J.Natn.Sci.Foundation Sri Lanka 2024 52 (1): 59 - 68
DOI: http://dx.doi.org/10.4038/jnsfsr.v52i1.11460

RESEARCH ARTICLE

Statistical Modelling

Comparison of methods for handling outliers in Cox regression model

N Alkan[1*], MC Pardo[2] and BB Alkan[3]

[1] *Department of Business Administration, Akdeniz University, Antalya, Turkey.*
[2] *Department of Statistics and Operational Research, Complutense University of Madrid, Spain.*
[3] *Department of Educational Sciences, Akdeniz University, Antalya, Turkey.*

Submitted: 14 November 2022; Revised: 09 November 2023; Accepted: 22 November 2023

Abstract: The Cox regression analysis is used to determine the relationship between a dependent variable and covariates in survival analysis involving censored data. The proportional hazards assumption is one of the most important assumptions of Cox regression. Outliers may have a strong influence on the Cox regression model's parameter estimates and lead to violation of the proportional hazard assumption. Therefore, having outliers in the data set is a problem for researchers. In this case, robust estimations are commonly used to infer the parameters in a more robust way. However, we explore a new approach consisting of considering an outlier as missing data and replacing it by the multiple imputation method. The aim of this study is to compare these two methods through simulation. Furthermore, an analysis of a lung cancer data set is considered for illustration. According to the results of the study carried out based on simulated data sets and a real data set, the multiple imputation method, which is a missing data analysis method, solves the problem of outliers better than the robust estimation method, as the outcome is closer to the results obtained through original data.

Keywords: Cox regression, multiple imputation, outliers, robust Cox regression.

INTRODUCTION

One of the most important assumptions of Cox regression is the proportional hazard assumption. Outliers in data could lead to violation of this assumption and it leads to the emergence of inaccurate estimates. Because outliers may have a strong influence on the model's parameter estimates, the outliers may be a problem for researchers. In many studies, data have been remodelled by deleting the observation with outliers. But in this case the data set to be analysed becomes smaller, compromising the statistical power of the study and eventually the reliability of its results. For this reason, it is not right to drop an observation just because it is an outlier, *e.g.*, individuals that lived too long or died too early when compared with others with the same clinical conditions.

Robust statistics have good performance of data sets with outliers and other small departures from model assumptions. Robust estimators are a modified class of regression parameters estimators (Bernarski, 1989). In Cox regression analysis, the partial likelihood function is used to estimate the parameters and the modification process is executed via this function. Recently, Farcomeni and Viviani (2011) proposed a modified Cox model that is fitted by trimming the smallest contributions to the partial likelihood.

Also another suggestion for the outlier problem is to use the multiple imputation method which is one of the missing data analysis methods (Alkan & Alkan, 2018). In this method, instead of deleting all the values of the observation with an outlier, only the outlier is deleted and this missing value is imputed by using multiple imputation method.

[*] Corresponding author (nesrin.alkan@gmail.com; https://orcid.org/0000-0003-1452-4780)

So, in this study, the solution of the problem that results from a violation of assumptions is discussed using two different methods. Firstly, the problem caused by outliers is transformed into a missing value problem and it is solved by the multiple imputation method, and then Cox regression analysis was applied to the completed data set. Secondly, robust estimates which give accurate results in case of deviations from the assumptions are obtained for Cox regression analysis. Therefore, our aim is to compare the results of Cox regression analysis after multiple imputation methods and the results of robust Cox regression analysis.

For this purpose, we carried out a simulation study. The simulations were made to determine how much outliers influence the parameter estimates. The simulation scenarios were created according to different censor ratios (20%, 40%, 60%), different outliers ratios (10%, 20% 30%), and sample size N = 50, 100. So the multiple imputation and robust Cox regression have been compared in different situations. Furthermore, we revisit the popular NCCTG lung cancer data to illustrate both methods.

MATERIALS AND METHODS

Cox regression

Cox regression analysis is used extensively in biological and medical studies in survival analysis involving censored data. In survival analysis, the Cox regression analysis is used to determine the relationship between dependent variable and covariates (Cox, 1972). The Cox regression model can be written as:

$$\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{X_i}) \qquad ...(1)$$

where $\lambda_i(t)$ represents the hazard function for the i-th individual, $\boldsymbol{\beta}'$ is the unknown parameter vector, $\mathbf{X_i}$ is the p-dimensional covariate vector for the i-th individual, and $\lambda_0(t)$ is called the baseline hazard function that represents the hazard when all the independent variables are equal to zero. The baseline hazard function is a non-parametric function that shows the change of hazard over time without considering the effects of specific covariates or independent variables (Hosmer & Lemeshow, 1999). This method estimates the parameters by maximizing the partial likelihood function (Kalbfleisch & Prentice, 1980). The partial likelihood is provided by the following equation:

$$\prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{\sum_{t_j \geq t_i}\exp(\boldsymbol{\beta}'\mathbf{X}_j)}\right]^{\delta_i} \qquad ...(2)$$

where t_i is the minimum of survival and censored time, $\delta_i = 0$ if censored and $\delta_i = 1$ if the event occurred. The proportional hazard assumption is the principal assumption of the Cox regression analysis**.** In this assumption, the hazard ratio (HR) of any two individuals is constant over the time axis in the model. Therefore, the Cox regression model is also known as a proportional hazard model. Furthermore, baseline hazard function is independent of the covariates. Reliable statistical inferences and estimates are obtained by providing this assumption (Kleinbaum & Klein, 1996).

In statistics, an outlier is an observation point that is different from the rest of the data. Outlier values in the dataset may have a great influence on parameter estimation. For this reason, model adequacy should be checked after the survival data set is modelled by Cox regression analysis (Xue & Schifano, 2017; Alonso & Pardo, 2020). Model diagnosis is one of the most important parts of the modelling process. Many diagnostic methods are based on the analysis of model residuals. One of the most well-known is the Schoenfeld residual analysis. The Schoenfeld residuals (1982) are the difference between the true value of the covariate and the average of weighted risk scores.

$$\hat{r}_k(\beta) = \mathbf{X_{(k)}} - \bar{x}(\beta, t_k), \qquad ...(3)$$

$$\bar{x}(\beta, t_k) = \frac{\sum_{i=1}^{n} Y_i(t_k)\exp(\boldsymbol{\beta}'\mathbf{X_i}(t_k))\mathbf{X_i}(t_k)}{\sum_{i=1}^{n} Y_i(t_k)\exp(\boldsymbol{\beta}'\mathbf{X_i}(t_k))} \qquad ...(4)$$

where $\bar{x}(\beta, t_k)$ is a weighted average of covariates over observations which are still at risk at time t_k. $\mathbf{X_{(k)}}$ is k-th covariate vector of a subject with event time t_k. $Y_i(t_k)$ indicates whether the i-th subject is still at risk at time t_k. The term $\exp(\beta' X_i(t_k))$ is the risk score for the i^{th} observation.

The approach to test the proportional hazard assumption in Schoenfeld (1982) is generalized by Grambsch and Therneau (1994). They defined the following function to test the proportional hazard assumption for each covariate j,

$$T_j(g) = \frac{\sum(g_j \hat{r}_{jk})^2}{D_{jj}} \qquad ...(5)$$

where g_j is an element of G_k which is a diagonal matrix and shows how the survival times should be transformed, $\hat{r}_{jk}$ is the j-th element of Schoenfeld residual, D_{jj} is an elementof $D = \sum G_k \hat{V}_k G_k^r - \left(\sum G_k \hat{V}_k\right)\left(\sum \hat{V}_k\right)^{-1}\left(\sum G_k \hat{V}_k\right)^T$ where $\hat{V}_k$ is the observed variance of $\hat{\beta}$ at the k –th time. This test statistic is used for testing the proportional hazard assumption of each covariate. The test statistic is distributed as $\chi^2_{(1),}$ if the proportional hazard assumption is provided (Xue & Schifano, 2017). Graphs of the Schoenfeld residuals against transformed time are used for checking violations of the proportional hazard assumption. If the residuals are around a horizontal line, the proportional hazard assumption is satisfied (Schoenfeld, 1982). Both statistical tests and graphical diagnostics which are based on the scaled Schoenfeld residuals are used to check proportional hazard assumption.

Multiple imputation method

Missing data may be encountered even in a well-planned and controlled study. If there is a missing value in the data, statistical power is reduced, biased estimates are produced and invalid results are obtained. Therefore, missing data is an important problem for researchers. Also statistical methods and software suppose that all variables in a model are complete. To solve the missing data problem, either the observation which has a missing value is deleted, or the value is imputed. The multiple imputation (MI) method, which has a better performance than other imputation methods, is a missing data analysis method (Alkan *et al*, 2013).

The multiple imputation method develops the Bayesian approaches to solve the problem of missing value in the data (Enders, 2010). In the multiple imputation method, for generating *m* complete data sets, each of the missing values is filled in *m* times. Standard statistical methods analyze the imputed data sets and combine the results from these analyses for the inference. Rubin (1987) outlined the formulas for combined parameter estimates, which are based on the arithmetic mean of the *m* complete data estimates. Our first proposal for handling the outlier problem is to consider each outlier as missing data as proposed Alkan and Alkan (2018).

Robust cox regression

The partial likelihood estimator β used for parameter estimation in the Cox regression is very sensitive to deviations from the model assumptions. Outlier values cause a violation of the most important assumption of the Cox regression. In such a case, unreliable, mis-established models can occur. For this reason Farcomeni and Viviani (2011) proposed the Robust Cox Regression for data sets with outliers. Bretagnolle and Huber-Carrol (1988) have shown that exclusion of the relevant covariate with outlier gives biased results. Reid and Crepeau (1985) and Bednarski (1989) have shown that even slight departures from the proportional hazard assumption lead to bias in the estimation of β. Bednarski (1993) showed how the proportional hazard estimator's prediction equation was modified to get robust estimates. Bednarski (1993) started by using equation (2) for this modification process and equation (6) was obtained

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\mathbf{X_i}-\frac{\sum_{\mathrm{t_j\geq t_i}}\boldsymbol{X}_\mathbf{j}\exp\left(\boldsymbol{\beta}'\mathbf{X_j}\right)}{\sum_{\mathrm{t_j\geq t_i}}\exp\left(\boldsymbol{\beta}'\mathbf{X_j}\right)}\right]\delta_{\mathrm{i}}=0 \qquad ...(6)$$

The estimator solving this equation is much affected by the large values of $\exp(\beta'\mathbf{X}_j)$. One way of reducing the effect of large values is to regulate equation (6) Equation (6) is modified using the funtion A(t, X), which is zero. Take a smooth non-negative function to obtain equation (7).

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A}(\mathrm{t_i},\mathrm{X_i})\left[\mathbf{X_i}-\frac{\sum_{\mathrm{t_j\geq t_i}}\mathrm{A}(\mathrm{t_i},\mathrm{X_j})\mathbf{x_j}\exp\left(\boldsymbol{\beta}'\mathbf{X_j}\right)}{\sum_{\mathrm{t_j\geq t_i}}\mathrm{A}(\mathrm{t_i},\mathrm{X_j})\exp\left(\boldsymbol{\beta}'\mathbf{X_j}\right)}\right]\delta_{\mathrm{i}}=0 \qquad ...(7)$$

The function A (t,X) is processed at two points. The outer sum is used to reduce the weight of uncensored observations which have large values of $\exp(\beta'\mathbf{X}_j)$. The A(t,X) in square brackets are computed for down-weighting all observations with relatively large values of $\boldsymbol{\beta}'\mathbf{X_j}$. Such double correction leads to the consistency of the estimator. The robust estimator of $\boldsymbol{\beta}$ is obtained by solving equation (7). Our second proposal for handling the outlier problem is to use robust estimation which traditionally addresses the consequences of having outliers in the data.

RESULTS AND DISCUSSION

In this study, the results of Cox regression analysis after multiple imputation methods and robust Cox regression analysis were compared for both simulation datasets and real datasets. In the simulation study, a Cox regression model with three covariates is considered. Standard normal distribution is assumed for the covariates. A Weibull distribution with scale parameter equal to 0.002 and shape parameter equal to 1 is used for baseline hazard function. A shape of 1 means a constant baseline hazard function. Censoring times are generated from a Weibull with scale parameter equal to 0.008 and shape parameter equal to 1 for a censoring proportion of 20%. And scale parameters of 0.004 and 0.002 are chosen to produce specific censoring proportions of 40% and 60%, respectively. The minimum of event time or censoring time was recorded as survival time. The true regression coefficients are fixed as β_1=1, β_2= −3, β_3= 2. Sample sizes n = 50 and 100 are selected.

In order to generate the data sets with outliers, some of the extreme values are given to the largest and smallest values of X_1 and X_2 in the simulated data sets. These values are evaluated as outliers. Changes are not made in the values of X_3. At this stage of the simulation study, different outlier proportions of 10%, 20%, and 30% are considered. In order to illustrate the use of multiple imputation method in the presence of outliers, the outlier values in the data are deleted and missing data sets are created. Cox regression analysis for clean simulated survival data sets, robust cox regression analysis for data sets with outliers, and multiple imputation method for missing data sets are used and estimation are repeated 500 times for each simulation. To determine how much outliers influence the parameter estimates and to compare the methods, Bias and RMSE are calculated as follows

$$\text{Bias} = \frac{\sum_{i=1}^{N}(\beta_i - \hat{\beta}_i)}{N}; \qquad ...(8)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\beta_i - \hat{\beta}_i)^2}{N-1}} \qquad ...(9)$$

where N = 500 and β_i is the parameter estimation of the Cox regression model which is calculated from the data set with no outliers. $\hat{\beta}_i$ is the parameter estimation of robust Cox regression and Cox regression after Multiple İmputation. So $\text{Bias}_{\text{Robust}}$ and $\text{Bias}_{\text{MI+Cox}}$ are calculated. Similar calculations are made for RMSE and the results are given in Table 1 and Table 2.

Table 1: Bias and RMSE for different outlier and censor ratios and N = 50 for each of the methods.

N	Censor ratio	Outlier ratio	Parameter	Bias MI+Cox	Bias robust Cox	RMSE MI+Cox	RMSE robust Cox
50	20%	10%	β_1	0.262	0.415	0.342	0.738
			β_2	0.174	0.399	0.265	0.685
			β_3	0.194	0.309	0.371	0.563
		20%	β_1	0.046	0.153	0.059	0.179
			β_2	0.108	0.293	0.159	0.341
			β_3	0.031	0.063	0.037	0.089
		30%	β_1	0.726	1.176	0.948	1.497
			β_2	1.743	2.432	2.359	3.042
			β_3	0.861	0.964	1.417	1.458
	40%	10%	β_1	0.174	0.416	0.249	0.545
			β_2	0.569	1.291	0.671	1.901
			β_3	0.187	0.487	0.234	0.731
		20%	β_1	0.229	0.765	0.293	0.912
			β_2	0.441	0.932	0.551	1.545
			β_3	0.397	0.534	0.495	0.869
		30%	β_1	0.320	0.883	0.459	0.948
			β_2	0.766	1.098	0.912	1.409
			β_3	0.685	0.446	0.801	0.509
	60%	10%	β_1	0.351	0.610	0.499	0.738
			β_2	0.499	0.509	1.067	0.702
			β_3	0.570	0.189	0.888	0.268
		20%	β_1	0.550	0.706	0.721	0.920
			β_2	0.605	1.269	0.725	1.717
			β_3	0.415	0.339	0.532	0.533
		30%	β_1	0.374	0.519	0.435	0.634
			β_2	0.815	1.086	1.123	1.452
			β_3	0.639	0.506	0.904	0.749

Table 2: Bias and RMSE for different outlier and censor ratios and N = 100 for each of the methods.

N	Censor ratio	Outlier ratio	Parameter	Bias MI+Cox	Bias robust Cox	RMSE MI+Cox	RMSE robust Cox
100	20%	10%	β_1	0.243	0.476	0.608	1.189
			β_2	0.422	0.512	1.056	1.281
			β_3	0.233	0.247	0.583	0.617
		20%	β_1	0.439	0.862	1.098	2.155
			β_2	0.932	1.359	2.33	3.397
			β_3	0.612	0.679	1.532	1.699
		30%	β_1	0.659	0.832	1.649	2.081
			β_2	1.351	2.295	3.379	5.739
			β_3	0.901	0.929	2.251	2.322
	40%	10%	β_1	0.281	0.456	0.702	1.140
			β_2	0.395	0.523	0.988	1.308
			β_3	0.262	0.262	0.654	0.656
		20%	β_1	0.502	0.717	1.256	1.792
			β_2	0.839	1.798	2.097	4.494
			β_3	0.612	0.724	1.530	1.811
		30%	β_1	0.636	0.687	1.589	1.717
			β_2	1.112	1.786	2.780	4.466
			β_3	0.696	0.677	1.742	1.692
	60%	10%	β_1	0.447	0.578	1.118	1.446
			β_2	0.648	1.244	1.620	3.109
			β_3	0.494	0.709	1.235	1.773
		20%	β_1	0.447	0.725	1.118	1.812
			β_2	0.654	1.531	1.635	3.828
			β_3	0.434	0.686	1.085	1.715
		30%	β_1	0.281	0.651	0.704	1.629
			β_2	0.638	1.244	1.595	3.109
			β_3	0.584	0.462	1.460	1.150

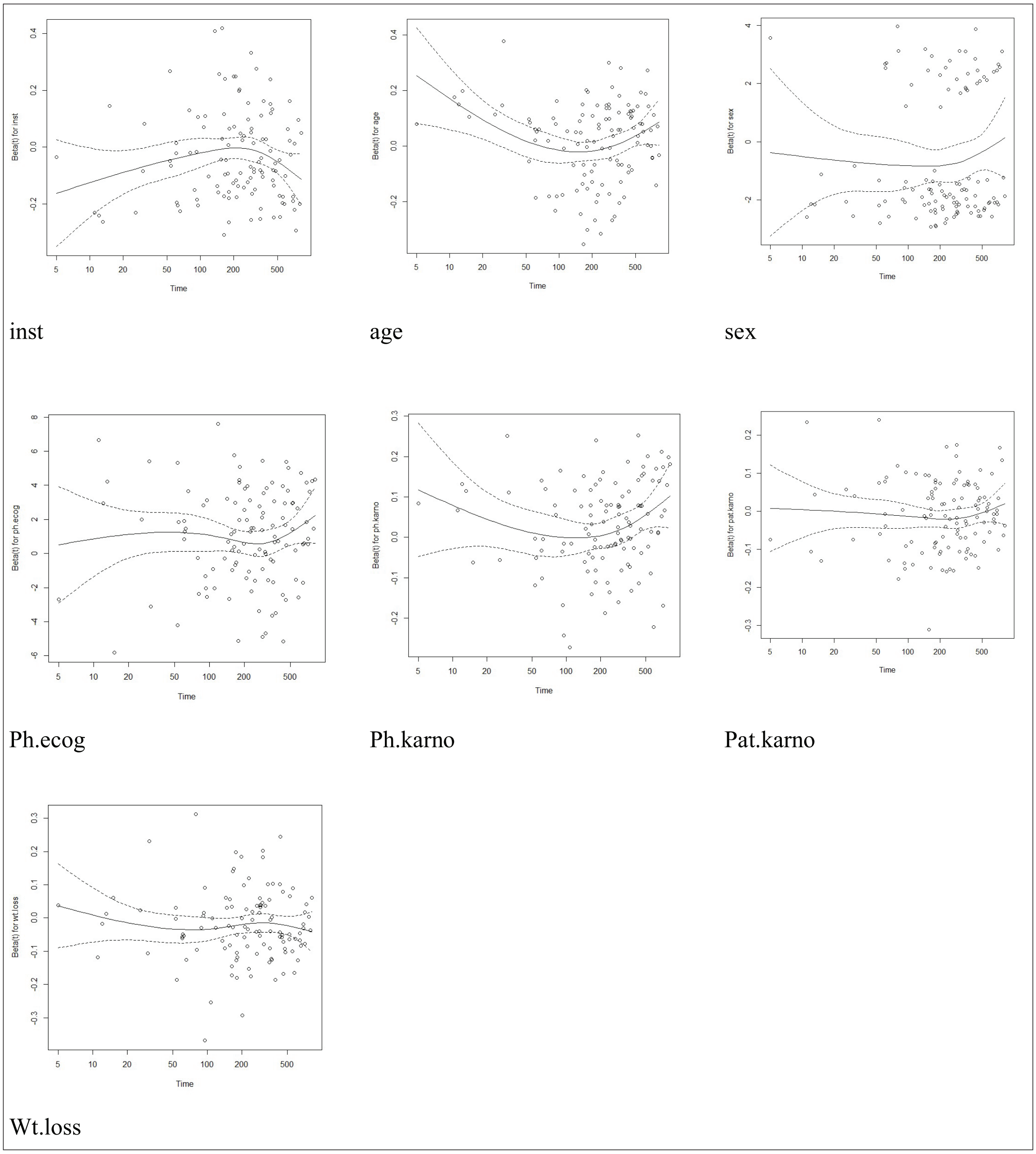
Bias and RMSE are calculated to determine how close to the true value the estimates are calculated for each method. When Table 1 and Table 2 are examined, both methods obtained estimates near the parameter value for all censors and outlier ratios. Also, it was observed that as the outlier ratio is increased from 10% to 30%, the RMSE and BIAS values calculated in both methods generally tend to increase. Although the increase in the outliers ratio affected the methods, both methods made good estimates. However, the BIAS and RMSE values of the estimates obtained with the MI+ Cox method were consistently smaller than those obtained from the robust method for the coefficients of X_1 and X_2 which are covariates with outliers in both N = 50 and N = 100. So, according to simulations, the multiple imputation method is a good alternative to robust methods in the presence of outliers.

In order to compare the proposed methods with a real data set, the NCCTG lung cancer data set from R library is used. There are 228 patients and 8 covariates in the data set; 61 observations were deleted due to missingness and as the result, N = 167 and number of events is 120 with a censored ratio of 28%. The prognostic factors in the data set are listed as follows: institution code (inst), age in years (age), sex, ECOG performance score (ph.ecog), Karnofsky performance score rated by physician (ph.karno), Karnofsky performance score as rated by patient (pat.

karno), calories consumed at meals (meal.cal), and weight loss in last six months (wt.loss) (Loprinzi *et al*, 1994).

Proportional hazard assumption is not provided for meal.cal which is one of the covariates in the data set. The aim of the study is to determine the method that gives the nearest estimation to the parameter estimation in the case where the assumption is achieved. Therefore, all variables in the data set that we will use as references will have to provide the assumption. For this reason, the meal.cal covariate is not included in the model.

As first, a residual analysis for original data is carried out and no significant outliers were found but some of the observations have potential outliers in wt.loss variables. To obtain an outlier data set, extreme values were given to the values of these observations and a data set containing 10% outliers was created.



inst age sex

Ph.ecog Ph.karno Pat.karno

Wt.loss

Figure 1: Graphs of Schoenfeld residuals for the original data set

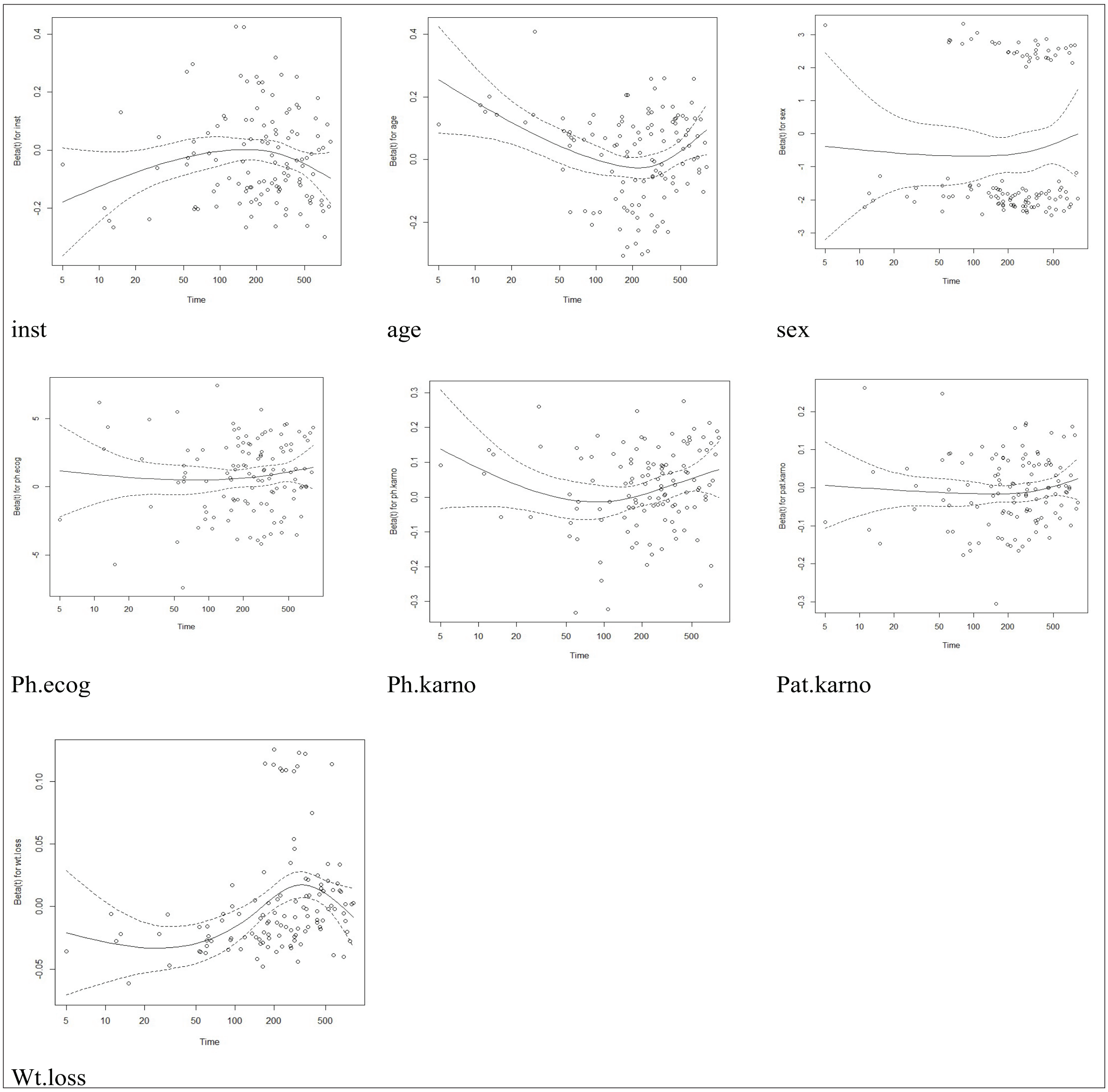
Testing of proportional hazard assumption for lung cancer data sets

The assumption test of the Cox regression analysis is performed using Schoenfeld residuals for the original data sets. The graphs of Schoenfeld residuals for each covariate of the original data set are given in Figure 1 and the statistical results of Schoenfeld residual analysis for the original data set are given Table 3.

Figure 1 shows that the residuals are randomly around a horizontal line for all of the covariates. In other words all variables provide proportional hazard assumption.

Table 3: Statistical results of Schoenfeld residual analysis for the original data set

	rho	chisq	p value
İnst	0.0298	0.125	0.723
Age	-0.0715	0.699	0.403
Sex	0.0613	0.416	0.519
Ph.ecog	0.0354	0.167	0.683
Ph.karno	0.0818	0.572	0.450
Pat.karno	0.0098	0.013	0.909
Wt.loss	-0.0283	0.114	0.736



inst

age

sex

Ph.ecog

Ph.karno

Pat.karno

Wt.loss

Figure 2: Graphs of Schoenfeld residuals for the data set with outlier

Table 3 contain the correlation between Schoenfeld residuals and transformed survival time, the test statistic which is given in equation (5) and the two sided p-value. From Table 3, we conclude that the same, proportional hazard assumption is provided for all variables (p > 0.05). Also the correlations are close to 0 so the assumption is provided. This shows that the original data set can be analyzed by Cox regression and the results obtained will be reliable.

The proportional hazard assumption has been tested after adding the outliers to the data set. Schoenfeld residual analysis graphs for each covariate are given in Figure 2. The statistical results of the residual analysis are also given in Table 4.

Table 4: Statistical results of Schoenfeld residual analysis for data set with outliers

	rho	chisq	p value
İnst	0.0151	0.0334	0.855
Age	-0.134	2.5459	0.111
Sex	0.0524	0.3151	0.574
Ph.ecog	0.0588	0.463	0.496
Ph.karno	0.0868	0.732	0.392
Pat.karno	0.050	0.350	0.554
Wt.loss	0.2844	11.821	0.00058

Figure 2 shows that in case of outliers in the data set, the assumption of proportional hazard for the wt.loss covariate is violated due to some of the residuals being scattered in a different way. Also, according to the statistical results in Table 4, the wt.loss covariate has been seen as not providing the assumption (p < 0,05).

The existence of outliers is a problem for researchers because they affect the proportional hazard assumption. In order to solve this problem, we propose to consider the outlier values as missing data and a value is assigned for each missing value by the multiple imputation method so that 5 completed data sets are obtained. The assumption test was performed with Schoenfeld residual analysis for each of the data sets which is completed with the multiple imputation method (MI), and the result was obtained for all data sets. The statistical result of Schoenfeld analysis for the first imputed data set is given in Table 5. Similar results were obtained for the other imputed data sets. That is, it was seen that the assumption provided for all imputed data sets.

According to Table 5, the proportional hazard assumption is satisfied for all variables (p > 0,05). That is, these results indicate that the problem of assumption violation caused by outliers can be solved by imputation.

Table 5: Statistical results of Schoenfeld residual analysis for imputed data set by multiple imputation

	rho	chisq	p value
İnst	0.0236	0.0824	0.774
Age	-0.0946	1.2585	0.262
Sex	0.0773	0.6798	0.410
Ph.ecog	0.0258	0.0917	0.762
Ph.karno	0.0860	0.6854	0.408
Pat.karno	0.0434	0.2663	0.606
Wt.loss	0.0695	0.7826	0.376

Statistical inference of each method for lung cancer data

As there are no outliers in the data set during the first phase of the study then the parameters are estimated by applying Cox regression analysis. These results are compared as original results and used as reference for other results.

In the second phase of the application, the assumption test of the Cox regression analysis is performed using the data set containing the outlier values. As a result, the outliers violate the proportional hazard assumption, and in this case the results are unreliable. Robust estimation can used to overcome this problem. Therefore robust Cox regression analysis is applied to lung cancer data with outliers.

Also, the multiple imputation method overcomes the problem caused by outliers. In the third phase of the application, firstly the outlier values in the data set are deleted and the missing data set is obtained. A value is assigned for each missing value by the multiple imputation method, so that 5 completed data sets are obtained. Cox regression is separately applied to the 5 completed data sets and the combined results are calculated. All results are given in Table 6.

Table 6: Parameter estimates

				Results for data set with outliers.					
	Original			Robust regression			Cox Regression with MI		
Parameters	β	SE	p value	β	SE	p value	β	SE	p value
inst.	-0.0304	0.013	0.0204	-0.0161	0.015	0.0285	-0.0302	0.013	0.0247
Age	0.0128	0.0117	0.2769	0.0065	0.0134	0.628	0.0134	0.012	0.2641
sex	-0.5669	0.2001	0.0046	-0.7041	0.2449	0.0041	-0.5699	0.201	0.0055
Ph.ecog	0.9073	0.2385	0.00014	0.7231	0.4397	0.0100	0.8571	0.239	0.001
Ph.karno	0.0266	0.0116	0.02214	0.0118	0.0202	0.5576	0.0234	0.0121	0.0480
Pat.karno	-0.0109	0.0080	0.173	-0.01272	0.0107	0.232	-0.0107	0.0079	0.1877
Wt.loss	-0.0167	0.0079	0.0346	-0.00306	0.0065	0.571	-0.0171	0.0083	0.0445

In order to compare the performances of robust Cox regression and Cox tegression with MI, Cox regression analysis results of the original data were used as a reference. Robust Cox regression analysis, which is recommended to be used safely in case of violation of the proportional hazard assumption, was applied to the data set containing 10% outliers. As a result, the analysis had similar results in terms of sign and magnitude to parameter estimations of the original data. Then, Cox regression with multiple imputation is applied to the outlier data set; the parameter estimates are very similar to the original estimates, as in the robust method. When these two methods were compared in terms of parameter estimates, the estimates of the Cox regression with MI method were found to be closer to the original estimates. Also, Cox regression with MI had a smaller standard error than robust Cox regression.

CONCLUSION

Outliers may lead to violation of the proportional hazard assumption, which is one of the most important assumptions of Cox regression. Therefore, outliers can have a great influence on parameter estimation. In such a case, the existence of outliers leads to incorrect results. For this reason, the outliers are a problem and there are different methods in the literature for solving this problem. Robust Cox regression is robust to outliers and can reduce the impact of outliers. Therefore, it is a recommended method as it gives safe results in the presence of outliers. The proposed method in this work is Cox regression with multiple imputation. In this method, the outliers are considered as missing data, and a value is assigned for each missing value by the multiple imputation method. Then the completed data set is analyzed by Cox regression. In this study, robust Cox regression and the Cox regression with multiple imputation methods are evaluated and their results are compared with original results.

According to the results of the study carried out on simulated data sets and a real data set, the multiple imputation method, which is a missing data analysis method, solves the problem of outliers better than the robust estimation method, due to results closer to the original results being obtained.

Acknowledgement

This study was supported by Sinop University with FEF-1901-18-16 project number and also partially by grant PID2022-137050NB-100 of the Spanish Ministry of Economy and Competitiveness.

REFERENCES

Alkan N. & Alkan B.B. (2018). A new approach for Cox regression analysis in the presence of outliers, *Süleyman Demirel University Journal of Natural and Applied Sciences* **22** (2): 637–643. DOI: https://doi.10.19113/sdufbed.37782

Alkan N., Terzi Y., Cengiz M.A. & Alkan B.B. (2013). Comparison of missing data analysis methods in Cox proportional hazard models. *Turkiye Klinikleri Journal of Biostatistic* **5**(2): 49–54.

Alonso R. & Pardo M.C. (2020). Assessing influence on the estimated coefficients efficiency in a Cox regression. *Journal of Statistical Computation and Simulation* **90**(7): 1216–1229. DOI: https://doi.org/10.1080/00949655.2020.1720986

Bednarski T. (1989). On sensitivity of Cox's estimators. *Statistics and Decisions* **7**: 215–228.

DOI: https://doi.org/10.1524/strm.1989.7.3.215

Bednarski T. (1993). Robust estimation in the Cox regression model. *Scandinavian Journal of Statistics* **20**:13–225.

Bretagnolle J. & Huber C. (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics* **15**: 125–138.

Cox D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society* **34**: 187–220. DOI: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Enders C.K. (2010). *Applied Missing Data Analysis,* pp.165-286. Guilford Pres, New York, USA.

Farcomeni A. & Viviani S. (2011), Robust estimation for the Cox regression model based on trimming. *Biometrical Journal* **53**(6): 956–73.
DOI: https://doi.org/10.1002/bimj.201100008

Grambsch P. & Therneau T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrik*a **81**: 515–526.
DOI: https://doi.org/10.1093/biomet/81.3.515

Hosmer D.W. & Lemeshow S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data.* John Wiley & Sons, Inc., Canada.

Kalbfleisch J.D. & Prentice R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley and Sons, New York, USA..

Kleinbaum D.G. & Klein M. (1996). *Survival Analysis, A Self Learning Text*. Springer, USA.

Loprinzi C.L., Laurie J.A., Wieand H.S., Krook E., Novotny P.J., Kugler J.W., Bartel J., Law M., Bateman M. & Klatt N.E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology* **12**(3): 601–607.
DOI: https://doi.org/10.1200/JCO.1994.12.3.601

Reid N. & Crepeau H. (1985). Influence function for proportional hazards regression. *Biometrika* **72**: 1–9.
DOI: https://doi.org/10.2307/2336329

Rubin D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons, New York, USA.

Schoenfeld D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**: 239–241.
DOI: https://doi.org/10.2307/2335876

Xue X. & Schifano E.D. (2017). Diagnostics for the Cox model. *Communications for Statistical Applications and Methods* **24**(6): pp. 583–604.
DOI: https://doi.org/10.29220/CSAM.2017.24.6.583