A STUDY OF SIGNIFICANCE IN VARIABLE IMPORTANCE REGRESSION MODELS

Dr. DHANANJAYA REDDY*

M. VENKATARAMANAIAH**

* Assistant Professor of Mathematics, Government Degree College, Puttur, A.P

Professor, Dept. of Statistics, S.V. University, Tirupati,A.P., India

Abstract: Regression analysis is the most useful statistical methods. But, many regression models are not suited for identification of the most important regresses and ranking of the regresses, article reviews in detail the various variable importance metrics for the linear model, particularly emphasizing variance decomposition metrics. This paper studies about the social demographic risk factors from cancer patients in southern part of India who were initially treated for first primary cancer stage 1 and were cancer free for at least 1 year after first primary cancer treatment; and applying the binary logistic regression model to estimate the probability of the second cancer occurrence and to determine the significant social demographic risk factors which affect the second cancer occurrence.

Key Words: binary logistic regression method, logistic regression model, maximum likelihood estimation, odds ratio, second cancer, variable importance methods.

INTRODUCTION

Linear model metrics are illustrated through the example analysis with non-linear parametric models, several principles from linear models have been adapted, and machine-learning methods have their own set of variable importance methods. Even though, there are many variable importance metrics, there is still no convincing theoretical basis for them, and all have a heuristic touch. A dichotomous variable is a type of variable that only takes on two possible values. Some examples of dichotomous variables include: Gender: Male or Female; Heads or Tails; Residential or Commercial [1, 2, 3, 4, 5].

This paper tried to determine the Logistic regression model to identify effective risk factors that cause the second cancer occurrence and also explain the relative risk for each studied covariate and its effect on the probability of the second cancer occurrence. The main objectives of the study are (i) to collect the information of second cancer patients from the records of various cancer hospitals and (ii) to identify the factors causing re-occurrence of the cancer using advanced statistical modeling techniques.

Survey was designed to investigate the social-demographic risk factors which affect the second cancer from cancer patients who were initially treated for first primary cancer stage I and were cancer free for at least 1 year after first primary cancer treatment. The ‘Second Cancer’ as a new primary cancer in a person with a history of another cancer or same cancer in another place in the body.

The collected information consists of risk factors for various types of cancers and patient socio-demographics like current age, age at first cancer diagnosed, gender, education, family history, family income, food habits, living area, marital status, mental stress, obesity, occupation, and smoking and undergoes radiation treatment.

THE BINARY LOGISTIC REGRESSION METHOD

The logistic regression allows predicting a discrete outcome such as group membership from a set of variables that may be continuous, discrete, dichotomous, or a mix. The objective of the study is to provide a focused introduction to the logistic regression model and its use in methods for modeling the relationship between a categorical outcome variable and a set of covariates. The concept of logistic regression model has been developed from a regression analysis point of view. The difference between logistic regression models from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous and the parameters in the model can be estimated by maximum likelihood estimation [6-12].

The logistic distribution with location parameter α is given by

$$F(x) = \exp [(x - \alpha) / \beta] / \{1 + \exp - [(x - \alpha) / \beta] \}^2; \qquad -\infty < x < \infty, \beta > 0$$

If the probability of an event occurring is p, the probability is (1 – p), then the corresponding logistic transformation or legit proportion is given by

$$\text{Legit } (p) = p / (1-p) \text{ , where } p = Pr(Y = 1)$$

$$1 - p = Pr(Y = 0)$$

As p tends to 0, Legit (p) tends to -∞ and as p tends to 1, Legit (p) tends to ∞. The function Legit (p) is a sigmoid curve that is symmetric about p = 0.5.

The logistic regression function is the legit transformation of p, where

$$\text{Legit } (p) = \ln (p/(1-p))$$

$$= \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

Where β_0 = the constant of the equation

β_i = the coefficient of the predicator variable i.

Odds Ratio (OR):

A measure of association named odds ratio by which the effect of the risk factor can be observed as well as relative risk or relative importance of a factor can be performed by comparing with reference group is defined as

$$OR = (p_1 / 1 - p_1) / (p_2 / 1 - p_2) \text{ where } p_1 = \exp(\beta_0 + \beta_1) / (1 + \exp(\beta_0 + \beta_1))$$

$$p_0 = \exp(\beta_0) / (1 + \exp(\beta_0 + \beta_1)$$

GOODNESS OF FIT TEST:

To assess the significance of the logistic regression coefficient Pseudo R^2test is used. Pseudo R^2 = 1- [Log-Likelihood (B) / Log-Likelihood (0)]; this value tends to be smaller than r-square and values of 0.2 to 0.4 are considered highly satisfactory.

ANALYSIS AND RESULTS

Data used for the analysis comprised of registered patients in different hospitals, South India, by different stages of cancer. These patients met the study assumptions were classified as

1. Has a first primary cancer stage 1
2. Has at least one year free cancer after first cancer treatment

The social demographic risk factors used are age at first cancer, gender, food, income, marital status, family history, smoking, education, mental stress, food habits, family income and obesity in addition to treatment by radiation.

The dependent variable used in the study was the classified variable second cancer , explanatory variable used in this study were, age at first cancer occurrence, gender (male – female), marital status (married – single), radiation treatment of first cancer (yes – no), family history of cancer (yes – no) , smoking (yes – no), obesity before first cancer (yes – no), Occupation (Employed – Non-Employee) , education (yes – no),food habits (nutrition diet – regular diet), mental stress (yes – no) and family income (high income – low income).

SPSS Version 20.0 software package is used for the analysis. From the table-1, we can find significant variables estimated coefficients (parameter estimates), standard error of the coefficients, Wald's Chi-Square values, p-values.

Table 1: The estimated significant coefficient, its S.E and Wald test for significant variables:

Parameter	DF	Estimate	Standard Error	Wald Chi-Sq	Pr>ChiSq
Intercept	1	-2.16	0.80	7.27	0.007
Gender	1	-1.07	0.33	10.68	0.0011
Family Ever Had Cancer	1	0.79	0.30	6.94	0.0084
Education	1	-0.44	0.24	3.32	0.0684
Smoke	1	1.90	0.30	39.09	<.0001
Stress	1	0.44	0.25	3.24	0.0717
Undergo Radiation Treatment	1	-0.80	0.26	9.64	0.0019
Age	1	0.04	0.02	5.38	0.0204

The coefficient estimates are used to estimate the probability of the second cancer occurrence as follows.

$P(Y = 1 / X) = e^{x} / (1 + e^{x})$ or $1 / (1 + \exp^{-z})$

In the above table, the variables undergo Radiation Treatment, Family history, Gender and Smoke are significant at 0.01 level of significance, and increasing age is significant at 0.05 levels. Stress and Education are significant at 0.10 level of significance. Now the model to predict the probability of second cancer occurrence is: P (Second Cancer occurrence) = – 2.154 – 0.7988 * Undergo Radiation Treatment + 0.7847 * Family History + 1.9019 * Smoke – 0.4433 * Education + 0.0378 * Age + 0.4405 * Mental Stress-1.0696* Gender.

Table2: Odds Ratios and 95% Confidence Intervals for Significant Covariates

Parameter	Point Estimate of odds ratio	95% Wald Confidence Limits	
Gender	0.34	0.181	0.652
Family Ever Had Cancer	2.19	1.223	3.929
Education	0.64	0.399	1.034
Smoke	6.70	3.690	12.160
Stress	1.55	0.962	2.509
Undergo Radiation Treatment	0.45	0.272	0.745
Age	1.04	1.006	1.072

Odds ratio from table (2) indicates that : Smokers are more susceptible to develop a second cancer than non – smokers; a patient with family history is more susceptible to develop second cancer ; stress and increasing age are also susceptible to develop a second cancer; educated patients are less susceptible to develop a second cancer than non- educated patients; treatment by radiation decreases the susceptibility of second cancer, and

gender is less susceptible to develop a second cancer that is male is less susceptible to develop a second cancer than female .

MODEL SUMMARY: pseudo R-square test by Cox-Snell and Max-rescaled R-Square values are given below

Table 3: Values of Test Statistics

Test Statistics	value
R-Square	0.2227
Max-rescaled R-Square	0.306
Percent Concordant	79%
Percent Discordant	21%

From the above table, it is observe that, the pseudo – R^2 which is exploited to study the goodness of fit of multiple logistic regression is found to be 0.2227, lies between 0.2 and 0.4 which indicates a very good fit of the model and also Max-rescaled R-Square is greater than 0.3 is considered as good fits for the model being tested.

CONCLUSION

The pseudo-R^2 is used to test the goodness fit of the model. By binary logistic regression model is constructed in order to find out the probabilities of occurring the second cancer by using the social demographic risk factors such as age at first cancer, Treatment by radiation, Family history, Marital status, Education, Income, Food, Obesity, Gender, mental stress, Occupation and Smoking. By analysis, the smoking habit, family history with cancer, increasing age and stress are highly influential factors for the occurrence of second cancer. On the other side undergo the radiation treatment, gender and education causes decrease the occurrence of second cancer.

REFERENCES

1. Afifi, A. A and Clarke. V (1990): Computer Aided. Multivariate Analysis (2nd ed.), New York, Van Nostrand Rein Hold
2. Amr I. Abdelrahman: “Applying Logistic Regression Model to The Second Primary Cancer Data”, unpublished article, Department of Statistics, Mathematics, and Insurance, Ain Shams University, Egypt
3. Ashour S, Abo Elfotouh S (2005): Presentation and statistical analysis using SPSSWIN. Second Part, Advanced Applied Statistics, Institute of Statistical Studies and Research. Cairo University, Egypt (in Arabic).
4. Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importnace of predictors in multiple regression. *Psychological Bulletin, 114*, 542-551.

5. Berkson, J (1955): " Maximum likelihood and Minimum Chi – Square estimates of the Logistic Function", Journal of American Association, Vol 50, PP : 130 – 162

6. COX, D. R (1958) : "The Regression Analysis of Binary Sequences", Journal of the Royal Statistical society series B (Methodological), Vol – XX, No. 2, PP : 215 – 242

7. Christersen, R (1997): Log Linear models and logistic regression, second Edition, Springer – Verlog, New York

8. Draper, N. R., and H. Smith; Applied Regression Analysis, 3d ed., John Wiley& Sons, New York, 1998

9. Homer, David and Leme show Stanley (2000): Applied Logistic Regression, Second Edition, A Wiley – Inter Science Publication, John Loiley & Sons, Inc.

10. Heinze, G.; Wallisch, C.; Dunkler, D. Variable selection–a review and recommendations for the practicing statistician. *Biom. J.* **2018**, *60*, 431–449

11. Johnson, J. W. (2000). A heuristic method for estimating the reltive weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*, 1-19.

12. Thomas, D.R.; Zhu, P.; Zumbo, B.D.; Dutta, S. On measuring the relative importance of explanatory variables in a logistic regression. *J. Mod. Appl. Stat. Methods* **2008**, 4-7.