Int. Journal of Math. Analysis, Vol. 7, 2013, no. 53, 2645 - 2654
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ijma.2013.38193

Several Types of Residuals in Cox Regression Model: An Empirical Study

Anwar Fitrianto*

Department of Mathematics, Faculty of Science/
Universiti Putra Malaysia, Malaysia
*anwarstat@gmail.com

Laboratory of Applied & Computational Statistics,
Institute for Mathematical Research,
Universiti Putra Malaysia

Department of Statistics, Faculty of Mathematics and Natural Sciences
Bogor Agricultural University, Indonesia

Rebecca Loo Ting Jiin

Department of Mathematics, Faculty of Science
Universiti Putra Malaysia, Malaysia

Abstract

There are several methods for calculating residual in survival analysis, especially in Cox regression model by which each method has specific use, such as goodness-of-fit, to identify possible outliers and influential observations, or in general to check necessary assumptions. In this article, we study four methods of residuals, namely Schoenfeld, Martingale, deviance, and score residuals and we applied those methods on cardiovascular data.

Keywords: Cox regression, goodness-of-fit, outliers, influential observations, residuals

1 Introduction

Adequacy of a fitted model needs to be assessed after a model has been assessed. Diagnostic procedures for model checking are known as essential parts of a modeling process. Many for this procedure are based on residuals. In survival analysis, especially when we build a Cox's proportional hazard model of Cox [2], few types of residuals can be considered for different purposes [6]. Several useful diagnostic tools which are based on residuals are (1) Schoenfeld residual for checking the proportional hazards assumption for a covariate, (2) Martingale residual used to examine overall test of the goodness-of-fit of a Cox model, (3) Deviance residual for detection of poorly predicted observations, and (4) Score residual for determination of influential observations.

Schoenfeld residual was purposed by Schoenfeld [5] as partial residual that is essential to interpretation of violation of the proportional hazards assumptions. Schoenfeld [5] mentioned that *i*th residual can be plotted against t_i to test the assumption in which residuals do not depend on time. He defined a partial residual as the different between the observed value of X_i and its conditional expectation given the risk set R_i. It can be written as the following equation:

$$\hat{r}_{ik} = X_{ik} - \hat{E}\left(X_{ik}|R_i\right) \tag{1}$$

in the vector $\hat{r}_i = \left(\hat{r}_{i1},\ldots,\hat{r}_{ip}\right)'$. He then expanded $E\left(X_{ik}|R_i\right)$ about $g\left(t_i\right)=0$ and obtain

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

A plot of $\hat{r}_{ik}$ versus t_i will be centered about 0 if proportional hazards holds $E\left(\hat{r}_i\right) \approx 0$. Kumar and Klefsjö [4] reviewed the graphical method purposed by Schoenfeld [5] for testing the assumptions of proportional hazards model. They found that the plots of the estimated partial residuals against the time should be randomly scattered around zero, since the conditional expectation of these residuals, given the risk sets, will be approximately zero and asymptotically uncorrelated. A large value of the estimate of the partial residual will indicate that the corresponding time to failure is more unlikely to be explained by the proportional hazards model.

Grambsch and Therneau [3] suggested scale the Schoenfeld residual by an estimator of its variance which yields a residual with greater diagnostic power than the unscaled residuals. The vector of scaled Schoenfeld residual is the product of the inverse of the covariance matrix times the vector of residuals. Let

$$\hat{r}_i^* = \left[\hat{Var}\left(\hat{r}_i\right)\right]^{-1}\hat{r}_i \tag{3}$$

be the scaled Schoenfeld residual. Then

$$E\left(\hat{r}_i^*\right) \approx g\left(t_i\right), \tag{4}$$

where the $\hat{r}_i$ is the partial residual at Equation (1) that was purposed by Schoenfeld [5].

Few years later, Barlow and Prentice [1] proposed another type of residual, which then was named as Martingale-based residual or Martingale residual. It is defined to as the following expression:

$$\hat{M}_i\left(t\right) = N_i\left(t\right) - \int_0^t Y_i\left(s\right)e^{\hat{\beta}' X_i(s)}d\hat{H}_0\left(s\right), \tag{5}$$

with $\hat{M}_i$ as a shorthand for $\hat{M}_i\left(\infty\right)$. The H_0 is the cumulative baseline hazard, and the $N_i\left(t\right)$ is the counting process for the *i*th subject in the set of indicates the number of observed events experienced over the time *t*. The residual at the Equation (5) reduces to more simple form as follows:

$$\hat{M}_i = \delta_i - \hat{H}_0\left(\tau_i\right)e^{\hat{\beta}' X_i}. \tag{6}$$

where τ_i denotes the observation time for subject *i* and δ_i known as the vital status for the Cox model, can identified as uncensored ($\delta_i = 1$) or censoring ($\delta_i = 0$). Martingale residuals compare the "observed" to "expected". The first part of Equation (6), δ_i is considered as observed number of events for *i*th observation with 1 or 0 while second part of this equation. The second part is $\hat{H}_0\left(\tau_i\right)e^{\hat{\beta}' X_i}$, which is the expected number of events for *i*th subject, accounting for censoring. So, the martingale residual is likely having the "excess" number of events and sum of these residuals which will be equal to 0.

According to Therneau et al. [6] one deficiency of the martingale residual M_i, particularly in the single event setting of Cox model, is that it is heavily skewed. This skewness distorts the appearance of a standard residual plot, makes them hard to use to identify outliers. Another disadvantage is it has maximum value of +1 and $-\infty$ for its minimum value.

Due to the drawback of the martingale residual, Therneau et al. [6] proposed deviance residual of the Cox regression. It is much more symmetrically distributed about zero and defined as

$$d_i = \text{sgn}\left(\hat{M}_i\right)\left[-2\left\{\hat{M}_i + \delta_i \log\left(\delta_i - \hat{M}_i\right)\right\}\right]^{\frac{1}{2}}, \quad (7)$$

where $\hat{M}_i$ is the martingale residual for the *i*th individual, and the function of $\text{sign}(.)$ is the sign function. Observations which correspond to relatively large deviance residuals are those that do not well fitted by the model, and then considered as outliers. The deviance residual transform the symmetrical of the martingale residuals so that the distribution of the deviance residual is better approximated by normal distribution than martingale residuals when censoring is minimal, let say < 25%.

The last residual type to be discussed in this paper is score residual. According to Therneau and Grambsch [7], it can be used in conjunction with the variance-covariance matrix of parameter estimates to approximate the iterative deletion process in parameter estimation. Moreover, score residual is quite useful in diagnosis of each subject's leverage on parameter estimates and the score residuals will sum to zero. $L_{ij}\left(\hat{\beta}\right)$ is the score residual for the *i*th subject on the *j*th covariate expressed as

$$L_{ij}\left(\hat{\beta}\right) = \int_0^{\infty} \left\{X_{ij}(t) - \overline{X}_j\left(\hat{\beta}, t\right)\right\} d\hat{M}_i(t), \quad (8)$$

where the $\overline{X}$ is defined as

$$\overline{X}_j(b,t) = \frac{\sum Y_i(t)\exp\left(b' X_i(t)\right) X_{ij}(t)}{\sum X_i(t)\exp\left(b' X_i(t)\right)}, \quad (9)$$

and is a function of time: the mean over the risk set at time *t*. This equation was evaluated at $\beta = b$. Large values of score residual imply large influence of the *i*th subject on the estimate of β_j, the coefficient of X_j. The score residual is not a single value of residual for each observation, but a set of values, one for each covariate in the fitted Cox regression model. These residuals also have property of zero for expected value and uncorrelated with one another.

2 Material and Methods

In this article, we use empirical data for cardiovascular study which is obtained from Serdang Hospital, Malaysia in which those observations are patients who experienced cardiovascular disease and stayed in Cardiothoracic

High Dependency Ward (CHDW) of Cardiology Department for a period to receive necessary medical treatment. The National Cardiovascular Database (NCVD) is a service supported by the Ministry of Health (MOH) of Malaysia to collect information about cardiovascular disease, which will enable us to know the incidence of this disease, to evaluate its risk factors and treatment in the country.

We will do model building using Cox regression for the cardiovascular hospitalization data. Then, from the obtained model, we do some residual analysis by graphical methods based on scaled Schoenfeld residual, martingale residual, deviance residual, and score residual. Also, we checked the violation of proportional hazards assumption by using numerical method then compared it with graphical method.

3 Results and Discussion

We conducted analysis for the Cox regression model prior to the residual analysis. Once we obtain a final reduced model for the Cox model [2], we can do some exploration about assumptions checking for the model. The stepwise model selection could identify that out of 70 covariates included in the analysis, 14 covariates have significant contribution to the hospitalization duration of the patients during the treatments in the CHDW. Then, we can plot scaled Schoenfeld residuals versus time. Out of 14 significant covariates in the Cox model, we choose arbitrarily variable heart rate to create a plot of scaled Schoenfeld residuals versus time to show proportional hazards assumption checking.

From Figure 1, we can see that scaled Schoenfeld residuals are uncorrelated with each other and scattered around zero. But, in order to minimize subjectivity of the graphical method, we also provide formal test for checking the independence assumption. It is done by checking the correlation of the scaled Schoenfeld residual of each significant covariates from Equation (3) with the *time*(of hospitalized), log(*time*) and $time^2$. Schoenfeld [5] claimed that the partial residuals must not depend on time. Hence, if there is no any correlation or association between the covariate and time, which means failed to rejected the H_0 (The null hypothesis is there is no correlation within the covariate and time), then there are no assumption violate in this model. Therefore, we compare the result of proportional hazards assumption checking by using the graphical and numerical methods.

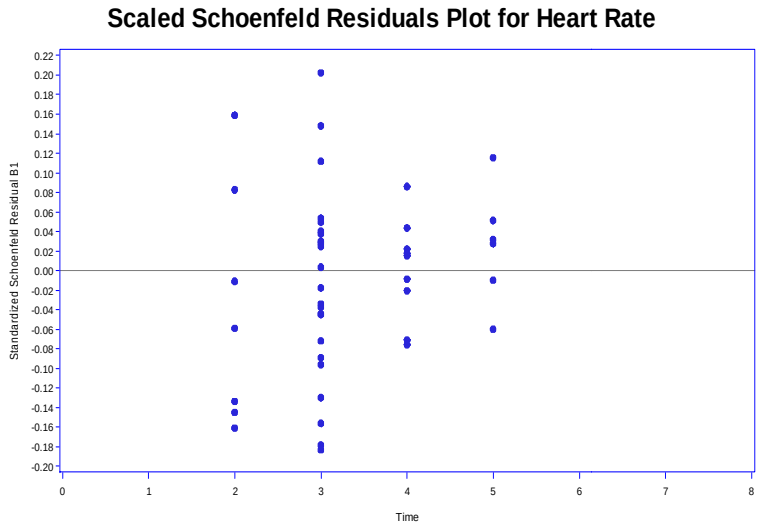

Figure 1. Scaled Schoenfeld Residuals Plot for Variable Heart Rate

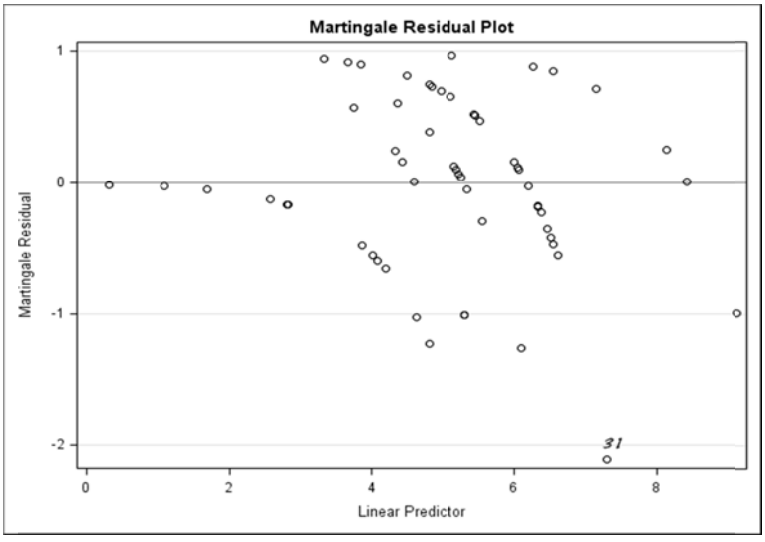

Figure 2. Martingale Residuals Plot

Table 1. Pearson Correlation for Scaled Schoenfeld Residuals of Significant Covariates with Time

Covariates	Time Variables		
	Ldays	**Days2**	**NDay**
Hypertension	-0.18963	-0.20389	-0.20153
p value	0.2122	0.1797	0.1843
Myocardial Infarction History	0.02603	0.05070	0.03874
p value	0.8652	0.7408	0.8005
Chronic Angina (Onset > 2 weeks)	0.14963	0.17617	0.16591
p value	0.3266	0.2470	0.2761
Chronic Lung Disease	0.08526	0.10126	0.09645
p value	0.5776	0.5081	0.5285
Renal Disease	-0.08188	-0.03491	-0.05768
p value	0.5928	0.8199	0.7067
Heart Rate	0.20160	0.20184	0.20387
p value	0.1842	0.1836	0.1792
Level of Good Cholesterol	0.00378	0.03503	0.02098
p value	0.9803	0.8193	0.8912
Level of Bad Cholesterol	-0.03031	-0.01393	-0.02323
p value	0.8433	0.9277	0.8796
Fasting Blood Glucose	-0.17504	-0.20490	-0.19386
p value	0.2501	0.1769	0.2020
ACE Inhibitor (During Admission)	-0.07286	-0.08638	-0.08164
p value	0.6343	0.5726	0.5939
Other Lipid Lowering Agent	0.11063	0.11237	0.11365
p value	0.4694	0.4624	0.4573
Calcium Antagonist (Pre Admission)	-0.06758	-0.09490	-0.08536
p value	0.6591	0.5352	0.5772
Insulin (After Discharge)	0.04753	0.04967	0.04914
p value	0.7566	0.7459	0.7485
Pulse Pressure	0.14346	0.14542	0.14785
p value	0.3471	0.3405	0.3324

Notes: Ldays = log(*time*), Days2 = $time^2$, and NDay = *time*

According to the Table 1, all the *p*-values are $> \alpha = 0.10$ in which we failed to reject the null hypotheses. It means there is no any correlation between each covariates of Cox model and time which implies that proportional hazards assumption is fulfilled.

The shape of martingale residuals plot in Figure 2 is skewed to reflect that there is only single event setting in the model. The martingale residuals plot shows an isolation point (with linear predictor score 7.30 and martingale residual -2.11) of patient number 31. This observation looks like an outlier in the

martingale residual plot. But, this observation is no longer distinguishable in the deviance plot. It suggests that there is no indication of a lack of fit for the model of that particular observation.

Figure 3 shows deviance residuals which consist of information about influential data and outliers. This plot is roughly symmetrically distributed around zero. When proportion of censoring is minimal (which is less than 25%), the distribution of the residual can be approximated by normal distribution. Since graphically they are approximately normally distributed, we can then move our attention to possible outliers. Under common definition, an observation is considered as an outlier if the value of the corresponding residual is outside the range of (-2.5, 2.5). Percentage of censoring of this cardiovascular hospitalization data was only 21.05%. As we can see, there are no any observations with deviance residuals less than -2.5 or greater than 2.5, so that there is no any outlier in this Cox model. It also means that there is no any observations have unusual large (large positive deviance residual) or small (large negative deviance residual) of hospitalization duration regarding their cardiovascular disease than expected hospitalization days.

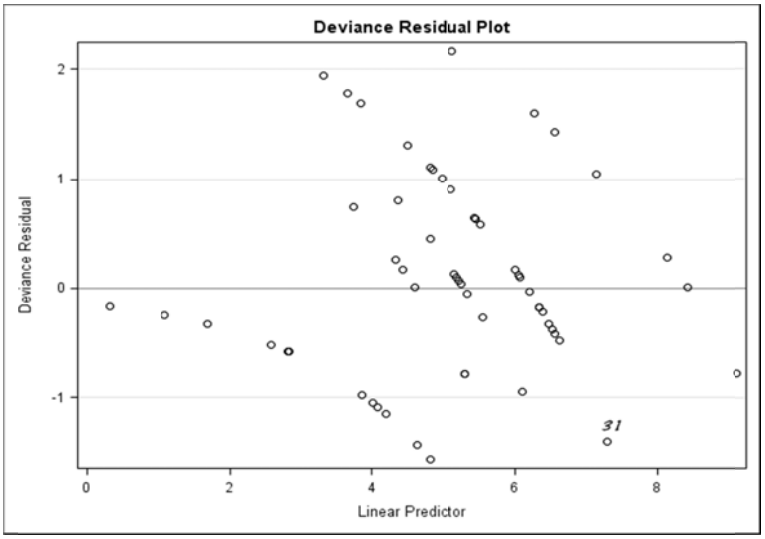

Figure 3. Deviance Residuals Plot

Meanwhile, a plot of score residuals against covariate is displayed in Figure 4. We need to have randomly scattered points, centered around zero if the fitted model is adequate. We use this plot in order to identify possible influential observations in the data since it has effects on model-based inferences. In general, influential observations are points that are isolated from all other points in the graph. There are no anomalies shown in Figure 4 of score residuals for patients' heart rate. However, the observations with the smallest score residual (-29.61) for heart rate is from patient 51 with a low heart rate of 53 was shown as an isolation point. This point could be determined as an observation that will affect the

estimation of coefficient for the parameter of heart rate, known as β_6 of the Cox regression in this study.

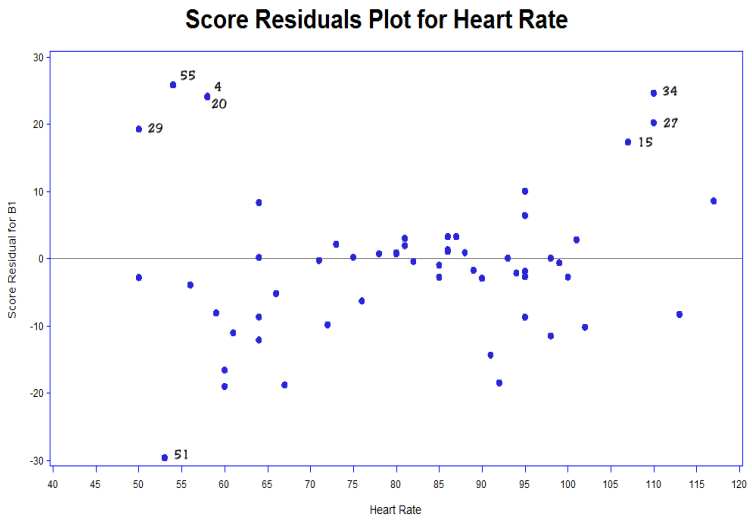

Figure 4. Score Residuals Plot for Covariate Heart Rate

Acknowledgements

The authors are grateful to University Grants Scheme by Universiti Putra Malaysia for awarding a research grant for supporting the research and to Department of Cardiology of the Serdang Hospital, Malaysia for providing necessary information including the data.

References

[1] W. E. Barlow, and R. L. Prentice, Residuals for relative risk regression, Biometrika, 75(1988), 65 – 74.

[2] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society. Series B (Methodological), 34(1972), 187 – 220.

[3] P. M. Grambsch and T. M. Therneau, Proportional hazards tests and diagnostics based on weighted residuals, Biometrika, 81(1994), 515 – 526.

[4] D. Kumar and B. Klefsjö, Proportional hazards model: A review, Reliability Engineering and System Safety, 44(1994), 177 – 188.

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

[6] T. M. Therneau, P. M. Grambsch, and T. R. Fleming, Martingale-based residuals for survival models, Biometrika. 77(1990), 147 – 160.

[7] T. M. Therneau and P. M.Grambsch, Modelling Survival Data: Extending the Cox Model, Springer, New York, 2000.

Received: August 6, 2013