Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

Survival Regression Model for Predicting Chronic Kidney Disease

Manoprabha M*, Balasubramaniam R, Deneshkumar V and Senthamarai Kannan K

Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti,
Tirunelveli – 627012, Tamilnadu, India.

manoprabhamurugan@gmail.com, bala.rcbe@gmail.com

ABSTRACT

Chronic kidney disease commonly known as end-stage kidney failure, describes the gradual loss of kidney function, which leads to dialysis or kidney transplant. This chronic disease has become a major cause of global morbidity and mortality. In this article, the covariates which affect the outcome of the patients are estimated using logistic regression. The prime goal is to accurately predict the target class for each case in the data.

Keywords: Logistic regression; Chronic kidney disease; prediction; Accuracy.

Introduction

Logistic regression has been used to analyse the survival data when it has binary outcomes. This survival regression technique has been applied to several investigations that examine relationship between risk factors and the event. This paper shows an illustration of the application of logistic regression. The purpose of this analysis is to assess the effects of multiple explanatory variables, which can be numerical or categorical, on the outcome variable.

Breslow (1975) analysed survival data under proportional hazards model. He reviewed the methodology for the statistical analysis of censored survival data which arise from a model in which the factors under investigation act multiplicatively on the hazard function of an underlying non-parametric survival distribution. This flexible approach provides computationally feasible solutions to the one-sample problem, multi-sample problem, regression with continuous covariates, regression in matched-pair designs, and evaluation of changes in treatment or prognostic status.

Whitehead (1980) fitted Cox regression model to survival data using GLIM. Methods of estimating the underlying survivor functions are discussed. The Poisson model which allows the use of GLIM is introduced and interpreted. Two different treatments of tied observations are mentioned, and their properties are compared with an example.

Robert D. Abbott (1985) has illustrated the application of logistic regression to survival analysis based on data from Framingham Heart study. Khaw and Barrett-Connor (1986) determined whether modifiable risk factors have a differential effect on cardiovascular risk in those with or without a family history of heart attack. Smoking was found to be the stronger predictor of cardiovascular disease.

Christensen (1987) performed a multivariate survival analysis using Cox's regression model. And illustrated a numerical analysis using this model. Goldfarb-Rumyantzev AS et al. (2003) generated a predictive algorithm for 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset. Logistic regression and tree based model were used as the predictive algorithm. The predicted probability of graft survival showed a strong correlation with the observed survival.

D.R. Cox (1972) outlined about the analysis of censored failure time and the conditional likelihood is obtained to infer the regression coefficients. The hazard function is taken as the function of the explanatory variables. Bradley Efron (1988) has discussed about logistic regression, survival analysis and Kaplan-Meier curve. He used the logistic regression to estimate the hazard rate and survival curve for censored data. This showed that parametric model can be used on censored data which provides both estimates and standard errors.

Katz and Hauck (1993) described about the time-dependent covariates and the application of proportional hazards regression for heart attack patients. This regression analysis can also be used to control for baseline differences between groups in nonrandomized studies and randomized clinical trials. Wannamethee et al., (1998) carried a prospective study on 7142 mean to examine the relationship between lifestyles and the likelihood of 15-year survival free of heart attack and diabetes. Cox predictive survival analysis is used to estimate the probability of survival.

Survival data are generally described and modelled in terms of survival function and hazard function. Survival function represents the probability that an individual survives beyond the specified time which can be non-parametrically estimated using KM method. Hazard function measures the risk of an event happening at a specified point in time (Clark et al., (2003)). Bewick et al. (2004) have reviewed on survival analysis and described about the Kaplan-Meier

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

method, log rank test and Cox's proportional hazards model. Some numerical examples have also been illustrated.

Ata and Sozer (2007) applied the Cox regression models with nonproportional hazards for lung cancer survival data. The Cox regression model, is widely used for the analysis of treatment and prognostic effects with censored survival data, makes the assumption of constant hazard ratio. In the violation of this assumption, different methods should be used to deal with non-proportionality of hazards. In this study, the stratified Cox regression model and extended Cox regression model, which uses time dependent covariate terms with fixed functions of time are discussed. The results are illustrated by an analysis of lung cancer data in order to compare these methods with respect to Cox regression model in the presence of nonproportional hazards.

David W. Hosmer et al. (2008) gave an interpretation for the fitted proportional hazards regression model. The inference is given from the estimated coefficient in the model. The estimated coefficient for a covariate represents the rate of change of a function of the dependent variable per-unit change in the covariate.

Luis Meira – Machado et al. (2008) has done a survival analysis on Stanford heart transplant patients and Galicia breast cancer patients focusing on estimating the transition probabilities and survival probabilities using multistate models. Here various models like Cox regression model, Cox semi Markov model have been compared with the multistate model. Based on this comparative study multistate Markov model yielded a new biological insight compared to other survival models.

Asil oztekin et al. (2009) constructed a model for predicting the graft survival for heart-lung transplantation patients. Some machine learning methods like neural network, decision tree and logistic regression were used for the classification analysis. Fine and Gray (2012) proposed a proportional hazard model for the sub distribution of the competing risk. The analysis of the competing risk involves modelling the cause specific hazard function but this model does not have a direct interpretation for the survival probability for the particular failure type. For overcoming this scenario cumulative incidence function is designed. Both models have been applied for a breast cancer dataset.

Karim et al. (2015) analysed the chronic disease conditions using logistic regression in Ghana. This study revealed that the occurrences of chronic disease conditions are associated with factors like age, sex, religion, ethnicity, marital status, and occupation, level of education and income levels.

Sergey krikov et al. (2007) predicted the kidney transplant survival using tree based model. They developed a model for predicting the probability of the graft survival at 1-, 3-, 5-, 7- and 10 years. Logistic regression was used for variable selection. The variable which shows significant result are included in the final tree based model. The performance of the model was tested using the ROC curve.

Shen et al. (2016) constructed prognostic nomograms for patients with resectable hepatocelluar carcinoma incorporating systemic inflammation and tumour characteristics. Cox model was used as the predition model. The risk factors with high hazard ratio were found out using this model. This model has the higher predictive power; it was assessed using the c-index. Kim and Li (2017) studied about the postoperative complications affecting survival after cardiac arrest in general surgery of 1352 patients. The associations between previous complications and mortality after cardiac arrest were assessed using Cox proportionalhazardmodels.

The unique feature of survival data is that not all patients experience the event at the end of the observation period, so the actual survival times for some patients are unknown. This refers to censoring which should be taken in to account for a valid inference. Survival time is mostly skewed and assumes that the data are normally distributed. Some nonparametric and semi parametric methods for survival analysis are reviewed by Schober and Vetter (2018).

Park et al. (2018) assessed the risk of chronic kidney disease in chronic HCV infected patients and the incidence reduction of CKD after receipt of HCV treatment using Cox regression model. They also evaluated the risk of MPGN and cryoglobulinemia in chronic HCV patients.

Methodology

Logistic Regression Model

Logistic regression was developed by statistician David Cox in 1958. It is a widely used statistical model that uses a logistic function to model a binary dependent variable. Logistic regression is a predictive model; it is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

Suppose we have an unbiased sample of n patients from a target population.

Let

$$d_i = \begin{cases} 1: if\ the\ i^{th} patient\ suffers\ some\ event\ of\ interest \\ 0: otherwise, and \end{cases}$$

x_i be a continuous covariates observed on the i^{th} patient.

The simple logistic regression model assumes that d_i has a Bernoulli distribution with

$E[d_i|x_i] = \pi[x_i]$=exp[$\alpha + \beta x_i$]/(1 + exp[$\alpha + \beta x_i$]), ... (1)

Where α and β are unknown parameters associated with the target population. Equivalently, we can rewrite the logistic regression model using

$$logit[\pi[x_i]] = \alpha + \beta x_i \quad \text{... (2)}$$

As

$$logit[E[d_i|x_i]] = \alpha + \beta x_i \quad \text{... (3)}$$

Logistic regression is an example of a generalized linear model. These models are defined by three attributes: the distribution of the model's random component, its linear predictor, and its link function. For logistic regression these are defined as follows.

1. The random component of the model is d_i, the patient's fate. In simple logistic regression, d_i has a Bernoulli distribution with expected value $E[d_i|x_i]$.
2. The linear predictor of the model is $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_U x_{iU}$.
3. The link function describes a functional relationship between the expected value of the random component and the linear predictor. Logistic regression uses the logit link function $logit[E[d_i|x_i]] = ln\frac{E[d_i|x_i]}{1-E[d_i|x_i]} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_U x_{iU}$ ... (4)

Result and Discussion

We have used the chronic kidney disease dataset for performing the predictive analysis. This dataset is a collection of 400 instances with 23 attributes. Here 1/3rd of the data has been used as the testing set for comparing the actual and predicted outcome. The predictive model used here is the logistic regression.

Table 1. Description about the attributes

Sl. No	Attribute	Description	Permissible values
1	age	Age	age in years
2	bp	Blood pressure	in mm/Hg
3	sg	Specific gravity	(1.005,1.010,1.015,1.020,1.025)
4	al	Albumin	(0,1,2,3,4,5)
5	su	Sugar	(0,1,2,3,4,5)
6	rbc	Red blood cells	normal, abnormal
7	pc	Pus cell	normal, abnormal
8	pcc	Pus cell clumps	Present, not present
9	sc	Serum creatinine	In mgs/dl
10	bgr	Blood glucose	In mgs/dl
11	bu	Blood urea	In mgs/dl
12	Sod	Sodium	In mEg/L
13	Pot	Potassium	In mEs/L

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

14	Hemo	Haemoglobin	In gms
15	Pcv	Packed cell	In cells/cumm
16	Wc	White blood cell	In cells/cumm
17	Rc	Red blood cell	Millions/cmm
18	Htn	Hypertension	Yes, no
19	Appet	Appetite	Good, poor
20	Pe	Pedal edema	Yes, no
21	Ane	Anaemia	Yes, no
22	Dm	Diabetes mellitus	Yes, no
23	Class	Class	Ckd, notckd

From table 1 observation 250 class ckd (i.e. disease) and the class notckd kidney disease). among the 400 are under the chronic kidney 150 are under (i.e. not chronic All the variables presented in the above table have been included in this study.

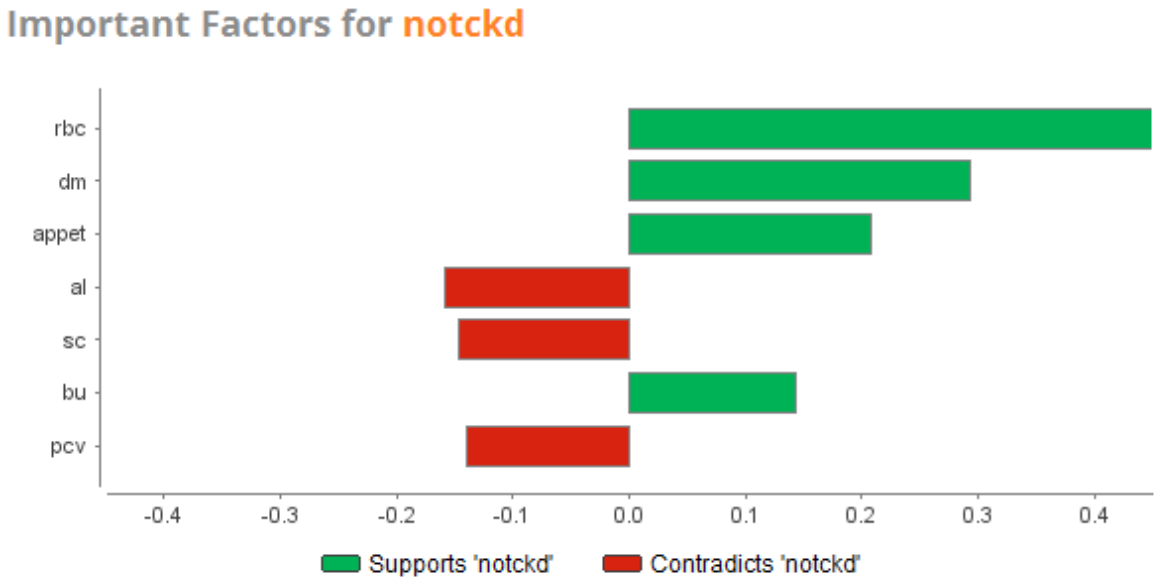

Figure 1. Important factors for prediction

Fig 1 shows that the factors al, sc and pcv are highly responsible for the patient to get chronic kidney disease. And if the factors rbc, dm, appet and bu are normal then the patients has lower chance to be in the class ckd.

Table 2. Predictors of the outcome identified by logistic regression

Variables	Coefficient	Std Coefficient	Std Error	P	Z
age	0.037	0.627	0.485	0.994	0.008
bp	0.052	0.709	0.575	0.993	0.009
sg	-804.096	-4.253	0.187	0.966	-0.043
al	6.091	8.206	0.833	0.942	0.073
su	-0.845	-0.897	0.960	0.993	-0.009
Rbc:abnormal	14.514	14.514	0.272	0.958	0.053
Pc:abnormal	-0.859	-0.859	0.281	0.998	-0.003
Pcc:present	-1.525	-1.525	0.319	0.996	-0.005
sc	1.274	4.757	0.396	0.974	0.032
bgr	0.024	1.738	0.152	0.987	0.016
bu	-0.073	-3.789	0.286	0.980	-0.025
Sod	-0076	-0.481	0.149	0.996	-0.009
Pot	0.073	0.262	0.436	0.999	0.002
Hemo	-0.616	-1.612	0.610	0.992	-0.010
Pcv	-1.525	-1.525	0.319	0.996	-0.005
Wc	-0.001	-1.641	0.030	0.984	-0.020

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

Rc	-4.352	-3.814	0.133	0.974	-0.033
Htn:no	-6.901	-6.901	0.231	0.976	-0.030
Appet:poor	9.024	9.024	0.202	0.964	0.045
Pe:yes	-0.859	-0.859	0.281	0.998	-0.003
Ane:yes	0.897	0.897	0.316	0.998	0.003
Dm:no	-12.731	-12.731	0.231	0.956	-0.055
Intercept	860.256	13.312	0.284	0.976	0.030

The coefficient, standard coefficient, p value, standard error and the Z value for each attribute has been presented in table 2. In logistic regression 0 is assigned automatically to the first category of the categorical variable and the model only estimates the coefficient value for the remaining category of that variable. For the variable age, the coefficient value 0.037 means that if the persons age is 1 unit more then he will have a 0.037 unit chance of having chronic kidney disease based on the p-value. The standard error 0.485 indicates the distance of the estimated slope from the true slope. Z-statistics 0.008 means that the predicted slope is going to be 0.008 units above the zero.

Logistic regression provides "odds" for an event. If an event has a probability p then odds of that event are $p/(1-p)$. Based on this, for a continuous variable like age the ratio is to be 1.0376 which means that for one unit change in age there is 1.0376 times increased change to be in the class ckd. And for categorical variable like pcc:present (i.e. pus cell clumps are present) the odds ratio is 0.217 which means that the patient with pcc present has 21% higher chance of getting chronic kidney disease than patient with pcc absent. Thus all the predictor variables follow the same way.

Table 3. Confusion matrix

	True notckd	true ckd
Predicted notckd	41	3
predicted ckd	2	68

Confusion matrix in table 3 is often used in analyzing the performance of the classification model that is used in the test data. Here 41+68 are the true positive and true negative values of the observations that are predicted correctly. Whereas 3+2 are the false positive and false negative values which shows contradiction with the actual outcome. Using these values the below results have been estimated.

Table 4. Evaluation result for prediction model

Accuracy	95.61%
AUC value	0.994
Classification Error	4.39%
Logistic regression classification report :	
Precision	97.42%
Recall	95.71%
F1-score	96.39%
Sensitivity	95.71%
Specificity	95.28%

The performance of the prediction model in table 4 has been evaluated which shows an accuracy of about 95.61% which means that our model has predicted 95% accurately. The precision, recall, f1-score, sensitivity and specificity all showed a good result for this model.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

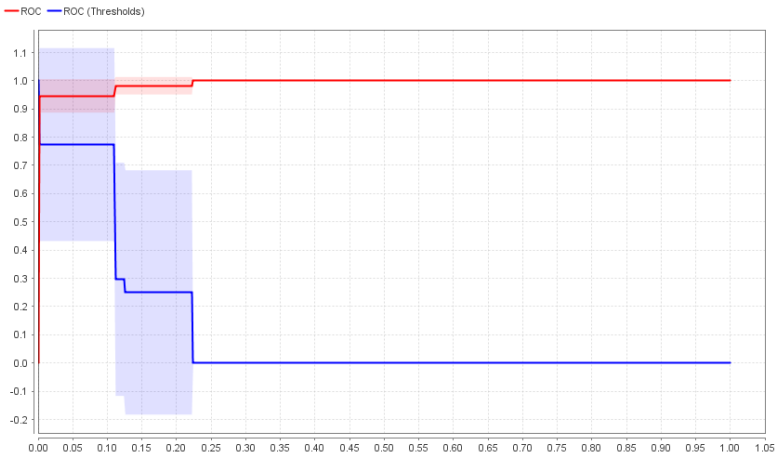

Figure 2. ROC curve for the prediction model

The ROC analysis is also performed using the prediction generated from the testing set. Fig 2 is constructed by plotting the sensitivity against the specificity for a predictive model using the probability threshold value between 0 and 1. The curve has larger value on Y axis and it travels across the top right indicates higher true positive values and lower false negative values. Area under the curve is used as a summary of the model, the score value ranges from 0 to 1. The area under the ROC curve was 0.99 which shows good prediction accuracy.

Conclusion

This study suggest in finding the significant covariates that affect the outcome of the patients. The performance of the logistic regression is assessed using the ROC curve. The area under the curve is 0.994 and the prediction accuracy is about 95%. The factors that support the patients to be in class notckd are rbc, dm, appet and bu. And the factors that contradict the patients to be in class notckd are al, sc and pcv. Thus from all these estimation logistic regression showed that it is the best survival regression model for predicting any clinical data with binary outcomes.

References

1. Abbott, R.D. (1985). *Logistic Regression in Survival Analysis*. American Journal of Epidemiology, 121(3), 465-471.
2. Ata, N & Sozer, MT 2007, '*Cox regression models with nonproportional hazards applied to lung cancer survival data*', Hacettepe Journal of Mathematics and Statistics, vol. 36, no. 2, pp. 157-167.
3. Billard, L & Dayananda, PWA 2014, '*A multi-stage compartmental model for HIV-infected individuals: I–Waiting time approach'*, Mathematical biosciences, vol. 249, pp. 92-101.
4. Christensen, E. (1987). *Multivariate Survival Analysis Using Cox's Regression Model.* American Association for the Study of Liver Diseases, 7(6), 1346-1358.
5. Clark, TG, Bradburn, MJ, Love, SB, Altman, DG 2003, *'Survival analysis part I: Basic concepts and first analyses'*, British Journal of Cancer, vol. 89, pp. 232-238.
6. Cox, D.R. (1972). *Regression Models and Life-tables.* Journal of the Royal Statistical Society. Series B, 34(2), 187-220.
7. Efron, B. (1988). *Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve.* Journal of the American Statistical Association, 83(402), 414-425.
8. Fine, JP & Gray, RJ 2012, '*A proportional hazards model for the sub distribution of a competing risk',* Journal of the American Statistical Association, vol. 94, no. 446, pp. 496-509.
9. Goldfarb-Rumyantzev, A.S., Scandling, J.D., Pappas, L., Smout, R.J., & Horn, S. (2003). *Prediction of 3-yr Cadaveric Graft Survival based on Pre-transplant Variables in a Large National Dataset*. Clinical Transplantation, 17, 485-497.
10. Hosmer, D.W., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, Second Edition, John Wiley & Sons, 92-131.
11. Karim, A, Saeed, BII, Darkwah, KF & Musah, AAI 2015, *'Analysis of chronic disease conditions in ghana using logistic regression'*, International Journal of Statistics and Applications, vol. 5, no. 4, pp. 133-140.
12. Katz, MH & Hauck, WW 1993, '*Proportional hazard (Cox) regression'*, Journal of General Internal Medicine, vol. 8, no. 12, pp. 702-711.
13. Kim, M & Li, G 2017, '*Postoperative complications affecting survival after cardiac arrest in general surgery*

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2153 – 2159
Received 05 March 2021; Accepted 01 April 2021.

patients', International Anesthesia Research society, pp. 1-7.
14. Krikov, S., Khan, A., Baird, B.C., Barenbaum, L.L., Leviatov, A., Koford, J.K., & Goldfarb-Rumyantzev, A.S. (2007). *Predicting Kidney Transplant Survival Using Tree-Based Modeling.* ASAIO Journal, 53, 592-600.
15. Meira-Machado, L, Una-Alvarez, JD, Cadarso-Suarez, C & Andersen, PK 2008, '*Multi-state models for the analysis of time-to-event data'*, Statistical Methods in Medical Research, vol. 18, pp. 195-222.
16. Oztekin, A., Delen, D., & Kong, Z. (2009). *Predicting the Graft Survival for Heart-Lung Transplantation Patients: An Integrated Data Mining Methodology.* International Journal of Medical Informatics, 78, e84-e96.
17. Park, H, Chen, C, Wang, W, Henry, L, Cook, RL & Nelson, DR 2018, *'Chronic Hepatitis C Virus (HCV) increases the risk of Chronic Kidney Disease (CKD) while effective HCV treatment decreases the incidence of CKD'*, Hepatology, vol. 67, no. 2, pp. 492-504.
18. Schober, P., & Vetter, T.R. (2018). *Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare.* International Anesthesia Research Society, 127(3), 792-798.
19. Shen, J, He, L, Li, C, Wen, T, Chen, W, Lu, C, Yan, L, Li, B & Yang, J 2016, '*Prognostic nomograms for patients with resectable hepatocelluar carcinoma incorporating systemic inflammation and tumor characteristics*', Oncotarget, vol. 7, no. 49, pp. 80783-80793.
20. Wannamethee, SG, Shaper, AG, Walker, M & Ebrahim, S 1998, '*Lifestyle and 15-year survival free of heart attack, stroke, and diabetes in middle-aged british men*', Archives of Internal Medicine, vol. 158, no. 22, pp. 2433-2440.
21. Whitehead, J 1980, '*Fitting Cox's regression model to survival data using GLIM'*, Journal of Royal Statistical Society. Series C (Applied Statistics), vol. 29, no. 3, pp. 268-275.
22. Khaw, K & Barrett-Connor, E 1986, '*Family history of heart attack: A modifiable risk factor', Circulation*, vol. 74, no. 2, pp. 239-244.
23. Breslow, NE 1975, *'Analysis of survival data under the proportional hazards model'*, International Statistical Review, vol. 43, no. 1, pp. 45-57.