, \gamma_{G-1}]'$ and $I_i = [I_{i1}, \dots, I_{i,G-1}]'$.

Let $\hat{\beta}_0$ be the estimate β under the null hypothesis $\gamma = 0$. The score test statistic for testing this null is,

$$\chi^2 = u(\hat{\beta}_0, 0)' \{ \text{var}[u(\beta, \gamma)] \}_{(\beta=\hat{\beta}_0, \gamma=0)}^{-1} u(\hat{\beta}_0, 0)$$

$$\text{where } \{ \text{var}[u(\beta, \gamma)] \}_{(\beta=\hat{\beta}_0, \gamma=0)} = - \frac{d u(\beta, \gamma)}{d[\beta, \gamma]},$$

evaluated at $(\beta = \hat{\beta}_0, \gamma = 0)$.

Since $u_1(\hat{\beta}_0, 0) = 0$ ($\hat{\beta}_0$ is the solution to $u_1(\hat{\beta}_0, 0) = 0$), the score statistic is actually based on the large sample distribution of $u_2(\hat{\beta}_0, 0)$. Using an algebraic identity, $u_2(\hat{\beta}_0, 0)$ can be expressed as

$$u_2(\hat{\beta}_0, 0) = \sum_{i=1}^n \int_0^\infty \{I_i - \bar{I}(s, \hat{\beta}_0, 0)\} dM_i(s, \hat{\beta}_0, 0) \quad \dots(5)$$

where

$$\begin{aligned} dM_i(s, \beta, \gamma) &= dN_i(s) - Y_i(s) \exp(\beta' Z_i + \gamma' I_i) \lambda_0(s) ds \\ &= dN_i(s) - E[dN_i(s) | N_i(u), Y_i(u), Z_i, I_i, u < s] \end{aligned}$$

and

$$dM_i(s, \hat{\beta}_0, \gamma) = dN_i(s) - Y_i(s) \exp(\hat{\beta}_0' Z_i) \hat{\lambda}_0(s) ds \quad \dots(6)$$

with

$$\hat{\lambda}_0(s) = \frac{\sum_{j=1}^n dN_j(s)}{\sum_{j=1}^n Y_j(s) \exp(\hat{\beta}_0' Z_j)} \quad \dots(7)$$

being the Breslow¹¹ estimate of the baseline hazard. Substituting (6) and (7) in (5), one can show that

$$\bar{I}(s, \hat{\beta}_0, 0) \sum_{i=1}^n dM_i(s, \hat{\beta}_0, 0) = 0$$

and thus that the g^{th} element of $u_2(\hat{\beta}_0, 0)$ equals,

$$\begin{aligned} u_{2g}(\hat{\beta}_0, 0) &= \sum_{i=1}^n I_{ig} \int_0^\infty dM_i(s, \hat{\beta}_0, 0) \\ &= \sum_{i=1}^n I_{ig} \int_0^\infty dN_i(s) - \sum_{i=1}^n I_{ig} \int_0^\infty Y_i(s) \exp(\hat{\beta}'_0 Z_i) \hat{\lambda}_0(s) ds \\ &= \sum_{i=1}^n I_{ig} \int_0^\infty dN_i(s) - \sum_{i=1}^n I_{ig} \int_0^\infty \hat{E}[dN_i(s) | N_i(u), Y_i(u), Z_i, I_i, u < s] \\ &= O_g - E_g \end{aligned} \quad \dots(8)$$

where O_g is the observed number of failures in region (group) g and E_g is the estimated expected number of failures [under model (1)] in region g . Alternatively, (8) can be expressed as $\sum_{i=1}^n I_{ig} \hat{M}_i$,

where

$$\hat{M}_i = \int_0^\infty \left\{ dN_i(s) - \hat{E}[dN_i(s) | N_i(u), Y_i(u), Z_i, I_i, u < s] \right\}$$

is the martingale residual given by model (1). Thus, our goodness-of-fit statistic is actually a function of the observed minus the estimated expected number of failures in each region or, equivalently, the martingale residuals in each region. Because of this fact, it can describe what is meant by large samples, i. e., the statistics have approximate chi-squared distribution with $(G-1)$ d.f. when the sample size is large.

Using the partition based on the percentiles of risk (percentiles of $\hat{\phi}_i$'s), the above statistic has little power to test whether the proportional hazard assumption is valid. However, this statistic can easily be extended to have power to detect non proportional hazards by using the approach of Schoenfeld².

There, in addition to partitioning the subject based on the percentiles of risk as explained above, the time axis should also be partitioned into say, τ intervals, which are consecutive and non-overlapping, containing approximately an equal number of subjects in each interval.

Accordingly, $(\tau - 1)$ indicators are defined,

$$I_{ik}^* = \begin{cases} 1 & \text{if } t_i \text{ is in region } k \\ 0 & \text{if otherwise,} \end{cases}$$

$k = 1, \dots, \tau - 1$. Then, in order to assess the goodness-of-fit along with testing PH assumption of model (1), we define the alternative Cox model including the $(\tau-1)$ $(G - 1)$ interaction terms as follows

$$\lambda(t | Z_i) = \lambda_0(t) \exp \left(\beta' Z_i + \sum_{g=1}^{G-1} \sum_{k=1}^{\tau-1} \gamma_{gk} I_{ig}^* I_{ik}^* \right) \quad \dots(9)$$

Then, as explained previously, the score test statistic can be used to compare model (1) and model (9). If model (1) is found as not significantly deviated from model (9) (i.e. the hazard is the same over covariate space as well as over time), then it can be decided that the goodness-of-fit and PH assumption hold for model (1).

A rough sample size criterion for the resulting goodness of fit statistic to be approximately chi-square, can be used in deciding the numbers G and τ . In particular, the $\tau \times G$ regions formed by the cross classification of the covariate and time partitions should be chosen in such a way that $6 \leq \tau G \leq D/5$, where D is the total number of failures in the data set.

Since the score statistic is similar to Pearson's chi-square for contingency tables (a function of observed and expected frequencies), it is suggested that all estimated expected counts E_{gk} in each cell representing $\tau \times G$ regions be greater than 1 and at least 80% should be greater than or equal to 5. For situations when this does not hold, the chi-square approximation for the score test may be poor. One possible solution to this is to use a smaller number of $\tau - G$ regions in the interval $[6, D/5]$, for which 80% of the E_{gk} 's are greater than or equal to 5.

Example

The data set used in this study is from a clinical trial on 238 heroin addicts following methadone maintenance treatment⁶, in which the outcome is the time until the addict terminates the treatment procedure. Thus, the endpoint of interest is not death, but termination of treatment. Some subjects were still in the clinic at the time these data were recorded and this is indicated by the variable "status", which is equal to 1 if the person had departed the clinic on completion of treatment (uncensored) and 0 otherwise (right censored). The measured possible explanatory variables for time to complete treatment are namely, maximum methadone dose ("dose", a continuous variable in units-milligrams), whether the addict had a criminal record ("prison", a categorical variable with two levels, 1 - have criminal records, 0 - no records) and the clinic in which the addict was being treated ("clinic", a categorical variable with two levels, 1 - clinic 1, 0 - clinic 2)- are recorded along with the termination times for all 238 individuals. The overall event rate for this data set is 0.63 (=number terminating treatment/238=150/238).

According to the main aim of this study, a PH model was fitted in order to determine the goodness-of fit of

the model by applying the techniques described in the previous section.

As a preliminary identification of proportionality of hazards or departure from it, log cumulative hazard plots based on Kaplan-Meier estimates are drawn for the three explanatory variables where the continuous variable “dose” is categorized into 3 levels <60, 60-79 and 80. Figures 1 is the log cumulative hazard plots for levels of variables “dose” “prison” and “clinic” respectively.

In Figure 1, the log cumulative hazard function (estimated by the Kaplan-Meier method) for the three levels of “dose” are approximately parallel. The log cumulative hazard function for the two levels of “prison” does not seem parallel and as it is depicted in the third graph of Figure 1, the distance between the log cumulative hazard function for the two levels of “clinic” increases over the log of survival time indicating non parallelism and the curves for the two levels are also seen to cross.

The next step is to fit the Cox PH model for the survival experience regressed on the explanatory variables. The explanatory variable – Methadone dose (“dose”) is the standardized methadone dose since it is observed that the dose varies by 10 to 15 units in the recorded data. Then, to fit the model, the PROC PHREG procedure of the SAS package was used and the following Cox PH model (parameter estimates are given in Table 1) is fitted using the forward selection criteria¹².

$$\lambda_i(t) = \lambda_0(t) \exp(\hat{\beta}_1^{prison} + \hat{\beta}_2 dose_i + \hat{\beta}_3^{clinic}) \quad \dots(10)$$

where $i, i = 1, \dots, 238$.

The fitted model (10) includes the categorical covariate “clinic” which showed non-proportional hazard when considered individually. Thus it is expected that the model will fail to hold valid the PH assumption unless the non-proportionality of hazards for “clinic” is cancelled out due to the adjustments of the other explanatory variables. However, without an appropriate residual analysis and goodness-of-fit tests, it is not appropriate to state anything about the validity of the fitted model.

Testing Validity of the Fitted Model:

The Cox-Snell residuals for model (10) are computed according to equation (2). The log-cumulative hazard plot of the Cox-Snell residuals depicted in Figure 2 is

used to identify any departure from a well fitted model that satisfies the proportional hazard assumption.

It is explained under a previous section that a straight line log cumulative hazard plot with unit slope and zero intercept will indicate that the PH assumption in the model holds. Figure 2 illustrates that there is a slight deviation of points from a straight line indicating that this is a borderline case. Thus, based on this plot it is difficult to come to a firm conclusion regarding departure from PHs.

The Schoenfeld residuals for each covariate “dose”, “prison” and “clinic” computed as given in equation (3) are observed and are plotted against time (time of termination from treatments). These plots are

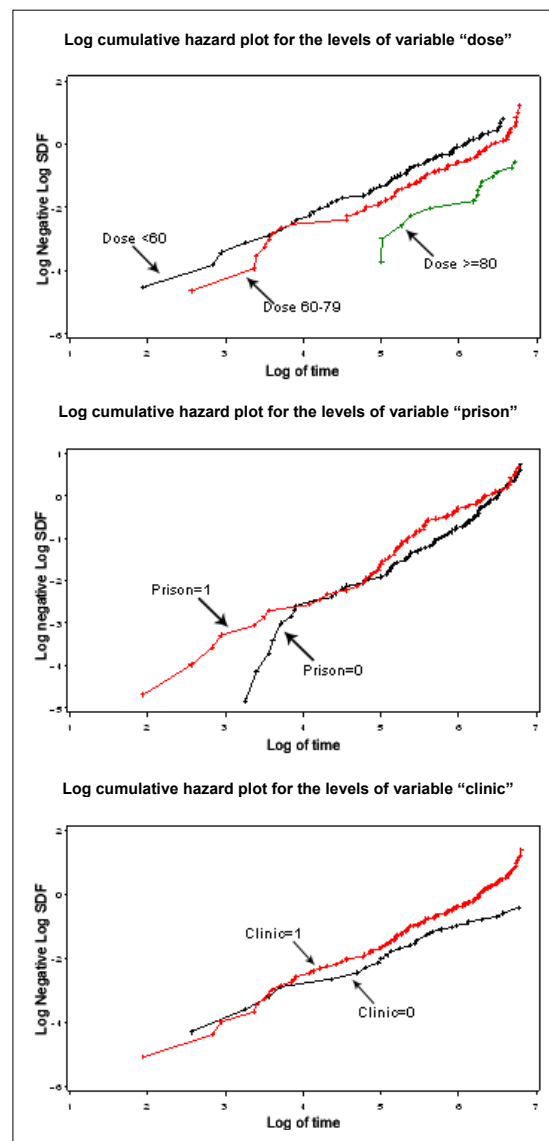


Figure 1: Log cumulative hazard plots for levels of the three variables

Table 1: Parameter estimates of model (10)

Variable	DF	Parameter estimate	Standard error	Chi-square confidence	Pr>ChiSq	Hazard ratio	95% Hazard confidence	ratio limits
prison (1)	1	0.32647	0.16722	3.8115	0.0509	1.386	0.999	1.924
dose	1	-0.5114	0.09218	30.7783	<.0001	0.6	0.501	0.718
clinic (1)	1	1.00875	0.21487	22.0411	<.0001	2.742	1.8	4.17

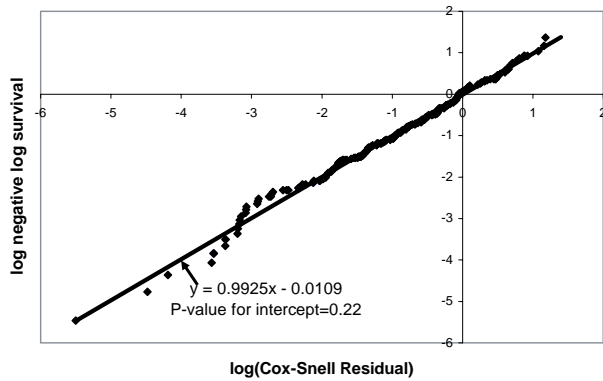


Figure 2: Log cumulative hazard plot of Cox-Snell residuals for model (10)

depicted in Figure 3 where fitted linear and quadratic regression lines for the Schoenfeld residuals are visible on the same plot.

According to Figure 3, the best fitted line for the Schoenfeld residuals for “dose” has a slope which is not significantly different from zero (since p-value for slope is $0.5454 \gg 0.05$), and the fitted quadratic line is balanced around the zero horizontal axis. The illustration in the plot of Schoenfeld residuals for “prison” leads to the same interpretation, where the fitted linear regression line has non-significant slope (since p-value for slope is $0.3653 \gg 0.05$), and the fitted quadratic line is balanced around the zero horizontal axis. Thus, it can be concluded that the covariates “dose” and “prison” do not violate the PH assumption of the fitted model (10) and are also not time dependent. However, the fitted linear regression line for the plot of Schoenfeld residuals for “clinic” (Third graph of Figure 3) has a slope which is significantly different from zero (p-value for slope is $0.0007 \ll 0.05$) and the fitted quadratic curve also highly deviates from the zero horizontal axis. This result indicates that the covariate “clinic” violates the PH assumption of the model (10) and hence leads to the conclusion that “clinic” is probably a time dependent covariate.

To justify the findings expressed in this section, the variable “clinic” can be tested for its time dependency. For that, it is tested whether there is a significant

Table 2: Observed and expected number of addicts terminated from treatment, predicted by model (10)

Region	Grouping by percentiles of $\hat{\phi}_i$	Time interval	Observed	Expected
1	1	1	5	13.64
2	2	1	13	20.81
3	3	1	18	30.06
4	4	1	14	23.96
5	5	1	11	27.37
6	1	2	9	1.27 *
7	2	2	16	3.19 *
8	3	2	17	5.29
9	4	2	20	9.59
10	5	2	27	14.83
			150	150

* Expected count less than 5

difference between the current model (10) and the model regressed on all the covariates in (10) plus the interaction term “time*clinic” [model (11)].

$$\lambda_i(t) = \lambda_0(t) \exp(\hat{\beta}_1^{prison} + \hat{\beta}_2^{dose_i} + \hat{\beta}_3^{clinic} + t * \hat{\beta}_4^{clinic}) \dots (11)$$

Under the null hypothesis that model (10) fits the data well, the deviance difference between model (10) and model (11) follows a Chi-square distribution with 1 d.f. The calculated difference in deviance between model (10) and (11) is 11.509 (p-value = $0.000693 \ll 0.05$). This indicates that the interaction term “time*clinic” is highly significant and hence it is confirmed that the variable “clinic” is a time dependent covariate and thus violates the PH assumption.

Testing the goodness-of-fit: To determine the goodness-of-fit of model (10), the 238 subjects are grouped into $G=5$ groups (approximately 48 individuals in each group), based on the percentiles of the partial likelihood estimates ($\hat{\phi}_i$) predicted by model (10). Simultaneously, the time axis is partitioned into $\tau=2$ intervals, $t \leq 367$ (days) and $t > 367$ (days), such that the 238 individuals are divided equally among each interval. Determining G and τ are done under restriction $6 \leq (\tau=2) (G=5) \leq (D=150)/5$, where $D=150$ is the total number of failures in the data set, and all estimated expected

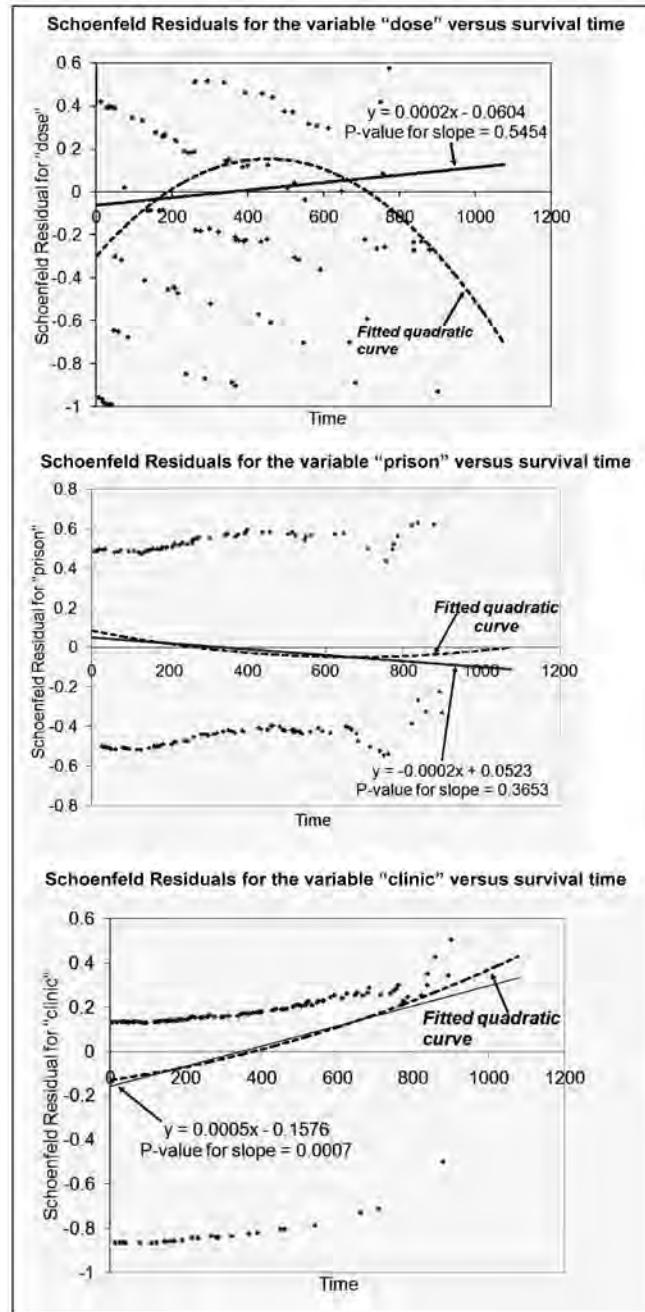


Figure 3: Schoenfeld residuals for each explanatory variable versus survival time

counts in each cell representing $(\tau = 2) \times (G = 5) = 10$ regions, are greater than 1 and at least 80% are greater than or equal to 5. The expected counts for the 10 regions, estimated by model (10) are illustrated in Table 2.

It can be seen from Table 2, that all expected counts are greater than 1 and 8 out of the 10 regions have expected counts greater than 5 (80%). Thus, this grouping is appropriate when using the Chi-square approximation for the goodness-of-fit test discussed in the following.

According to the grouping defined above, the indicator variables I_{ig} [indicating that the i^{th} individual belongs to g^{th} region partitioned based on the percentiles of the partial likelihood estimates ($\hat{\varphi}_i$)] where $g = 1, \dots, 4$ and I_{ik}^* (indicating that the i^{th} individual belongs to k^{th} time interval) where $k = 1$, are defined.

Then, the alternative Cox model given below is fitted,

$$\lambda(t|Z_i) = \lambda_0(t) \exp\left(\hat{\beta}_1^* \text{prison} + \hat{\beta}_2^* \text{dose}_i + \hat{\beta}_3^* \text{clinic} + \sum_{g=1}^4 \sum_{k=1}^1 \gamma_{gk} I_{ig} * I_{ik}^*\right) \dots (12)$$

The score test statistic for model (10) is 56.27 associated with 3 d.f. and that of the alternative model (12) is 159.19 with 7 d.f. Thus the score goodness-of-fit statistic for the fitted model (10) is $\chi^2=102.93 (=159.19-56.27)$ with 4 ($=7-3$) d.f. (p -value <0.001) indicating that the alternative model (12) is significantly different from model (10) and hence the goodness-of-fit fails for model (10). Since this test is powerful in testing the PH assumption too, it can be concluded that the PH assumption in the fitted model (10) is violated.

DISCUSSION

The Cox PH model is the most popular method of examining the effect of explanatory variables on survival. However, it requires the assumption of proportional hazards between strata formed by the combinations of levels of the different explanatory variables. Thus, when fitting a PH model it is vital to assess the assumption of proportionality. There are numerous methodologies in the literature^{2-5, 8-10} for checking the assumption of PHs.

Kaplan-Meier survival estimates in graphical format are useful in preliminary identification of proportional hazards for levels of categorical variables taken individually. However, this method is cumbersome when there are several explanatory variables. Also it is a univariate method and does not adjust for other covariates. In this case, more advanced techniques are required. One may fit proportional hazards regressed on several explanatory variables (PH models) in order to identify combined effects of several covariates on the relative risk. However, in the fitted PH model, the explanatory variables included in the model should satisfy the restriction that the relative risk is proportional over the time for different levels of covariates (i.e. PH assumption). If this requirement is present in the fitted PH model then the assumption of PH is not violated.

This paper was basically written with the objective of illustrating the usefulness of a new global goodness-of-fit test and discussing a number of established approaches in determining the validity of a fitted Cox PH model. In this paper this new test and a number of established methods for testing the PH assumption are examined by way of an example. Simultaneously developing a software programme in SAS to use in applying these techniques was a secondary objective of this paper. The methods are applied to a data set taken from a clinical trial on heroin addicts following methadone maintenance treatment⁶, in which the time until the addict terminates the treatment procedure is measured. Three possible covariates namely maximum methadone dose (“dose”), whether the addict had a criminal record (“prison”) and the clinic in which

the addict was been treated (“clinic”)- that were suspected to influence the termination time were recorded for 238 heroin addicts.

Prior to model fitting, the conventional descriptive method, in identifying proportional hazards on each categorical covariate taken individually, namely Kaplan-Meier survival curves, are used on data. This identified that the two levels of “clinic” and also the two levels of “prison” do not have proportional hazards, hinting that these variables may violate the PH assumption when it is included in the model. The variables “dose” showed approximate proportional hazards. However, a firm conclusion was not made about this result since a univariate method is not reliable when dealing with several explanatory variables.

A Cox PH model was then fitted to the data, using forward selection procedure that ended up including all 3 explanatory variables into the model.

The Log cumulative hazard plot of Cox-Snell residuals suggested that the goodness-of-fit of the model is borderline, as the points deviate slightly from a straight line. Then, the Schoenfeld residuals for each covariate were studied. This indicated that the variable “clinic” violates the PH assumption of the fitted model while the other two covariates “prison” and “dose” do not. According to Schoenfeld⁵, one of the reasons that a covariate violates PH assumption is when it is time dependent.

The Cox PH model for the hazard of death at time t for the i^{th} of n individuals in a study can be expressed as equation (1) where $\beta'Z_i = \sum_{j=1}^n \beta_j x_{ji}$. This model can be generalized to the situation when one or more explanatory variables are time dependent by writing $x_{ji}(t)$ in place of x_{ji} in equation (1). There are several simple ways to extend the Cox model to include a time dependent covariate¹³. More complicated forms of relationships between the outcome and explanatory variable over time are discussed by Fisher and Lin¹⁴. These authors explain the way in choosing the functional form of the time dependence of the covariate which they mention is usually not self evident but may be suggested by biological understanding. Some popular functional forms are time lagged covariates, moving weighted average of value over a lag time period, linear form, Splines, Piecewise constant functions with the covariate etc., to name a few. In our study the simple linear form was selected as the choice of a complex functional form raises the possibility of too much modeling and great over-fitting of the data set. Also it is known that when the PH assumption does

*Corresponding author

not hold this is a good choice, though it is somewhat restricted. Thus, using this approach it is found that the variable “clinic” is time dependent.

As the final approach, the global goodness-of-fit test for Cox PH models, proposed by Schoenfeld ², that has power to detect the insufficiency of covariates in describing the relative risks and the assumption of PH, was applied to the fitted model. Here an alternative model which contains variables that indicate partitions of relative risk (hazard ratios) over covariate space and the time space, where if these indicator variable are found significant in the model, implies that the covariates are not sufficient or the PH assumption does not hold. In other words, to perform this, an alternative model that contains all the covariates in our fitted model plus the indicator variables that indicate the partitioned regions had to be compared with the fitted model using the score test statistic which has chi-squared distribution when our model has been correctly specified. Accordingly, the 238 subjects are grouped into 5 groups based on the percentiles of the partial likelihood estimates predicted by the fitted model and the time axis is partitioned into 2 intervals ($t \leq 367$ days and $t > 367$ days). Thus, altogether there were $5 \times 2 = 10$ regions and this partitioning did not violate the requirement for a Chi-square distribution of the score statistic since all estimated expected counts (expected number of addicts that terminate treatments) in the 10 regions are greater than 1 and 80% are greater than or equal to 5. Then, the alternative model [model (12)] is fitted as described previously and it was found that there is a highly significant difference between that [model (12)] and our fitted model [model (10)]. Thus, with this result, it is determined that the goodness-of-fit along with the PH assumption is not satisfied for the fitted PH model. Finally, it is concluded that the PH assumption does not hold for the fitted model due to the inclusion of time varying covariate – “clinic” – and hence does not satisfy the global goodness-of-fit.

Another important finding of this analysis is that most of the other techniques were subjective and were unable in borderline cases to confirm the lack-of-fit or violation of PH assumption due to time dependent

covariates. The only objective method of all methods examined in this paper is the global goodness-of-fit test proposed by Schoenfeld ². Therefore, it is recommended as the most reliable method in validating the PH model.

References

1. Cox D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*. **34**(2):187-220.
2. Schoenfeld D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**(1):145-153.
3. Moreau T., O'Quigley J. & Lellouch J. (1986). On D. Schoenfeld's approach for testing the proportional hazards assumption. *Biometrika* **73**(2): 513-515.
4. Cox D.R. & Snell E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* **30**:248-275.
5. Schoenfeld D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**(1):239-241.
6. Caplehorn J. & Bell J. (1991). Methadone dosage and the retention of patients in maintenance treatment. *The Medical Journal of Australia* **155**(1):60-61.
7. Kaplan E.L. & Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**: 457-481.
8. Parzen M. & Lipsitz S.R. (1999). A global goodness-of-fit statistic for Cox regression models. *Biometrika* **55**(2):580-584.
9. Hosmer D.W. & Lemeshow S. (1980). *Applied Logistic Regression*. Wiley, New York.
10. Moore D.S. & Spruill M.C. (1975). Unified large sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics* **3**(3): 599-616.
11. Breslow N.E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review* **43**(1):45-57.
12. Agresti A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.
13. Collett D. (1994). *Modelling Survival Data in Medical Research*. Chapman and Hall, London.
14. Fisher L.D. & Lin D.Y. (1999). Time-dependent covariates in the Cox proportional hazards regression model. *Annual Review of Public Health* **20**: 145-57.

Annex**SAS CODES:**

```

/*importing the data set*/
PROC IMPORT OUT= WORK.heroindata
  DATAFILE= "C:\survival\heroin.csv"
  DBMS=CSV REPLACE;
  GETNAMES=YES;
  DATAROW=2;
RUN;

data heroin;
  set heroindata;
  if clinic=1 then clin=1; if (clinic ne 1) then clin=0;

  /*categorizing dose to use only in Kaplan-Meier
  plots*/
  if dose<60 then dose_cat=1; else if dose>59 & dose
  <80 then dose_cat=2; else dose_cat=3;
run;

/*Drawing Kaplan-Meier curves for the three
variables*/
proc lifetest data=heroin plots=(s,lls);
time time*status(0);
strata dose_cat;
run;

proc lifetest data=heroin plots=(s,lls);
time time*status(0);
strata clin;
run;

proc lifetest data=heroin plots=(s,lls);
time time*status(0);
strata prison;
run;

/*standardizing the variable "dose"*/
proc stdize data=heroin out=heroin2;
  var dose;
run;
/*Fitting the best model selected using the Forward
selection procedure (model (10))*/
proc phreg data=heroin2;
  model time*status(0)=prison dose clin;
  output out=diagout survival=svr resmart=mr;
run;
/*svr=estimated survival function, mr=martingale
residual*/

data diagno;
  set diagout;
  csr=-1*log(svr);/*computing Cox-Snell residual*/
run;
/*plotting Cox-Snell residuals versus survival time*/

```

```

proc gplot data=diagno;
plot csr*time;
symbol1 v=dot i=disjoin;
run;

/*plotting the log cumulative hazard plot of Cox-Snell
residual with it's best fitted straight line*/
proc lifetest outsurv=resid;
  time csr*status(0);
run;

data csres;
  set resid;
  llp=log(-log(survival));
  lcsr=log(csr);
run;
proc reg data=csres;
  model llp=lcsr;
run;
data csres2;
  set csres;
  yhat= -0.01089+ 0.99254*lcsr;
run;
proc gplot data=csres2;
  plot llp*lcsr yhat*lcsr/overlay;
run;

/*Observing Schoenfeld residuals for dose*/
proc phreg data=heroin2;
  model time*status(0)=dose clin prison/ rl;
  output out=sres RESSCH=sconr;
run;

data sconan;
  set heroin2; set sres;
  /*fitting straight line for the Schoenfeld residuals for
  dose*/

  proc reg data=sconan;
  model sconr=time;
  run;

  data plotres3;
  set sconan;
  psconr= -0.06045 + (0.0001749*time);
  /*Drawing the fitted straight line and the plots in the
  same graph*/

  proc gplot data=plotres3;
  plot sconr*time psconr*time/overlay;
  run;

/* Observing Schoenfeld residuals for prison*/
proc phreg data=heroin2;
  model time*status(0)=prison clin dose/ rl;
  output out=sres RESSCH=scon;
run;

```

```

data sconan;
set heroin2; set sres;
/*fitting straight line for the Schoenfeld residuals for
prison*/

  proc reg data=sconan;
  model sconr=time;
  run;

  data plotres2;
  set sconan;
  psconr= 0.05227 + (-0.00015132*time);
  run;
/*Drawing the fitted straight line and the plots in the
same graph*/
  proc gplot data=plotres;
  plot sconr*time psconr*time/overlay;
  run;

/* Observing Schoenfeld residuals for clinic*/
proc phreg data=heroin2;
  model time*status(0)=clin prison dose/ rl;
  output out=sres RESSCH=sconr;
run;

data sconan;
set heroin2; set sres;
/*fitting straight line for the Schoenfeld residuals for
clinic*/
  proc reg data=sconan;
  model sconr=time;
  run;

  data plotres;
  set sconan;
  psconr= -0.15762 + (0.00045630*time);
/*Drawing the fitted straight line and the plots in the
same graph*/
  proc gplot data=plotres;
  plot sconr*time psconr*time/overlay;
  run;

/*to check whether "clinic" is time dependent*/
proc phreg data=heroin2;
  /*model with the interaction "time*clinic"*/
  (model (11))/
  model time*status(0)=prison dose clin cl_t/ rl;
  cl_t=clin*time;
run;

****GOODNESS-OF-FIT TEST PROCEDURE****/
proc sort data=heroin;
by time;
run;

proc stdize data=heroin out=heroin2;
  var dose;
  run;

```

```

proc phreg data=heroin2;
  model time*status(0)=prison dose clin;
  OUTPUT out=resout xbeta=xb resmart=rm;
run; /*xb=linear predictor, rm=martingale residual*/

```

```

data test1;

```

```

set heroin2;
set resout;
phi=exp(xb); /*computing the partial likelihood
estimate*/
proc sort data=test1;
by phi;
run;

```

```

data test2;

```

```

set test1;
c=_n_;
/*partitioning the data into 5 groups according to the
percentiles of partial likelihood estimate*/
if c<239 then I1=1; else I1=0;
if c<191 then I2=1; else I2=0;
if c<144 then I3=1; else I3=0;
if c<96 then I4=1; else I4=0;
if c<49 then I5=1; else I5=0;
run;

```

```

data test3;

```

```

set test2;
proc sort data=test3;
by time;
run;

```

```

data test4;

```

```

set test3;
c2=_n_;
/*partitioning the time space in to 2 groups*/
if c2<120 then T1=1; else T1=0;
/*defining interaction terms that indicate the 10
regions*/
int1=I1*T1; int2=I2*T1; int3=I3*T1; int4=I4*T1;
run;

```

```

/*TO COMPARE THE CURRENT MODEL (model
(10)) WITH THE FITTED ALTERNATIVE MODEL
(model (12))*/
/*current model*/

```

```

proc phreg data=heroin2;

```

```

  model time*status(0)=prison dose clin;
  run;
/*alternative model*/
proc phreg data=test4;
  model time*status(0)=prison dose clin int1-int4;
  run;

```