ISSN (Online): 2455-7838
SJIF Impact Factor (2017): 5.705

EPRA International Journal of

Research & Development

(IJRD)

Monthly Peer Reviewed & Indexed International Online Journal

Volume: 3, Issue:6, June 2018

Published By :
EPRA Journals

SJIF Impact Factor: 5.705 Volume: 3 | Issue: 6 | June | 2018 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)

VISUAL EXAMINATION OF OUTLIERS EFFECTS ON DIFFERENT RESIDUAL ANALYSIS METHODS

Nesrin ALKAN[1]
[1,]Department of Statistics, Sinop University, Sinop, Turkey

B. Baris ALKAN
[2] Department of Statistics, Sinop University, Sinop, Turkey

ABSTRACT

The presence of the outliers cause to be a violation of the proportional hazard assumption, which is one of the most important assumptions of the Cox regression analysis. For this reason, the existence of outliers in the data set is a problem for researchers. In survival analysis, it is very important to determine outliers in the data set. The determination of outliers is based on analysis of residuals. Residual types most commonly used in survival analysis are known as Cox-Snell, Martingale, Deviation and Schoenfeld residuals. Since residual analysis yields graphical results, interpretation of the graph requires a special experience. In this study, the changes of residual graphics obtained from these methods are visually examined in the presence of outliers. Thus, it has been visually shown that different residual analysis method should be used for different purposes such as model adaptation, detecting the outliers and assumption control.

KEYWORDS: *Outliers, residual analysis, cox regression*

1. INTRODUCTION

Researchers are often interested in the comparison of different treatment groups in clinical and epidemiological studies. People in groups may have additional attributes. For instance, people may have many features such as demographic variables, physiological variables and behavioural variables. These variables are named independent variables or covariates. Cox regression analysis is used to determine the relationship between dependent variable and covariates. Also Cox regression analysis is the most commonly used method for modelling these data (Cox, 1972).

Partial likelihood estimator which is used for parameter estimation in Cox regression is highly sensitive to deviations from the model (Bednarski, 1989). For this reason, the proportional hazard assumption must be checked. Many statistical analyses are sensitive to the violation of basic assumptions. The existence of outliers causes the violation of the assumptions. Outliers that differ from the rest of the data when compared to the other data cause a violation of the most important assumption of Cox regression (Hawkins, D. M., 1980). In such a case unreliable, inaccurate established models are emerging. For this reason, the big problem for researchers is that there is an outlier in the data set. In Survival analysis, it is very important to determine outliers in the data set. The determination of outliers is based on analysis of residuals. In Survival analysis, there are various types of residuals which are used for different purposes and control the adequacy of the Cox regression model. The most commonly used residual types in survival analysis are Cox-Snell residuals, Martingale residuals, Deviation residuals and Schoenfeld residuals.

In this study, the Cox regression model is investigated in case of violation of assumptions. Also it is aimed to examine how the graphs of the residual analysis change if there are outliers in the data set. For this purpose, survival data of 174 lung cancer patients taken from Ondokuz Mayis University, Faculty of Medicine are used in our study and the assumptions for the data are checked with the residual analysis methods and then outliers in the data set have been examined. Outliers are not observed in the original data and the assumptions have been provided. In this study, in case of violation of assumptions, to examine how the residual analysis chart has changed extreme values are given some values in the data set and a data set with 10% outliers is created. Thus, it is possible to compare visually the results obtained from the outlier data set with the

results obtained from the original data. R 3.3.3 with survival library was used for all graphs obtained in the application.

2. COX REGRESSION ANALYSIS

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. The Cox regression model may be written as:

$$h(t;\mathbf{x}_i)=h_0(t)\exp(\beta' \mathbf{x}_i)$$

where $h(t;x_i)$ represents the hazard function, β' is the unknown parameter vector, x_i is the covariate vector, $h_0(t)$, is called the baseline hazard function (Hosmer and Lemeshow, 1999). This method uses the partial likelihood to estimate the parameters and parameter estimates in the method are obtained by maximizing partial likelihood function (Kalbfleisch and Prentice, 1980). The partial likelihood is provided by the following equation:

$$\prod_{i=1}^{n}\left[\frac{\exp(\beta' x_i)}{\sum_{t_j\geq t_i}\exp(\beta' x_j)}\right]^{\delta_i} \quad (1)$$

where t_i, minimum of survival and censored time, δ_i is 1 in the case of death and 0 in the other case. The principal assumption of the Cox regression analysis is the proportional hazard assumption**.** This assumption is that 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. Reliable statistical inferences and estimates are obtained by providing this assumption.

3. OUTLIERS

In statistics, an outlier is an observation point that is differ from the rest of the data. Outlier values in the dataset may have a great influence on parameter estimation (Alkan and Alkan, 2018, Farcomeni and Viviani, 2011). For this reason, model adequacy should be checked after the survival data set is modeled by Cox Regression analysis. The diagnostic methods used for model control are the most important part of the modeling process. These diagnostic methods base on the analysis of model residuals (Nardi and Schemper, 1999). The analysis of residuals is an effective way in uncovering the different types of models insufficiency. Survival analysis has several types of residuals that can be used for different purposes that control the adequacy of the Cox regression model. The most commonly used residual types in survival analysis are Cox-Snell residuals, Martingale residuals, Deviation residuals and Schoenfeld residuals. We can summarize these methods as follows.

Cox-Snell Residuals

Cox-Snell residuals are used for model conformity in Cox regression analysis. Cox Residuals for observations are given following equation.

$$r_j = \hat{H}_o(t_j)\exp\{\sum_{k=1}^{p} x_{jk}\beta_k\}\ , \qquad j=1,2,\ldots,n$$

where β_k represents regression coefficients, x_{jk} indicates covariates. $\hat{H}_o(t_j)$ is the cumulative hazard ratio estimator. The graph of these residuals gives visual information about the suitability of the model. If the Cox regression model is appropriate, residuals fall on a 45 degree sloped line (Collet, 1994).

Martingale Residuals

Martingale residuals are regarded as difference between observed and expected value and defined as

$$m_j= \delta_j - r_j$$

where δ_j event status of j-th observation r_j is Cox-Snell residuals. The most important feature of Martingale waste is that its totals and its average are zero. Moreover, the covariance between two residuals is also zero. The Martingale residual graph is drawn separately for each explanatory variable. Martingale residuals indicate whether the variables need any transformation for model conformity. If the transformation in the model is favorable, the points on the graph will be linear (Gillespie,2006).

Deviance Residuals

Deviance residuals are the result of conversion of martingale residuals. Martingale residuals take values between -∞ and 1. Deviance residuals can be defined as

$$d_j = sign(m_j)\{-2[m_j + \delta_j\log(\delta_j - m_j)]\}^{\frac{1}{2}}$$

where δ_j event status of j-th observation m_j represents Martingale residuals. If the Martingale residuals are equal to zero, the Deviation residuals become zero. Martingale residual show a rather skewed distribution. Deviation residuals obtained after the transformation of the Martingale residuals are more normally-shaped distribution than the Martingale. This allows for easier interpretation of residual graphics. Outliers in the dataset have residuals that are different from other residuals. Therefore, Deviance residuals can be used as a graphical tool to detect outliers (Gillespie,2006).

Schoenfeld Residuals

The Schoenfeld residuals are the difference between the true value of the covariate and the average of weighted risk scores. Schoenfeld residuals which are used to control the validity of the proportional hazard assumption are plotted against time (Schoenfeld, 1982). If the residuals are around a horizontal line, the proportional hazard assumption is provided. Schoenfeld residuals for k-th covariate and i-th unit is as the following equation.

$$\hat{r}_{ki=}\delta_i\left[X_{ki} - \frac{\sum_{j\in R(t_i)} X_{kj}exp(\hat{\beta}'X_j)}{\sum_{j\in R(t_i)} exp(\hat{\beta}'X_j)}\right]$$

where δ_j event status of j-th observation and $R(t_i)$ denotes all observations that are at risk. Hosmer ve Lemeshow (1999) suggests that the scaled Schoenfeld residuals graph be used for the

proportional hazard assumption. Graphs plotted against time for each covariate to find outliers by scaled Schoenfeld residuals based on the covariance matrix of regression coefficients. It can be written as the following equation.

$$\hat{r}_{ki}^{*} = m \sum_{i=1}^{p} V_{ki} \hat{r}_{ki}$$

where m represents total number of dead individuals, and V represents the covariance matrix estimated from the regression coefficients.

4. APPLICATION

In this study, it is aimed to investigate how the graphs of residual analysis change if there is an outlier in the data set. For this purpose, survival data of 174 lung cancer patients were used from Ondokuz Mayis University, Faculty of Medicine. In the data set, 36 of 174 patients still live when the study ends, so they are censored patients. As first, residual analysis for original data and no significant outliers were found but to obtain an outlier data set, extreme values were given to the some observations and a data set containing 10% outliers was created. Residual analysis is applied for the data set with outlier. To check whether the model is suitable, the Cox-Snell residuals are drawn for both dataset with the outlier value in Figure 1. When the graphs given in Figure 1 are examined, it is seen that the residuals are on a 45 degree inclined line, so that they are suitable for both models but the residuals in the outlier data set are slightly larger.

Scaled Schoenfeld residual analysis for the control of the proportional hazard assumption was made for both the original dataset and the dataset containing the outliers, and the change in the graphs was examined. The graphs are given in Figure 2.

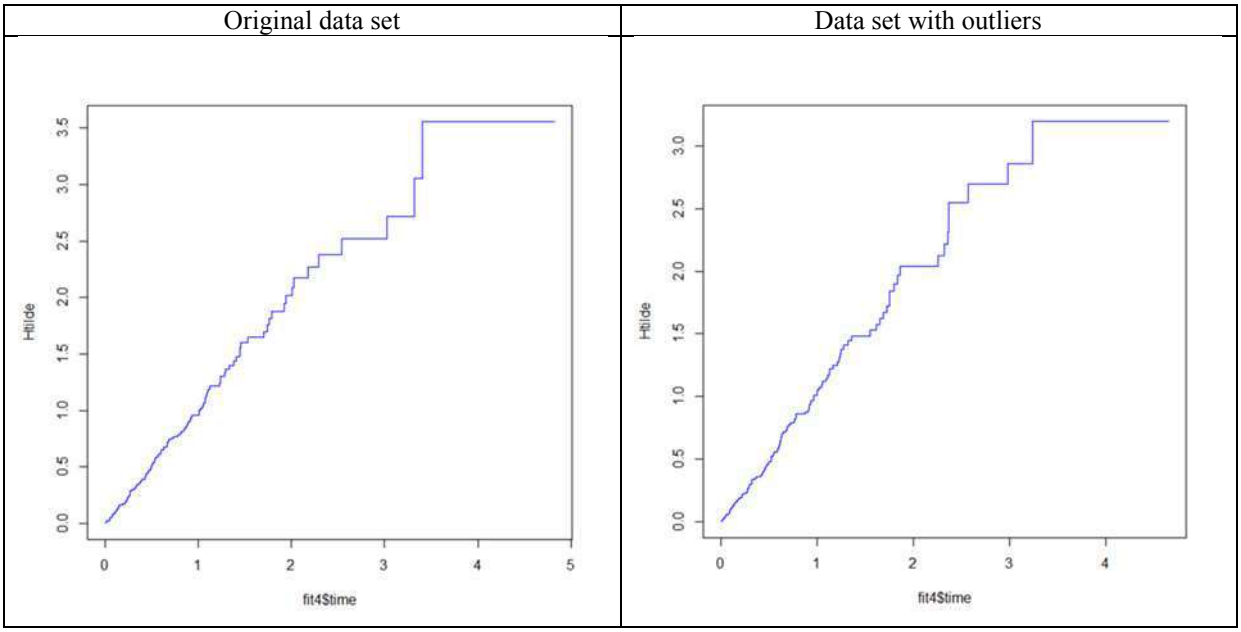


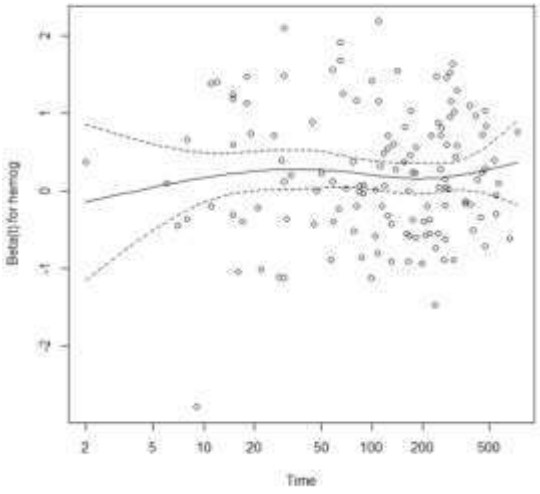
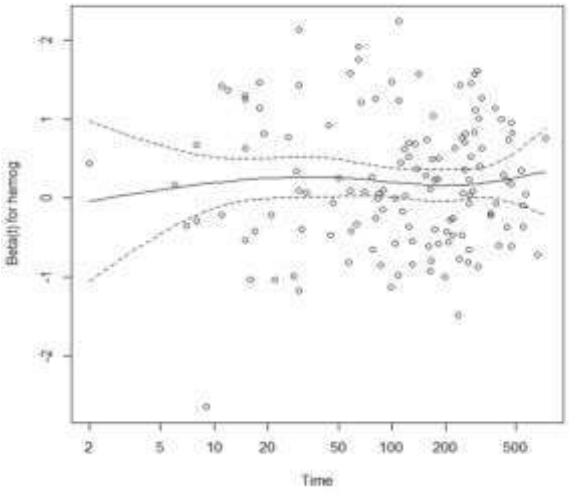
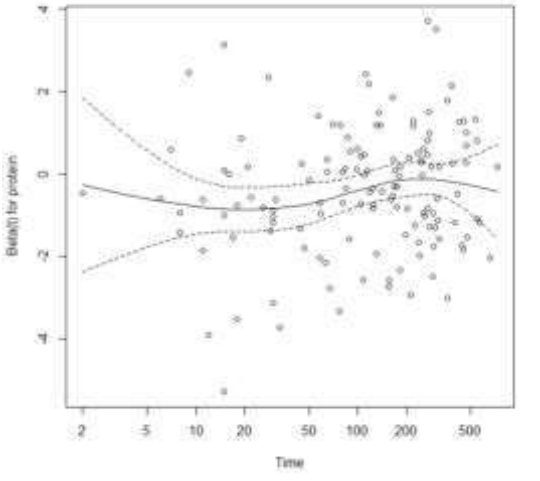
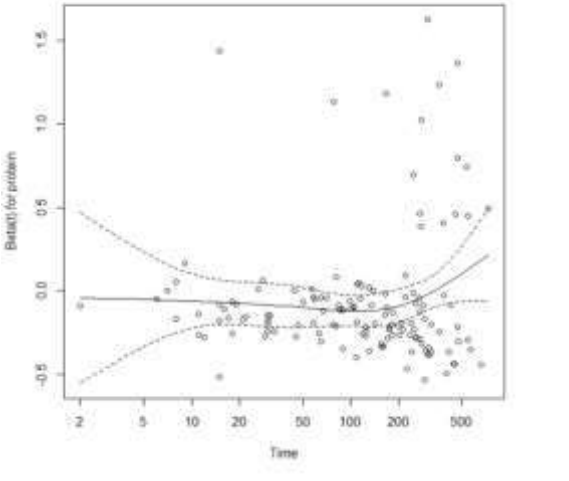
Figure 1: Cox-Snell residuals graphics

When the graphs obtained for the original data set given in Figure 2 are examined, residuals coincidentally extend around a horizontal line for hemoglobin, protein and albumin covariates. That is, all covariates provide a proportional hazard assumption. However in case of outliers in the data set, the assumption of a proportional hazard for hemoglobin and albumin covariates is provided. However, in the protein covariate, some of the residuals are scattered in a different way and it has been seen as not providing the assumption.

Deviance residuals are found both for the original data set and for the outlier value data set to detect outliers found in the data set. The graphs are given in Figure 3. According to Figure 3, the outliers in the protein covariate changed the shape of the graph. The graphs for the other variables are almost identical, whereas the outliers in the protein covariate are clearly visible in deviance residuals graph.

The graphs shown in Figure 4 were obtained to show the effect of outliers on the Martingale residuals and the change in the graph. According to the graph in Figure 4, the outliers in the protein variable changed the shape of the graph.

Consequently, the effect of outliers in the protein variable on residual analysis graphs is seen very clearly. So we can say that, residual analysis are very beneficial methods for the detection of outliers and control of assumptions. In other words residual analysis is used for different purposes such as model adaptation, detecting the outliers and assumption control.

Original data set	Data set with outliers
	
	

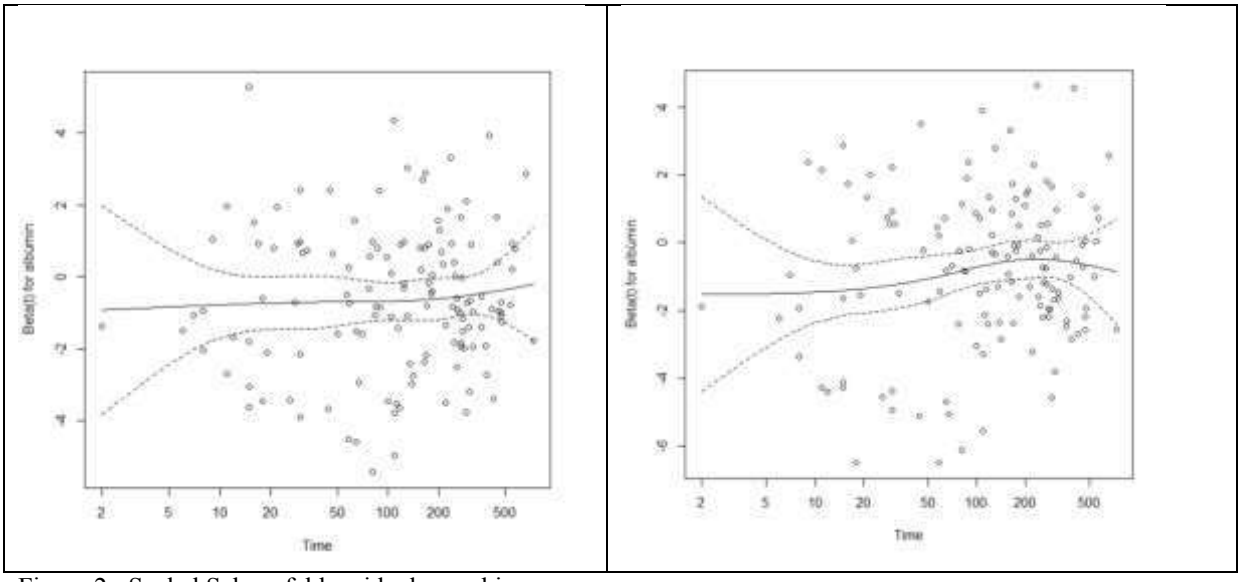
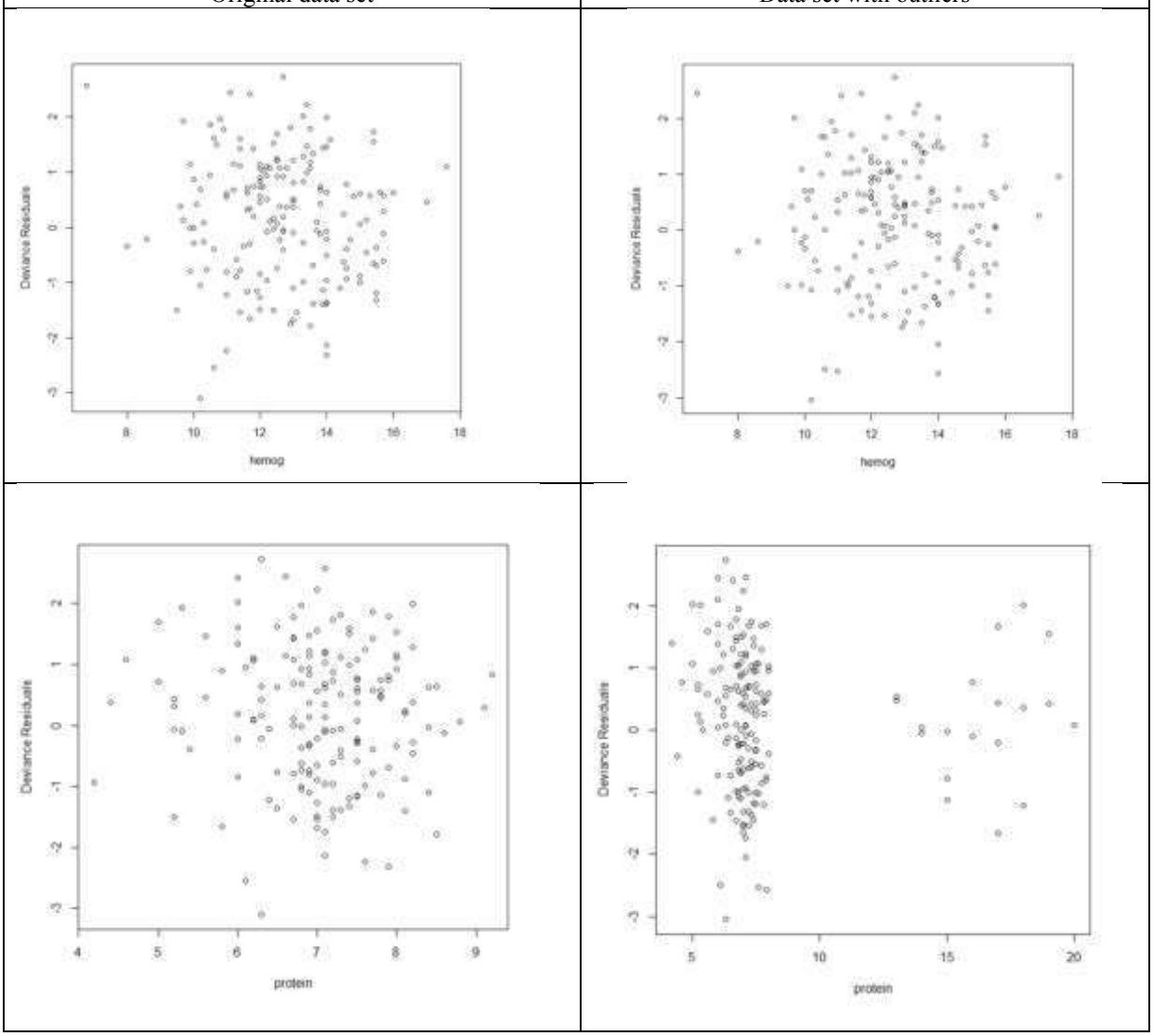


Figure 2 : Scaled Schoenfeld residuals graphics

Original data set	Data set with outliers



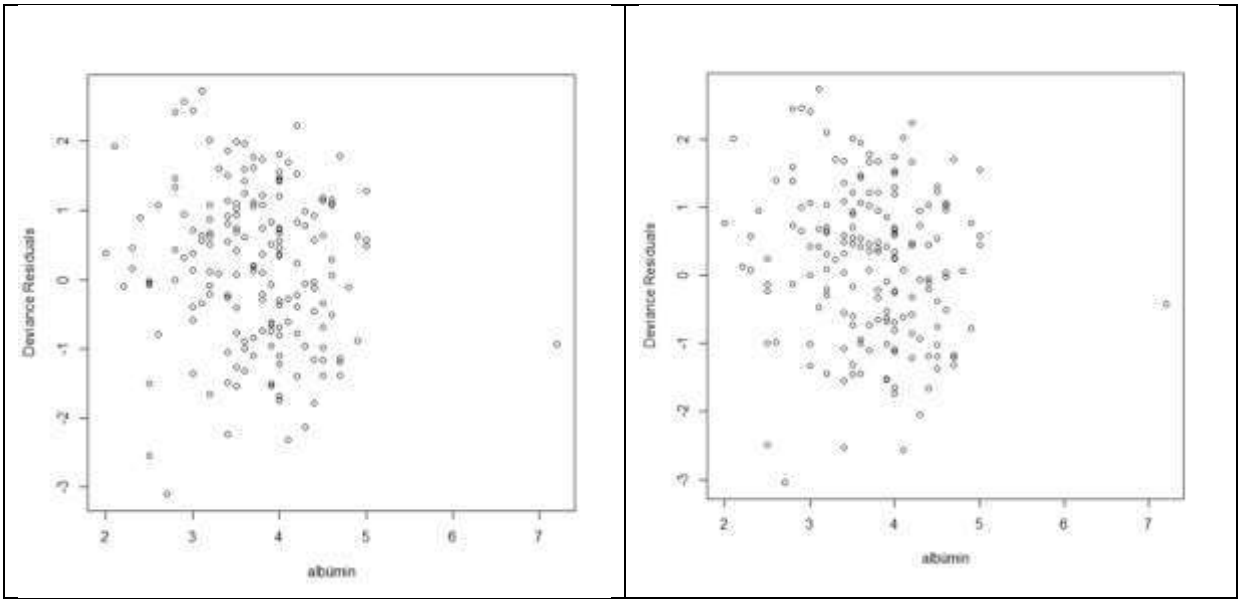
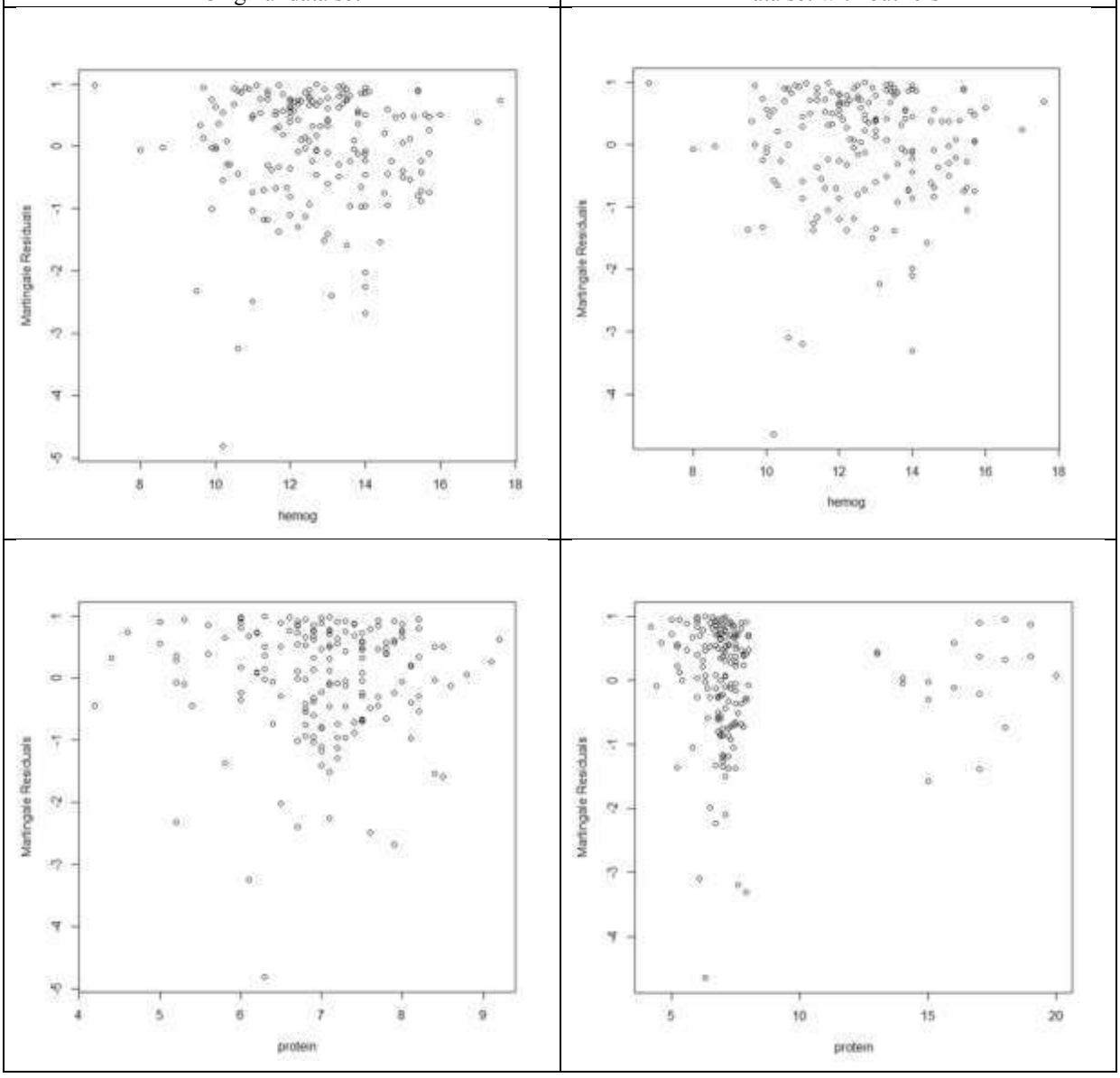


Figure 3: Deviance residuals graphic

Original data set	Data set with outliers



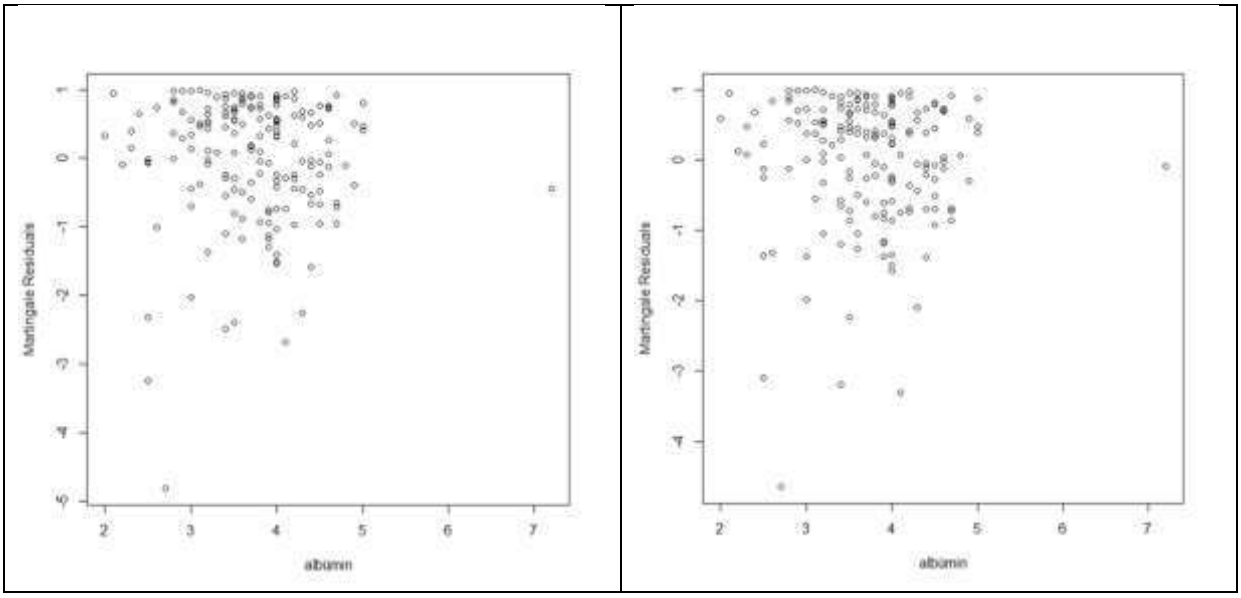


Figure 4: Martingale residuals graphic

5. CONCLUSION

In this study, Cox regression analysis, outliers and residual analysis methods are examined. Outliers that may cause a violation of the proportional hazard assumption, which is one of the most important assumptions of the Cox regression can have a great influence on the parameter estimation. In such a case, the presence of the outlier leads to inaccurate results. Residual analysis is very important for the detection of outliers and the control of assumptions. The most commonly residual analysis methods that used in the survival analysis have been examined as visually. Residual analyses are used for identification of different types of model inadequacies. Namely, Cox Snell for model conformity, Martingale for covariate conformity for fit model, Deviance for determination the outliers, Schoenfeld for control of the assumption are used. Since residual analysis yields graphical results, interpretation of the graph requires experience. In order to create a guide for inexperienced researchers, this study was designed. Each of these methods was first applied to 174 patients with lung cancer as original data. As result, graphs for the original data are obtained in which case the assumptions are provided and no outliers. Then these methods can be applied to the data set containing the outliers. Thus, it has been visually shown that different residual analysis method should be used for different purposes such as model adaptation, detecting the outliers and assumption control.

REFERENCES

1. *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. (in press)*
2. *Bednarski, T. (1989). On sensitivity of Cox's estimator. Statistics and Decisions 7, 215–228.*
3. *Cox, D.R., 1972. Regression models and life tables. Journal of the Royal Statistical Society, 34, 187-220.*
4. *Farcomeni, A. and Viviani, S. (2011). Robust estimation for the cox regression model based on trimming. Biometrical Journal, 53(6):956–973.*
5. *Gillespie, B., 2006. Checking Assumptions in The Cox Proportional Hazards Regression Model, Midwest SAS Users Group Dearnborn, Michigan.*
6. *Hawkins, D. M. (1980). Identification of Outliers. Chapman and Hall, London.*
7. *Hosmer D.W., & Lemeshow S. 1999. Applied survival analysis: regression modeling of time to event data. Wiley, John Wiley&Sons, Incorporated, Canada.*
8. *Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model, Biometrika , 69, 239-241.*
9. *Nardi, A., Schemper, M. 1999. New residuals for Cox regression and their application to outlier screening. Biometrics, Jun;55 (2): 523-9.*