Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Robustifying Cox - Regression Model Estimation Using M - estimators with application to breast cancer patients

Salwa Salah Aldean Qassim Haidari and Bashar A. Al-Talib

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received March 30, 2023
Accepted June 11 , 2023
Available online, December 1, 2023

Keywords:
Cox regression
Robust Regression
Robust weight
Outliers
Huber's
M-Estimate

Correspondence:
Salwa Salah Aldean Qassim Haidari
salwa.20csp111@student.uomosul.edu.iq

Abstract

This paper focused on estimating the survival time of real data for breast cancer patients in Nineveh province for the period between 2007- 2013. Robust estimation formulas were proposed and dealt with the Cox regression model in survival analysis. Determine the degree of hazard faced by women infected with this disease. Where it was proposed to use some Robust weights, and some classical variance estimators were replaced with Robust estimators to get an efficient estimation of the model, and also the suggestion of Robust weight functions. The Huber weight function was the best and was applied with the three templates to get the best model for the person of the variables that influence the occurrence of the event.

DOI: 10.33899/IQJOSS.2023.0181221

1. Introduction

The Cox regression model is one of the important models proposed by the scientist Cox (1972), and it is one of the methods that fits the case of the dependent variable that will be bi-response, as is the case in the data of cancer patients (a binary variable related to time). Existence of outliers in the data whose primary objective is to estimate the parameters of a model that represents information about the majority of the data. The weighted estimate in the Cox regression model proposed by (Schemper et.al. 2009) has advantages over the proposed models in the case of commensurate hazards and this model is not limited by the distribution of survival time.

2. Cox Regression Model

The Cox Regression Model is one of the most important and most common models in survival analysis models that deal with cases in which the time variable that precedes the occurrence of a specific event is important in analyzing the phenomenon of which the study is concerned. This method has several advantages, the most important of which is that it is considered It is one of the modern methods characterized by the accuracy of its results, as well as the ease of dealing with survival data that appears when time is taken into account. (Bradic, 2011), (Ali & Qadir, 2022).

The original Cox regression model was proposed by Cox (1972). If T was a continuous random variable, the basic model would be as follows:

$$h(t|X_i) = h_0(t)\exp(X_i^T\beta) \quad (1)$$

$h(t|X_i)$: It represents the conditional hazard of the occurrence of the event at time t for the items that have the vector of the explanatory variables X_i^T

$h_0(t)$: The baseline hazard function that depends on time and corresponds to the vector of the explanatory variables

β: is a vertical vector with dimension p×1 of the unknown parameters.The Partial likelihood method is used to estimate the unknown parameters.

X_i: is a class vector with dimension 1×p of the explained variables.

$\exp(X_i^T\beta)$: It is the relative hazard (Hazard Ratio) that does not depend on time, that is, the effect of the variables explained by the increase or decrease in hazard is constant and does not change according to the change of a time point t. Also, the ratio between any two rates of hazard is constant and not dependent on time. (Al-Saqal,2020) (Al-Kafrani,2015).

3. The Concept of Robust Regression

The estimation of the parameters of the normal regression model in the presence of outliers is inefficient; Because there will be a mismatch between the data of the subject of the study and the basic hypotheses that must be available in the model. Because of this, the traditional methods will lose their good properties for estimating the parameters of the studied model. Therefore, alternative statistical methods have been found that are more resistant to the presence of outliers, and these are called Robust Methods. These estimates resulting from this alternative method are called Robust Estimators, and they are insensitive to outliers (Al-Obeidi, 2015).

Robust-statistics is the least affected method by outliers. It is a procedure that seeks to identify outliers and reduce their impact on parameter estimations. It also works on its residual analysis and reduces the weights of outliers (Low-weighted), or completely removes outliers in general. The assigned weights must be studied. For each view, an estimate of the views that were largely excluded, and an assessment of whether these views were significant in the analysis. Robust regression can be defined as:

An estimator retains many of the desirable properties of estimates when some of the regression model's assumptions are violated. A general class of statistical procedures designed to reduce the sensitivity of estimating regression model parameters to failure to meet model assumptions. We can also say that the estimator or statistical procedure is Robust if it provides useful information (Hamoudat, 2020).

4. Concept of Outliers

Sometimes researchers may encounter many statistical problems, some of which may be clear and others not clear, so the researcher will find himself in need of new methods that enable him to organize the course of the experiment by making the resulting error as small as possible and at the same time obtaining an unbiased estimate. for the amount he is looking for.

The idea of studying outliers began with simple ideas based on intuition and guesswork (Bhar & Cupta, 2001).

The issue of outliers has been discussed in many studies and researche, because these values have an impact on the accuracy and integrity of the statistical data and the accuracy of the results to be achieved, and outliers are defined as observations that deviate greatly from other observations, that is, they are not consistent with the rest of the observations of the group for any of the variables. a particular phenomenon (Al-Baqaal, 2017).

Also, outlier values are defined as those observations that seem illogical and show a significant deviation from the rest of the components of the sample in which these observations were found. It was reported (Barnett, 1978) that the outlier observation in a group of data is that observation that appears illogical when compared to other The data set Many researchers have defined outlier values, but all definitions flow into one concept, which is that outlier observation is an observation that is inconsistent with the rest of the observations (Keller & Brian, 2000).

Rousseeuw also showed that outliers can be detected by looking at the error boxes, but this belief is wrong. When the extreme values are leverage points, they can approach the regression line, meaning that their error is very small, while the rest of the error values are large. Although these points are good, and from this the importance of diagnosing outliers appears as an important step in the analysis and decision-making process and represents one of the general goals in data analysis (Barnett & Lewis, 1978).

The appearance of the outlier's vision is due to several reasons, which are:

The outlier appears when the data reverts to a heavy tailed distribution.

Outliers appear when we have a contaminating distribution.

Hawkins (1980) explained that the data come from two types of distributions:

The first: the basic distribution, which generates new data.

The second: the contaminating distribution, which generates outlier values.

These values arise as a result of errors in measurement, reading, recording, or sampling errors arising from poor sample selection and poor representation of the community (Al-Dabbagh, 2020).

5. Outliers Detection

Many researchers have been concerned in their research with the issue of outliers and how they affect the accuracy of the results and study them to reach the best decisions. The newly discovered methods have varied in the field of outliers and their identification. One of these methods is scatter plots, which is the most common method. As well as the box drawing method.

A distinction is made between two types of methods for detecting outliers, including Univariate Methods, which will take each variable separately, and multiple methods, which will take into account correlations between variables in the same data set. There are also other methods for detecting outliers, including the histogram, which is a very common graphic format (Hammodat, 2020).

Tukey also introduced a method for organizing interval-scaled data called stem-and-leaf plotting. It is an alternative to the Histogram. The presence of outliers or outliers in the observations of the explanatory variables or the response variable affects the estimates of the model parameters, as well as the selection of the variables affecting the regression model and the associated statistics (Dan & Ijeoma, 2013).

6. Robust M-Estimators

There are many robust methods available for estimating the model parameters, and it is one of the modern robust techniques with good characteristics, the most important of which is the Robust M estimators (M-Estimators), as it has been recently paid attention to by many researchers due to its efficiency, and its use in estimating the model parameters in the absence of subject The distribution of errors for the normal distribution, that is, the observations of the response variable do not follow the normal distribution, as it works to give less weight to the outlier observation to reduce its impact and use the iteration method in the calculation, which leads to reducing the effect of self-correlation (Huber, 1973).

This method is summarized in minimizing the effect of the large residual values, i.e. minimizing the sum of the squares of the errors, bearing in mind that the estimator M corresponds to the Maximum Likelihood estimates because the function ρ(.) links the potential function to choose an appropriate distribution of the residuals, and that the residual error is:

$$e_i = y_i - \hat{y}_i \quad (2)$$

Let us assume that a data set $(x_1, x_2, \dots, x_n)$ represents a random sample that follows a continuous distribution with a probability density function $f(x, \theta)$ as θ represents the distribution parameter and the parameter θ can be estimated using the greatest possible method and the partial likelihood formula for the greatest possibility function is as follows:

$$lnL(\theta) = \sum_{i=1}^{n} lnf(x_i - \theta) = -\sum_{i=1}^{n} \rho(x_i - \theta) \quad (3)$$

$$\rho(x) = -lnf(x) \quad (4)$$

Where $lnL(\theta)$ is maximized and in terms of the function $\rho(.)$

$$\sum_{i=1}^{n} \varphi(x_i - \theta) = 0 \quad (5)$$

$$\varphi(x) = \rho'(x) = \frac{\partial \rho(x)}{\partial x} = \frac{f'(x)}{f(x)} \quad (6)$$

where $\varphi(x)$ is called the effect function and represents the partial derivative of the function ρ

$$\sum_{i=1}^{n} \varphi \frac{(x_i - \theta)}{\sigma} = 0 \quad (7)$$

$$\sigma = K(\delta)$$

Since:

K: an integer.

δ: an initial estimate of the measurement parameter.

And the possibility function for a random sample $(x_1, x_2, \dots, x_n)$ is in the following form:

$$L(\mu, \delta) = \prod_{i=1}^{n} \frac{1}{\delta} f\left(\frac{x_i - \hat{\theta}}{\delta}\right) \quad (8)$$

And estimates of the greatest possibility θˆ , δˆ to μ , δ are chosen to maximize this possibility, and when deriving the natural logarithm of the function of the greatest possibility we get

$$\sum_{i=1}^{n} \varphi \left[\frac{(x_i - \theta)}{\hat{\delta}}\right] = 0 \quad (9)$$

(al-obedi_2015).

7. Some Robust weight functions

a- Huber function

The Robust estimators are based on the Huber weight function, which has arithmetic advantages, but is sensitive to points of attraction, which are parabolas in the vicinity of zero and increase linearly at the $|x| <$ c level, where an efficiency of 95% is obtained when the errors are distributed normally with a constant. , and that the Huber weight function is:

$$\psi_{Huber}(e_{is}, c) = \begin{cases} 1 & if, |e_{is}| \leq c \\ \frac{c}{|e_{is}|} & otherwise \end{cases} \quad (10)$$

As c takes the default value c = 1.345, as the cut-off constant for each function is used to modify the efficiency of the resulting estimators for specific distributions, (S represents the standard deviation of errors), and smaller values than the constant c are more resistant to extreme values, but at the expense of lower efficiency when distributed Naturally, the tuning constant generally gives reasonable high efficiency in the case of normal, and sometimes in applications we need to estimate the standard deviation to be used in the results, and that the value of the cut-off constant ranges from one standard deviation to two standard deviations for the values of observations or errors. standard for errors. Sometimes it is recommended to use Huber's estimator in almost all cases (Al-Obeidi, 2015).

b- Hampel function

$$\Psi_{Hampel(e_i,c)} = \begin{cases} 1 & if \quad |e_{is}| \le a \\ \frac{a}{|e_{is}|} & if \quad a < |e_{is}| \le b \\ \frac{a(c-|e_{is}|)}{|e_{is}|(a-b)} & if \quad b < |e_{is}| \le c \\ 0 & if \quad |e_{is}| > c \end{cases} \tag{11}$$

As the default values for tuning constants are a=2, b=4 and c=8.

c- Bisquare function

Also called Tukey Beaton or Tukey Biweight, it reaches 95% efficiency when the errors are normally distributed.

$$\Psi_{Bisquare(e_i,c)} = \begin{cases} \left[1 - \left(\frac{e_{is}}{c}\right)^2\right]^2 & if \quad |e_{is}| \le c \\ 0 & if \quad 0 \end{cases} \tag{12}$$

Where c defaults to c=4.685.

8. The Average Hazard Ratio (AHR)

Hazards are expected to change according to survival analysis often over time and hazard ratio, so statistical methods will require determining time or taking the mean value over time relativity. There are 3 equations for the average hazard ratio

$$sAHR = \int \frac{h_1(t)}{h_0(t)} w(t) f(t) dt \tag{13}$$

$$gAHR = exp\left[\int Log\left(\frac{h_1(t)}{h_0(t)}\right) w(t) f(t) dt\right. \tag{14}$$

$$AHR = \frac{\int (h_1(t)/h(t)) w(t) f(t) dt}{\int (h_0(t)/h(t)) w(t) f(t) dt} \tag{15}$$

Where AHR is the definition of average Hazard Ratio,
sAHR stands for Simple Average Hazard Ratio,
gAHR for Geometric Average Hazard Ratio,

The weight function w(t) was chosen to reflect the relative importance related to the hazard ratios in different periods, the most used values are w(t)=1 and w(t)=s(t).

Survival function or equivalent, the proportion of individuals affected by the hazard ratio at t. (Schemper, 2009).

9. Average Ratio Estimate (ARE)

The consequences of the assumptions violating the Cox proportional hazards model are discussed and options for proportionality are reviewed to deal with the non-proportional Cox model, where an additional option for analysis is proposed, which produces weighted estimates at the time points in which events occur. This procedure can be considered as generalizations of the tests for multiple covariates variables, in the same manner that the proportional hazards model represents a generalization of the log-rank test and its advantages are represented in the estimates of the average hazard ratios also for the covariates with the Non-proportional and especially convergent hazard functions. (Breslow,1974)

Through an empirical study, it was found that the average hazard ratios are very close to the accurate calculations of the average hazard ratios, as defined by.

suppose that $T_i,\ C_i\ and\ Z_i(.),\ i = 1,\ \dots,\ n.$ A random sample from the distribution T,C,Z(.) that satisfies the model

$$\lambda(t|Z(t)) = \lambda_o(t) exp\{\beta(t) Z(t)\} \tag{16}$$

where T is the observation time (failure) of the random variable. The time of the common variable Z(.) is a predictable process. For each value of i we note $X_i = min(T_i,\ C_i)\ \ \delta_i = \mathrm{I}(T_i \le C_i)$ (Prentice,1978).

10. Study data and its variables

The data were obtained from Hazem Al-Hafiz Hospital for Cancer Cancers in Nineveh Governorate, and that the research population represents patients with breast cancer who were diagnosed with the disease in the period from 2007 to 2013. Their total number is 246 patients. The data of this study are the dates of disease diagnosis until the date of death or the date of the last follow-up of the patient in 2013 AD, and the follow-up period was calculated in months from the date of diagnosis of the disease until death or the date of the last follow-up, which is the survival time. Survival which is given a value of (0) when the patient is dead and a value of (1) if the patient is alive

This variable is a descriptive variable whose mathematical relationship with the rest of the variables is built by the Cox regression model and as follows:

$$h(t) = h_o(t).\exp(\beta_1 X_1 + \cdots + \beta_p X_p)$$

That is ،the variable adopted in the above equation is h(t)(the severity function) represents the variable of a function in terms of the original response variable(Y) since h(t)represents the death rate, The illustrative variables are as follows:

1-Age:Represents age in years and is a qualitative variable

2-Presentation :It is a descriptive variable

3- Tumor Site :A descriptive variable that takes (1) if it is from the right side ،and takes (2) if it is from the left side" Left."

4- Mass Site :Descriptive variable

5- Lymph Node :A descriptive variable that takes the following values:

(0)If negative

(1)If positive

6- Metastasis :a descriptive variable

7- Estrogen :This variable takes one of the following values :(1) :Yes :(0) No

8- Progesterone :It takes one of the following values :(1) :Yes (0) No

9- Her2 is an immunoma generator on the surface of cells :It takes one of the following values :(0) No (1) There is" Yes."

10- DXT :Deep X-ray therapy has one of the following values" :(0) :not treated" :(1) ،"treated."

11- Hormonal Therapy :

" (0)not treated."

" (1)treated."

Month :The time variable on which Cox's regression method depends ،calculated from the beginning of the diagnosis of the injury and ending with knowledge of the analysis of survival (occurrence of the event) which is death in our case.

11. Applied side

Cox regression model with Huber weight function and Robustness to outlier data

After we analyzed the data using R ,We note from Tables (1, 2, 3) that when applying the three templates (PH-AHR-ARE) in the absence of Robustness and applying the weight function (Huber) to the variables, it turns out that in the template (PH) and (AHR) We showed a significant variable of the progesterone hormone, and its probability value was (0.000000), meaning that (sig = 0.00000 < 0.05) in both templates, which means that this variable is significant, meaning that it effect the occurrence of the death event.

Observing Wald's statistics, it became clear to us that the (PH) template had the largest value, as it was (1087.745), meaning that this template was more efficient than the rest of the templates in determining the variables affecting the model.

Table 1: Template model (PH) with weighted estimation, outliers, and Robustness

	coef	se(coef)	exp(coef)	lower 0.95	upper 0.95	Z	p
Age	3.264560e-02	8.893815e-02	1.033184e+00	8.679076e-01	1.229935e+00	0.3670596	0.7135746
Presentation	-1.066291e-01	1.011800e-01	8.988590e-01	7.371689e-01	1.096014e+00	-1.0538552	0.2919492
TumorSite	-7.056361e-02	1.330493e-01	9.318685e-01	7.179643e-01	1.209501e+00	-0.5303569	0.5958645
MassSite	-2.096512e-02	4.503349e-02	9.792531e-01	8.965250e-01	1.069615e+00	-0.4655450	0.6415412
LymphNode	1.549980e-01	3.473722e-01	1.167656e+00	5.910592e-01	2.306739e+00	0.4462014	0.6554518
Metastisis	-4.250971e-02	7.340607e-02	9.583812e-01	8.299557e-01	1.106679e+00	-0.5791035	0.5625194
Estrogen	3.360517e+13	1.750889e+15	Inf	0.000000e+00	Inf	0.0191932	0.9846870
Progesterone	3.352044e+01	1.680864e+00	3.611938e+14	1.339635e+13	9.738546e+15	19.9423868	0.0000000
Her2	-2.432878e-01	3.194519e-01	7.840458e-01	4.192022e-01	1.466423e+00	-0.7615790	0.4463113

DXT	-2.493506e-01	3.521785e-01	7.793067e-01	3.907811e-01	1.554115e+00	-0.7080234	0.4789307
HormonalTherapy	-3.360517e+13	1.750889e+15	0.000000e+00	0.000000e+00	Inf	0.0191932	0.9846870
Wald Chi-square	1087.745						
df	11						
P-value	0						

Table 2: AHR template model with weighted estimation, outliers, and Robustness

	coef	se(coef)	exp(coef)	lower 0.95	upper 0.95	z	p
Age	1.296394e-01	1.199621e-01	1.138418e+00	8.998907e-01	1.440169e+00	1.080669899	0.2798440
Presentation	-6.972579e-02	1.492868e-01	9.326495e-01	6.960578e-01	1.249659e+00	-0.467059231	0.6404575
TumorSite	-1.770558e-01	1.910350e-01	8.377331e-01	5.760980e-01	1.218190e+00	-0.926823970	0.3540179
MassSite	-3.099635e-02	6.064499e-02	9.694791e-01	8.608299e-01	1.091841e+00	-0.511111507	0.6092730
LymphNode	1.388320e-03	4.698770e-01	1.001389e+00	3.986970e-01	2.515145e+00	0.002954646	0.9976425
Metastisis	-3.935879e-02	1.007381e-01	9.614057e-01	7.891477e-01	1.171265e+00	-0.390704152	0.6960159
Estrogen	8.557650e+14	2.382990e+15	Inf	0.000000e+00	Inf	0.359113996	0.7195098
Progesterone	3.549821e+01	3.105173e+00	2.610211e+15	5.936734e+12	1.147635e+18	11.431958054	0.0000000
Her2	-6.247518e-01	4.694392e-01	5.353943e-01	2.133469e-01	1.343572e+00	-1.330847150	0.1832393
DXT	-1.630334e-01	4.734777e-01	8.495628e-01	3.358695e-01	2.148921e+00	-0.344331632	0.7305969
HormonalTherapy	-8.557650e+14	2.382990e+15	0.000000e+00	0.000000e+00	Inf	-0.359113996	0.7195098
Wald Chi-square	198.0365						
df	11						
P-value	0						

Table 3: Template model (ARE) with weighted estimation, outliers, and Robustness

	coef	se(coef)	exp(coef)	lower 0.95	upper 0.95	z	p
Age	2.576991e-02	8.522381e-02	1.026105e+00	8.682586e-01	1.212647e+00	0.30237927	0.7623630
Presentation	-1.150697e-01	1.078077e-01	8.913040e-01	-7.215390e-01	1.101012e+00	-1.06736087	0.2858089
TumorSite	-6.504123e-02	1.281974e-01	9.370288e-01	7.288382e-01	1.204689e+00	-0.50735220	0.6119077
MassSite	-2.397881e-02	4.647393e-02	9.763064e-01	8.913073e-01	1.069411e+00	-0.51596273	0.6058804
LymphNode	1.507345e-01	3.433470e-01	1.162688e+00	5.932061e-01	2.278876e+00	0.43901497	0.6606507
Metastisis	-4.005471e-02	7.183455e-02	9.607369e-01	8.345623e-01	1.105987e+00	-0.55759673	0.5771198
Estrogen	-5.418376e+13	1.717920e+15	0.000000e+00	0.000000e+00	Inf	-0.03154033	0.9748386
Progesterone	4.437392e+01	4.191355e+01	1.867878e+19	3.930732e-17	8.876129e+54	1.05870116	0.2897359
Her2	-1.862637e-01	3.241103e-01	8.300547e-01	4.397679e-01	1.566714e+00	-0.57469242	0.5654993
DXT	-2.439481e-01	3.534906e-01	7.835283e-01	3.918889e-01	1.566558e+00	-0.69011206	0.4901237
HormonalTherapy	5.418376e+13	1.717920e+15	Inf	0.000000e+00	Inf	0.03154033	0.9748386
Wald Chi-square	9.738883						
df	11						
P-value	0.5540166						

Proposed algorithm for fortifying Cox's regression model

In this thesis, an algorithm was attempted to fortify the estimation process in the Cox regression model to identify the main factors affecting the dependent variable (dwell time in our data), and this was done by following the following series of steps:

1. The use of some templates in estimation, where the three templates (PH-AHR-ARE) were applied.
2. The use of suggested fortified weights based on some weight functions for M estimators, where the Huber, Hampel and Bisquare functions are used by entering them as weights on the estimated target function.
3. Adopting the Huber function to give Robusty to data and compare the results obtained with all the estimated models.
4. To confirm the reliability of the estimated model, Bootstrap samples are generated from the data.
5. Selection of the most efficient models estimated through the significance of the parameters and the values of the parent's statistics.

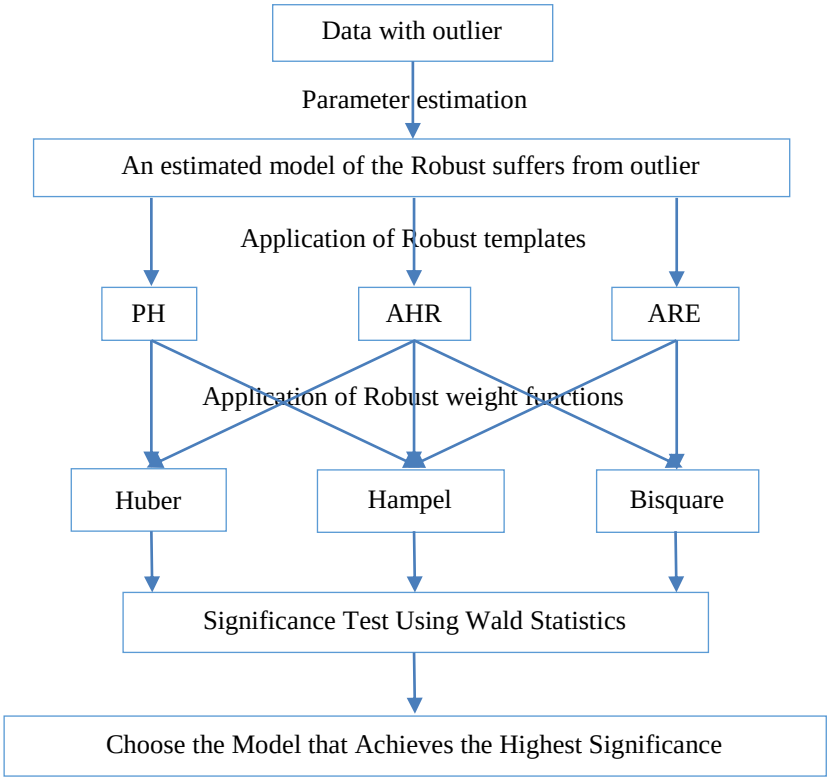

Figure 1: The proposed algorithm to Robust the Cox regression model

The results of the proposed likelihood to immunize the Cox regression model

The results of the analysis indicated that when applying the three templates (PH-AHR-ARE) in the case of outlier data, only the disease progression variable appeared in the absence of significant Robust to the variables, and its probability value was (0.04473447), meaning that (sig = 0.04473447 < 0.05) It effect the model, but in the case of immunity to the variables, the disease progression variable (the way the disease progressed) and the progesterone variable appeared significant, and the probability value for them was respectively (0.02479655, 0.000000, 0.03495496) mean that (sig=0.02749655<0.05) (sig=0.000000<0.05) , (sig=0.03495496<0.05) effect the occurrence of the event). As for when applying the Cox regression model of the Robust weighted by functions (Huber, Hampel, Bisquare) on the data that contain outlier values, it was found that the Huber function is the function that gave good results, as the progesterone hormone variable showed us a significant in the case of Robustness to the variables, which means that the progesterone hormone The most influential variable on the occurrence of the event (death).

It has been found from conducting all the analyzes of the Wald statistic that in the case of data without outliers, the results of the template (ARE) appeared in the case of a greater value of Robust, which amounted to (739.0424), but in the case of data with outliers and in the presence of Robustness, the Wald statistic for the (PH) test was greater It amounted to (938.2368), and in the case of applying weight to outlier data, a larger (PH) test also appeared, and its value was (1087.745).

And by making comparisons using (P-Value) and (Wald) test, we found that the template (PH) in the presence of outlier values and Robust weights is the best among all the templates that were tested, as it produced a model with the highest significance.

Table 4: The proposed method for immunizing the Cox regression model

template	The test	with outlier And without Robust	with outlier And with Robust	With outlier, without Robust and with weight	With outlier, with Robust and with weight
PH	Wald Chi-Square	15.31619	938.2368	0.1076209	1087.745
	P-Value	Presentation 0.04473447	Presentation 0.02479655 Progesterone 0.00000000		Progesterone 0.00000000
AHR	Wald Chi-Square	6.123973	750.6761	0.2705897	198.0365
	P-Value		Progesterone 0.00000000		Progesterone 0.00000000
ARE	Wald Chi-Square	14.91192	568.3403	0.163119	9.738883
	P-Value		Presentation 0.03495496 Progesterone 0.00000000		

Conclusions

1. When applying the three templates (PH-AHR-ARE), when the data were outlier, only the disease progression variable appeared to be significant, Progesterone has a significant effect on the incidence of death (survival time) for patients.
2. When applying the weighted Cox regression model of the hippocampus (Huber) on the data that contains outlier values, the progesterone hormone variable appeared to be significant in the case of Robust to the variables, which means that the progesterone hormone was the most influential variable on the occurrence of the event (death).
3. It became clear to us from conducting all the analyzes of the father's statistic that in the case of non-outlier data, the template (ARE) showed the largest value in the case of Robust and amounted to (739.0424), while in the case of data with outlier values and the presence of Robustness, the father's statistic for the template (PH) was greater and amounted to (938.2368), and in the case of applying the weight to outlier data, the template also showed a larger (PH) and its value was (1087.745).

Acknowledgment

The authors are very grateful to the University of Mosul, College of Computer Science and Mathematics for their provided facilities, which helped improve this work's quality.

Conflict of interest

The author has no conflict of interest.

References

1. Al-baqaal, Salih Muayad Shaker, (2017). "Robust Proposed Methods for Mean Analysis of Regression Models and Comparison with Ordinary Least Squares Estimators Using Simulations". PhD thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.
2. Al-Dabbagh, Zainab Tawfiq Hamed, (2020). "Virtuous Penalty Methods for Parameter Estimation and Selection of Variables in a Linear Regression Model". PhD thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.
3. Ali, T., & Qadir, J.R. (2022). Using Wavelet Shrinkage in the Cox Proportional Hazards Regressing Model (Simulation study). *Iraqi Journal of Statistical Sciences*, 19(1), 17-29.
4. Al-Kafrani, Shamel Munawer Aziz, (2015). "The use of Cox regression methods in survival analysis applied to breast cancer patients in Nineveh Governorate". Master Thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.
5. Al-Obeidi, Nada Nizar Muhammad, (2015). "Proposed weighted methods for two-stage cluster immunity and estimation of regression models (cluster regression) by applying data on thalassemia patients in Nineveh province". Master Thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.

6. Al-Saqal, Uday Essam Sultan (2020). “Adapted Penal Possibility Method for Selecting Variables in Cox's Regression Model”. Master Thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.
7. Barnett, V. & Lewis, T., (1978). "Outliers in statistical data". John Wiley and Sons, New York.
8. Bhar, L. & Gupta, V., (2001). "A useful statistic for studying outliers in Experimental Designs". *The Indian Journal of Statistic*, New Delhi, Volume 63, Series B pt. 3, pp. 338-350.
9. Bradic, J., Fan, J., & Jiang, J. (2011). "Regularization foe cox's proportional hazard model with NP-Dimensionality". *The annals of Statistics*, Vol. 39, No. 6, pp. 3092-3120. Doi.10.1214/11-A0s911.
10. Breslow, N.E. "Covariance analysis of censored survival data", Biometrics, 30, 89-99. 1974.
11. Cox DR. (1972). "Regression Models and Life Tables". (with Discussion). *Journal of the Royal statistical society*, series B; Vol. 34, No. 2, pp. 187-220.
12. Dan, E., & Ijeoma, O., (2013). "Statistical analysis Method of Detecting outliers in Aunivariate data in a regression analysis model". *International Journal of Education and Research*. Vol. 1, No. 5.
13. Fox, J. and Weisberg, S. AN “R companion to Applied Regression”. Sage Thous and Oaks, CA, Second edition. 2011.
14. Hammoudat, Alaa Abdel Sattar, (2020).” Using some wavelet reduction techniques and robust methods to improve the efficiency of generalized additive model estimators”. PhD thesis, College of Computer Science and Mathematics, Department of Statistics and Informatics, University of Mosul.
15. Hawkins, D.M. (1980). "Identification of outliers". Original published by Chapman & Hall. Research and Statistics Council for Scientific and Industrial Research, South Africa.
16. Huber, P. (1973). "Robust Regression Asymptotics, conjectures, and Monte Carlo". *Ann, Statistics*, Vol.1, No. 5, pp. 799-821.
17. Kalbfleisch, J.D., Prentice, R.L. “Estimation of the average hazard ratio”. Biometrika 1981; 68: 105-112. Doi: 10.1093/biomet/68.1.105
18. Keller, G. & Brian Warrack, (2000). "Statistic for management and Economics". 5th Edition Duxburg, Thomson Learning U.S.A.
19. Prentice, R.L.” Linear rank Test with right censored data” Biometrika, 65,pp.167-179. 1978.
20. Schemper, M., Wakounig, S., Heinze, G. "The estimation of average hazard ratios by weighted Cox-Regression" statistics in medicine, 28: 2473-2489. 2009.

تحصين تقدير نموذج انحدار كوكس باستخدام مقدرات M مع تطبيقها على مرضى سرطان الثدي

سلوى صلاح الدين قاسم حيدري و بشار عبد العزيز مجيد الطالب

قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة: ركز هذا البحث على تقدير وقت البقاء لبيانات حقيقية لمرضى سرطان الثدي في محافظة نينوى للفترة من 2007 إلى 2013. تم اقتراح صيغ تقدير حصينة مع نموذج انحدار كوكس في تحليل البقاء وتحديد درجة الخطورة التي تواجهها المرأة المصابة بهذا المرض. حيث تم اقتراح استخدام بعض الأوزان الحصينة، وتم استبدال بعض مقدرات التباين التقليدية بمقدرات حصينة للوصول إلى تقدير فعال للنموذج، وكذلك اقتراح دوال وزن حصينة. كانت دالة الوزن هوبر هي الأفضل وتم تطبيقها مع القوالب الثلاثة للوصول إلى أفضل أنموذج شخص المتغيرات التي تؤثر على حدوث الحدث.

الكلمات المفتاحية: انحدار كوكس،الانحدار الحصين،الاوزان الحصينة،القيم المتطرفة، هوبر،تقدير M.