Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 583–604

https://doi.org/10.29220/CSAM.2017.24.6.583
Print ISSN 2287-7843 / Online ISSN 2383-4757

Diagnostics for the Cox model

Yishu Xue[a], Elizabeth D. Schifano[1,a]

[a]Department of Statistics, University of Connecticut, USA

Abstract

The most popular regression model for the analysis of time-to-event data is the Cox proportional hazards model. While the model specifies a parametric relationship between the hazard function and the predictor variables, there is no specification regarding the form of the baseline hazard function. A critical assumption of the Cox model, however, is the proportional hazards assumption: when the predictor variables do not vary over time, the hazard ratio comparing any two observations is constant with respect to time. Therefore, to perform credible estimation and inference, one must first assess whether the proportional hazards assumption is reasonable. As with other regression techniques, it is also essential to examine whether appropriate functional forms of the predictor variables have been used, and whether there are any outlying or influential observations. This article reviews diagnostic methods for assessing goodness-of-fit for the Cox proportional hazards model. We illustrate these methods with a case-study using available R functions, and provide complete R code for a simulated example as a supplement.

Keywords: deviance residuals, martingale residuals, proportional hazards, Schoenfeld residuals, score residuals

1. Introduction

The proportional hazards model proposed by Cox (1972) has been widely used in modeling censored survival data. It is both easy to implement and easy to interpret, usually making it the biostatistician's first model to attempt when faced with time-to-event or survival data. In addition, its usage has extended to fields beyond biostatistics, such as predicting bank failures in finance (Lane *et al.*, 1986), estimating customer attrition probability in insurance (Kang and Han, 2004), identifying determinants for duration of unemployment in labor market research (Kupets, 2006), and modeling time until a policy is adopted in political science (Jones and Branton, 2005). It has been deemed one of the "breakthroughs in statistics" (Kotz and Johnson, 1992), and has been cited over 46,256 times.

Due to its pervasive applicability, before taking the results from a fitted Cox model as valid, one should address a few questions: is the proportional hazards assumption satisfied? Are the functional forms of the variables appropriate? Are there any outliers or influential observations? To answer these questions, multiple methods have been proposed, many of which rely on different types of residuals of the model.

In this article, we review methods for assessing goodness-of-fit of the Cox proportional hazards model. The rest of this article is organized as follows. In Section 2, we present the notation and important quantities in estimating a Cox model. In Section 3, we discuss the potential problems, and review the literature on tests and graphical methods proposed to identify them. The diagnostic

[1] Corresponding author: Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120 Storrs, Connecticut 06269, USA. E-mail: elizabeth.schifano@uconn.edu

Published 30 November 2017 / journal homepage: http://csam.or.kr

plots and tests for each problem are illustrated with an application regarding dental clinic visits using existing R software in Section 4. We conclude with a brief discussion and suggestions for possible remedies when nonproportionality is identified.

2. Preliminaries

Let T_i^* be the true time to event for individual i, $i = 1 \ldots, n$. Define T_i to be the minimum of T_i^* and censoring time C_i, which is a time beyond which individual i cannot be observed, i.e., $T_i = \min(T_i^*, C_i)$. We assume C_i is independent of T_i^* for each $i = 1, \ldots, n$. Define δ_i as an indicator equal to 1 if T_i^* is observed for individual i and 0 if the observation is censored, i.e., $\delta_i = I(T_i^* < C_i)$. Additionally, let X_i be the p-dimensional vector of covariates for the i^{th} individual. Assume we observe (T_i, δ_i, X_i) independently for $i = 1, \ldots, n$.

The Cox model as given in the 1972 manuscript specifies the hazard for individual i as

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' X_i), \tag{2.1}$$

where λ_0 is an unspecified, nonnegative function called the baseline hazard, and β is a p-dimensional vector of coefficients. The model is also known as the proportional hazards model, since the hazard ratio for two subjects with covariate vectors X_i and X_j is

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(\beta' X_i)}{\lambda_0(t) \exp(\beta' X_j)} = \exp\{\beta'(X_i - X_j)\},$$

which is independent of time. In the case of a single binary predictor, β is the logarithm of the hazard ratio between the corresponding two subgroups of the data.

The Cox model has subsequently been extended to incorporate time-dependent covariates. We will use the notation $X_i(t)$ to allow for the possibility of time-dependent covariates. In this case, the hazard ratio for two subjects with covariate vectors $X_i(t)$ and $X_j(t)$ becomes

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp\{\beta' X_i(t)\}}{\lambda_0(t) \exp\{\beta' X_j(t)\}} = \exp\{\beta'(X_i(t) - X_j(t))\},$$

such that the model assumes the covariates have proportionate effects on the hazard function over time (see, e.g., Fisher and Lin, 1999).

Estimation of β is based on the partial likelihood introduced by Cox (1972) and later formulated by Cox (1975):

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\{\beta' X_i(t)\}}{\sum_{j=1}^{n} \exp\{\beta' X_j(t)\} Y_j(t)} \right\}^{\delta_i}, \tag{2.2}$$

where δ_i is the censoring indicator, and $Y_j(t) = I(T_j > t)$ indicates whether the j^{th} subject is still at risk (alive) at time t.

Assume there are no ties in the failure time data. Let $N_i(t)$ be the number of events for subject i at time t, which is either 0 or 1, and define

$$dN_i(t) = I(T_i \in [t, t + \Delta t), \delta_i = 1),$$

where Δt is chosen to be sufficiently small so that $\sum_{i=1}^{n} dN_i(t)$ is either 0 or 1 at any time t. Using the counting process formulation of Fleming and Harrington (1991), the partial likelihood (2.2) can be alternatively written as

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t)\gamma_i(\beta, t)}{\sum_j Y_j(t)\gamma_j(\beta, t)} \right\}^{dN_i(t)}, \tag{2.3}$$

where $\gamma_i(\beta, t)$ is the risk score for subject i, $\gamma_i(\beta, t) = \exp\{\beta' X_i(t)\} \equiv \gamma_i(t)$. We will use this form for further derivations. Taking the logarithm of (2.3), we have the log partial likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left[Y_i(t)\beta' X_i(t) - \log \sum_j Y_j(t)\gamma_j(t) \right] dN_i(t). \tag{2.4}$$

Differentiating (2.4) with respect to β yields the $p \times 1$ score vector $U(\beta)$:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [X_i(s) - \bar{x}(\beta, s)] dN_i(s), \tag{2.5}$$

where $\bar{x}(\beta, s)$ is a weighted average of X over those observations still at risk at time s with weights $Y_i(s)\gamma_i(s)$:

$$\bar{x}(\beta, s) = \frac{\sum_{i=1}^{n} Y_i(s)\gamma_i(s)X_i(s)}{\sum_{i=1}^{n} Y_i(s)\gamma_i(s)}. \tag{2.6}$$

The $p \times p$ information matrix is the negative second derivative of (2.4), given by

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} V(\beta, s) dN_i(s), \tag{2.7}$$

where $V(\beta, s)$ is the weighted variance of X at time s:

$$V(\beta, s) = \frac{\sum_i Y_i(s)\gamma_i(s)[X_i(s) - \bar{x}(\beta, s)]'[X_i(s) - \bar{x}(\beta, s)]}{\sum_i Y_i(s)\gamma_i(s)}.$$

We obtain the maximum partial likelihood estimator by solving the partial likelihood equation

$$U\left(\hat{\beta}\right) = 0. \tag{2.8}$$

The solution $\hat{\beta}$ is consistent and asymptotically normal with mean β, the true parameter vector, and variance $\{\mathcal{E}\mathcal{I}(\beta)\}^{-1}$, the inverse of the expected information matrix. While the expected information matrix is unknown, the observed information matrix $\mathcal{I}(\hat{\beta})$ is available as

$$\mathcal{I}(\hat{\beta}) = -\left. \frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta'} \right|_{\beta=\hat{\beta}}. \tag{2.9}$$

We may use its inverse, $\mathcal{I}^{-1}(\hat{\beta})$, as the estimated variance of $\hat{\beta}$.

3. Goodness-of-fit diagnostics for the Cox model

Most diagnostic procedures for the Cox model are based on its residuals. There are multiple types of residuals defined for the Cox model, and they can often serve different purposes in model diagnostics. In this section, we review methods to identify (1) violations of the proportional hazards assumption, (2) appropriate functional forms of covariates, (3) outlying observations, and (4) influential observations. We demonstrate the usage of many of these methods in Section 4, and provide implementation details in the Supplemental Materials.

3.1. Proportional hazards assumption

3.1.1. Diagnostics based on Schoenfeld residuals

Schoenfeld (1980) proposed a chi-squared goodness-of-fit test statistic for the proportional hazards regression model which utilized a residual of the form *Expected - Observed*. The formal definition and its properties were later discussed in Schoenfeld (1982).

Let d denote the total number of events, and let $X_{(k)}$, $k = 1, \ldots, d$ be the covariate vector of a subject with an event at the k^{th} event time t_k. Further let R_k denote the risk set at time t_k, which is the set of all individuals who are still alive ("at risk") at t_k. The Schoenfeld residual is defined as

$$r_k\left(\hat{\beta}\right) = X_{(k)} - E\left(X_{(k)} | R_k\right), \quad k = 1, \ldots, d \tag{3.1}$$

which, when there are no tied event times, is indeed $r_k(\beta) = X_{(k)} - \bar{x}(\beta, t_k)$, where $\bar{x}(\beta, s)$ as given in Equation (2.6) is evaluated at t_k.

In practice, we replace β with $\hat{\beta}$ and obtain $\hat{r}_k$. If the proportional hazards assumption holds, $E(\hat{r}_k) \doteq 0$. Therefore, a plot of Schoenfeld residuals against event times will approximately scatter around 0.

Moreau *et al.* (1985, 1986) proposed a test statistic for goodness-of-fit of the Cox model, with the alternative model being one having time-varying coefficients. In the case of fitting a model with a single covariate in several levels, the statistic is of a sum of quadratic expressions, and reduces to the statistic in Schoenfeld (1980) for two-level problems, but is computationally simpler.

Grambsch and Therneau (1994) generalized the approach in Schoenfeld (1982) to test the proportional hazards assumption. Assuming the true hazard function is of the time-varying form

$$\lambda_i(t) = \lambda_0(t) \exp\left[\{\beta + G(t)\theta\}' X_i(t)\right], \tag{3.2}$$

where $G(t)$ is a diagonal matrix with jj element $g_j(t)$, they showed that the test statistic

$$T(G) = \left(\sum G_k \hat{r}_k\right)^T D^{-1} \left(\sum G_k \hat{r}_k\right) \tag{3.3}$$

with

$$D = \sum G_k \hat{V}_k G_k^T - \left(\sum G_k \hat{V}_k\right)\left(\sum \hat{V}_k\right)^{-1} \left(\sum G_k \hat{V}_k\right)^T,$$

where $\hat{V}_k$ is the observed variance of $\hat{\beta}$ at time t_k, has an asymptotic χ^2 distribution with p degrees of freedom. They suggested using the average variance matrix $\bar{V} = \mathcal{I}^{-1}(\hat{\beta})/d$ to approximate $\hat{V}_k$, since most often, the variance matrix of $X(t)$ changes slowly and the approximation makes little difference to the estimates. They also pointed out that the tests in other previous works fall under this framework with different choices of $G(t)$. Table 1 summarizes the related publications and the form of $G(t)$ they used. The form of $G(t)$ is diagonal for all the articles so we refer to a univariate $g(t)$.

Table 1: Articles and their functional forms of $G(t)$ falling under the framework of Grambsch and Therneau (1994)

Article	$g(t)$
Cox (1972), Gill and Schumacher (1987), Chappell (1992)	A specified function of time
Schoenfeld (1980), Moreau *et al.* (1985), O'Quigley and Pessione (1989)	Piecewise constant on non-overlapping time intervals with the constants and intervals predetermined
Harrell (1986)	$g(t) = \bar{N}(t-)$, tests the correlation between the rank of the event times and the Schoenfeld residuals
Lin (1991)	The proposed test is equivalent to $g(t) = t$ when the maximizer of a weighted partial likelihood, $\hat{\beta}_w$, is based on a one-step Newton-Raphson algorithm staring from $\hat{\beta}$
Nagelkerke *et al.* (1984)	Let $g_j(t_1) = 0$ and $g_j(k+1) = a_j^2 \hat{r}_{jk}$, $j = 1, \ldots, p$ to test for the serial correlation of the Schoenfeld residuals, where a_j is the weight of the j^{th} covariate;

The most common choices of $g(t)$ are the identity transformation, the rank transformation, the log transformation, and the left-continuous version of the Kaplan-Meier (KM) estimated survival curve of t (i.e., $1 - \text{KM}(t-)$). In addition, the cox.zph function in the **survival** package (Therneau, 2017) allows for user-defined forms of $g(t)$ in obtaining $T(G)$. The function also provides a χ^2 test for each covariate j as

$$T_j(g) = \frac{\left(\sum g_j \hat{r}_{jk}\right)^2}{D_{jj}},$$

where g_j and D_{jj} are the jj elements of $G(t)$ and D, respectively, and $\hat{r}_{jk}$ is the j^{th} element of $\hat{r}_k$. The test statistic will have a χ_1^2 distribution if the proportional hazards assumption is satisfied.

Park and Hendry (2015) showed that the decision of time transformations can have profound implications for the conclusions reached. In addition, they suggested that prior to fitting the model, practitioners should first determine the levels of censoring in their data, as in some cases an alternative model might be more appropriate than the Cox model. Exploratory graphical analysis, such as histograms, should be used to see if there are any outlying survival times. If there are few outliers, the test of Grambsch and Therneau (1994) should be done using the untransformed time. Otherwise, the rank transformation is a better choice. They showed using simulations that, with low levels of censoring, the rank and the KM transformation perform approximately equally well. When the level of censoring increases, the rank transformation tends to outperform the KM and natural log transformations.

Keele (2010) pointed out that, while the test of Therneau and Grambsch has been widely used as it is easy to conduct and interpret, application of the test requires some care due to it being sensitive to several forms of misspecification. Omitted predictors, omitted interactions and nonlinear covariate functional forms can all significantly affect the test result. The paper also emphasized the importance of correcting the functional form for continuous covariates before checking for nonproportionality (see Section 3.2).

Winnett and Sasieni (2001) discussed situations in which the approach of Grambsch and Therneau (1994) might provide misleading estimates of time-varying coefficients and presented an example using Mayo clinic lung cancer data. They also suggested using a compromise between $\hat{V}_k$ and $\bar{V}$ for such situations, such as a smoothed version of $\hat{V}_k$.

Despite the fact that the test of Grambsch and Therneau (1994) allows for time-dependent covariates, Grant *et al.* (2014) showed using simulation that its performance, when there are indeed time-dependent covariates, is highly unstable and its power depends largely upon factors that are unknown in practice, such as when the hazard ratio changes, and by how much it changes. Grant *et al.* (2014) focused on the identity, log, rank, and KM transformations for $g(t)$ in their simulations, and concluded that this instability suggests limited value of the test in (3.3) in the presence of time-dependent covariates in real-world applications. Fisher and Lin (1999) suggests the approach of Lin (1991) for time-dependent covariates, but note that the approach can be sensitive to the choice of weight function. Fisher and Lin (1999) also cites the approach of Wei (1984), which is based on the score process. Please see Wei (1984) for further details.

Xue *et al.* (2013) extended Schoenfeld residuals to case-cohort studies in epidemiological studies of rare disease and defined case-cohort Schoenfeld residuals as the difference of the covariate value and its mean, conditioned on the case-cohort risk set $\tilde{R}_i(t)$. They also made proper adjustments to the KM estimating procedure by taking into account the influence of each cohort on the increment of the cumulative hazard. They also proposed a test of proportionality based on the correlation between their modified Schoenfeld residuals and $g(t)$, where g could be the identity, rank, or KM transformation. If proportionality holds for a covariate, the correlation should be close to 0. Large values of correlation, however, are often indications of nonproportionality.

3.1.2. Diagnostics based on Cox-Snell residuals

Another residual that assists in evaluating the proportional hazards assumption is the Cox-Snell residual. Cox and Snell (1968) provided a general definition of residuals instead of limiting the scope to only linear models. Kay (1977) used the methods in Cox and Snell (1968) to derive the residuals for the proportional hazards regression model. The Cox-Snell residual for the i^{th} observation is defined as:

$$\hat{e}_i = \hat{\Lambda}_0(T_i)\exp\left\{\hat{\beta}'X_i\right\}, \quad i = 1, \ldots, n, \tag{3.4}$$

where $\hat{\Lambda}_0$ is the estimated cumulative baseline hazard, which can be obtained using the method of Breslow (1974) as

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

It was concluded that if the model was correctly specified and no observation was censored, the residuals should approximately exhibit the properties of a random sample of size n from a unit exponential distribution. This can be checked using an exponential Quantile-Quantile plot. Crowley and Hu (1977) used heart transplant survival data to illustrate the usage of Cox-Snell residuals. When censoring is present, however, the residuals are no longer approximately unit exponential.

3.1.3. Diagnostics based on martingale residuals

The martingale residual, which is a slight modification of Cox-Snell residual, also assists in assessing proportionality. It was first discussed by Lagakos (1981) and later by Barlow and Prentice (1988). Further work was done by Therneau *et al.* (1990). The martingale residual process is defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\exp\left\{\hat{\beta}'X_i(s)\right\} d\hat{\Lambda}_0(s), \quad i = 1, \ldots, n, \tag{3.5}$$

where $N_i(t)$ and $Y_i(s)$ are defined in Section 2.

The martingale residual is defined as the martingale residual process at the end of the study, i.e.,

$$\hat{M}_i = \delta_i - \int_0^\infty Y_i(s) \exp\left\{\hat{\beta}' X_i(s)\right\} d\hat{\Lambda}_0(s). \tag{3.6}$$

Asymptotically, $E(\hat{M}_i) = 0$ and $\mathrm{Cov}(\hat{M}_i, \hat{M}_j) = 0$ for $i \neq j$.

Lin *et al.* (1993) presented a procedure that used cumulative sums of martingale-based residuals, which have been sorted in advance by the order of follow-up time and/or value of a covariate. They considered the process

$$W(z) = \sum_{i=1}^{n} I\left(\hat{\beta}' X_i < z\right) \hat{M}_i. \tag{3.7}$$

The process $W(z)$ will be an approximate Gaussian process and fluctuate around 0 if the Cox model has been correctly specified. One can perform more formal tests to assess normality (e.g., Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling).

Grønnesby and Borgan (1996) concluded that when β is one-dimensional, (3.7) only checks the coding of the covariate. However, when β is of higher dimension, $W(z)$ cannot detect whether the effects of covariates vanish with time or not. Grønnesby and Borgan (1996) grouped the individuals after their linear predictions, i.e., replaced $I(\hat{\beta}' X_i < z)$ with $I(\hat{\beta}' X_i \in \Omega_l)$ in (3.7) for some interval Ω_l, which usually is a quartile group. This is equivalent to introducing the $g \times n$ grouping matrix Q, where g is the number of intervals and $Q_{l,i} = I(\hat{\beta}' X_i \in \Omega_l)$. Given the asymptotic distribution of the estimated martingale residuals, the grouped martingale residual process, $H(\cdot) = Q\hat{M}(\cdot)$, once properly normalized, converges to a mean zero multivariate Gaussian process. Then with $\hat{\Sigma}(t)$, such that $\hat{\Sigma}_{ij}$ is an estimate of the covariance between $H_{\Omega_i}(t)$ and $H_{\Omega_j}(t)$, the test statistic

$$T_C(t) = (H_{\Omega_1}(t), \ldots, H_{\Omega_g}(t))\hat{\Sigma}^{-1}(t)(H_{\Omega_1}(t), \ldots, H_{\Omega_g}(t))'$$

has an approximate χ^2_{g-1} distribution when the proportional hazards assumption holds.

Marzec and Marzec (1997a) established the asymptotic behavior of processes based on sums of weighted martingale-transformed residuals. They developed Kolmogorov-Smirnov and Cramér-von Mises types of omnibus tests using the fact that, in special cases, they appear to be transformed Brownian motions or Brownian bridges. As the derivation is complicated, please see Marzec and Marzec (1997a) for further details.

3.1.4. Graphical methods

In addition to formal tests, graphical methods to assess the proportional hazards assumption for categorical predictors have been developed by Cox (1979) and Arjas (1988). Hess (1995) summarized these methods and their extensions, including (1) plotting the Cox model's estimated survival curves $\hat{S}(t)$ against nonparametric (e.g., KM) estimates; (2) plotting the estimated cumulative hazard functions $-\log \hat{S}(t)$ against time and checking if their ratio is constant for any given t; (3) plotting the cumulative hazard functions against each other and checking if the slope is constant; (4) plotting the logarithm of the cumulative hazard functions, $\log(-\log \hat{S}(t))$, against time and checking if the curves are approximately parallel; (5) plotting the differences in the log cumulative hazard functions against time and checking if the curve of the differences are approximately constant; and (6) plotting the Schoenfeld residuals against time and checking for changes in patterns of scattering.

The aforementioned graphical methods all have one common limitation: they only apply to categorical predictors that have only a few levels. If a predictor has many levels or is continuous, the survival curves and cumulative hazard functions would no longer be informative. Therneau and Grambsch (2000) suggested plotting the cumulative Schoenfeld residuals ordered by survival times against survival times. If the proportional hazards assumption holds, the cumulative sum should be a random walk starting and ending at 0. These plots, however, can be difficult to read.

3.2. Functional forms

Martingale residuals, defined in Equation (3.6), play an important role in functional form diagnostics. Barlow and Prentice (1988) provided more detailed discussion and illustrated that plots of such residuals may provide insight to the choice of model form. Therneau *et al.* (1990) discussed the usage of martingale residuals in investigating the functional form of covariates. To examine a particular covariate, they suggest fitting a proportional hazards model omitting that covariate and computing the martingale residuals $\hat{M}_i$ as given in Equation (3.6). Then a smoothed plot of $\hat{M}_i$ versus the omitted covariate often gives approximately the correct functional form of the covariate (e.g., linear, quadratic) to place in the exponent of a Cox model. They also pointed out, however, that this plot did not work well when dealing with large covariate effects, and that it requires the covariate of interest to be uncorrelated with other covariates in the model.

Henderson and Milner (1991) noticed that plots of the martingale residuals against time, although useful, can exhibit systematic patterns which are not *a priori* predictable even when the model fails. They suggested two amendment approaches and gave an example for illustration. One approach was to superimpose the estimated mean when plotting residuals, which enables comparison between the observed patterns and the expected patterns. The other approach was to subtract the conditional expected value from each observed residual and scale it using its standard deviation, which could be consistently estimated from the data according to Barlow and Prentice (1988). Then the standardized residuals, when plotted, should be randomly scattered if the model is appropriate.

Grambsch (1995) proposed two aspects from which the martingale residual plot in Therneau *et al.* (1990) can be improved. One aspect is to modify the martingale plot for counting process data because of the close relationship between counting process models and Poisson regression. Suppose Z is the variable of interest. If a monotonic relationship between Z and the hazard $\lambda_i(t)$ is expected, a log-linear form is often adequate. The model

$$\lambda_i(t) = \exp\left(\sum_{j=1}^{p-1} \beta_j f_i(X_{ij}) + \alpha Z_i\right) \lambda_0(t), \quad i = 1, \ldots, n,$$

is fitted, and the expected count for the i^{th} individual is

$$\hat{E}_i = \int_0^{T_i} \exp\left(\sum_{j=1}^{p-1} \hat{\beta}_j f_j(X_{ij}) + \hat{\alpha} Z_i\right) \hat{\lambda}_0(t) dt,$$

where $\hat{\lambda}_0(t)$ is the estimated baseline hazard and $\hat{\alpha}$ is the estimated parameter for Z. The martingale residual in this case would be $\hat{M}_i = \delta_i - \hat{E}_i$, and the generalized linear model (GLM) partial residual is given by

$$\frac{\hat{M}_i}{\hat{E}_i} + \hat{\alpha} Z_i.$$

McCullagh and Nelder (1983) recommended plotting the partial residual against Z as an informal check for the correctness of the guess for functional form.

The other aspect mentioned by Grambsch (1995) comes from the penalized likelihood approach of Hastie and Tibshirani (1993). They assumed that the functional form of covariate X_j is an unknown, smooth function f_j, and proposed the alternative formulation

$$\lambda(t) = \lambda_0(t)\exp\left(\sum_{j=1}^{p} f_j(X_j)\right),$$

which enables estimation of all functional forms at the same time. To avoid overfitting, they maximized the penalized partial likelihood with penalty $\sum_{j=1}^{p} \nu_j \int f_j''(s)^2 ds$, where $\nu_j \geq 0,\ j = 1,\dots,p$, are smoothing parameters that can be tuned. Both approaches lead to approximately the same solution, but the latter is computationally more complex since the optimization is done within the kernel of the partial likelihood.

3.3. Outlying observations

A plot of martingale residuals against the linear prediction $\hat{\beta}'X(t)$ or the risk score $\gamma_i(t)$ defined in Section 2 often helps to identify the observations who have died too soon or lived too long, based on the assumed model. Nevertheless, having a range of $(-\infty, 1]$, the martingale residual is often heavily skewed, and may be misleading. Therneau *et al.* (1990) used a liver disease data set to demonstrate these scenarios, where the martingale residual plot indicated that some observations died too soon while in actuality they were not outliers at all. They pointed out that it is a favorable practice to transform the residuals to a more normal shaped distribution to help assess the prediction accuracies for individual subjects.

Inspired by the deviance residuals for GLM in McCullagh and Nelder (1983), Therneau *et al.* (1990) introduced the deviance residual for a Cox model:

$$d_i = \operatorname{sgn}\left(\hat{M}_i\right)\left[-2\left\{\hat{M}_i + \delta_i \log\left(\delta_i - \hat{M}_i\right)\right\}\right]^{\frac{1}{2}}, \tag{3.8}$$

where δ_i is again the censoring indicator for subject i. From the functional form it is apparent that the deviance residual is essentially a transformation of the martingale residual. Therneau *et al.* (1990) concluded that with less than 25% of censoring, the deviance residual is approximately normally distributed. With censoring rates greater than 40%, too many points will lie near 0 and make the distribution not normal, but the set of residuals is still symmetrized. Plotting d_i against $\hat{\beta}'X_i$ or $\exp(\hat{\beta}'X_i)$ will help identify potential outliers which have deviance residuals with too large absolute values.

Noticing that deviance residuals do not have a reference distribution and the normal approximation can sometimes be unsatisfactory (Fleming and Harrington, 1991), Nardi and Schemper (1999) proposed two new types of residuals: (i) the log-odds residual $L_i = \log[S_i(T_i)/\{1 - S_i(T_i)\}]$ and (ii) normal deviate residual $\eta_i = \Phi^{-1}\{S_i(T_i)\}$, $i = 1,\dots,n$, where Φ^{-1} is the inverse normal cumulative distribution function. Assuming $S_i(\cdot)$ is known, the sampling distribution for L_i is logistic with $E(L_i) = 0$ and $\text{var}(L_i) = \pi^2/3$, and standard normal for η_i. In practice, we use the predicted survival for observation i, $\hat{S}_i(T_i)$, to calculate $\hat{L}_i$ and $\hat{\eta}_i$, which converge in probability to L_i and η_i, respectively. Based on simulations, they concluded the performances of these two residuals when identifying outliers are better than that of the deviance residual since they are both unimodal, and the empirical distribution of deviance residual often becomes bimodal because of censoring. They suggested that one can use the quantiles of the normal distribution, ±1.64 and ±1.96, and of the logistic distribution ±2.94 and ±3.66, to help identify potential outliers.

3.4. Influential observations

The score vector U defined in Equation (2.5) is of great importance in influential diagnostics. Again, using the counting process formulation, the score residual for the i^{th} individual is defined to be

$$r_{Ui}\left(\hat{\beta}\right) = \int_0^\infty \left[X_i - \bar{x}\left(\hat{\beta}, s\right)\right] d\hat{M}_i(s), \quad i = 1, \ldots, n, \tag{3.9}$$

where $\bar{x}(\hat{\beta}, s)$ is the $\bar{x}(\beta, s)$ defined in Equation (2.6) evaluated at $\beta = \hat{\beta}$, and $\hat{M}_i(s)$ is defined in Equation (3.5).

In studying the influence of one observation, a general practice is to delete that observation, fit the model again, and compare the parameter estimates with those of the model fit on the complete data. Nevertheless, the Cox model is conceptually different from linear or generalized linear models in that it involves both parametric and nonparametric estimation. Therefore, an observation could be influential in terms of more than just regression coefficients. We review measures of both in this section.

3.4.1. Influence on regression coefficients

Cain and Lange (1984) presented a method for approximating the influence of individual cases on the Cox model's parameter estimates. Let $\hat{\beta}$ be the value of β that maximizes the partial likelihood (2.3) and $\hat{\beta}_{(i)}$ denote the estimate of β when observation i is deleted. They approximated $\hat{\beta} - \hat{\beta}_{(i)}$ by assigning to observation i weight w_i. Suppose $w_j = 1$ for any $j \neq i$. Then $\hat{\beta}$ can be regarded as a function of w_i and we have $\hat{\beta}(1) = \hat{\beta}$ and $\hat{\beta}(0) = \hat{\beta}_{(i)}$. The first-order Taylor series expansion about $w_i = 1$ gives:

$$\hat{\beta} - \hat{\beta}_{(i)} \simeq \frac{\partial \hat{\beta}}{\partial w_i}, \quad i = 1, \ldots, n,$$

where $\partial\hat{\beta}/\partial w_i$ is evaluated at $w_i = 1$. They evaluated the derivative treating the score vector U in Equation (2.5) as a function of $\hat{\beta}$ and w_i, and obtained:

$$\frac{\partial U}{\partial \hat{\beta}} \frac{\partial \hat{\beta}}{\partial w_i} + \frac{\partial U}{\partial w_i} = 0.$$

Notice that $\partial U/\partial\hat{\beta}$ is the negative observed information matrix defined in Equation (2.9). Hence we obtain

$$\frac{\partial \hat{\beta}}{\partial w_i} = \mathcal{I}^{-1}\left(\hat{\beta}\right) \frac{\partial U}{\partial w_i}. \tag{3.10}$$

The partial derivative $\partial U/\partial w_i$, when evaluated at $w_i = 1$, becomes exactly the score residual r_{Ui} in Equation (3.9). Therefore

$$\left(\frac{\partial \hat{\beta}}{\partial w_i}\right)_{w_i=1} = \mathcal{I}^{-1}\left(\hat{\beta}\right) r_{Ui}.$$

Let D be the $n \times p$ matrix with the i^{th} row being $\hat{\beta} - \hat{\beta}_{(i)}$, and r_U be the $n \times p$ matrix with the i^{th} row being the vector of score residuals for observation i. Then the above approximation, put into matrix form, becomes

$$D = r_U \mathcal{I}^{-1}\left(\hat{\beta}\right). \tag{3.11}$$

We call D the matrix of *dfbeta* residuals. When we divide D_{ij} by the observed standard deviation of $\hat{\beta}_i$, which is the square root of the i^{th} diagonal element of the inverse observed information matrix $\mathcal{I}^{-1}(\hat{\beta})$, we get D_S, the matrix of *dfbetas* residuals. Conventionally, the i^{th} observation is considered to be influential if $D_{S\,ij} > 1$ for small to medium datasets, and if $D_{S\,ij} > 2/\sqrt{n}$ for large datasets.

Reid and Crépeau (1985) presented influence functions for the Cox model to identify possible influential observations and gave the same statistic (3.11) as in Cain and Lange (1984).

Storer and Crowley (1985) pointed out that a good estimate of $\hat{\beta} - \hat{\beta}_{(i)}$ can also be obtained using an augmented regression model. The design matrix is augmented using a binary indicator variable for the i^{th} observation and taking a single Newton-Raphson step towards the fit of the augmented model gives the estimate of change in β. This estimate, they argued, is easy to compute.

3.4.2. Overall influence

Pettitt and Daud (1989) discussed the disadvantages of the approaches that try to approximate $\hat{\beta} - \hat{\beta}_{(i)}$. They concluded that only using single-case deletion statistics may cause some cases to be masked, i.e., the deleted observation may influence the value of the test statistic enough so that an actual outlier is not declared as outlier. They suggested changing the weights of each observation, and studying the change in the likelihood caused by this perturbation. They adopted the approach of Cook (1986) and defined the likelihood displacement to be

$$\mathrm{LD}(w) = 2\left[\ell\left(\hat{\beta}\right) - \ell\left(\hat{\beta}(w)\right)\right], \tag{3.12}$$

where $\hat{\beta}(w)$ maximizes the weighted partial likelihood

$$\mathrm{PL}_w(\beta) = \prod_{i=1}^{n} \frac{\exp\left(\beta' X_i(t)\delta_i w_i\right)}{\left[\sum_{j\in R_i} w_j \exp\left(\beta' X_j(t)\right)\right]^{\delta_i w_i}}. \tag{3.13}$$

The weighting scheme of $w_i = 1, i \neq j$ and $w_j = 0$ in Cain and Lange (1984) is an appropriate and specific case. Using second-order approximation, we have

$$\ell\left(\hat{\beta}\right) - \ell\left(\hat{\beta}(w)\right) \approx \frac{1}{2}\left[\hat{\beta} - \hat{\beta}(w)\right]^T \mathcal{I}\left(\hat{\beta}\right)\left[\hat{\beta} - \hat{\beta}(w)\right].$$

Let $U_w(\beta)$ be the score function corresponding to the weighted partial log-likelihood. With another approximation that

$$\frac{\partial\hat{\beta}(w)}{\partial w^T} = \mathcal{I}^{-1}\left(\hat{\beta}\right)\frac{\partial U_w(\beta)}{\partial w^T},$$

which is essentially the matrix form of Equation (3.10), LD(w) reduces to

$$\mathrm{LD}(w) \approx (w_0 - w)^T r_U \mathcal{I}^{-1}\left(\hat{\beta}\right) r_U^T (w_0 - w), \tag{3.14}$$

where w_0 is a vector of 1's and r_U is the score residual matrix. The approach of Cook (1986) looks for an unit-length $l_{n\times 1}$ that maximizes $l^T B l$, where $B = r_U \mathcal{I}^{-1}(\hat{\beta}) r_U^T$. The maximum $\xi_{\max}$ is the largest eigenvalue of B, and is attained when $l_{\max}$ is the corresponding eigenvector. Cook (1986) concluded that $\xi_{\max} > 1$ indicates notable local sensitivity, and that a locally influential observation must be globally influential, although the reverse is not necessarily true.

Weissfeld (1990) adopted the idea of Cook (1986) to measure the change in likelihood function by computing its curvature. Originally, in Cook's work, the change could be caused by perturbations in the score vector or the covariates. Weissfeld (1990) proposed for the Cox model three ways to perturb the data: weighting the observations in the log partial likelihood using a vector $\boldsymbol{w}$ of weights, adding a vector to the vector of censoring indicators $(\delta_1, \ldots, \delta_n)$, and adding a scaled weight vector $\boldsymbol{w}$ to the covariates, where the scale is usually the standard deviation of the corresponding coefficient. Then take $\ddot{F} = \Delta^T \mathcal{I}^{-1}(\hat{\beta})\Delta$, where $\mathcal{I}^{-1}(\hat{\beta})$ is the inverse of the observed information matrix in Equation (2.9) and Δ is the partial derivative matrix of the score vector U to the weights, which takes different forms for the three perturbation schemes. The maximum eigenvalue of $\ddot{F}$, $C_{\max}$, is informative in that large or small values point to possible influential observations. It was concluded that perturbation of the covariates is useful for locating observations that influence the estimated coefficients, and the other two pertubations will help detect observations that may impact the results of likelihood ratio tests. It was also indicated that the proposed approach is capable of detecting influential observations caused by masking.

Barlow (1997) proposed a modification of the method in Pettitt and Daud (1989). Their approach replaces $\mathcal{I}(\hat{\beta})$ in Equation (3.14) using the inverse of the robust covariance matrix in Lin and Wei (1989). The substitution, upon further derivation, provides a scalar measure of influence with known mean to be the ratio of number of events and number of observations, and range of (0,1). The approach can also be generalized to include designs with multiple failures and to case-cohort designs. They illustrated the usage of this method by plotting the calculated influence measure against the covariate of interest, and visually looking for any particularly influential observations.

In addition to traditional delete-one approaches, Wei and Kosorok (2000) developed case interaction influence measures for unmasking observations masked by other observations in the Cox model. They proposed the following statistic to assess the joint influence of observations i and j:

$$\begin{aligned}
-\left(\hat{\beta} - \hat{\beta}_{(j)} - \hat{\beta}_{(i)} + \hat{\beta}_{(i,j)}\right) &= \left(\hat{\beta}_{(i)} - \hat{\beta}_{(i,j)}\right) - \left(\hat{\beta} - \hat{\beta}_{(j)}\right) \\
&= \left(\hat{\beta}_{(j)} - \hat{\beta}_{(i,j)}\right) - \left(\hat{\beta} - \hat{\beta}_{(i)}\right) \\
&= \left(\hat{\beta} - \hat{\beta}_{(i,j)}\right) - \left\{\left(\hat{\beta} - \hat{\beta}_{(i)}\right) + \left(\hat{\beta} - \hat{\beta}_{(j)}\right)\right\},
\end{aligned}$$

where $\hat{\beta} - \hat{\beta}_{(i,j)}$ and $\hat{\beta}_{(i)} - \hat{\beta}_{(i,j)}$ are related to the joint influence and conditional influence in Lawrance (1995). On one hand, if the value of the test statistic is small, we conclude that the parameter estimate is not significantly influenced by the deletion of one observation, with or without incorporating the other observation in estimation. A large value, on the other hand, would imply that the joint influence of these two observations is significantly different from the sum of their individual influences, and the identified pairs need further investigation. In cases where two moderately influential observations have substantial joint influence, or where two individually influential observations have little joint influence, however, their proposed diagnostic cannot identify them.

Zhu *et al.* (2015) investigated case-deletion measures, conditional martingale residuals, and score residuals for the Cox model with missing covariate values. They proposed the Q-distance to examine the effects of deleting individual observations on the estimates of finite-dimensional and infinite-dimensional parameters. They also addressed the problem of quantifying influence by introducing a detection probability of being influential for each observation and for any case-deletion measure. A large value of detection probability is an indicator of being influential. The forms and derivation of the Q-distance and the detection probability are complicated; the interested reader should see Zhu *et al.* (2015) for full details.

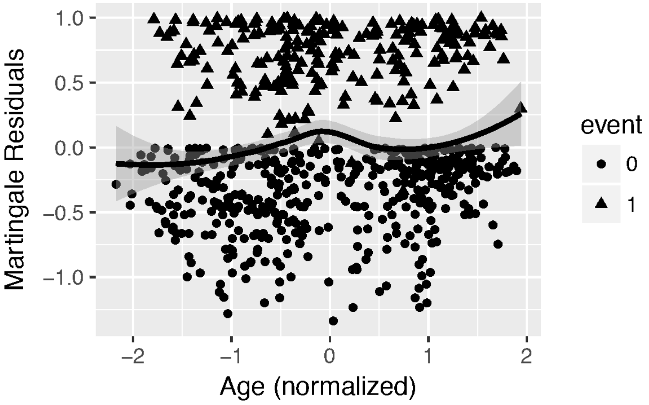

Figure 1: *Plot of martingale residual of the model excluding Age against Age.*

4. A case study

The dental restoration longevity data, provided by the University of Iowa College of Dentistry's Geriatric and Special Needs (SPEC) Clinic (see Caplan *et al.*, 2017), is used as a case study to demonstrate the diagnostic methods of the Cox model. For this analysis, electronic data was obtained during the 5-year period from 1/1/1995–12/31/1999. The health record numbers were scrambled by IT personnel to ensure that no Personal Health Information was included. Subsequently, the Institutional Review Board at the University of Iowa declared that this project is exempt from Human Subjects Review, due to the anonymous nature of the data.

We identified 697 unique patients who went to the SPEC Clinic to treat their molars upon their first visit and received restoration in amalgam, composite, or glass ionomer. The follow-up of their visits began on the date of restoration. Any restoration that was replaced with another intracoronal or extracoronal restoration, accessed for endodontic therapy, or extracted was deemed to have undergone an event. If the restoration results in an event, the event date would become the end of follow-up. Restorations that did not incur an event are considered censored up to the date of the patient's second-to-last visit to any College of Dentistry's clinic. Among the 697 patients, 228 experienced an event during the follow-up, giving a censoring rate of 67.3%.

We considered the following covariates: Gender, Age at receiving restoration (centered and scaled), Occupation (Faculty, Non-faculty) and Size (Small, Medium, Large). Analysis was performed using the **survival** package in R, and figures were produced using the `survminer` (Kassambara and Kosinski, 2017), `ggplot2` (Wickham, 2009) and `ggfortify` (Tang *et al.*, 2016) packages. R code for analysis is available in our supplementary material.

4.1. Functional form

As suggested in Section 3.2, we should determine the appropriate form of covariates to include in the model before testing for proportionality. Age is the only continuous covariate whose form needs to be assessed. We use the methods of Therneau *et al.* (1990): fit a model excluding Age and obtain its martingale residuals. The martingale residuals are plotted against Age in Figure 1. We also superimpose the loess pointwise confidence band. The curvy behavior of the loess fit indicates that we should consider higher orders of Age.

Table 2: Proportionality test results for Model 1 and Model 2

	Model 1		Model 2	
	χ^2 Stat	p-value	χ^2 Stat	p-value
Male	0.586	0.444	0.429	0.513
Age	4.029	0.045	3.555	0.059
Age^2	-	-	0.638	0.424
Non-Faculty	0.429	0.513	0.558	0.455
SizeMedium	1.560	0.212	1.711	0.191
SizeSmall	0.298	0.585	0.416	0.519
Global	6.788	0.237	6.932	0.327

Table 3: Cox regression results for tooth restoration failure for the Model 2

	Estimate	exp(Estimate)	Standard error	Z Stat	p-value
Male	−0.221	0.802	0.137	−1.612	0.107
Age	0.206	1.228	0.0076	2.709	0.007
Age^2	−0.092	0.912	0.085	−1.075	0.282
Non-Faculty	0.116	1.123	0.146	0.795	0.427
SizeMedium	−0.140	0.869	0.165	−0.850	0.395
SizeSmall	−0.510	0.601	0.169	−3.018	0.003

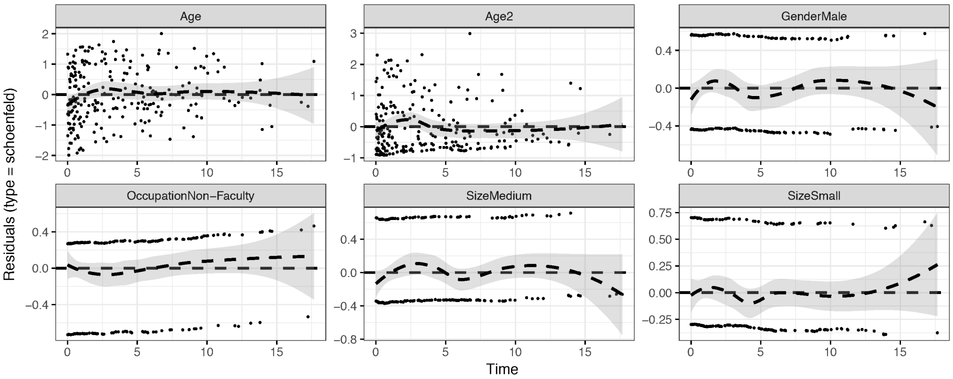

Figure 2: *Schoenfeld residuals for each covariate against survival time.*

4.2. Proportional hazards

As suggested in Section 4.1, we consider including the square of Age (Age^2) in the model. We fit two models: one with only linear Age effects (Model 1) and another model with linear and quadratic Age effects (Model 2) to assess the improvement to the model when correcting the functional form. To assess the proportional hazards assumption, we used the cox.zph function from the **survival** package to obtain both the individual χ^2_1 statistics for each covariate and the global χ^2_p statistic for each model. The test results are summarized in Table 2. While both models passed the global test, Age in Model 1 did not pass the individual test at the 0.05 level. In Model 2, however, both Age and Age^2 pass the individual test of proportionality at the 0.05 level.

The parameter estimates of Model 2 are summarized in Table 3. Restorations for males tend to fail later than for females, while restorations for older patients tend to fail sooner. Compared to large restorations, medium and small restorations are less likely to fail.

As mentioned in Section 3.1.1, when the proportional hazards assumption holds, the Schoenfeld

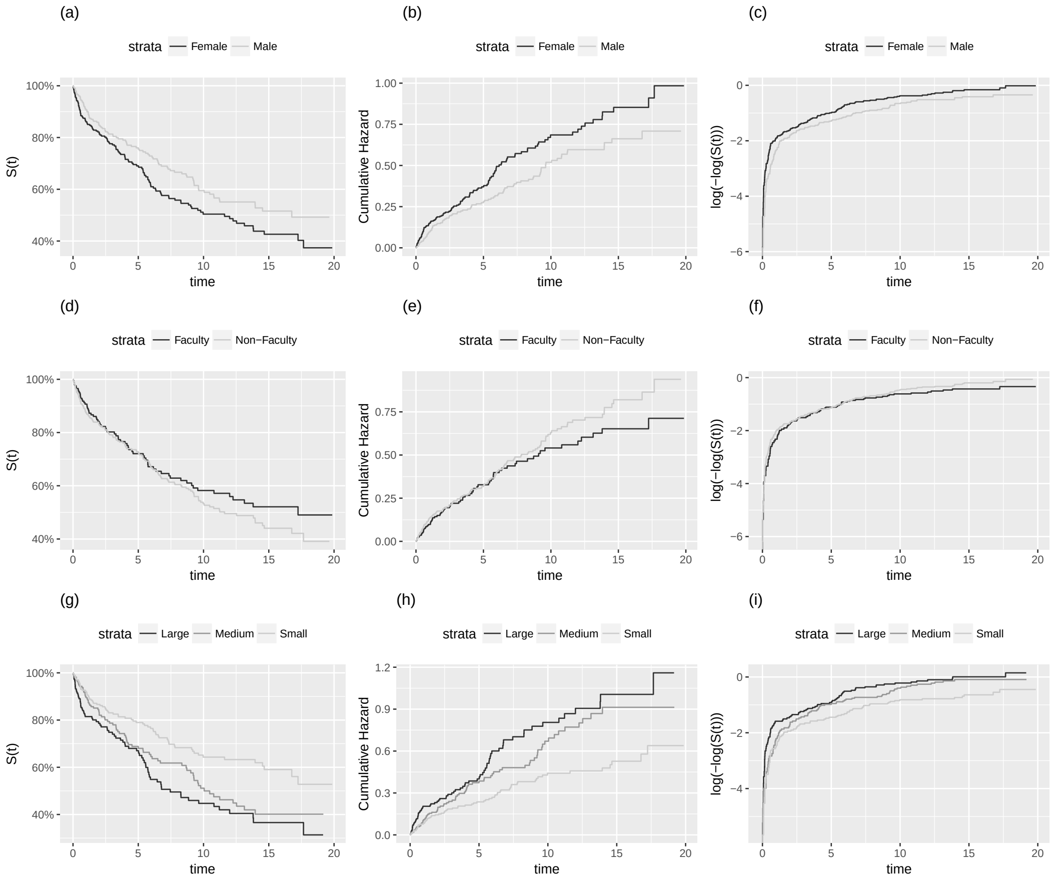

Figure 3: *Estimated survival curves, cumulative hazards and log-log transformed survival curves for categorical covariates. The first row is for Gender, the second row is for Occupation, and the third row is for Size.*

residuals will be close to zero. Therefore a plot of the Schoenfeld residuals against survival times would be informative (Schoenfeld, 1982). Figure 2 shows the plots for Model 2. For all six covariates, the smoothed pointwise confidence bands are all around 0, which again confirms that there is no obvious evidence against the proportional hazards assumption.

For the three categorical covariates (Gender, Occupation, and Size), we also utilize the graphical methods in Section 3.1.4 to check the proportional hazards assumption. For each covariate, we plot the estimated survival curves ($\hat{S}(t)$), the cumulative hazards ($-\log\hat{S}(t)$) and the log-log transformed survival ($\log(-\log\hat{S}(t))$) against survival times in Figure 3. The three plots for Gender indicate that the hazards of the two gender strata are proportional, but the lack of large discrepancy indicates that this proportional effect is not significant. Similarly, the ignorable discrepancy between the two occupation strata tells the same story. The three plots for restoration size strata, however, are more informative, in that although the proportionality effect is small between SizeLarge and SizeMedium, it is highly significant between SizeLarge and SizeSmall.

As mentioned in Section 3.1, the martingale residual can be used to graphically assess the proportional hazards assumption as well. We plot the cumulative sum of martingale residuals ordered by

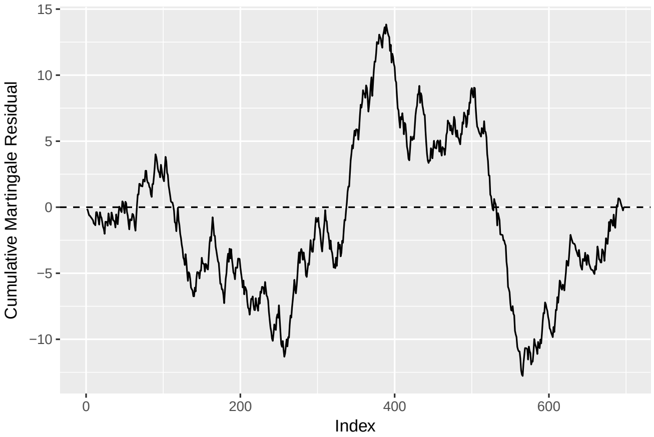

Figure 4: *Cumulative sum of martingale residuals of Model 2, ordered by Age.*

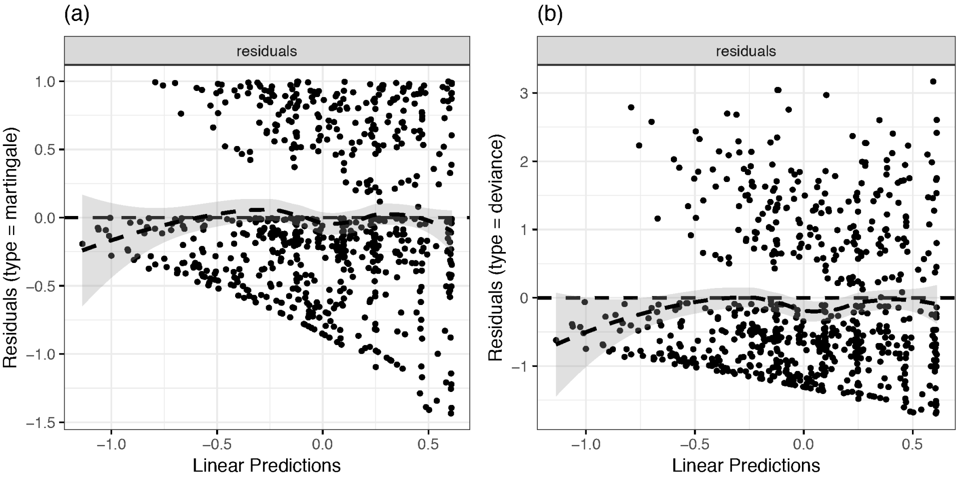

Figure 5: *Plot of (a) martingale and (b) deviance residuals of Model 2.*

Age in Figure 4. The curve fluctuates around zero as expected.

4.3. Outlying observations

As suggested in Section 3.3, both the martingale residual and the deviance residual are useful for identifying outlying observations, but the deviance residual is less skewed and therefore more useful. We plot both residuals against the linear predictions, $\hat{\beta}'X$, in Figure 5. In Figure 5(a), the martingale residuals do not vary much against the linear predictions, and fail to identify any outlying observations. Using ±1.96 as thresholds, the deviance residuals plotted in Figure 5(b) identify 34 potential outliers. Upon further investigation, these subjects turned out to be much younger than other subjects (46.6 vs. 55.1) but their restorations failed very soon. Due to the high censoring rate, however, the normal-approximation-based thresholds may not be appropriate.

We also use the log-odds residual and the normal deviate residual discussed in Section 3.3 to

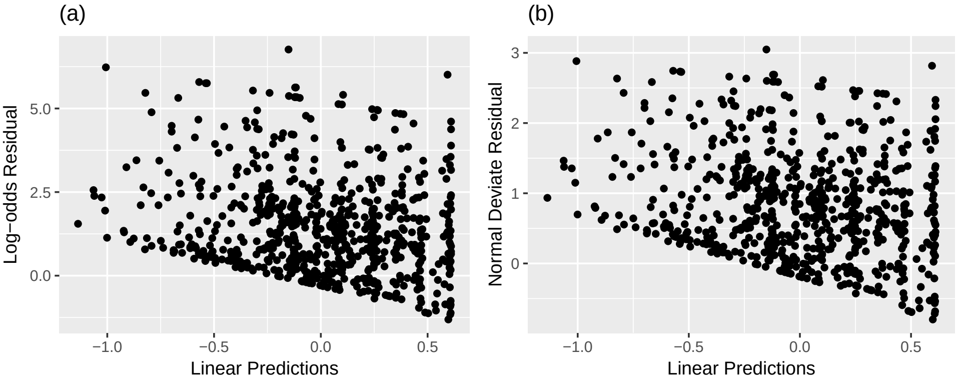

Figure 6: *Plot of (a) log-odds and (b) normal deviate residuals of Model 2.*

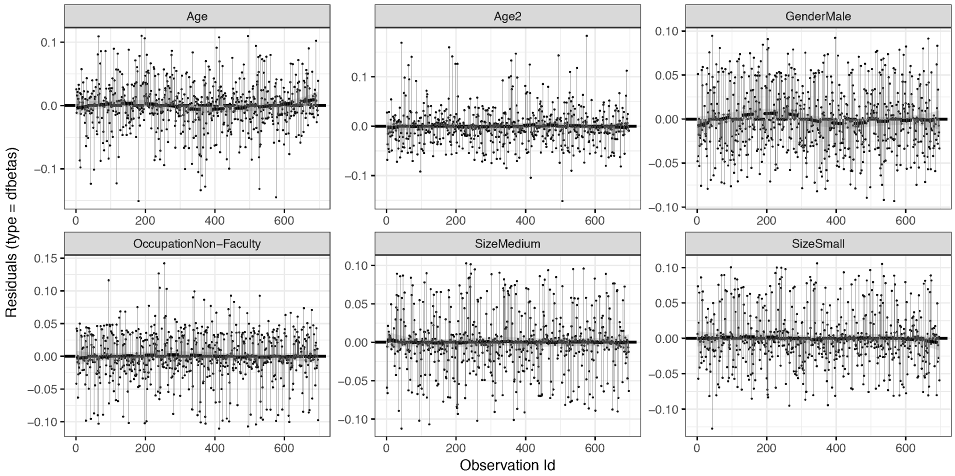

Figure 7: *Dfbetas residuals for covariates of Model 2.*

look for potential outliers. Both the log-odds residual in Figure 6(a) and the normal deviate residual in Figure 6(b) identify the same set of 67 potential outliers, which is bigger than the set of outliers identified by the deviance residual. This set, however, still consists of younger individuals (51.7 vs. 55.1) whose restorations failed very soon.

4.4. Influential observations

We use the methods in Section 3.4 to perform influential diagnostics. We first look at influence of observations on parameter estimates and plot the *dfbetas* residuals in Figure 7. As illustrated, no observation caused any parameter change of more than 15% of that parameter's standard error. Considering that there are 697 observations, we can conclude there are no significantly influential observations.

We also present the likelihood displacement approach in Figure 8. The absence of particularly large likelihood displacements further confirms our conclusion from Figure 7.

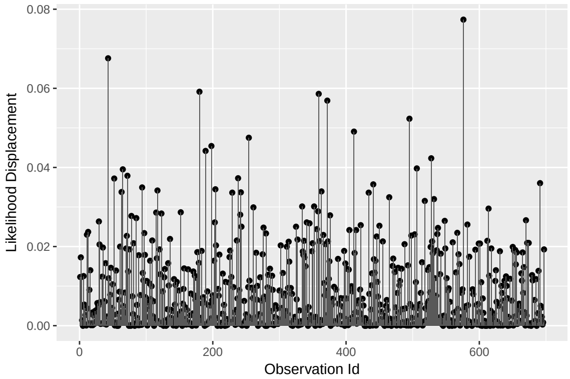

Figure 8: *Likelihood displacement caused by dropping each observation.*

5. Discussion

With such wide usage across a variety of disciplines, the importance of Cox regression for modeling time-to-event data cannot be overstated. As a consequence, one must consider the validity of the results from such an analysis before making any conclusions. This article presents a review of diagnostic methods for the Cox proportional hazards model, including methods for identifying violations of the proportional hazards assumption, appropriate functional forms of continuous covariates, outlying observations, and influential observations. Using a non-linear functional form of covariates can often improve model fit, while detected outlying or influential observations should be investigated further before taking any action.

Violations of the proportional hazards assumption can be addressed in several ways, the most common include the use of time-varying coefficients and stratified models. Time-varying coefficients, which are left out in this article, make a major source of nonproportionality and are often the alternatives to test against when assessing the proportional hazards assumption. More flexible models that incorporate this effect have been studied by Murphy and Sen (1991), Hastie and Tibshirani (1993), Verweij and van Houwelingen (1995), Sargent (1997), Marzec and Marzec (1997b), Cai and Sun (2003), Tian *et al.* (2005), Fan *et al.* (2006), and more recently by Chen *et al.* (2012). In practice, the graphical tools in the **survival** package enable us to check if there are any time-varying coefficients (Therneau *et al.*, 2017). Another popular approach for addressing non-proportionality is using a stratified Cox model. In this case, it is assumed that individuals in different strata have different baseline hazard functions, but all other predictor variables satisfy the proportional hazards assumption within each stratum. Related chapters can be found in Therneau and Grambsch (2000), Kalbfleisch and Prentice (2002), Lawless (2003), and Collett (2015).

The Cox model has also been extended to the analysis of interval-censored survival data. Such models have been studied by Finkelstein (1986), Farrington (2000), Goggins and Finkelstein (2000) and recently Heller (2010). In particular, Farrington (2000) proposed the counterparts to the Cox-Snell, martingale, deviance, and Schoenfeld residuals and illustrated their usage in model diagnostics under the interval-censored framework.

Acknowledgements

The authors are very grateful to Professor Jun Yan for his helpful discussions and comments on an earlier version of this manuscript.

References

Arjas E (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model, *Journal of the American Statistical Association*, **83**, 204–212.

Barlow WE (1997). Global measures of local influence for proportional hazards regression models, *Biometrics*, **53**, 1157–1162.

Barlow WE and Prentice RL (1988). Residuals for relative risk regression, *Biometrika*, **75**, 65–74.

Breslow N (1974). Covariance analysis of censored survival data, *Biometrics*, **30**, 89–99.

Cai Z and Sun Y (2003). Local linear estimation for time-dependent coefficients in Cox's regression models, *Scandinavian Journal of Statistics*, **30**, 93–111.

Cain KC and Lange NT (1984). Approximate case influence for the proportional hazards regression model with censored data, *Biometrics*, **40**, 493–499.

Caplan DJ, Li Y, Wang W, *et al.* (2017). Restoration longevity among geriatric and adult special needs patients, *bioRxiv*, https://doi.org/10.1101/202069.

Chappell R (1992). A note on linear rank tests and Gill and Schumacher's tests of proportionality, *Biometrika*, **79**, 199–201.

Chen K, Lin H, and Zhou Y (2012). Efficient estimation for the Cox model with varying coefficients, *Biometrika*, **99**, 379–392.

Collett D (2015). *Modelling Survival Data in Medical Research*, CRC press, London.

Cook RD (1986). Assessment of local influence, *Journal of the Royal Statistical Society Series B (Methodological)*, **48**, 133–169.

Cox DR (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society Series B (Methodological)*, **34**, 187–220.

Cox DR (1975). Partial likelihood, *Biometrika*, **62**, 269–276.

Cox DR (1979). A note on the graphical analysis of survival data, *Biometrika*, **66**, 188–190.

Cox DR and Snell EJ (1968). A general definition of residuals, *Journal of the Royal Statistical Society Series B (Methodological)*, **30**, 248–275.

Crowley J and Hu M (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association*, **72**, 27–36.

Fan J, Lin H, and Zhou Y (2006). Local partial-likelihood estimation for lifetime data, *The Annals of Statistics*, **34**, 290–325.

Farrington CP (2000). Residuals for proportional hazards models with interval-censored survival data, *Biometrics*, **56**, 473–482.

Finkelstein DM (1986). A proportional hazards model for interval-censored failure time data, *Biometrics*, **42**, 845–854.

Fisher LD and Lin DY (1999). Time-dependent covariates in the Cox proportional-hazards regression model, *Annual Review of Public Health*, **20**, 145–157.

Fleming TR and Harrington DP (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

Gill R and Schumacher M (1987). A simple test of the proportional hazards assumption, *Biometrika*, **74**, 289–300.

Goggins WB and Finkelstein DM (2000). A proportional hazards model for multivariate interval-censored failure time data, *Biometrics*, **56**, 940–943.

Grambsch PM (1995). Goodness-of-fit and diagnostics for proportional hazards regression models. In *Recent Advances in Clinical Trial Design and Analysis* (pp. 95–112), Springer, Boston.

Grambsch PM and Therneau TM (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, **81**, 515–526.

Grant S, Chen YQ, and May S (2014). Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates, *Lifetime Data Analysis*, **20**, 355–368.

Grønnesby JK and Borgan Ø (1996). A method for checking regression models in survival analysis based on the risk score, *Lifetime Data Analysis*, **2**, 315–328.

Harrell FE (1986). The PHGLM procedure. In *SUGI Supplemental Library Users Guide* (5th ed, pp. 437–466), SAS Institute Inc., Cary.

Hastie T and Tibshirani R (1993). Varying-coefficient models, *Journal of the Royal Statistical Society Series B (Methodological)*, **55**, 757–796.

Heller G (2010). Proportional hazards regression with interval censored data using an inverse probability weight, *Lifetime Data Analysis*, **17**, 373–385.

Henderson R and Milner A (1991). On residual plots for relative risk regression, *Biometrika*, **78**, 631–636.

Hess KR (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression, *Statistics in Medicine*, **14**, 1707–1723.

Jones BS and Branton RP (2005). Beyond logit and probit: Cox duration models of single, repeating, and competing events for state policy adoption, *State Politics & Policy Quarterly*, **5**, 420–443.

Kalbfleisch JD and Prentice RL (2002). *The Statistical Analysis of Failure Time Data* (2nd ed), John Wiley & Sons, New York.

Kang H and Han ST (2004). Prediction of the probability of customer attrition by using Cox regression, *Communications for Statistical Applications and Methods*, **11**, 227–233.

Kassambara A and Kosinski M (2017). *survminer: Drawing Survival Curves Using "ggplot2"* (R package version 0.4.0).

Kay R (1977). Proportional hazard regression models and the analysis of censored survival data, *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **26**, 227–237.

Keele L (2010). Proportionally difficult: testing for nonproportional hazards in Cox models, *Political Analysis*, **18**, 189–205.

Kotz S and Johnson NL (1992). *Breakthrough in Statistics: Volume I, Foundations and Basic Theory*, Springer-Verlag, Berlin.

Kupets O (2006). Determinants of unemployment duration in Ukraine, *Journal of Comparative Economics*, **34**, 228–247.

Lagakos SW (1981). The graphical evaluation of explanatory variables in proportional hazard regression models, *Biometrika*, **68**, 93–98.

Lane WR, Looney SW, and Wansley JW (1986). An application of the Cox proportional hazards model to bank failure, *Journal of Banking & Finance*, **10**, 511–531.

Lawless JF (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed), John Wiley & Sons, New York.

Lawrance AJ (1995). Deletion influence and masking in regression, *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 181–189.

Lin DY (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *Journal of the American Statistical Association*, **86**, 725–728.

Lin DY and Wei LJ (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association*, **84**, 1074–1078.

Lin DY, Wei LJ, and Ying Z (1993). Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika*, **80**, 557–572.

Marzec L and Marzec P (1997a). Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models, *The Annals of Statistics*, **25**, 683–714.

Marzec L and Marzec P (1997b). On fitting Cox's regression model with time-dependent coefficients, *Biometrika*, **84**, 901–908.

McCullagh P and Nelder JA (1983). *Generalized Linear Models*, Chapman & Hall, London.

Moreau T, O'Quigley J, and Lellouch J (1986). On D. Schoenfeld's approach for testing the proportional hazards assumption, *Biometrika*, **73**, 513–515.

Moreau T, O'Quigley J, and Mesbah M (1985). A global goodness-of-fit statistic for the proportional hazards model, *Applied Statistics*, **34**, 212–218.

Murphy SA and Sen PK (1991). Time-dependent coefficients in a Cox-type regression model, *Stochastic Processes and Their Applications*, **39**, 153–180.

Nagelkerke NJD, Oosting J, and Hart AAM (1984). A simple test for goodness of fit of Cox's proportional hazards model, *Biometrics*, **40**, 483–486.

Nardi A and Schemper M (1999). New residuals for Cox regression and their application to outlier screening, *Biometrics*, **55**, 523–529.

O'Quigley J and Pessione F (1989). Score tests for homogeneity of regression effect in the proportional hazards model, *Biometrics*, **45**, 135–144.

Park S and Hendry DJ (2015). Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses, *American Journal of Political Science*, **59**, 1072–1087.

Pettitt AN and Daud IB (1989). Case-weighted measures of influence for proportional hazards regression, *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **38**, 51–67.

Reid N and Crépeau H (1985). Influence functions for proportional hazards regression, *Biometrika*, **72**, 1–9.

Sargent DJ (1997). A flexible approach to time-varying coefficients in the Cox regression setting, *Lifetime Data Analysis*, **3**, 13.

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

Schoenfeld D (1982). Partial residuals for the proportional hazards regression model, *Biometrika*, **69**, 239–241.

Storer BE and Crowley J (1985). A diagnostic for Cox regression and general conditional likelihoods, *Journal of the American Statistical Association*, **80**, 139–147.

Tang Y, Horikoshi M, and Li W (2016). ggfortify: Unified interface to visualize statistical result of popular R packages, *The R Journal*, **8**, 478–489.

Therneau TM (2017). *A Package for Survival Analysis in S*, (Version 2.41-3).

Therneau T, Crowson C, and Atkinson E (2017). Using time dependent covariates and time dependent coefficients in the Cox model, Retrieved November 10, 2017, from: ftp://ftp.br.debian.org/CRAN /web/packages/survival/vignettes/timedep.pdf

Therneau TM and Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, Berlin.

Therneau TM, Grambsch PM, and Fleming TR (1990). Martingale-based residuals for survival models, *Biometrika*, **77**, 147–160.

Tian L, Zucker D, and Wei LJ (2005). On the Cox model with time-varying regression coefficients, *Journal of the American Statistical Association*, **100**, 172–183.

Verweij PJM and van Houwelingen HC (1995). Time-dependent effects of fixed covariates in Cox regression, *Biometrics*, **51**, 1550–1556.

Wei LJ (1984). Testing goodness of fit for proportional hazards model with censored observations, *Journal of the American Statistical Association*, **79**, 649–652.

Wei WH and Kosorok MR (2000). Masking unmasked in the proportional hazards model, *Biometrics*,

56, 991–995.

Weissfeld LA (1990). Influence diagnostics for the proportional hazards model, *Statistics & Probability Letters*, **10**, 411–417.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York.

Winnett A and Sasieni P (2001). Miscellanea. A note on scaled Schoenfeld residuals for the proportional hazards model, *Biometrika*, **88**, 565–571.

Xue X, Xie X, Gunter M, *et al.* (2013). Testing the proportional hazards assumption in case-cohort analysis, *BMC Medical Research Methodology*, **13**, 88.

Zhu H, Ibrahim JG, and Chen MH (2015). Diagnostic measures for the Cox regression model with missing covariates, *Biometrika*, **102**, 907–923.

Received November 9, 2017; Revised November 13, 2017; Accepted November 13, 2017