Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (6): 1033-1038, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.1033.1038

Identify Attributable Variables and Interactions in Breast Cancer

[1]Yong Xu, [2]James Kepner and [1]Chris P. Tsokos
[1]Department of Mathematics and Statistics, Radford University, VA, 24142, USA
[2]250 Williams Street, Suite 600, American Cancer Society, Atlanta, GA 30303, USA

Abstract: The object of the present study is to develop a statistical model for breast cancer tumor size prediction for United States patients based on real uncensored data. When we simulate breast cancer tumor size, most of time these tumor sizes are randomly generated. We want to construct a statistical model to generate these tumor sizes as close as possible to the real patients' data given other related information. We accomplish the objective by developing a high quality statistical model that identifies the significant attributable variables and interactions. We rank these contributing entities according to their percentage contribution to breast cancer tumor growth. This proposed statistical model can also be used to conduct surface response analysis to identify the necessary restrictions on the significant attributable variables and their interactions to minimize the size of the breast tumor.

Key words: Statistical modeling, survival analysis, tumor size simulation, breast cancer

INTRODUCTION

National Cancer Institute (2010) defines breast cancer as following: Cancer that forms in tissues of the breast, usually the ducts (tubes that carry milk to the nipple) and lobules (glands that make milk). It occurs in both men and women, although male breast cancer is rare.

The proposed model that we are developing includes individual variables, interactions and higher order variables if applicable. In developing the statistical model, the response variable is the tumor size at diagnosis for breast cancer patients. We have identified 26 possible attributable variables for breast cancer, denoted, X1, X2, ..., X26. For example, X1 stands for patient ID and X2 stands for the patient's age at diagnosis. In this study, we would like to find the relation between the tumor size and all other attributable variables. We cannot use survival time to predict the tumor size since death time happens after the tumor is detected. Therefore, we exclude the variable survival time (x25) and the censoring indicator function vss (x26) in the first part of study. Thus, we have only 24 variables left to construct our statistical model.

In the present analysis, we used real data from the Surveillance Epidemiology and End Results (SEER) Program. SEER collects and compiles information on incidence, survival and prevalence from specific geographic areas representing about 26 percent of the U.S. population plus cancer mortality for the entire U.S.

The proposed statistical model is useful in predicting the tumor size given data for the attributable variables. It is statistically evaluated using R square, R square adjusted, the PRESS statistic and several types of residual analyses. Finally, its usefulness is illustrated by utilizing different combinations of the attributable variables.

In addition, the attributable variables are ranked according to their contributions to accurately estimate a patient's tumor size.

HISTORICAL REVIEW

Survival analysis is used more and more in many areas. Many researchers have contributed to this subject. The accelerated life testing model was studied by Qiu and Tsokos (2000). A semi-parametric accelerated failure model was introduced by Shang and Jeremy (2002). An analytical approach on cure rate model based on uncensored data was discussed by Uddin *et al.* (2006). Using decision tree for competing risks for breast cancer was discussed by Ibrahim *et al.* (2008). There are several researches that had been done to determine the factors that are contribute to the relapse time of the breast cancer (Eleni and Gabriel, 2008; Habibi *et al.*, 2008; Freedman *et al.*, 2009; Brawley, 2009).

DEVELOPMENT OF THE NONLINEAR STATISTICAL MODEL

We randomly extract 155 uncensored breast cancer patients' information from SEER data base. The data was obtained from 2000 to 2006. We want to develop a statistical model with full information instead of censored; therefore, we will use the 155 uncensored patient's

Corresponding Author: Yong Xu, Department of Mathematics and Statistics, Radford University, VA, 24142, USA

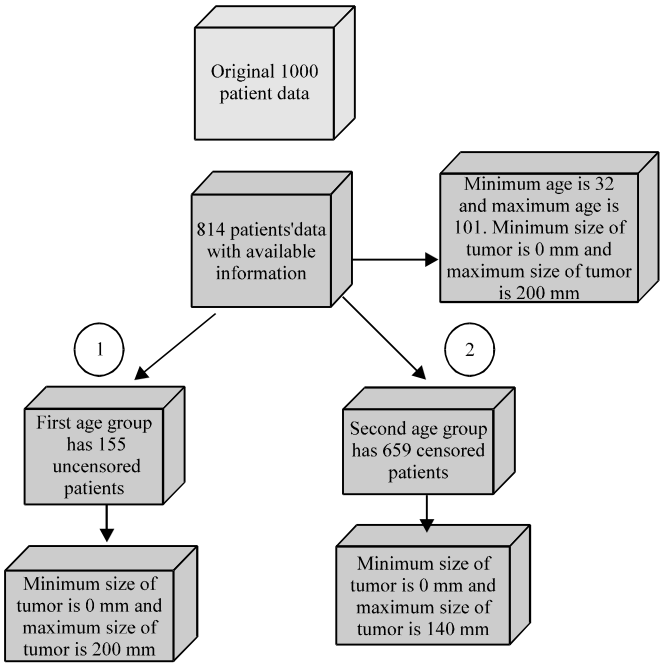

Fig. 1: Breast cancer data tree diagram

Table 1: List of attributable variables

Name	Full name of variables
X1	Patient ID (id)
X2	Age at diagnosis (age)
X3	Year of Birth (birthy)
X4	Birth Place (birthp)
X5	Sequence Number Central (snc)
X6	Month of diagnosis (month)
X7	Year of diagnosis (year)
X8	Primary Site (ps)
X9	Laterality (la)
X10	Histologic Type ICDO3 (ht)
X11	Behavior Code ICDO3 (bc)
X12	Type of Reporting Source (trs)
X13	RXSumm SurgPrimSite (rxps)
X14	RXSumm Radiation (rxr)
X15	RXSumm RadtoCNS (rxcns)
X16	Age Recode Year olds (ager)
X17	Site Recode (sr)
X18	CSS chema (css)
X19	AJCC stage3 rdedition (ajcc)
X20	First malignant primary indicator (findi)
X21	State-county recode (scr)
X22	Race (race)
X23	Cause of Death to SEER site recode (cod)
X24	Sex
X25	Survival time recode (survtime)
X26	Vital Status recode (vss)

Within parenthesis are short form of variables

information to construct our statistical model. The data tree to develop a statistical model taking into consideration diagram is shown in Fig. 1.

We proceed the twenty four attributable variables (Table 1). The form of the statistical model is given by tumor size as a function of x1, x2,… , x24. Note that some of the variable's values are obtained after the tumor size is recorded. In our analysis all the patients in the data base have breast cancer. We utilize the values of the tumor size once the patient has gone through a diagnostic process. Thus, the general statistical form of the proposed model with all possible attributable variables and interactions will be of the form in Eq. 1.

$$TS = \alpha_0+\alpha_1A_1+\alpha_2A_2+...+\alpha_iA_i+\beta_1B_1+\beta_2B_2+...+\beta_jB_j+\varepsilon \quad (1)$$

Here, TS stands for tumor size, the α and β are the coefficients and A is the first order term of the attributable variables and B are the possible interactions and higher order terms. The objective is to develop the most representative estimate of the above model based on available data.

One of the basic underlying assumptions in formulating an estimate of the above statistical model is that the response variable should be Gaussian distributed. Unfortunately, in the present form that is not the case. This fact is clearly demonstrated by the QQ plot shown in Fig. 2.

Furthermore, the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) with the necessary calculation of the test statistic W = 0.7437 and p-value = 3.787e-15 is additional evidence that the tumor size does not follow normal probability distribution. We proceed in utilizing

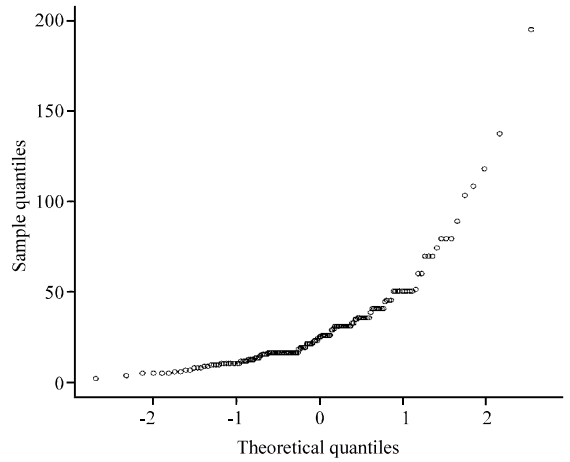

Fig. 2: QQ plot for testing normality for original data

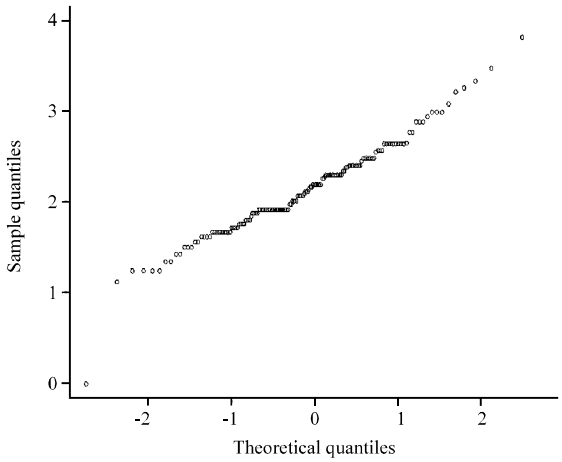

Fig. 3: QQ plot for testing normality for transformed data

the Box-Cox transformation (Box and Cox, 1964) to the tumor size to determine if such a filter will modify the given data to follow the normal distribution so that we can proceed to formulate the proposed statistical model. Applying the Box-Cox transformation results in the statistical information presented in Table 2. One tumor size data's value is zero and Box-Cox transformation can only apply to a positive data set. Therefore, we use 0.00000000000001 to replace this zero value so we can perform Box-Cox transformation.

Therefore, we decide to use the transformed tumor size as our response and we redo the Box-Cox transformation test with the QQ plot to see if the transformed data will follow Gaussian probability distribution. With the Box-Cox results for transformed data presented in Table 2, we can conclude that the transformed data follows the Gaussian probability distribution. Also the QQ plot in Fig. 3 supports the fact that the transformed tumor size follows the Gaussian probability distribution. Since the transformed power is only 0.2659, we decide the logarithmic filter will be appropriate in this situation for the transformed tumor size data.

Thus, we can proceed to estimate the coefficients of the attributable variables for the filtered transformed tumor size data to obtain the coefficient of all possible interactions and at the same time determine the significant contributions of both attributable variables and interactions.

We begin with the previously defined twenty-four attributable variables x1, x2,... , x24 and the two hundred and seventy-six first degree 2nd order interaction between each pair and the two thousand and twenty-four first degree 3rd order interactions between any three variables. We did not consider any 4th and higher order interactions. Since we already have a lot more terms than

Table 2: Box-cox transformation for normality for the original transformed data

			Wald	
Data	Est. power	SE	Power = 0	Power = 1
Original	0.2659	0.0339	7.8445	-21.6563
Transformed	1	0.1275	7.8444	-2e-04

patients in the data itself, we utilized sub modeling skills and two way selection procedures to construct our model. To develop the models, initially we start building our model with a total of two thousand and three hundred and twenty four terms that include initial contribution of the attributable variables and the described interactions. More than thirty candidate models were constructed.

During statistical analysis in the estimation process, we found only three of the twenty-four attributable variables were significant contributors. The significantly contributing and interaction variables are RXR(X14), RXPS(X13) and AJCC(X19). However, SNC(X5), HT(X10), themselves individually do not significantly contribute to the response variables but when they interact with other variables they do significantly contribute to the response variable. Therefore, we still keep them in our final model. There are thirty-one missing values in the variable AJCC. We use the mean of the rest of the data value in the variable AJCC to replace the NA value in order to perform prediction of the model. Thus, the results of estimation of Eq. 1 are given by Eq. 2 as follows:

$$\ln(\hat{TS}^{.2659}) = 2.7\text{-}2.09\times10^{-2}X_5 - 2.14\times10^{-4}X_{10} + 3.28\times10^{-3}X_{13} \\ -3.3X_{14}+2.72\times10^{-3}X_{19} + .24X_{14}\times X_{19} + 1.91\times10^{-4}X_5 \\ \times X_{10}\times X_{14} - 1.19\times10^{-1}X_5\times X_{14}\times X_{19} \quad (2)$$

We will utilize the initial transformation that we used to transform the response data to get the result in

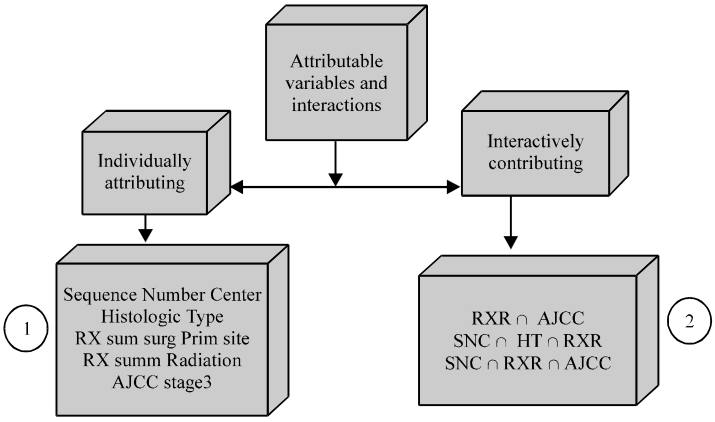

Fig. 4: Breast Cancer Attributable Variable Diagram

Eq. 3 by taking the exponential on both sides of the 2 and then take 3.61's power on both sides of the Eq. 2:

$$\begin{aligned}\hat{TS} = (\exp(&2.7\text{-}2.09\times10^{-2}X_5 - 2.14\times10^{-4}X_{10} + 3.28\times10^{-3}X_{13}\\ &\text{-}3.3X_{14} + 2.72\times10^{-3}X_{19} + .24X_{14}\times X_{19} + 1.91\times10^{-4}X_5\\ &\times X_{10}\times X_{14} - 1.19\times10^{-1}X_5\times X_{14}\times X_{19}))^{3.76}\end{aligned} \quad (3)$$

The proposed statistical model's high quality has been evidenced by R square and R square adjusted, which are the key criteria of evaluating such models. The regression sum of squares (SSR) also called the explained sum of squares is the variation that is explained by the regression model. The Sum of Squared Errors (SSE), also called the residual sum of squares, is the variation that is left unexplained. The total sum of squares (SST) is proportional to the sample variance and equals the sum of SSR and SSE. The coefficient of determination R^2 is defined as the proportion of the total sum of squares that is explained by the model. That is R^2 = SSR/SST. It provides an overall measure of how well the model fits the data. R-square adjusted will adjust for the degrees of freedom in the model and it works better when there are a substantial number of parameters in the model. R-square adjusted is given by:

$$R^2_{adj} = 1 - \frac{SSE/df(e)}{SST/df(t)}$$

Here df(e) means the degree of freedom of the SSE and df(t) means the degree of freedom of SST.

The prediction of residual error sum of squares (PRESS) statistics evaluate how good the estimation is each time a data point is removed. The PRESS statistic is defined by:

$$PRESS = \sum\left(\frac{(y-\hat{y})}{(1-H_{ii})}\right)^2$$

(Allen, 1971, 1974) and Hii is the diagonal elements of the projection matrix. Therefore, we should choose the model with the smallest PRESS statistics from all candidate models.

For our final model the R squared is 0.88 and R squared adjusted is 0.87. Both R squared value and R squared adjusted value are high (close to 90%) and these two are very close to each other. This shows our model's R squared increase is not due to the increase of the parameters' estimates, but rather the good quality of the proposed model to predict tumor size given values of the identified attributable variables. Secondly, the PRESS statistics' results support the fact that the proposed model is of high quality. Table 3 shows the best three models based on the PRESS statistic out of total thirty-six models and it is clear that the best model is number 36 which is the final model.

Furthermore, R square and R square adjusted are calculated for those 36 models which are of interest but the proposed model still gives the best possible estimates of the tumor size for breast cancer in SEER's data.

In Table 4, all the important attributable variables and interactions are given. For example, X5 and X10 are not significant by themselves, only in combination with the others. We summarize the attributable variables individually and interactively in the following schematic network as showed in Fig. 4.

The ranks of the most important attributable variables with respect to their contribution to estimating tumor size are given in Table 5. This is the interaction among the sequence number central (SNC), RXsumm Radiation (RXR) and AJCC stage3 rdedition (AJCC).

Table 3: PRESS statistics for best three models

Model No.	PRESS value	Rank of the model
31	96.73797	3
33	98.4167	2
36	104.8218	1

Table 4: List of attributable variables

No	Individual variables	Name of individual variables
1	X5	Sequence Number Central
2	X10	Histologic Type ICDO3
3	X13	RXSumm SurgPrimSite
4	X14	RXSumm Radiation
5	X19	AJCC stage3 rdeditionInteractions
6	X14:X19	RXR∩AJCC
7	X5:X10:X14	SNC∩HT∩RXR
8	X5:X14:X19	SNC∩RXR∩AJCC

Table 5: Rank of variable according to contributions

Rank	Variables
1	X5:X14:X19
2	X14:X19
3	X19
4	X5:X10:X14
5	X5
6	X14
7	X13
8	X10

VALIDATION OF THE PROPOSED MODEL

We use two methods to validate the model. The first method is to use the proposed model to calculate the predicted value for each tumor size and then calculate the residuals. A residual is defined as the original value minus the predicted value. Table 6 shows the last ten residuals out of the total one hundred fifty-five residuals.

The mean of the residuals is -0.0286, variance of the residuals is 1.588, standard deviation is 1.26 and standard error of the residuals is 0.1012.

The second method we will utilize is called cross validation. We construct our model using only the data left and the constructed model will be same structure as our proposed model with only coefficients being different. Quality of the model will be tested using three settings.

We first randomly divide the data into two datasets of the same size. We use one of the datasets to construct the model and then use the resulting model to predict the values in the other dataset. Then switch the two data sets and repeat the procedure. The mean of all residuals turned out to be 1.0652916.

Next, Dataset was divided into six small data sets and use five of them to construct the model and validate the model using the sixth one. Repeat the same procedure for each of the six small datasets. The mean of all residuals was 0.1318486.

Finally, Divide the dataset into 155 datasets and use all 154 datasets to construct the model and validate the model using the one left out. Repeat the procedure 155 times. Table 6 shows the last ten residuals out of the total one hundred fifty-five residuals.

Table 6: Residual analysis for original data and cross validation

No	Original data residual values	Cross validation residual values
146	1.3264266	0.13328523
147	1.0579828	0.08479963
148	0.9756659	0.07219656
149	1.7362950	0.36801567
150	0.9643773	0.07578824
151	1.2427113	0.11712640
152	1.3705402	0.14230504
153	1.3640997	0.14117093
154	1.6072370	0.19570259
155	1.9573079	0.29023850

The mean of the residuals was 0.634, the variance of the residuals was 42.89, standard deviation of the residuals was 6.55 and standard error of the residuals was 0.53.

USEFULNESS OF THE PROPOSED STATISTICAL MODEL

We can conclude from our extensive statistical analysis that there are only three significant attributable variables to the tumor size for breast cancer namely, RXR(X14), RXPS(X13) and AJCC(X19). As for SNC(X5), HT(X10). They themselves individually do not significantly contribute to the response variables; however, when they interact with other variables, they do significantly contribute to the response variable. Furthermore, we also tested two thousand and three hundred possible interactions of the attributable variables and we found three interactions to significantly contribute to tumor size for breast cancer.

This model is useful for a number of reasons:

- One can also use the proposed model to generate various scenarios of the breast cancer tumor size as a function of different values of the subjective entities for data simulation purpose.
- It can be used to identify the significant attributable variables
- It identifies the significant interactions of these attributable variables
- The significant contributions to the breast cancer tumor size are ranked
- A confidence interval for the tumor size can be constructed with parametric analysis. By obtaining the (1-α)% confidence limits for the response, we can describe how confident we are that our estimate is close to the actual tumor size

- The model as shown in equation 3 can be used to perform surface response analysis to place the restrictions on the significant attributable variables and interactions to minimize the breast cancer tumor size. We can also put restrictions on the variables to minimize the response of the tumor size by nonlinear control with (1-α)% confidence limits

CONCLUSIONS AND DISCUSSION

In the present study, parametric analysis was performed to estimate tumor size for breast cancer patients for data simulation purpose. The initial measurement of tumor size was collected from the SEER database. Those data do not follow normal probability distribution. Using the standard Box-Cox transformation, the SEER tumor size data became approximately normally distributed. We developed a nonlinear statistical model (nonlinear in terms of the power and logarithm of the response variable). Through the process of developing the statistical model, we found only four variables, namely, rxr(X14), rxps(X13) and ajcc (X19) and three interactions that significantly contribute to the tumor size. The proposed statistical model was evaluated using the R-square, R-square adjusted, PRESS statistics and three cross validation methods all of which support the high quality of the developed statistical model. This model can be used to obtain a good estimate of tumor size knowing the four significantly attributable variables and three interaction terms.

We validate this model on four different data sets. Since this type of model is strongly data driven so we need to make the necessary modification. For example, the Box-Cox transformation's power estimation needs to be changed according to the data set. The specific model's interaction and estimation of the coefficients need to be recalculated. After the necessary modifications these four data sets fit the model as following. The model fits one of the data sets pretty well with high R-square more than 90. The model fit two data sets not very well with R-square around 30 and one data set pretty bad with R-square around 10. Considering the complexity and random behavior of the breast cancer and the limitation of our available data set, we are not surprised with the result. We wish to make the update of the model based on large scale data set for the future study.

REFERENCES

Allen, D.M., 1971. The prediction sum of squares as a criterion for selecting predictor variables. Technical Report No. 23, Department of Statistics, University of Kentucky.

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16: 125-127.

Box, G.E.P. and D.R. Cox, 1964. An analysis of transformation. J. Roy. Stat. Soci., 26: 211-252.

Brawley, O.W., 2009. Is race really a negative prognostic factor for cancer. J. Nat. Cancer Instit., 101: 970-971.

Eleni, A. and N.H. Gabriel, 2008. Prognostic factors in metastatic breast cancer successes and challenges toward individualized therapy. J. Clin. Oncol., 26: 3360-3662.

Freedman, R.A., Y. He, E.P. Winer and N.L. Keating, 2009. Trends in racial and age disparities in definitive local therapy of early-stage breast cancer. J. Clin. Oncol., 27: 713-719.

Habibi, G., S. Leung, J.H. Law, K. Gelmon and H. Masoudi *et al*., 2008. Redefining prognostic factors for breast cancer: YB-1 is a stronger predictor of relapse and disease-specific survival than estrogen receptor or HER-2 across all tumor subtyples. BCR., 10: 86-86.

Ibrahim, N.A., A. Kudus, I. Daud and M.R. Abu-Bakar, 2008. Decision tree for competing risks survival probability in breast cancer study. World Acad. Sci. Eng. Technol., 38: 15-19.

National Cancer Institue, 2010. Breast cancer. http://www.cancer.gov/cancertopics/types/breast.

Qiu, P. and C.P. Tsokos, 2000. Accelerated life-testing model building with Box-Cox transformation. Sankhya Idian J. Stat., 62: 223-235.

Shang, L.C. and T.M.G. Jeremy, 2002. A semi-parametric accelerated failure time cure model. Stat. Med., 21: 3235-3247.

Shapiro, S.S. and M.B. Wilk, 1965. An analysis of variance test for normality (Complete samples). Biometrika, 52: 591-611.

Taj Uddin, M., M.N. Islam and Q.I.U. Ibrahim, 2006. An analytical approach on cure rate estimation based on uncensored data. J. Applied Sci., 6: 548-552.