Sakarya Tıp Dergisi
Sakarya Med J

E-ISSN: 4146-409X
Yayıncı / Publisher: Sakarya Üniversitesi

Cilt/Vol. 14, Sayı/No. 1, 20-27, 2024
DOI: http://doi.org/10.31832/smj.1390306

Research Article/ Araştırma Makalesi

Use of Residuals and Rank Product in Detection of Outlier in Survival Analysis with Crimean-Congo Hemorrhagic Fever Data

Osman Demir[1], Ünal Erkorkmaz[2]*

[1] Tokat Gaziosmanpasa University Faculty of Medicine, Department of Biostatistics, Tokat, Türkiye
mosmandemir@hotmail.com

[2] Sakarya University Faculty of Medicine, Department of Biostatistics, Sakarya, Türkiye
uerkorkmaz@sakarya.edu.tr

*Corresponding Author

Received Date:
01.12.2023
Accepted Date:
19.02.2024
Available Online Date:
15.03.2024

Purpose: Survival analysis is a statistical method used in many fields, especially in the field of health. It involves modeling the relationship between the survival time of individuals after a treatment or procedure and the event called response. The presence of outliers in the data may cause biased parameter estimations of the established models. Also, this situation causes the proportional hazards assumption to be violated especially in Cox regression analysis. Outlier(s) are identified with the help of residuals, Bootstrap Hypothesis test and Rank product test.

Method: In R.4.0.3 software, outlier(s) are determined on a clinical dataset by the Schoenfeld residual, Martingale residual, Deviance residual method and Bootstrap Hypothesis test (BHT) based on Concordance index, and Rank product test.

Results: After the cox regression established by the backward stepwise and robust cox regression, it was observed that the established models did not fit. So, the outlier(s) determined by the methods mentioned.

Conclusion: It was decided that only one observation could be excluded from the study. As in the survival data, in many data types, outliers can be detected and further analyzes can be applied by using the methods mentioned.

Keywords: Outliers, Residuals, Concordance index, Rank product

1.INTRODUCTION

Survival analysis is a process that involves modeling relationships with time in the occurrence of the event. Clinically, it involves examining the relationships between an individual's survival time after a particular treatment or procedure and the event occurring in response. At the same time, the effect of independent variables on survival time can be modeled. The occurrence of the event usually occurs as death. If the event has not occurred, those individuals are included in the study as censored. Cox regression analysis is frequently used to examine the effect of independent variables on survival time.[1] For this analysis, which is a semi-parametric model, the variables in the model must satisfy the proportional hazards assumption. This assumption means that the hazard ratio is constant over time. The presence of outliers in the variables in the data indicates that the Cox regression coefficients deviate from the true value. Therefore, it causes wrong findings in parameter estimation.[2] There are studies in the literature on outlier detection in wide areas such as normal data, multivariate normal data, censored data, negative data, time series data, gene expression data.[3] As in these studies, the evaluation of the adequacy of the established models has an important place in the diagnostic procedures. A large part of this process includes the evaluation of residuals. There are many residual methods in the literature. In this study, Schoenfeld residual, Martingale residual, Deviation residual and Concordance index based Bootstrap Hypothesis methods are used.

Cite as: Demir O, Erkorkmaz U. Use of Residuals and Rank Product in Detection of Outlier in Survival Analysis with Crimean-Congo Hemorrhagic Fever Data. Sakarya Med J 2024; 14(1): 20-27 DOI: 10.31832/smj.1390306

In the study, effective observations are obtained with the rank product test by using the results of the residuals and concordance c-index-based Bootsrap Hypothesis test in outlier detection. Outlier(s) are be identified with the help of residuals and Rank product test.

2.MATERIALS and METHODS

Study Selection: With the help of the data obtained by Aktaş et al.[4], who worked on 209 patients diagnosed with CCHF (Crimean-Congo Hemorrhagic Fever) between May 2010 and September 2015 in Tokat State Hospital. The study data was approved by the Tokat Gaziosmanpaşa University Clinical Research Ethics Committee. The data set consists of clinical information on 48 covariates of 209 patients.[4] Analyses were performed using the packages "survival"[5], "coxrobust"[6], "BCSOD"[7], "qvalue"[8] in R 4.0.3[9] software.

2.1.Cox regression model

Cox regression model, which is a frequently used method in survival analysis, examines the relationship between survival time as the dependent variable and one or more independent variables on which the effect is investigate.[1,10]

$$h(t,X) = h_0(t)\exp(\beta' X), \qquad (1)$$

In the equation, $\beta=(\beta_1,\dots,\beta_p)$ are the unknown regression coefficients, h_0 (t) is baseline hazard and $X=(X_1,\dots,X_p)$ is the covariate vector.[11] Although Cox regression analysis is a frequently used model, the Cox robust regression model is recommended because the presence of outliers causes large changes in parameter estimations.

2.2.Cox robust regression model

This model is obtained by weighting the partial likelihood function in the Cox regression model.[12,13]

Let m(t,X) be a weight function, where m_ij=m(t_i,X_i) ve m_i=m_ii=m(t_i,X_i) are the weight 1≤i≤j≤n. Robust state of partial likelihood function for parameter estimation is

$$\sum_{i=1}^{n} m_i \delta_i \left[X_i - \frac{\sum_{j\geq i} m_{ij} \exp(X_j^T \beta) X_j}{\sum_{j \geq i} m_{ij} \exp(X_j^T \beta)} \right] = 0 \qquad (2)$$

This model reduces the contribution of outliers to the model in parameter estimation.[14]

2.3.Outlier detection methods in survival analysis

2.3.1.Residuals

In the detection of outliers that have a significant effect on parameter estimation, it is important to use residuals to reveal whether the established model meets the assumptions.

2.3.2.Schoenfeld

Schoenfeld residuals, also known as a,score residual, are used to test the proportional hazards assumption in the Cox regression model. This type of residual has a set of values for each independent variable in the model, rather than one value for each observation.[15] To test he assumption that Schoenfeld residuals do not depend on time, Schoenfeld stated that the ith residual can be plotted against t_i to test the assumption that the residuals are not time dependent. Schoenfeld residual is

$$\hat{r}_{(i)} = X_i - \frac{\sum_{j\in R_i} X_j \mathrm{e}^{(\hat{\beta}^T X_j)}}{\sum_{j \in R_i} \mathrm{e}^{(\hat{\beta}^T X_j)}} \qquad (3)$$

Where t_i is ith survival time and X_i is covariate vector and R_i is risk set.

Kumar and Klesjö found that the partial residuals estimated against time should be randomly distributed around 0. Therefore these residuals are summed to zero.[16]

2.3.3.Martingale

Barlow and Prentice[17] proposed the type of residual named Martingale-based residual or Martingale residual. Martingale residual for ith individual is

$$\widehat{M}_i = \delta_i - \widehat{H}_0(t_i)\exp(\hat{\beta}^T X_i) \qquad (4)$$

Where δ_i is event. Martingale residuals take values between -∞ and 1. It shows an asymmetrical distribution. A value close to 1 indicates a shorter than expected survival time, a large negative value indicates a long survival time.[18,19]

2.3.4.Deviance

Deviance residuals proposed by Therneau, Grambch and Fleming[20] were converted from Martingale residuals. These residuals are given as

$$d_i = sign(\widehat{M}_i)\sqrt{2}\left[-\widehat{M}_i - \delta_i \log(\delta_i - \widehat{M}_i)\right]^{1/2} \qquad (5)$$

They are distributed symmetrically around zero.

2.3.5.The Concordance c-index

This method proposed by Harrell et al.[21] to demonstrate the performance of survival analyses. It measures the probability of a higher prediction in the individual in whom the event occurred for the first time. This statistics, which is sensitive to outliers, measures how well the predicted values fit with the rank-ordered response variables[14]. The error rate is calculated as 1-c and c represents the Harrell concordance index. Error rates range from 0 to 1, with a value of 0 indicating the best accuracy. There are 3 alternative methods for outlier detection in survival analysis using the c index: (1) One-Step Deletion, (2) Bootstrap Hypothesis test and (3) Dual Bootstrap Hypothesis test.

Bootstrap Hypothesis test which will be used in this study, tests concordance variation over bootstrap samples without ith individual.

Hypotheses for ith. observation are given as

$$H_0: \delta C_i \leq 0,\ H_1: \delta C_i > 0 \qquad (6)$$

Where 〖δC〗_i=C_(i-) -C_all, C_(i-) is the c-index of model establised without i. individual ve C_all is the c-index of model with all variables.

The smallness of the p values obtained from the hypotheses indicates the observation is outlying.

2.3.6.Rank Product Test

Rank product test is a method used to derive an overall conclusion from the findings obtained from the methods used to identify outliers. This method, which is a non-parametric statistical method, was first used in meta-analysis and microarray studies.[22,23] In this method, the aim is to provide a unified definition with the ranking obtained from the methods used.[24]

Let n, m be the number of individuals and the outlier detection method, respectively. Let *Pij*,

be the outlyingness of ith individual for jth method, with 1≤i≤n and 1≤j≤m.

The deviance rank is given as

$$R_{ij} = rank(P_{ij}), \quad 1 \leq R_{ij} \leq n. \qquad (7)$$

For each method, the lowest ranks obtained indicate more outliers than the others. After obtaining ranks for each method, the rank product is defined as

$$RP_i = \prod_{j=1}^{m} R_{ij}. \qquad (8)$$

To determine the statistical significance of 〖RP〗_i, the permutation approaach[23], logarithm approach[25], and exact probability[26] are used. The algorithm in this study produces accurate approximate p values based on the geometric mean of the upper and lower bounds, defined recursively. Since more than one test is performed here, the problem of increasing type I error in multiple tests is encountered. For this problem, false discovery rate (FDR), which is less conservative than the Bonferroni correction, is preferred.[27] The FDR, the expected rate of false positives among all significant tests, ranks the p-values in ascending order and divides them by percentiles. FDR is determined by the q-value.

3.RESULTS

It was aimed to create an application area in determining residual value with the help of the data obtained by Aktaş et al.[4], who worked on 209 patients diagnosed with CCHF (Crimean-Congo Hemorrhagic Fever) between May 2010 and September 2015 in Tokat State Hospital. The study data was approved by the Tokat Gaziosmanpaşa University Clinical Research Ethics Committee. The data set consists of clinical information on 48 covariates of 209 patients.[4]

Analyses were performed using the packages "survival", "coxrobust", "BCSOD", "qvalue" in R 4.0.39 software. The dimensionality reduction was performed on the data set using backward stepwise method in Cox regression analysis.

In the data set, the variables Gender, Treatment, Fibrinogen, Alp (Alkaline Phosphatase), D_bil (Direct bilirubin), Ldh (Lactate dehydrogenase), T_bil (Total bilirubin), Mono (Monocytes), Hgb (Hemoglobin), Inr (International normalized ratio), Aptt (Activated partial thromboplastin time), Ferritin obtained after backward stepwise method in Cox regression analysis were included in the model. In R 4.0.3, the package "survival" is used to obtain Cox regression model, Schoenfeld Residuals, Martingale residuals, deviance residuals. The package "coxrobust" is used to obtain Cox robust regression model. The package "BCSOD" is used to perform Bootstrap Hypothesis test based on Concordance c-index. Finally, the package "qvalue" is used to obtain q-values.

Descriptive statistics for the variables to be used in the model are given in Table 1.

Firstly, we modeled the Cox regression model using the function "coxph" in the library "survival". A robust method of Cox regression was used to show consistency with previous Cox regression analysis results. The library "coxrobust" package was used for robust cox regression model (Table 2).

From the result, the results are not consistent with the previous cox regression model. In Cox robust model, gender and treatment variables are statistically significant for a 5% level of significance (Table 2).

First, proportional hazards assumption, which is the Cox regression model assumption, needs to be tested. We tested proportional hazards assumption using the function "cox.zph" in library "survival".

Figure 1 show that the proportional hazards assumption of the established model is met. The p value for all variables is above 0.05. So, the proportional hazards hypothesis is not violated.

Table 1.

General distribution of variables in the model

Qualitative variables		n (%)	
Prognosis	Alive	181 (86.6)	
	Death	28 (13.4)	
Treatment	Support treat.	182 (87.1)	
	Support treat+Antiviral	27 (12.9)	
Gender	Female	82 (39.2)	
	Male	127 (60.8)	
Quantitative variables		Mean±SD	Median [Q1-Q3]
Fibrinogen		289.89±92.01	279[234-356]
Alp		131.88±103.31	95[66-158]
D_bil		0.61±1.44	0.2[0.13-0.39]
Ldh		976.31±1001.08	583[362-1201]
T_bil		0.95±1.47	0.49[0.32-0.82]
Mono		0.28±0.27	0.17[0.09-0.35]
Hgb		12.99±2.16	13.23[11.9-14.6]
Inr		1.25±0.45	1.14[0.98-1.37]
Aptt		50.79±23.82	42[35.2-60]
Ferritin		6790.29±12754.84	2000[646-4432]

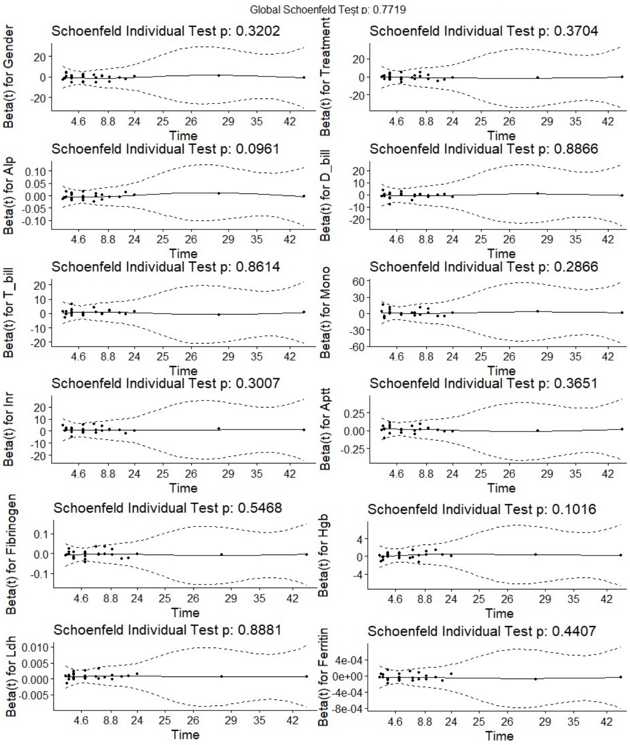
Table 2.

The results on the Cox Regression, Robust Cox Regression and Final Cox Regression

Variables	Cox Regression			Robust Cox Regression			Final Cox Regression		
	coef	HR	p	coef	HR	p	coef	HR	p
Gender	-0.746	0.525	0.156	-1.739	0.843	0.039	-1.198	0.567	0.035
Treatment	-0.650	0.615	0.291	-2.319	0.877	0.008	-1.748	0.752	0.020
Fibrinogen	-0.005	0.003	0.086	-0.006	0.005	0.223	-0.008	0.003	0.012
Alp	-0.003	0.002	0.146	-0.009	0.004	0.021	-0.006	0.003	0.023
D_bill	-0.592	0.456	0.194	-1.074	1.110	0.332	-0.859	0.499	0.085
Ldh	0.001	0.000	<0.001	0.001	0.001	0.020	0.001	0.000	<0.001
T_bill	0.642	0.383	0.094	1.056	1.040	0.311	0.878	0.430	0.041
Mono	0.992	1.011	0.327	1.889	1.950	0.333	2.027	1.069	0.058
Hgb	0.207	0.125	0.099	0.225	0.138	0.104	0.202	0.117	0.084
Inr	0.884	0.460	0.055	1.535	1.080	0.155	1.439	0.620	0.020
Aptt	0.018	0.008	0.017	0.050	0.014	<0.001	0.032	0.009	<0.001
Ferritin	0.000	0.000	0.009	0.000	0.000	0.046	0.000	0.000	<0.001

Figure 1.

Schoenfeld residual plot for independent variables



Martingale and deviance residuals were used for outlier detection. The function "resid" was used for martingale and deviance residuals. From martingale and deviance residuals, there are 6 common outliers. These outliers can also be see in the figure 2.

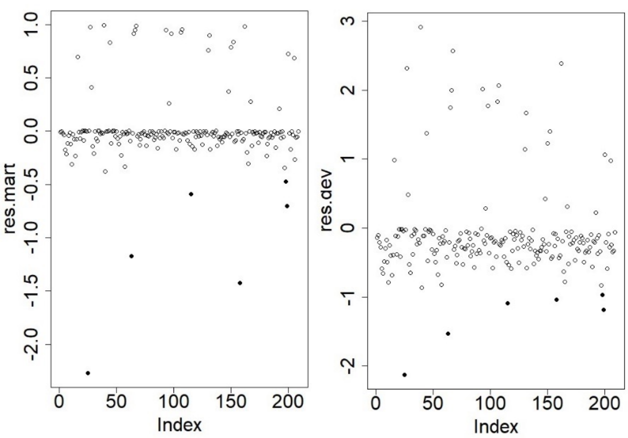


Figure 2.

Martingale and deviance residuals

The other outlier detection method is the bootstrap hypothesis (BHT) based on the concordance c-index. The package "BCSOD" was used for BHT. Here the bootstrap number is 1000. The lowest p-values in table indicate outliers (Table 3).

Table 3.

Result of BHT on concordance c-index

id	exp infl	max	p-value
67	0.039	0.103	0.089
93	0.022	0.108	0.240
63	0.021	0.110	0.252
66	0.020	0.109	0.268
25	0.021	0.103	0.269

The top outliers are given in the table 3. Finally, to obtain an overall result, rank product test was used. In rank product test, we performed the algorithms p-values and q-values, respectively. p-values are obtained with the function "rankprodbounds". q-values are obtained with the package "qvalue". If we combine the results obtained, we obtain Table 4.

Table 4.

The results of the rank product test

id	rank_ martingale	rank_ deviance	rank_ bht	p values	q values
25	1	5	25	0.0002	0.0481
5	19	35	5	0.0097	0.4047
6	16	32	6	0.0089	0.4047
11	10	28	11	0.0090	0.4047
63	3	13	63	0.0071	0.4047
14	14	30	14	0.0169	0.5196
40	7	25	40	0.0199	0.5196
158	2	21	158	0.0189	0.5196
39	209	1	39	0.0229	0.5322
7	37	53	7	0.0367	0.5944
57	9	27	57	0.0370	0.5944
115	5	19	115	0.0299	0.5944
199	4	17	199	0.0362	0.5944
30	17	33	30	0.0438	0.6537
10	36	52	10	0.0479	0.6680
1	142	163	1	0.0572	0.7036

Since 25th observation has the smallest significant q-value, further analysis can be made by excluding this observation from the study. The Cox regression model after the 25th observation eliminated is given in Table 2. After eliminating an outlier, there is an increase in the number of significant variables in final cox regression model.

4.CONCLUSION

According to the results obtained from the clinical data set, the outliers detected according to martingale residual method, deviance residual method and BHT based on concordance c-index. As a general approach combining these methods, rank product test was used. In our clinical data set, there is one outlier. After eliminating the 25th observation, there was an increase in the number of significant variables in cox regression model. There are different residual methods in the literature to be used for the outlier detection.[14,19,28] In the rank product method, the aim is to provide a unified definition with the ranking obtained from the methods used. The rank product test can be applied by using these different methods. Presence of outlier causes the proportional hazards assumption to be violated especially in Cox regression analysis. In order to avoid this situation, it is recommended to use these methods for outlier detection. With this application, especially in survival data, it will find application in different disciplines.

Acknowledgments

The study data was approved by the Tokat Gaziosmanpaşa University Clinical Research Ethics Committee (Project number:16-KAEK-020).

Disclosure Statement

No potential conflict of interest was reported by the authors.

Author Contribution Statement

Concept/Design: OD. Analysis/Interpretation: OD, ÜE. Data Acquisition: OD, Writing: OD, ÜE. Revision and Correction: OD, ÜE. Final Approval: ÜE, OD.

References

1. Cox DR. Regression models and life-tables. Journal of Royal Statistical Society. 1972;34(2):187-202.
2. Pinto JD, Carvalho AM, Vinga S. Outlier detection in survival analysis based on the concordance c-index. SCITEPRESS-Science and Technology Publications, Lda; 2015:75-82.
3. Eo S-H, Hong S-M, Cho H. Identification of outlying observations with quantile regression for censored data. arXiv preprint arXiv:14047710. 2014.
4. Aktas T, Aktas F, Ozmen C, Ozmen Z, Kaya T, Demir O. Mean Platelet Volume (mpv): A New Predictor of Pulmonary Findings and Survival in Cchf Patients? Acta Medica Mediterranea. 2017;33(2):183-190.
5. Therneau TM. A Package for Survival Analysis in R. R package version 3.2-11. 2021.
6. Bednarski T, Borowicz F, Scogin S. Coxrobust: Fit Robustly Proportional Hazards Regression Model. 2022.
7. Pinto J. BCSOD: This packages provides 6 methods to perform outlier detection in survival context.. R package version 1.0. 2015.
8. Storey John D, Bass AJ, Dabney A, Robinson D. Qvalue: Q-value estimation for false discovery rate control. R package version 2.15.0. 2017.
9. Team RC. R: A language and environment for statistical computing. 2013.
10. Androulakis E, Koukouvinos C, Mylona K, Vonta F. A real survival analysis application via variable selection methods for Cox's proportional hazards model. Journal of Applied Statistics. 2010;37(8):1399-1406.
11. Cox DR, Oakes D. Analysis of survival data. Chapman and Hall, London; 1984.
12. Bednarski T. Robust estimation in Cox's regression model. Scandinavian Journal of Statistics. 1993;20:213-225.
13. Minder CE, Bednarski T. A robust method for proportional hazards regression. Statistics in medicine. 1996;15(10):1033-1047.
14. Carrasquinha E, Veríssimo A, Vinga S. Consensus outlier detection in survival analysis using the rank product test. bioRxiv. 2018:421917.
15. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239-241.
16. Kumar D, Klefsjö B. Proportional hazards model: A review. Reliability Engineering & System Safety. 1994;44(2):177-188.
17. Barlow WE, Prentice RL. Residuals for relative risk re-

gression. Biometrika. 1988;75(1):65-74.

18. Carrasquinha E, Veríssimo A, Lopes MB, Vinga S. Variable selection and outlier detection in regularized survival models: Application to melanoma gene expression data. Springer; 2018:431-440.
19. Karasoy D, Tuncer N. Outliers in survival analysis. Alphanumeric journal. 2015;3(2):139-152.
20. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika. 1990;77(1):147-160.
21. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-2546.
22. Caldas J, Vinga S. Global meta-analysis of transcriptomics studies. PLoS One. 2014;9(2):e89318.
23. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank. Products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters. 2004;573(1-3):83-92.
24. Carrasquinha E, Veríssimo A, Lopes MB, Vinga S. Identification of influential observations in high-dimensional cancer survival data through the rank product test. BioData Mining. 2018;11(1):1.
25. A. KJ. Comments on the rank product method for analyzing replicated experiments. FEBS Letters. 2010;584(5):941.
26. Eisinga R, Breitling R, Heskes T. The exact probability distribution of the rank product statistics for replicated experiments. FEBS letters. 2013;587(6):677-682.
27. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):479-498.
28. Halabi S, Dutta S, Wu Y, Liu A. Score and deviance residuals based on the full likelihood approach in survival analysis. Pharmaceutical statistics. 2020;19(6):940-954.