Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 49 (2) (2020), 869 – 886
DOI : 10.15672/hujms.630402

Research Article

Visual research on the trustability of classical variable selection methods in Cox regression

Nihal Ata Tutkun*, Yasemin Kayhan Atilgan

Department of Statistics, Hacettepe University, Ankara, Turkey

Abstract

Multivariate models such as the Cox regression model, if developed carefully, are powerful tools for making prognostic prediction which are frequently used in studies of clinical outcomes. Many applications require a large number of variables to be modelled by using a relatively small patient sample. Determination of the important variables in a model is critical to understand the behaviour of phenomena as the independent variables contribute the most to the outcome. From a practical perspective, a small subset of independent variables are usually selected from a large data set without the loss of any predictive efficiency. Automatic variable selection algorithms in scientific studies are commonly used for obtaining interpretable and practically applicable models. However, the careless use of these methods may lead to statistical problems. The performance of the generated models may be poor due to the violation of assumption, omission of the important variables, problems of overfitting, and the problem of multicollinearity and outliers. In order to enhance the accuracy of a model, it is essential to explore the data and its main characteristics before making any statistical inference. This study suggests an approach for acquiring a trustworthy model selection procedure for survival data by performing classical variables selection methods, accompanied by a graphical visualization method, namely robust coplot. Thus, it enables us to investigate the discrimination of observations, clusters of the variables and clusters of the observations that are highly characterized by a particular variable in a one graph. We present an application of combined method, as an integral part of statistical modelling, on survival data on multiple myeloma to show how coplot results are used in automatic variable selection algorithm in Cox regression model-building.

Mathematics Subject Classification (2010). 62-07,62-09,62N01,62P10

Keywords. Cox regression model, graphical visualization,multidimensional scaling, robust coplot, variable selection

*Corresponding Author.
Email addresses: nihalata@hacettepe.edu.tr (N. A. Tutkun), ykayhan@hacettepe.edu.tr (Y. K. Atilgan)
Received: 07.10.2019; Accepted: 11.03.2020

1. Introduction

Many events of life, whether scientific, environmental or social have multiple and specific reasons and these reasons are usually connected to one another. A multivariate model, which is a statistical analysis tool, allows us to determine the relative contributions of different independent variables to an outcome. The strength of multivariate model is its ability to determine how multiple independent variables, which are connected to one another, influence an outcome. Clinical studies, in particular, are in need of a multivariate model because most researches have been done on a prognosis which is usually determined by a large number of variables [13, 15]. It is generally an unknown fact that variables are significantly connected to the outcome and thus, they should be included in the generated model. Researchers may often collect data from a large scale of candidate demographic and clinical variables for the purpose of an accurate determination of a subset of variables that explains the predicted variable best. Identification of the best subset among the many variables which will be included in a model so-called variable selection procedure is a critical part of building a model. If the sample size is not sufficiently large, the number of independent variables should be decreased in the analysis [15]. Even though an increase in the number of observations is more desirable, it may not be possible in the process of analysis. Besides, if two variables are highly correlated with each other, the model may not be reliable to assess the independent impact of each variable on the outcome. This problem is named as multicollinearity and it requires only one entry of the very highly related variables to the model [9]. In order to determine the correlation structure between variables, one may need to evaluate a correlation coefficient matrix with all the proposed independent variables. The problem with a correlation matrix is that it only evaluates the relationship between two variables, without the adjustment of the other variables [15]. As another approach, multiple bivariate comparisons can be performed. However, it requires too many couple comparisons and bivariate associations of two independent variables and it cannot reflect the simultaneous contribution of a number of independent variable to the outcome.

In the literature, several approaches are used for decreasing the number of independent variables. For example, one may exclude the variables that are uncorrelated with the outcome variable; one may combine two or more strongly/moderately correlated variables into a single variable; one may use variable selection algorithms to exclude the variables that have minimal impact on the outcome variable etc. However, these approaches might have some drawbacks [9, 15]. In the factor analysis, the number of independent variables can be reduced without omitting a variable. This method turns the cluster of variables into a factor and the original variables which are the major interest of medically oriented data analysis would be lost. Obtained factors may not be useful or interpretable for the clinician because one can measure the importance of a factor from the outcome but cannot measure the importance of any other variables from the outcome.

Automatic variable selection procedure is another approach for decreasing the number of independent variables in the analysis. This procedure helps us decide which independent variables will be included in a multivariate model and this model determines the minimum number of independent variable which are necessary to estimate the outcome accurately. Most of the statistical software packages present a diverse range of variable selection techniques such as backward, forward and stepwise selection. Automatic variable selection algorithms are available in any statistical software package and they are commonly used for obtaining interpretable models that are practically applicable in scientific studies [14]. However, the careless use of these methods may lead to statistical problems [22]. In these selection techniques, selection or deletion of variables continues to evaluate each variable for the way it improves the fit of the model. It is natural to think that these selection procedures eventually produce the same subset of variables. When the selection

algorithms present the same subset of variables to the researcher, it may be seen as a sign of a trustworthy model; however, this is not always the case. Moreover, highly correlated independent variables in the model, might make it impossible to solve the equation without deleting one or more variables from the analysis. Apart from the above mentioned multicollinearity problem particularly in clinical studies, continuous independent variables may have nonlinear relations with the outcome. Modelling a nonlinear relation with a linear model would not be desirable. Determining of such variables is also crucial while using the automatic variable selection algorithms, because the researcher can keep the variable in the model by a simple transformation and this variable may provide valuable information. Detection of the possible observation(s) that does not follow a similar pattern with the majority of the data is also another important issue in the modelling process since a model generated by the selection algorithms may be negatively influenced by the outliers. Lastly, determining the possible cluster(s) of observations and a variable that defines this cluster(s) can be also informative in modelling process. Due to the situations described above, researchers may require a simple pre-examination of the data to prevent such unexpected statistical problem before the application of variable selection techniques into the empirical multidimensional data.

The main contribution of the present study is to display the benefits that comes from combining the two methods that exist in both variable selection and multidimensional graphical representation fields. In order to build precise models that can explain the amount of predicted variable with minimum number of predictor variables without the disadvantages of outlier observations and multicollinearity problem, robust coplot which is a data visualization technique, is used as an auxiliary technique. Making a visual investigation of the multidimensional data by robust coplot method before having further statistical inferences provides the researcher information about the relations among all independent variables which are all taken from the outcome, the relation between each independent variable and the outcome variable, the cluster of observations, the cluster of variables and suspicious observation. Additionally, robust coplot presents the correlation coefficients of the variables which enables us to measure how strong the linear relations between variable, and the data are. If the correlation coefficient of an independent variable is low, this variable has little contribution to the model or has a non-linear correlation. As a result, it should be expected that the relevant variable should not be included in the Cox regression model.

Many multivariate statistical methods analyze the observations and the variables separately. In robust coplot method, clusters of variables, clusters of observations and the characterization of observations can be seen in one graph [3]. Among a wide spectrum of graphical techniques which are available for the management of multidimensional dataset, coplot method has attracted much more attention for various purposes from a wide range of areas in recent years [3]. Studies that focus on robust coplot - an approach developed for reducing the impact of outliers - and a more flexible software, called RobCoP, which can produce coplot and robust coplot graphs are also available in the literature [4]. Robust coplot method is specifically convenient for visualizing and interpreting clinical data set. In contrast to many other multivariate methods such as principal component, cluster and factor analysis which produce the composite of variables, coplot uses original variables, and representation and interpretation of the original variables and observations and these are more crucial and meaningful in clinical studies [6].

In life sciences, the data set may consist of many variables and the decision of which variables should be in the model poses as a difficult and confusing problem. Furthermore, the outliers or the multicollinearity problems negatively affect the choosing of the correct model. Conventional variable selection techniques based on information criterion such as AIC (Akaike [1]), BIC (Schwarz[23]), and Cp (Mallows[19]), are widely used for selecting an appropriate model. AIC and BIC are also used in survival analysis. The leading

researchers studying on information criterion are Tibshirani [25], Faraggi and Simon [11], Volinsky and Raftery [26], Fan and Li [10], Liang [17]. Although these criteria work well and are efficiently implemented in well-developed statistical softwares such as R and SAS [17] for existing models, for new-developed models they should be inferred theoretically and added to the package programs. It is useful for researchers to initially examine the data which is independent from the model. In this study, robust coplot analysis is used as a preliminary examination of the data before building one of the most popular survival models i.e. Cox regression model (CRM). The aim of this study is to compare the results acquired from using this approach with conventional variable selection methods in CRM.

The rest of the paper is organized as follows: In Section 2, robust coplot method and CRM are briefly explained. Multiple myeloma data set is described in Section 3. In addition to the results from robust coplot findings, the findings from conventional variable selection methods, and the comparison of the obtained results are discussed. The paper is concluded with discussions in Section 4.

2. Materials and methods

2.1. Cox regression model

The most common approach for modelling the effects of variable on survival is Cox regression model as it takes into account the effect of censored observations into account [8]. In survival analysis, regression models for survival data is traditionally based on CRM. The effect of the variables on survival acts multiplicatively on some unknown baseline hazard rate which makes it difficult to model variable effects as they change over time. Although the model is based on the assumption of proportional hazards, no particular form of probability distribution is assumed for the survival times. The model is therefore referred as a semi-parametric model [2].

The data based on a sample of size n, consists of (t_i, d_i, x_i), $i = 1, ..., n$ where t_i is the time on study for the i^{th} individual, d_i is the event indicator (d_i=1 if the event has occurred and d_i=0 if the lifetime is censored) and x_i is the vector of variables for the i^{th} individual. Hazard function for CRM is given by

$$h_i(t) = h_0(t) \exp(\beta' x_i)$$

where h_0(t) is the baseline hazard function and β is a $p \times 1$ vector of unknown parameters [7]. The ordered death times are denoted by $t_1 < \ldots < t_k$ and the set of individuals who are at risk at the time t_j are denoted by R(t_j), so R(t_j) is the set of individuals who are alive and uncensored at a time prior to t_j. Then, the likelihood for CRM is given by

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta' x_j)}{\sum_{\ell \in R(t_j)} \exp(\beta' x_\ell)}$$

where x_j is the vector of variables for the individual who dies at the j^{th} ordered death time, t_j [7]. Newton-Raphson iteration is most commonly used algorithm for the estimation of regression coefficients.

Regression coefficients (or transforms thereof such as $exp(.)$) are easily interpretable and this makes CRM popular in life sciences. The assumption of the model is linearity, that is expected outcome value is thought to be modeled by a linear combination of independent variables, and additivity, that is the effects of the independent variables can be added [1]. In a setting with several independent variables, the fundamental interpretation of a regression coefficient β_j in a linear predictor model is the expected change in outcome (or log odds or log hazard) if X_j changes by one unit and all other variables are held constant. Consequently, β_j measures the conditional effect of X_j. A single true model and

the correct model specification can be rarely assumed. This implies that the interpretation of β_j changes if the set of independent variables in the model changes and X_j is correlated with the other independent variables [14].

When there is few amount of independent variables in data set, building and developing a model is much more easier. However, in routine work, which variables should be included in a model is priorly not known and we often have the candidate variables within the range of 10-30. This number is often too large to be considered in a statistical model [14]. Also, the number of possible models that are required to be fitted is computationally time consuming.

In multivariate Cox regression modelling, the selection of variables and the fit of final model is very important. Since all comments are made according to the final model and the assessment of obtained results are crucial in life sciences. There are numerous variable selection methods based on significance and/or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof [14].

In practical applications, the first and most common applicable approach is the use of automatic routines based on forward selection, backward elimination or stepwise procedures for variable selection that are used. Collett [7]recommends using a likelihood ratio test for all variable inclusion/exclusion decisions. Iterated testing between the models yields forward selection (FS) or backward elimination (BE) variable selection algorithms, depending on whether one starts with an empty model or with a model that all independent variables are considered upfront. Most statisticians prefer backward elimination (BE) over forward selection (FS), especially when collinearity is present [20]. However, when models can become complex, for example in the context of highdimensional data, then FS is still possible[1]. The second easily applicable approach is the use of significance criteria which are applied to include or exclude independent variables from a model and select a model from a set of plausible models.

2.2. Robust coplot method

Typically, clinical studies require the analysis of multidimensional data which include many clinical, demographic, socioeconomic, and interested outcome variables collected from patients. Robust coplot graph, based on two superimposed graphs, is a simple picture of a multidimensional data set. The first graph shows the embedding of n observations into two- dimensional space. This representation conserves relative distance between the observations which means two observations that are close to each other in p dimensions are embedding closely in two dimensions. The second graph consists of p vectors that represent the variables, and reflects the relations among the variables. This method provides a simultaneous investigation of the relationship patterns between both observations and variables in a dataset. When the data contain outliers, obtained results from robust coplot are unaffected by these outliers. Robust coplot output is mainly generated with three steps [3].

At the first step, for the purpose of treating the different scale variable equally, the data matrix X_{nxp} is normalized into Z_{nxp}. The elements of standardized data matrix are deviations from column median, med(.) which is divided by their median absolute deviation value, $MAD(x_j) = 1.4826med\left(|x_j - med\left(x_j\right)|\right)$ as follows:

$$z_{ij} = \frac{x_{ij} - med\left(x_j\right)}{MAD\left(x_j\right)}$$

where z_{ij} is the i^{th} row and j^{th} column element of the standardized matrix Z_{nxp}, x_j is the j^{th} column of the data matrix X_{nxp}. After the data matrix is standardized in a robust way, multidimensional scaling embeddings of p-dimensional n observations into two-dimensional space are determined at the second step.

At the second step, multidimensional embedding of the data set are formed. Robust multidimensional scaling (RMDS) is used for visualing dispersion of n observations into two-dimensional space [12]. RMDS uses the outlier aware cost function given in the following equation;

$$f(O,Y) = \sum_{i<j} [\delta_{ij} - d_{ij}(Y) - o_{ij}]^2 + \lambda \sum_{i<j} |o_{ij}|$$

where, δ_{ij} is the dissimilarity metric among i^{th} and j^{th} row of the Z_{nxp}, Y_{nx2} is the coordinate matrix for two-dimensional space, i^{th} row, j^{th} column element of the outlier matrix, O, is $o_{ij} = sgn(d_{ij} - d_{ij}(Y))\, max(0, |d_{ij} - d_{ij}(Y)| - 1/2)$ which defines the outlier variable, and $\lambda > 0$ is the parameter that controls the number of presumed outliers in the dissimilarity matrix[3]. Kruskall stress value,(σ), is used as a measure for deciding how good the fit of the configuration of n observations obtained by RMDS is.

At the last step, p vectors are drawn on the graph which is obtained in the second step. Each variable is denoted by a vector acquired from the center of gravity of the n observations. The direction and the magnitude of the vector are determined by the median absolute deviation correlation coefficient (MADCC) robust against the outliers [24]. Direction of the each vector is chosen in a way that the correlation between the original values of the corresponding variable and their projections on the chosen vector at maximum value. The degree that shows how good a vectors fit is assessed by correlation value. A decision should be made to keep or delete the variables that do not fit the graphical representation, in other words, variables that have low correlation values before further statistical analysis. The magnitude of the vector is proportional to the evaluated correlation value. Additionally, observations with high value in this vector are located in the graph where the vector points to. MADCC is defined as follows:

$$\rho_{j,MADCC} = \frac{MAD^2(u_j) - MAD^2(k_j)}{MAD^2(u_j) + MAD^2(k_j)}$$

where, u_j and k_j are the robust principal variables which are defined as follows:

where, z_j is the j^{th} column of Z_{nxp}, and v_j represents the projection values of all points in the MDS graph on the jth variable vector for a specific direction.

In the robust coplot representation, observations are colorized according to the selected categorical variable in order to understand the available observations discrimination better. The outcome of the robust coplot analysis gives the following inferences about multidimensional data. Two highly correlated variables are represented by two vectors that are close in the same direction, and if the correlation of the variables is negative, the corresponding vectors will lie in opposite directions. Two uncorrelated variables are represented by two vectors which are perpendicular to each other. Observations which are highly characterized by a specific variable are embedded close to each other and this mass is placed in the same direction as the variables vector. Possible outlier observation(s) is embedded far from the mass of observations [24].

3. Results and discussion

3.1. A descriptive look at the dataset

In this study, multiple myeloma data set of Krall et al. [16] is used for illustration. Krall, et. al. [16] analyzed the data from a study on multiple myeloma and the study researches 65 patients which are treated with alkylating agents. Out of 65 patients 48 patients died in the study and 17 of them survived. The variable time represents the survival time (survtime) in months, started from the time of diagnosis. The censoring variable consists of two values 0 and 1, that indicate whether the patient is alive or dead, respectively. Therefore, the censoring rate is 26% and the type of censoring is right.

Table 1. Descriptive statistics .

	X01	**X02**	**X05**	**X07**	**X09**	**X10**	**X11**	**X12**	**X14**	**X15**	**X16**
Mean	1.39	10.2	60.15	3.76	1.55	6.74	30.35	3.62	8.6	5.22	10.12
Std. Error of Mean	0.039	0.317	1.282	0.03	0.045	0.781	2.485	0.746	0.279	0.272	0.225
Std. Deviation	0.313	2.558	10.334	0.242	0.364	6.3	20.032	6.012	2.249	2.19	1.816
Skewness	0.872	-0.304	0.062	2.156	-0.792	0.925	-0.332	2.337	0.834	0.913	1.928
Std. Error of Skewness	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297
Kurtosis	0.436	-0.501	-0.642	9.115	0.202	0.239	-1.218	5.114	1.778	0.524	5.486
Std. Error of Kurtosis	0.586	0.586	0.586	0.586	0.586	0.586	0.586	0.586	0.586	0.586	0.586
Minimum	1	5	38	3	0.477	0	0	0	4	2	7
Maximum	2	15	82	5	2	24	68	27	17	12	18
Q1	1.15	8.8	51	3.63	1.35	1	10.5	0	7	3	9
Q2	1.32	10.2	60	3.73	1.62	6	35	1	9	5	10
Q3	1.58	12.05	67.5	3.87	1.85	10	49.5	4	10	7	10.5

The data about myeloma consists of 16 demographic and clinical variables which are as follows: X01: Log BUN at diagnosis, X02: Hemoglobin at diagnosis, X03: Platelets at diagnosis (0: abnormal, 1: normal), X04: Infections at diagnosis (0: none, 1: present), X05: Age at diagnosis (complete years), X06: Sex (1: male, 2: female), X07: Log WBC at diagnosis, X08: Fractures at diagnosis (0: none, 1: present), X09: Log ZBII at diagnosis (log % of plasma cells in bone marrow), X10: % Lyiaphocytes in peripieral blood at diagnosis, X11: % Myeloid cells in peripheral blood at diagnosis, X12: Proteinuria at diagnosis, X13: Bence Jone protein in urine at diagnosis(1: present, 2: none), X14: Total serum protein at diagnosis, X15: Serum globin (gm%) at diagnosis, X16: Serum calcium (mgm%) at diagnosis.

We use 11 variables which are continuous, survival time and censoring variable. Robust coplot representation of categorical variables are not meaningful. When working with categorical variables, standardization of data matrix with the median and MAD estimators may result in an undefined Z matrix. Additionally, PCC and MADCC are the correlation coefficients that measures the relation between two continuous variables. Thus, categorical variables are used as color coded variables and the results are presented in Figure 3 (a) and (b). The use of categorical variables by means of color coding is quite helpful in identifying clusters of observations and possible outliers. Since categorical variables are quite common in clinical studies, one may want to include these variables in the Cox regression model as independent variables. In such cases, the researcher may use the classical coplot method instead of robust coplot [6].

Descriptive statistics for the variables used in this application are given in Table 1.

3.2. Robust coplot findings

Robust coplot method is used for revealing the relations among a set of variables in order to have an idea about the variables which are the backbone of CRM. Before starting the classical variable selection procedures, robust coplot graph is used for deciding possible variables to be eliminated. Robust coplot software, RobCop, which is publicly available (see Atilgan [24] for details about software), is used for visualing myeloma data and exploring potential associations.

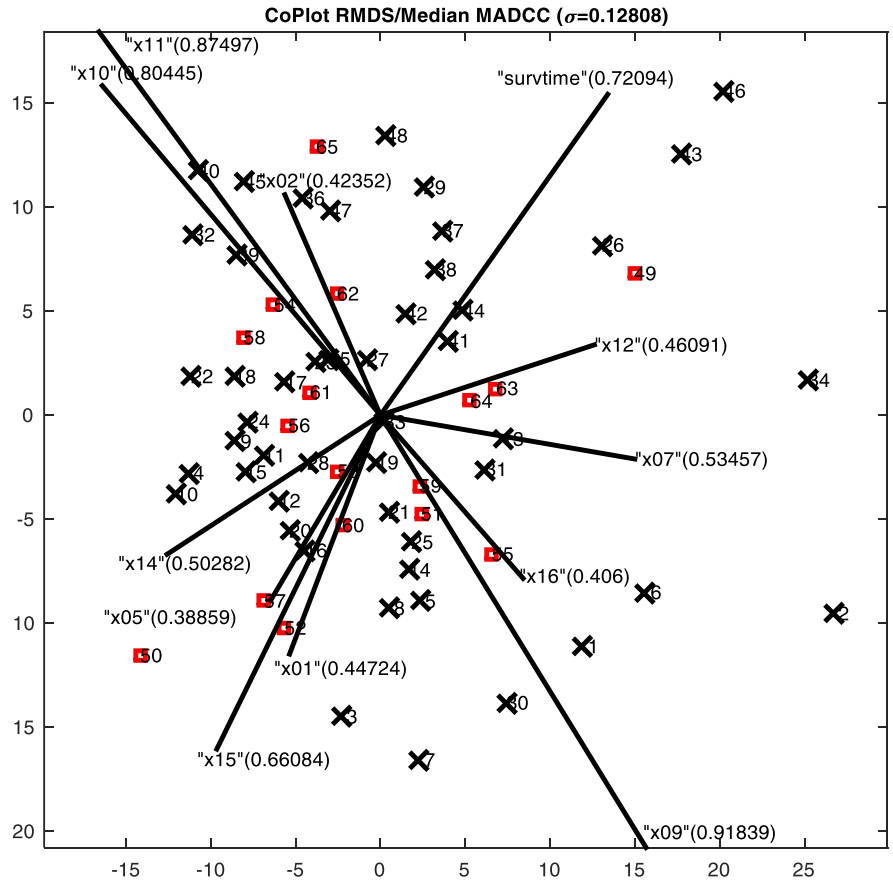


Figure 1. Initial robust coplot map of myeloma data

Figure 1 is the robust coplot graph of myeloma data and it includes all variables in the data set. Each vector in Figure 1 corresponds to a variable. Observations, since an extreme outlier, censoring value, in order to display the spread of observations on the reduced two-dimensional space. Failed observations are colored as the black cross, while censored observations as the red square.

The Kruskall stress value of MDS fit is found as 0.128 which is close to fair fit, and Figure 1 is available for interpretation. MADCC correlation values of each variables are seen in parenthetical in Figure 1.

X01, X02, X05, X12, and X16 are found to be low correlated variables (evaluated correlation coefficient values are lower than 0.50). Removal of a variable from the data set requires a redraw of robust coplot graph because this procedure affects the previous steps. Different variable combinations produce different graphs. Therefore, one-by-one subtraction is performed instead of subtracting all of the variables at once. All combinations of eliminated variable(s) are tried. This enables us to decide the redundant variables, before starting the model building process for myeloma data.

Figure 1 helps us to find the variables that have high positive or negative correlations with the other variables. For example, X10 and X11 are highly positively variables, and X02 is also moderately positively correlated with these variables whereas X09 and X11 are highly negatively correlated variables. Survtime and X01, X05, and X15 are in high negative correlation. Survtime and X12 are in high positive correlation. The angles between the survtime and X10, X11 and X16 are close to 90 degree, it is implicated that

Table 2. Angles between the 12 variables in myeloma data.

	X01	**X02**	**X05**	**X07**	**X09**	**X10**	**X11**	**X12**	**X14**	**X15**	**X16**	**Survtime**
X01	0											
X02	127	0										
X05	11	116	0									
X07	107	126	118	0								
X09	62	171	73	45	0							
X10	109	18	98	144	171	0						
X11	113	14	102	140	175	4	0					
X12	130	103	141	23	68	121	117	0				
X14	37	90	26	144	99	72	76	167	0			
X15	6	121	5	113	68	103	107	136	31	0		
X16	72	161	83	35	10	179	175	58	109	78	0	
Survtime	164	69	175	57	102	87	83	34	159	170	92	0

survtime and these variables to be uncorrelated. The angles of the variable vectors which are relative to the other variable vectors are given in Table 2. It is obvious that the angle between the two vectors is an indicator of the correlations between their corresponding variables.

After all variables are used for generating the robust coplot graph, we run the robust coplot method for myeloma data set in several times. Low correlation variables are removed from the analysis one at a time, and we run the method on different combinations on the remaining data to see which variables are stable. It is found out that some of the low correlation variables consistently have low correlations. Instead of giving all of the steps, we present the removed variables in this paper.

Four variables had correlations below 0.50, X12, X16, X07, and X05, have been removed from the analysis one at a time, respectively. Removing these variables had no major effect on previous robust coplot findings, namely, the association patterns among the remaining variables which have high MADCC values, but the correlation values of remaining variables have increased. Figure 2 presents the robust coplot graph of reduced data set.

The Kruskall stress value of this graph is 0.117, and the correlations of remaining variables are higher than 0.60. These are usually considered as acceptable goodness-of-fit values [3].

Figure 2 demonstrates that the correlations between X01, X14 and X15 variables are positive and high. This variable cluster is highly negatively correlated with survtime. It may be assumed that the variables that grow together are nearly duplicated to provide information for making inference about survtime. The variables X10 and X11 are positively highly correlated variables, and these variables are nearly ortogonal to the variable survtime. Additionally, variable X09 is also nearly ortogonal to the variable survtime. The contents of information in the uncorrelated variables with survtime explain the variation in the amount of survtime which is aspected as low. Since robust coplot analyzes observations and variables simultaneously, clusters of observations and their nature are able to be seen. Extreme outlier(s) does not appear in myeloma data set, and survived and non-survived patients are not clustered. The projection of a point on a vector should be proportional to its distance from the corresponding variable's average, where higher than the average is in the direction of the vector and vice versa. For instance, observation 50 have higher values in X01, X14, and X15, but small value in survtime.

Two models are inferred from myeloma data with the help of robust coplot findings. First, all variables are considered for constituting a CRM. Afterwards, reduced data set are

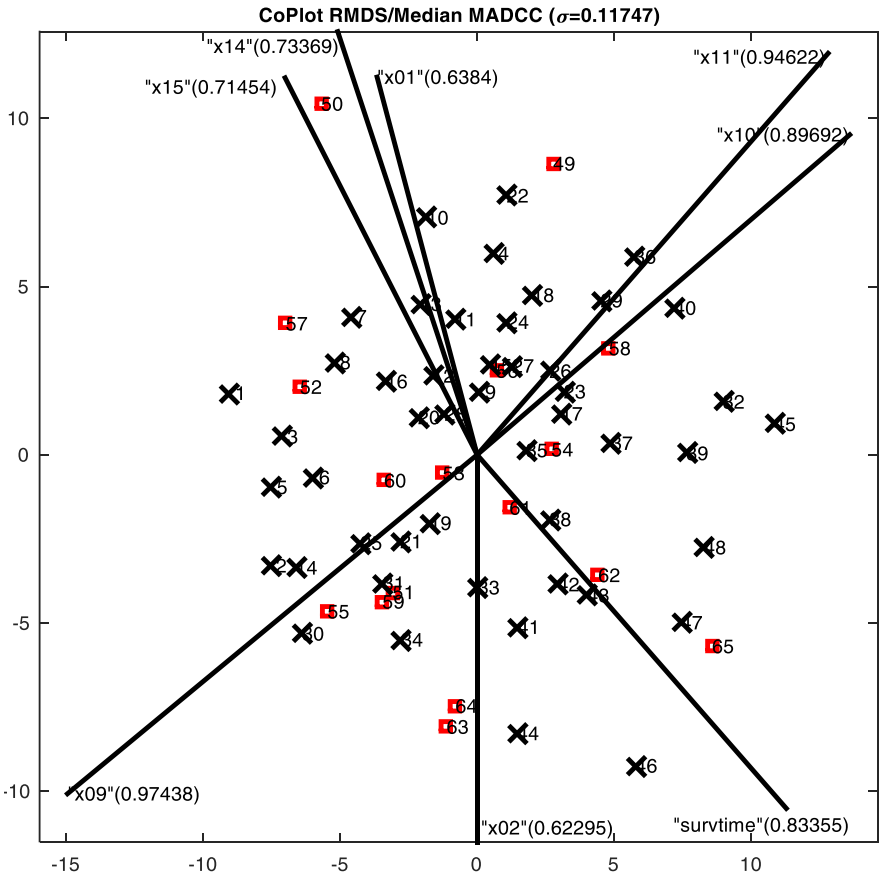


Figure 2. Robust coplot map of reduced myeloma data

considered for building a CRM. These two processes are compared and we have achieved the following findings; Variables which have low MADCC values are eliminated from the CRM. Variables which are orthogonal with the survtime are eliminated from the CRM. One of the variables that are highly correlated with each other is eliminated from the CRM during the variable selection procedures. Furthermore, the final models inferred from the full data set and the reduced data set are the same.

3.3. Cox regression model findings

First of all, univariate CRM is applied to the data set for the purpose of seeing the univariate effects of the variables on survival. The obtained results are given in Table 3

The results show that the Log BUN (X01) and Hemoglobin (X02) at diagnosis are important variables as they affect the risk of death whereas the others are not significiant at a 95% confidence level. The estimated hazard of Log BUN is 5.912 that means, the risk of death increases a 5.912 unit for 1-unit increase in log BUN. There is a 0.117% decrease in risk of death for one unit increase in Hemoglobin at diagnosis (HR=0.883).

Then the two scenarios are set up according to the results of robust coplot analysis. First scenario examines CRM with variables X01, X02, X05, X07, X09, X10, X11, X12, X14, X15, and X16. Second scenario builds the full model with variables X01, X02, X09, X10, X11, X12, X14, and X15.

Table 3. The results of univariate Cox regression model

	Hazard Ratio	**Std. Error**	**95% Confidence Interval**		**z**	**P>|z|**
X01	5.912	3.583	1.803	19.391	2.93	0.003
X02	0.883	0.049	0.791	0.985	-2.22	0.026
X05	0.998	0.016	0.967	1.03	-0.11	0.912
X07	2.415	1.756	0.581	10.04	1.21	0.225
X09	1.39	0.602	0.594	3.25	0.76	0.448
X10	0.983	0.023	0.939	1.029	-0.74	0.462
X11	0.997	0.008	0.982	1.013	-0.34	0.736
X12	1.008	0.021	0.968	1.05	0.38	0.705
X14	1.093	0.074	0.957	1.248	1.32	0.188
X15	1.067	0.076	0.927	1.227	0.9	0.368
X16	1.11	0.112	0.912	1.352	1.04	0.298

Table 4. The summary of multivariate Cox regression model

	Cox regresion model	**Variables in the model**	**AIC**	**BIC**
Scenario 1	Full model	X01 X02 X05 X07 X09 X10 X11 X12 X14 X15 X16	310.3952	334.3134
	Stepwise selection	X01 X02	301.3336	305.6824
	Forward selection	X01 X02	301.3336	305.6824
	Backward selection	X01 X02	301.3336	305.6824
Scenario 2	Full model	X01 X02 X09 X10 X11 X14 X15	305.5589	305.5589
	Stepwise selection	X01 X02	301.3336	305.6824
	Forward selection	X01 X02	301.3336	305.6824
	Backward selection	X01 X02	301.3336	305.6824

We have run the full model and the models with variable selection procedures. The model selection criteria are given in Table 4. The results of multivariate CRM without the variable selection are given in Table 5 and the results of multivariate CRM with the variable selection are given in Table 6.

For Scenario 1, the model selection criteria lead us to use the variables X01 and X02 for the final model, and these results are concordant with the univariate CRMs. For this scenario, the variable selection for the data set is consistent with robust coplot findings.

For Scenario 2, X01 and X02 are statistically significant in univariate and multivariate CRM without variable selection. Additionally the variable selection procedures lead us to use the variables X01 and X02 in the final model. In summary, the results of second scenario are same with the first scenario and robust coplot findings in terms of variable selection.

Consequently, regarding the myeloma data set, Log BUN (X01) and Hemoglobin (X02) at diagnosis are determined as important variables which affect the risk of death. The estimated hazard of Log BUN is 5.474 that means the risk of death increases a 5.912 unit (HR=exp(1.7)=5.474) for a 1-unit increase in log BUN. There is a 0.11% decrease in risk of death for a one unit increase in Hemoglobin at diagnosis (HR=exp(-0.118)=0.89).

Table 5. The results of multivariate Cox regression model without variable selection

		Coef.	Std. Error	95% Conf. Interval		Hazard Ratio	Std.	95% Conf. Interval		z	P>|z|
Scenario 1	X01	1.741	0.676	0.415	3.067	5.703	3.857	1.515	21.467	2.57	0.01
	X02	-0.181	0.068	-0.314	-0.048	0.835	0.057	0.731	0.953	-2.66	0.008
	X05	-0.008	0.021	-0.049	0.034	0.992	0.021	0.952	1.034	-0.37	0.715
	X07	0.773	0.713	-0.624	2.17	2.167	1.545	0.536	8.761	1.09	0.278
	X09	0.293	0.875	-1.422	2.009	1.341	1.174	0.241	7.456	0.34	0.737
	X10	-0.042	0.035	-0.111	0.027	0.959	0.034	0.895	1.027	-1.2	0.23
	X11	0.005	0.017	-0.027	0.038	1.005	0.017	0.973	1.039	0.32	0.748
	X12	0.032	0.029	-0.024	0.089	1.033	0.03	0.976	1.093	1.12	0.262
	X14	0.288	0.168	-0.041	0.616	1.333	0.223	0.96	1.852	1.72	0.086
	X15	-0.18	0.154	-0.481	0.121	0.835	0.128	0.618	1.129	-1.17	0.242
	X16	0.027	0.13	-0.228	0.281	1.027	0.133	0.796	1.325	0.2	0.838
Scenario 2	X01	1.819	0.658	0.529	3.109	6.166	4.058	1.697	22.397	2.76	0.006
	X02	-0.158	0.065	-0.285	-0.031	0.854	0.055	0.752	0.969	-2.45	0.014
	X09	0.116	0.818	-1.487	1.718	1.123	0.918	0.226	5.575	0.14	0.887
	X10	-0.025	0.031	-0.087	0.037	0.975	0.031	0.917	1.037	-0.79	0.428
	X11	-0.005	0.014	-0.033	0.024	0.995	0.014	0.967	1.024	-0.33	0.743
	X14	0.287	0.142	0.01	0.565	1.333	0.189	1.01	1.759	2.03	0.042
	X15	-0.209	0.143	-0.49	0.072	0.811	0.116	0.613	1.074	-1.46	0.144

Table 6. The results of multivariate Cox regression model with variable selection

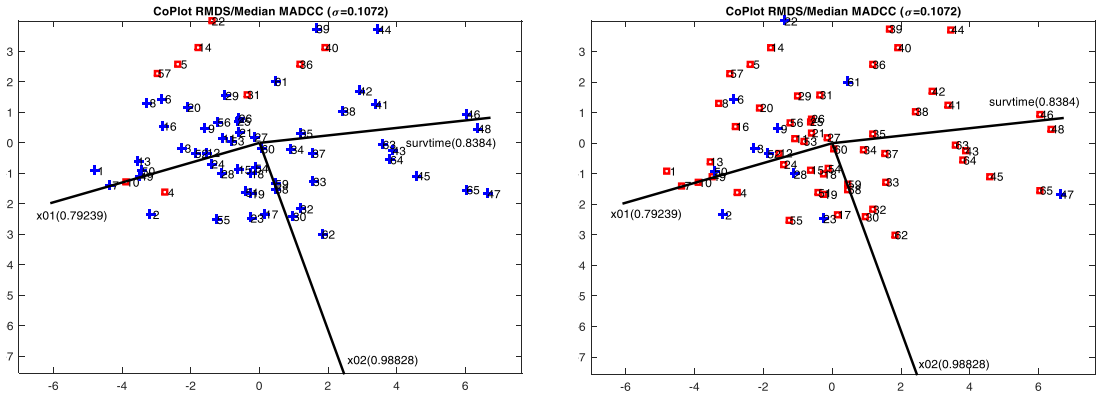
		Coef.	Std. Error	95% Conf. Interval		z	P>|z|
Scenario 1	X01	1.7	0.613	0.498	2.901	2.77	0.006
	X02	-0.118	0.058	-0.231	-0.005	-2.05	0.041
Scenario 2	X01	1.7	0.613	0.498	2.901	2.77	0.006
	X02	-0.118	0.058	-0.231	-0.005	-2.05	0.041

3.4. Findings from color coding variables

Robust coplot graphs of obtained model (categorical variables; Censoring variable, X03, X04, X06, X13) are drawn according to different color coded variables. The purpose of this is in doing so was to reveal possible clusters of observations that share common characteristics in groups.

Colorization of observations based on variable, censoring variable, sex and Bence Jone protein in urine at diagnosis do not form cluster patterns, obtained graphs are given in Supplementary Figure 3(c-d-e).

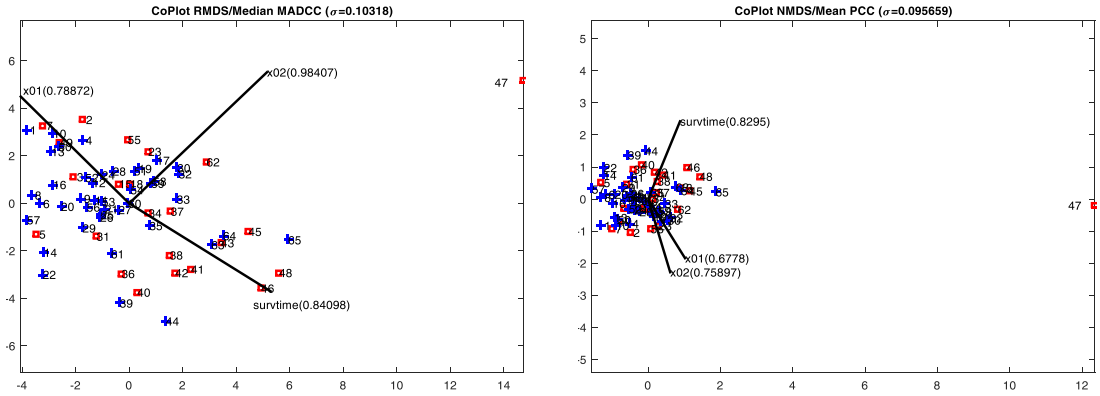
However, it can be easily seen from Figure 3(a) that patients whose Platelets at diagnosis are abnormal have lower Hemoglobin values. For example, hemoglobin values of observations 22 and 14 are 5.5 and 5.1 respectively. Log BUN values of observations 10 and 4 are high and these two observations are embedded in the same direction as the X01 vector. Observation 40 has the highest survtime value among the patients whose Platelets at diagnosis are abnormal.



(a) Platelets at diagnos is color coded variable red square:0, blue plus:1

(b) Infections at diagnosis is color coded variable red square:0, blue plus:1

Figure 3. Colorization of observations



(a) Robust Coplot Map

(b) Classical Coplot Map

Figure 4. Coplot analysis of contaminated (47) data set

A visual inspection of Figure 3(b) displays that patients who have infections at diagnosis have lower survival time. However, observation 47 does not follow this manner. A separate examination of this patients conditions might be beneficial.

It is mentioned before that when fitting a CRM, one single outlier is enough to make the estimator take values arbitrarily far from their true value, and lack of robustness of CRM is widely discussed in the literature [5, 18, 21]. Identifying and removing outlying observations is crucial for providing more accurate relations between variables and survival time. An emphasis on even one single strong outlier, in classical parameter estimates, might poorly effect the classical variables selection algorithms. For instance, the Log BUN value of observation 47 is incorrectly typed as 4.3222 instead of 1.3222 and thus, obtained data set is called as contaminated data set. Graphs of robust coplot and classical coplot are given in Figure 4(a) and (b) respectively.

The classical coplot produces a degenerated graphical representation of the multidimensional data set when the underlying dataset contains one strong outlier. A single

observation poorly affects the associations between the variables, and it may lead to obtain slightly different coefficient estimates. On the other hand, robust coplot graph preserves the correct relations between variables and survival time, and indicated suspicious observation for further examinations.

CRMs are applied for contaminated data set. The renewed results are given in Supplementary Table (7-8-9). The size of hazard for Log BUN (X01) severely decreases for the both scenarios, and it also becomes insignificant. The selection methods suggest a final model that contains only Hemoglobin (X02). There is a 0.12% decrease in the risk of death for one unit increase in Hemoglobin at diagnosis (HR=exp(-0.124)=0.88). The variable selection methods propose the same model that robust coplot findings suggest.

4. Conclusion

In model building studies, the researcher encounters with a large number of variables and deals with the question of which of these variables should be included in a model. The answer cannot be known immediately because the observed variables are often highly correlated with the interested outcome variable. Thus, another question of whether there is any need to put all these variables in the model building process comes up. A common approach for reducing the number of candidate variables is to apply firstly data reduction techniques such as principal component analysis for the purpose of defining smaller set of uncorrelated variables and then, automatic applying automatic selection procedures for the purpose of reducing the risk of overfitting and multicollinearity. However, in medical studies, interpretation of component variables may be unreasonable. Consequently, a more preferable approach is to rely on classical variable selection methods, expert knowledge and clinical judgment during the process of deciding on primarily important variables that must be included in the model. In the case of absent or limited of expert knowledge, visual representation tools of multivariate data are very useful for better understanding the underlying structure of the data in an unsupervised way.

This manuscript presents the results of application of two statistical methods to analyze survival data and shows the usefulness of the adapted approach in the context of Cox regression model. Robust coplot outcome gives pragmatic recommendations on duplicated variables for the researchers, low correlated variables, and the variables which have no relations with the predicted variable. Due to the multidimensional and complex nature of the survival data, identifying outlier(s) is not an easy task. However, considering the contaminated dataset example, it is illustrated that preliminary examination of the data set by robust coplot leads to the identification of suspicious observations. Although the methods are standard methods; their combination is new and has potential advantages over the classical ways of analyzing such data. However, as a future work, to explicitly show the usefulness of the approach a simulation experiment should be conducted for different censoring types and also censoring rates.

Acknowledgment. We would like to thank the referees for their valuable reviews and highly appreciate the comments and suggestions, which significantly contributed to improving the quality of the publication.

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control AC **19**, 716-723, 1974.

[2] N. Ata and M.T. Sozer, *Cox regression models with nonproportional hazards applied to lung cancer survival data*, Hacet. J. Math. Stat. **36** (2), 157-167, 2007.

[3] Y.K. Atilgan, *Robust coplot analysis*, Comm. Statist. Simulation Comput. **45** (5), 1763-1775, 2016.

[4] Y.K. Atilgan and E.L. Atilgan, *RobCoP: A Matlab Package for Robust CoPlot Analysis*, Open Journal of Statistics **7**, 23-35, 2017.
[5] T. Bednarski, *On sensitivity of Coxs estimator*, Statistics and Decisions **7**, 215-228, 1989.
[6] D.M. Bravata, K.G. Shojania, I. Oklin and A. Raveh *A tool for visualizing multivariate data in medicine*, Stat. Med. **27** (12), 2234-2247, 2007.
[7] D. Collett,*Modeling Survival Data in Medical Research*, 2nd Ed. New York: Chapman @ Hall/ CRS A CRC Press Company, 2003.
[8] D.R.Cox, *Regression Models and Life Tables*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **34** (2), 187-220, 1972.
[9] S. Derksen and H.J. Keselman, *Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables*, Brit. J. Math. Stat. Psy. **45** (2), 265-282, 1992.
[10] J. Fan and R. Li, *Variable selection for Cox's proportional hazards model and frailty model*, Ann. Statist. **3**, 74-99, 2002.
[11] D. Faraggi and R. Simon, *Bayesian variable selection method for censored survival data*, Biometrics **54**, 1475-1485, 1998.
[12] P.A. Forero and G.B. Giannakis, *Robust multi-dimensional scaling via outlier sparsity control*, Robust multi-dimensional scaling via outlier sparsity control, 1183-1187, 2011.
[13] Jr F. Harrell and K.L. Lee, *Regression Modelling Strategies for Improved Prognostic Prediction*, Stat. Med. **3**, 143-152, 1984.
[14] G. Heinze, C. Wallisch and D. Dunkler, *Variable selection - A review and recommendations for the practicing statistician*, Biom J. **60** (3), 431-449, 2018.
[15] M.H. Katz,*Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers*, Third Edition, Cambridge University Press, New York, 2011.
[16] J.M. Krall, V.A. Uthoff and J.B. Harley, *A step-up procedure for selecting variables associated with survival*, Biometrics **31**, 49-57, 1975.
[17] H. Liang, and G. Zou, *Improved AIC selection strategy for survival analysis*, Comput. Statist. Data Anal. **52** (5), 2538-2548, 2008.
[18] A. Nardi and M. Schemper, *New residuals for Cox regression and their application to outlierscreening*, Biometrics **55**, 523-529, 1999.
[19] C.L. Mallows Nardi and M. Schemper, *Some comments on Cp*, Technometrics **15**, 661-675, 1973.
[20] N. Mantel, *Why stepdown procedures in variable selection*, Technometrics **12**, 621-625, 1970.
[21] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, New York: Wiley Interscience, 1987.
[22] K.L. Sainani, *Multivariate regression: The pitfalls of automated variable selection*, Am. J. Phys. Med. Rehabil. **5**, 791-794, 2013.
[23] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6**, 461-464, 1978.
[24] G. Shevlyakov and P. Smirnov, *Robust estimation of the correlation coefficient: an attempt of survey*, Austrian J. Stat. **40**, 147-156, 2011.
[25] R. Tibshirani, *The lasso method for variable selection in the Cox model*, Stat. Med. **16**, 385-395, 1997.
[26] C.T. Volinsky and A.E. Raftery, *Bayesian information criterion for censored survival models*, Biometrics **56**, 256-262, 2000.

SUPPLEMENTARY MATERIAL

Table 7. The summary of multivariate Cox regression model for the contaminated data set

	Cox regresion model	Variables in the model	AIC	BIC
Scenario 1	Full model	X01 X02 X05 X07 X09 X10 X11 X12 X14 X15 X16	316.1785	340.0968
	Stepwise selection	X02	306.5977	308.7721
	Forward selection	X02	306.5977	308.7721
	Backward selection	X02	306.5977	308.7721
Scenario 2	Full model	X01 X02 X09 X10 X11 X14 X15	312.857	328.0777
	Stepwise selection	X02	306.5977	308.7721
	Forward selection	X02	306.5977	308.7721
	Backward selection	X02	306.5977	308.7721

Table 8. The results of multivariate Cox regression model without a variable selection for the contaminated data set

		Coef.	Std. Error	95% Conf. Interval		Hazard Ratio	Std. Error	95% Conf. Interval		z	P>|z|
Scenario 1	**X01**	0.242	0.278	-0.302	0.787	1.274	0.354	0.739	2.196	0.87	0.383
	X02	-0.209	0.07	-0.347	-0.071	0.811	0.057	0.707	0.931	-2.97	0.003
	X05	0	0.021	-0.041	0.04	1	0.021	0.96	1.041	-0.02	0.987
	X07	1.28	0.757	-0.204	2.763	3.595	2.721	0.816	15.85	1.69	0.091
	X09	0.431	0.857	-1.249	2.111	1.539	1.319	0.287	8.254	0.5	0.615
	X10	-0.042	0.036	-0.112	0.028	0.959	0.034	0.894	1.028	-1.18	0.238
	X11	0.014	0.017	-0.019	0.047	1.014	0.017	0.981	1.048	0.83	0.408
	X12	0.037	0.029	-0.019	0.093	1.038	0.03	0.981	1.097	1.29	0.196
	X14	0.288	0.172	-0.049	0.625	1.334	0.229	0.952	1.869	1.67	0.094
	X15	-0.201	0.166	-0.527	0.125	0.818	0.136	0.591	1.133	-1.21	0.226
	X16	0.08	0.131	-0.177	0.336	1.083	0.142	0.838	1.4	0.61	0.542
Scenario 2	**X01**	0.138	0.273	-0.396	0.673	1.148	0.313	0.673	1.959	0.51	0.612
	X02	-0.178	0.067	-0.309	-0.046	0.837	0.056	0.734	0.955	-2.65	0.008
	X09	0.187	0.833	-1.446	1.82	1.206	1.004	0.236	6.17	0.22	0.822
	X10	-0.035	0.032	-0.097	0.028	0.966	0.031	0.907	1.028	-1.08	0.279
	X11	0.003	0.015	-0.026	0.033	1.003	0.015	0.974	1.034	0.22	0.822
	X14	0.307	0.158	-0.002	0.615	1.359	0.214	0.998	1.851	1.95	0.052
	X15	-0.227	0.166	-0.552	0.098	0.797	0.132	0.576	1.103	-1.37	0.172

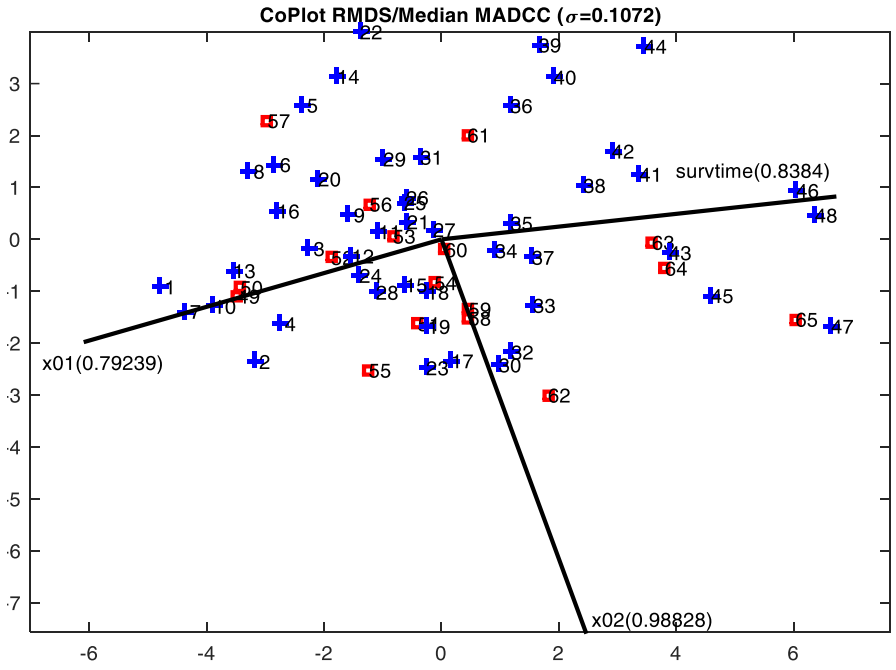


Figure 5. Color coded variable for censoring; red square:0, blue plus:1

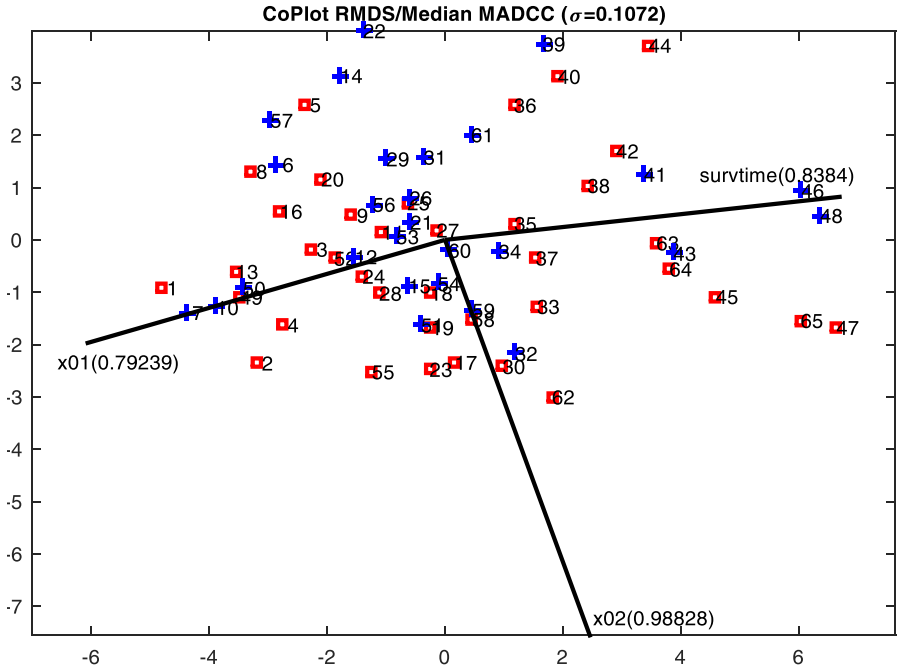


Figure 6. Color coded variable for sex; red square:0, blue plus:1

Table 9. TThe results of multivariate Cox regression model with a variable selection for the contaminated data set

		Coef.	Std. Error	95% Conf. Interval		z	P>|z|
Scenario 1	**X02**	-0.124	0.056	-2.22	0.026	-0.234	-0.015
Scenario 2	**X02**	-0.124	0.056	-2.22	0.026	-0.234	-0.015

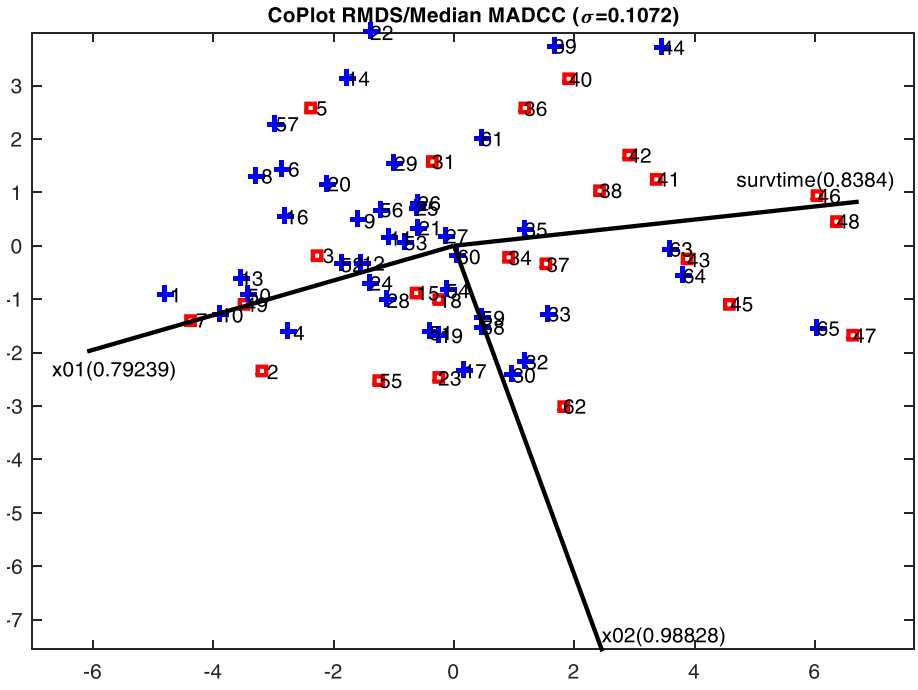


Figure 7. Color coded variable for Bence Jone protein in urine at diagnosis; red square:1, blue plus:2