Journal of the Faculty of Environmental Science and Technology,
Okayama University
Vol.9, No.1, pp.37-44, February 2004

A Numerical Study on Statistical Diagnostics in Cox Proportional Hazards Models for Survival Data Analysis

Jimin SUNG* Yutaka TANAKA†

(Received December 2, 2003)

Summary

There have been proposed so far many methods of statistical diagnostics in Cox regression for checking the goodness of the estimated model or checking the adequacy of the data. The former type contains the checking of the overall goodness of fit, the validity of the assumption of proportional hazards and the proper functional forms of the effects of covariates. While the latter type contains the checking whether there exist singly and/or jointly influential observations in the data set. In the present paper we study numerically the performances of various methods of diagnostics including our method of influence analysis for multiple-case diagnostics (Sung and Tanaka, 2003) by analyzing a real data set of lung cancer patients.

Key words: Cox regression, Influence function, Local influence, Influential Subsets, Cox-Snell residuals, Martingale residual, Deviance residual

1 Introduction

Cox proportional hazards model or shortly Cox regression plays an important role in survival analysis, in particular, in the comparison of treatment effects after adjusting the effects of other relevant covariates when the dependent variable is obtained as survival time. Statistical diagnostics is, needless to say, very important in model building. In Cox regression it consists of the assessment of the overall goodness of fit or the precision of the prediction, the validity checking of the assumption of proportional hazards, the search for proper functional forms of covariates and influence analysis. Major techniques are given in recent books such as Therneau and Grambsch (2000), Everitt and Rabe-Hesketh (2001), Lee and Wang (2003), Klein and Moeschberger (2003), and Tableman and Kim (2003). In the present paper we apply various techniques of statistical diagnostics to a set of real data and study their performances numerically. Emphasis is placed on graphical techniques including our method of influence (Sung and Tanaka, 2003).

2 Cox Proportional Hazards Model

Cox (1972) proposed a model which is called proportional hazards model. It is described as below. Let $(t_i, \delta_i, \mathbf{Z_i})$ be an observation vector of individual i for $i = 1, \cdots, n$, where t_i indicates the death or censored time, δ_i the dummy variable to denote death ($\delta_i = 1$) or censored ($\delta_i = 0$), $\mathbf{Z}_i = [Z_{i1}, ..., Z_{ip}]^T$ the covariate values of individual i. Then the hazard function is expressed by

$$h_i(t) = h_0(t) \exp(\beta^T \mathbf{Z}_i),$$

where $h_0(t)$ indicates so-called baseline hazard and it is assumed that the baseline hazard function is the same for all individuals in the study. It is

*Graduate School of Natural Science and Technology, Okayama University, Tsushima, Okayama 700-8530, Japan. sungjm@ems.okayama-u.ac.jp

†Department of Environmental and Mathematical Science, Okayama University, Tsushima, Okayama 700-8530, Japan. tanaka@ems.okayama-u.ac.jp

called proportional hazards because, if we look at two individuals with covariate values $\mathbf{Z}$ and $\mathbf{Z}^*$, the ratio of their hazards rates is constant without depending on time t.

3 Parameter Estimation

Cox (1972) proposed a method of maximum partial likelihood, where the partial likelihood is formed by multiplying conditional probabilities P(individual i dies at t_i | one death at t_i). Note that the partial likelihood does not contain the baseline hazard function. When there are ties between death times, they are incorporated into partial likelihood using Breslow's (1974) method. Then, when we introduce case-weight w_i for influence analysis, the partial log-likelihood is expressed as

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{D}\sum_{k=1}^{p}\beta_k \sum_{l\in\Delta_i} w_l Z_{lk} \\ &- \sum_{i=1}^{D}\sum_{l\in\Delta_i} w_l \log\Big[\sum_{j\in R_i} w_j \exp(\sum_{k=1}^{p}\beta_k Z_{jk})\Big],\end{aligned}$$

where $t_{(i)}$'s ($t_{(1)} < t_{(2)} < \cdots < t_{(D)}$) indicate the distinct death times, d_i the number of individuals who died at time $t_{(i)}$, Δ_i the subset of individuals who died at $t_{(i)}$, and R_i the risk set at $t_{(i)}$, and w_j the case-weight for individual j. The regression coefficients can be obtained by solving the likelihood equation

$$U_h(\beta) = \partial\ell(\beta)/\partial\beta_h = 0, \quad h = 1, 2, \cdots, p \ ,$$

formulated by differentiating partial log-likelihood instead of full log-likelihood function, and the precision of the estimates can be evaluated with the observed information matrix $\mathbf{I}(\hat{\beta})$ defined as

$$\begin{aligned}\mathbf{I}(\hat{\beta}) &= [I_{gh}(\hat{\beta})]_{p\times p} \ , \\ I_{gh}(\beta) &= \frac{\partial \mathbf{U}_h(\beta)}{\partial\beta_g} = \frac{\partial^2 l(\beta)}{\partial\beta_g\partial\beta_h}.\end{aligned}$$

4 Statistical Diagnostics

4.1 Model Checking

4.1.1 Goodness of Fit

There are several kinds of residuals in Cox proportional hazards models. Among them Cox-Snell residuals and deviance residuals can be used for assessing the overall fit and detecting outliers, respectively. The Cox-Snell residuals are defined by

$$r_{ci} = \hat{H}_0(t_i)\exp\big(\sum_{k=1}^{p} Z_{ik}\hat{\beta}_k\big),$$

where $\hat{H}_0(t)$ is Breslow's estimator of the baseline hazard rate giving by

$$\hat{H}_0(t) = \sum_{t_i\le t} \frac{d_i}{\sum_{j\in R(t_i)} \exp\big(\sum_{k=1}^{p} Z_{jk}\hat{\beta}_k\big)}$$

If the final proportional hazards model is correct and the estimated regression coefficients are close to the true values, the Cox-Snell residuals r_{ci}'s can be regarded as a sample from a unit exponential distribution, and therefore, the plot of $\hat{H}(r_{ci})$ against r_{ci} should be a 45°-line through the origin (Klein and Moeschberger, 2003). The deviance residuals are defined by

$$D_i = sign(\hat{M}_i)\sqrt{-2(\hat{M}_i + \delta_i)\log(\delta_i - \hat{M}_i)}$$

where $\hat{M}_i (= \delta_i - r_{ci})$ is the i-th martingale residual, δ_i being the dummy variable to indicate censored or uncensored. It is known that the deviance residuals are symmetrically distributed about zero when the fitted model is adequate, and individuals with large positive or negative deviance residuals are poorly predicted by the model.

4.1.2 Assumption of Proportional Hazards

Suppose we are interested in checking the assumption of proportionality of the hazard rates of individuals with different values of a covariate Z_1 after adjusting for all other relevant covariates Z^*. We assume that there is no term of interaction between Z_1 and any of the remaining covariates. Then, we stratify the covariate Z_1 into K disjoint strata $G_1, \cdots, G_K$, and fit a Cox model with covariate Z^* to each stratified data set, and estimate cumulative baseline hazard rates $\hat{H}_g(t)$, $g = 1, \cdots, G$. If the assumption of proportional hazards holds, the plots of $\log\hat{H}_g(t)$ versus t should be approximately parallel. Therefore, the assumption can be checked by the plots of $\log\hat{H}_g(t)$ for $g = 1, 2, \cdots, G$.

4.1.3 Functional Form of a Covariate

Suppose that the covariate vector Z is partitioned into a single covariate Z_1, for which we do not know what functional form to use, and a covariate vector Z^*, for which we know the proper function forms. We assume that we need not consider any interaction term between Z_1 and Z^*. Then our optimal Cox model can be written as

$$H(t|Z_1, Z^*) = H_0(t)\exp(\beta^{*T} Z^*)\exp(f(Z_1)).$$

The functional form $f(Z_1)$ can be formed in the following steps :

1. Fit a Cox model to the data with covariates Z^*, and compute the martingale residuals, $\hat{M}_j$, $j = 1, \cdots, n$.
2. Plot $\hat{M}_j$ versus Z_{j1} for $j = 1, \cdots, n$.
3. Apply an appropriate smoothing technique such as Lowess (Cleveland,1979) to the plot. Then the smoothed-fitted curve gives an indication of the function f.

4.2 Influence Analysis

Methods of influence analysis have been proposed in Cox proportional hazards and related models by Cain and Lange(1984), Wei and Su(1999), Wei and Korosok(2000). The former two derived influence functions for regression coefficients in the proportional hazards models and the models with somewhat generalized hazards functions, respectively, and the last one proposed a method of multiple-case diagnostics based on pairwise deletion and pairwise differentiation. Tanaka and his coworkers (see, Tanaka, 1994; Tanaka and Zhang, 1999) proposed a general procedure of influence analysis including multiple-case diagnostics as well as single-case diagnostics in general statistical modeling. Their method is to reduce the dimension by applying PCA with metric V^{-1}, when V indicates an asymptotic covariance matrix of the estimated parameters, to the influence functions and search for individuals which are located far from the origin and on similar directions from the origin for the purpose of multiple-case diagnostics. It is known that their multiple-case diagnostics is closely related to that of Cook's local influence. In Sung and Tanaka (2003), we proposed a method of multiple-case diagnostics in Cox regression models with censored observations based on the above general procedure. The general procedure is described as follows.

1. Compute the influence function vector $\dfrac{\partial\hat{\beta}}{\partial w_i}$, for $i = 1, 2, \cdots, n$
2. (Single-case diagnostics) Summarize the influence function vector into scalar influence measures, from various aspects such as the influence on the estimate, on its precision, and on the goodness-of-fit. Find individuals with large values of the measures.
3. (Multiple-case diagnostics) Search for subsets of individuals whose members are individually influential and have similar influence patterns by using PCA with metric $[\widehat{acov}(\hat{\beta})]^{-1}$.
4. Confirm the influence of single or multiple individuals by omitting them.

4.2.1 Influence Function

The influence of each individual on the estimate $\hat{\beta}$ can be evaluated with the partial differential coefficient of $\hat{\beta}$ with respect to w_j, i.e., $\partial\hat{\beta}/\partial w_j, j = 1, \cdots, n$, and it provides an approximation to $\hat{\beta} - \hat{\beta}_{(j)}$, where $\beta_{(j)}$ indicates the estimates for β based on the sample without individual j. We shall call this partial differential coefficient the influence function. It is also called dfbeta as in the case of ordinary regression (see, Tableman and Kim, 2004, p.165). Application of the differentiation of implicit function yields

$$\partial\hat{\beta}/\partial w_j = \Big[-\frac{\partial U}{\partial\beta}\Big]^{-1}_{\underline{w}_0}\frac{\partial U}{\partial w_j},$$

where the term in brackets is the observed information matrix. As shown in Cain and Lange (1984) the differential coefficient of the score function $\partial U/\partial w_j$ is given by

$$\begin{aligned}\frac{\partial U_h}{\partial w_j}\Big|_{\underline{w}_0} &= \delta_j\Big[Z_{jh} - \sum_{i=1}^{D}\hat{E}(Z_h|\dot{R_i})\Big] \\ &- \sum_{i:j\in R_i}^{D} d_i\frac{\exp(\sum_{m=1}^{p}\hat{\beta}_m Z_{jm})}{\sum_{i\in R_k}\exp(\sum_{m=1}^{p}\hat{\beta}_m Z_{km})} \\ &\times [Z_{jh} - \hat{E}(Z_h|R_i, j\in R_i)],\end{aligned}$$

where

$$\hat{E}(Z_h|R_i) = \frac{\sum_{i\in R_j} Z_{jh}\exp(\sum_{m=1}^{p}\hat{\beta}_m Z_{jm})}{\sum_{i\in R_j}\exp(\sum_{m=1}^{p}\hat{\beta}_m Z_{jm})}.$$

The equation of $\partial U/\partial w_j$ shows that the change in the score vector U due to changes in w_j consists of the sum of two components. The first component is included only if individual j died and is the difference between the covariates for case j and the weighted average of covariates for all individuals in the risk set R_i. This term is called the partial residual of Schoenfeld (1982). The second measures the combined effect that changes in w_j have upon all the risk sets that include individual j (Cain and Lange, 1984).

4.2.2 Single-Case Diagnostics

For single-case diagnostics, we compute Cook's D for each individual to study the influence on the regression coefficients. We can regard the individuals with large values of D as singly influential observations. The Cook's D is defined by

$$D_i = (\partial\hat{\beta}/\partial w_i)^T V^{-1}(\partial\hat{\beta}/\partial w_i),$$

where V is an estimate for the asymptotic variance-covariance matrix of $\hat{\beta}$. Here we define $V = [\mathbf{I}(\hat{\beta})]^{-1}$ using the observed information matrix .

4.2.3 Multiple-Case Diagnostics

The basic idea of the multiple-case diagnostics in the general procedure is to reduce the dimension by applying PCA with metric $\mathbf{I}(\hat{\beta})$ to the influence function $\partial\hat{\beta}/\partial w_j, j = 1,\cdots,n$, search for subsets of individuals which are located far from the origin and on similar directions in the space of dominant principal components. If such subsets are found, we regard them as candidates for influential subsets of individuals. More precisely, we compute the eigenvalues λ_j and the associated eigenvectors $\underline{a}_j$ of the eigenproblem

$$\Big[\frac{1}{n}\sum_{k=1}^{n}(\frac{\partial\hat{\beta}}{\partial w_k})(\frac{\partial\hat{\beta}}{\partial w_k})^T - \lambda V\Big]\underline{a} = \underline{0},$$

draw a scatter matrix of dominant PCs, i.e., $u_{jk} = \underline{a}_j^T\left(\partial\hat{\beta}/\partial w_k\right)$, $k = 1,\cdots,q$, and search for subsets of individual described above.

5 Numerical Example

5.1 Data Set

For illustration, we analyze a data set for 137 patients with lung cancer on a randomized clinical tiral conducted by the Veteran's Administration, which is taken from Kalbfleisch and Prentice (1980, pp. 223-224). In this data set 128 observations are uncensored and 9 are censored. The data set consists of the information on survival time, an indicator for censoring, prior therapy(prior), Karnofsky performance status(kps), age, month since diagnosis(diag), cell-type and treatment of test or standard chemotherapy(therapy). The cell type is composed of four categories such as squamous, small, adeno and large. These four types of cell are expressed by using dummy variables. The major purpose of the trial is to test the treatment effect after adjusting the effects of covariates. The Kaplan-Meier survival curves for the standard and test groups are shown in Figure 1, and the log-rank test statistics for the difference of these two curves, which is computed by neglecting all covariates, 0.0082(p-value : 0.9277) shows that it is not statistically significant. It is noted that, though it is not statistically significant, the two survival functions cross with each other and the test group seems to live longer than the standard group.

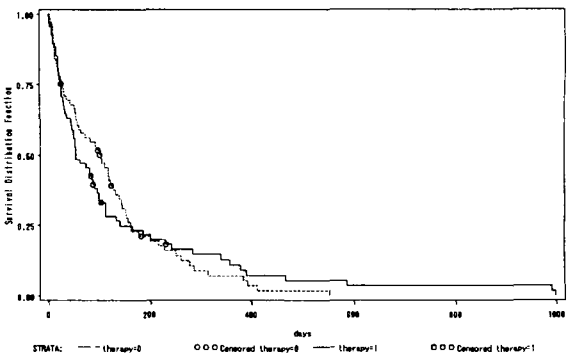

Figure 1 Estimated survival function for type of therapy(0=standard, 1=test)

5.2 Model Fitting

A Cox regression model with all main effects of the covariates is fitted to the data set. The result

is shown in Table 1. The therapy is not statistically significant, while three covariates kps, small

Table 1 Maximum Likelihood Estimation for a Cox regression model with all covariates

Variable	Coef	z	P
kps	-0.03282	-5.95801	0.000
diag	0.00008	0.00895	0.99
age	-0.00871	-0.93612	0.35
prior	0.07159	0.30817	0.76
squam	-0.40129	-1.41955	0.16
small	0.46027	1.72890	0.084
adeno	0.79478	2.62408	0.0087
therapy	0.29461	1.41945	0.16

Table 2 The result of backward procedure of variable selection with AIC values

Step	Variable	AIC
Step1	therapy,prior,kps,diag age,squam,small,adeno	966.359
Step2	therapy,prior,kps,age squam,small,adeno	966.359
Step3	therapy,kps,age squam,small,adeno	966.476
Step4	therapy,kps,squam small,adeno	964.359
Step5	therapy,kps,small,adeno	961.275
Step6	therapy,kps,adeno	967.130
Step7	therapy,kps	973.244
Step8	therapy	1013.760

Table 3 Estimated best Cox regression model

Variable	Coef	z	P
therapy	0.2099	1.06	0.29
kps	-0.0306	-6.00	0.000
small	0.6378	2.83	0.005
adeno	0.9798	3.76	0.0002

and adeno are significant at 0.10 level. As our major purpose is to assess the effect of therapy, we search for the best model among the models which contain therapy as an independent variable, and under such a condition we apply backward elimination procedure of variable selection to the Cox regression model shown in Table 1. Table 2 shows the AIC value of each step of the procedure, and as the result, the model of step 5 is selected as the best model. Precise information of the fitted best model is shown in Table 3. It is noted

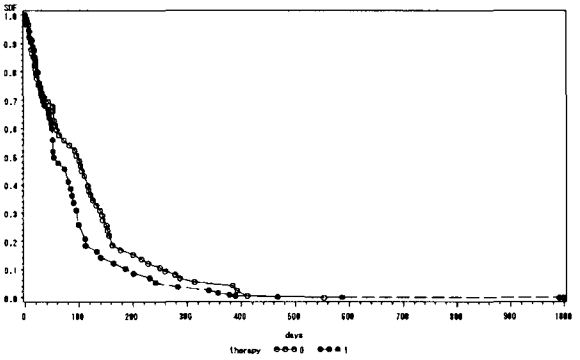

Figure 2 Estimated survival functions for two types of therapy(0=standard, 1=test) in final model

that the effect of therapy is not significant also in this model. Figure 2 shows the estimated survival functions for the standard therapy (coded " 0 " .) and the test therapy (coded " 1 ").

5.3 Diagnostics (1) : Model Checking

5.3.1 Goodness of Fit

As explained in Section 4 Cox-Snell residuals can be used for assessing the goodness of fit of the model. Let r_{ci} and $\hat{H}(r_{ci})$ denote the Cox-Snell residual for the i-th individual and the estimated

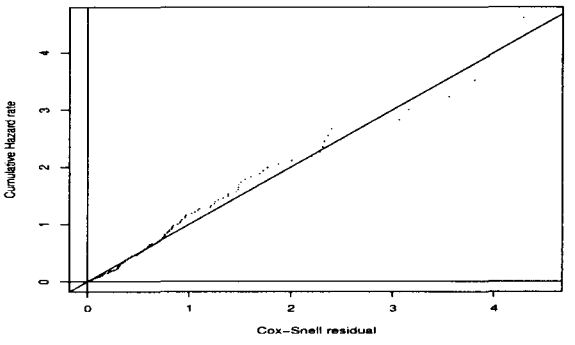

Figure 3 Cox-Snell residuals plot for the fitted final model

cumulative hazard for r_{ci}, respectively. Then, as given in Section 4 the plot of $\hat{H}(r_{ci})$ versus r_{ci} should be a 45°-line through the origin, if the model is correct and the estimates $\hat{\beta}$'s are close

to the true values. Figure 3 gives the plot for the final model. Overall the residuals fall on a straight line with an intercept zero and a slope one. We see from Figure 3 that the final model provides a reasonable fit to the data. It is also

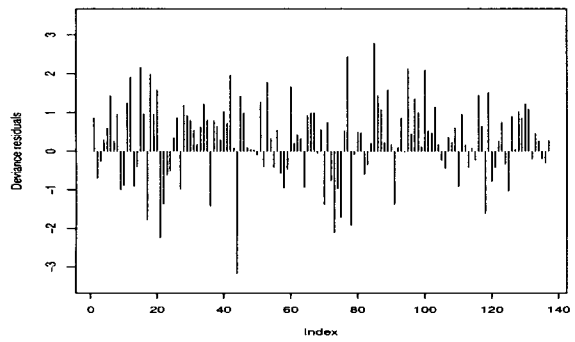

Figure 4 Deviance residual for the fitted final model

known that deviance residuals are symmetrically distributed about zero when the fitted model is adequate. Figure 4 shows the deviance residuals of all individuals. It is noted that the distribution is approximately symmetric and there exists no clearly outlying observation.

5.3.2 Assumption of Proportional Hazards

Let us apply a graphical method to check whether the assumption of proportional hazards is valid or not. When this assumption holds, the plot of $\log \hat{H}_g(t)$ for $g = 1, \cdots, G$ are parallel, where

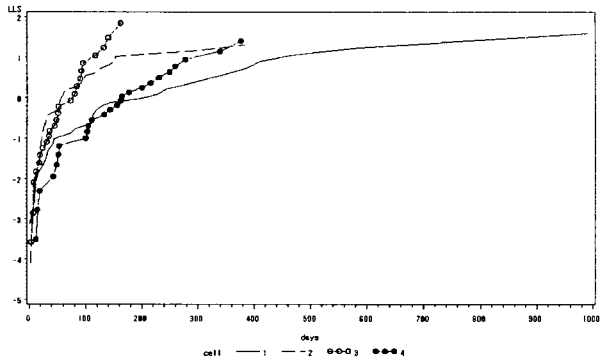

Figure 5 Plot of $\log \hat{H}_g(t)$ versus survival time (1: squam, 2 : small, 3: adeno, 4 : tall)

$\hat{H}_g(t)$ is the estimated cumulative baseline hazard rate in the g-th stratum when we stratify an important covariate (or a set of covariates) into G strata and fit a Cox regression model separately to each stratum. Figure 5 shows the plot of $\log \hat{H}_g(t)$ against survival time for four cell types("squam", "small", "adeno" and "tall"). It is not clear whether these four curves are parallel or not.

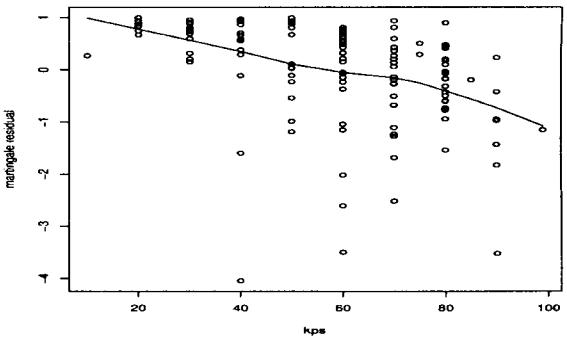

Figure 6 Martingale residual plot for the covariate kps

5.3.3 Functional Form of a Covariate

To study functional form of a covariate martingale residuals can be used just like partial regression plot or partial residual plot in ordinary regression analysis. In our example only "kps" is a quantitative variable among the three significant covariates. So we try to search for an appropriate functional form of the effect of kps. Martingale residuals for covariate kps are plotted against the kps values along with a Lowess smooth curve. Figure 6 shows the results. It seems that the effect of kps can be approximated well with a linear function.

5.4 Influence Analysis

We applied our influence analysis to the obtained model to investigate if there exist singly or jointly influential observations. Figure 7 shows the index plot of Cook's D and Figure 8 gives the scatter plot of the first and second principal components obtained by PCA with metrix V^{-1}, where the eigenvalues are 1.976, 1.001, 0.629, 0.561, in order of their agnitudes. Looking these figures we can find that there may exist two singly influential observa-

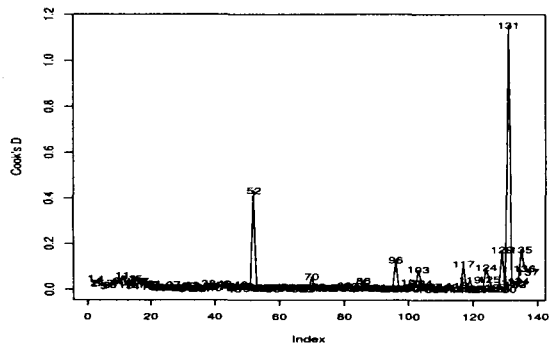

Figure 7 Single-case diagnostics : Index plot of Cook'D

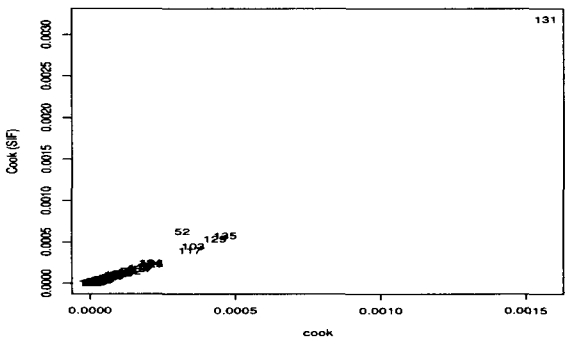

Figure 9 The vailidty of Influence Analysis

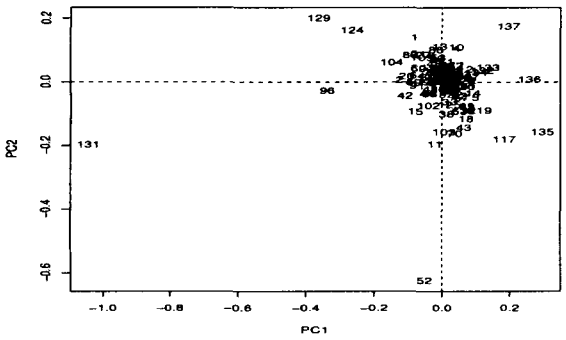

Figure 8 Multiple-case diagnostics : Scatter plot of the first and second principal components of influence functions

tions, i.e., individuals #131 and #52. The scatter plot is drawn for Cook's D's which are computed in two different manners. One is based on $\partial\hat{\beta}/\partial w_i$ as in Section 4.2.1, and the other based on SIF_i instead of $\partial\hat{\beta}/\partial w_i$, where

$$SIF = \hat{\beta} - \hat{\beta}_{(i)} \quad , i = 1, 2, \cdots, n \quad ,$$

$\hat{\beta}_{(i)}$ indicating the $\hat{\beta}$ without i-th individual. The resulting scatter plot is given in Figure 9.

6 Concluding Remarks

In the present paper we applied various methods of statistical diagnostics to a data set of 137 patients of lung cancer and studied their performances. For functional form of a covariate the martingale residual plot with a Lowess smooth clearly suggested that the proper functional form of a quantitative covariate is a linear function. Concerning the assumption of proportional hazards the stratified plot of $\log \hat{H}_g(t)$ could not suggest whether the curves are parallel or not. The result of our influence analysis suggest that there are two singly influential observations but no jointly influential observations. It is also suggested that the use of the partial derivative $\partial\hat{\beta}/\partial w_i$ gives a good approximate for $\hat{\beta} - \hat{\beta}_{(i)}$, the difference of the estimates of the sample with and without the i-th individuals. Looking all the above results we may say that the methods we studied are useful in statistical diagnostics for modeling the Cox regression.

References

[1] Breslow, N. E. (1974) Covariance analysis of censored survival data. *Biometrics*, 30, 89-99.

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

[3] Cleveland, W. S. (1979) Robust Locally Weighted Regression and Smoothing Scatter Plots. *Journal of the American Statistical Association*, 74, 829-836.

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

[5] Everitt, B. and Rabe-HeskethS. (2001) *Analyzing Medical Data Using S-Plus (Statistics for Biology and Health)* . Springer.

[6] Lee, E. T. and Wang, J. W. (2003) *Statistical Methods for Survival Data Analysis.* Wiley.

[7] Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data.* Wiley.

[8] Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis.* Springer.

[9] 成 祉旻, 田中 豊. (2003) Cox 比例ハザードモデルに対する複数観測値診断の一方法. 2003 年度統計関連学会, 19-20

[10] Tableman, M. and Kim, J. S. (2004) *Survival Analysis using S : Analysis of Time-to-Event Data.* Chapman.

[11] Tanaka, Y. (1994) Recent Advance in Sensitivity Analysis in Multivariate Methods. *J. Jpn. Soc. Comp. Statist.*, 7, 1-25.

[12] Tanaka, Y. and Zhang, F. (1999) R-mode and Q-mode Influence Analysis and Local Influence Approach. *Comp. Statist. Data Anal.*, 31, 2325-2347.

[13] Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data-Extending the Cox Model.* Springer.

[14] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990) Martingale-Based Residuals for Survival Models. *Biometrika*, 77, 147-160.

[15] Wei, H. W. and Kosorok, R. M. (2000) Masking unmasked in the proportional hazards model. *Biometrics*, 56, 991-995,.

[16] Wei, H. W. and Su, J. S. (1999) Model choice and influential cases for survival studies. *Biometrics*, 55, 1295-1299,.