Vol.10.Issue.4.2022 (Oct-Dec)

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

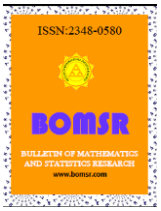
BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

SOME CONSIDERATIONS FOR VARIABLE IMPORTANCE IN REGRESSION MODELS

Dr. S. Mohan Prabhu
Assistant Professor and Head, Department of Statistics,
Muthayammal College of Arts and Science (A Unit of Vanetra Group),
Rasipuram, Namakkal, Tamilnadu-637408, India.
Email:mohanprabhusu@gmail.com
DOI:10.33329/bomsr.10.4.50



ABSTRACT

In this paper to give the effect of demographic variables on second cancer patients in this study using supervised machine learning methodologies. The social and demographic risk factors for developing a second malignancy were studied in this study. The following algorithms were investigated: RD-RANSAC, k-Nearest Neighbor, Nave Bayes, Random Forest, and Neural Network. The socio-demographic risk factors retrieved from the 1000 samples used are Gender, Undergo Radiation Treatment, Family Ever Had Cancer, Smoke, Diet & Exercise, Employment, Marital Status, Education, Income, Stress, Area, Age (in intervals), Age at First Cancer (in intervals), and Obesity.

Key words: cancer cell, k-nearest neighbor, nave bayes, random forest, neural network.

AMS Subject Classification: 46L53, 62C10, 14F05, 92B20.

1. INTRODUCTION

Regression is a term created by Francis Galton. The average height of children born to parents of a specific height tended to shift or "regress" toward the average height of the population as a whole, Galton observed in a well-known research. This was true even though tall parents had tall children and short parents had short children. In other words, children born to parents who are extremely tall or short tend to be closer to the demographic average in height.

Karl Parson, a Galton buddy who collected over a thousand records of family group heights, confirmed the universal regression law. Tall and short sons "regressed" toward the average height of all males, according to his studies. He noticed that the average height of sons of tall dads was lower than that of their fathers, whereas the average height of sons of short fathers was greater than that of their fathers. According to Galton, this was a "regression to mediocrity."

Regression analysis, which has various applications in the applied sciences, is the most popular topic in mathematical statistics. Regression analysis is used in a variety of domains, including engineering, physical and chemical sciences, economics, management, life and biological sciences, and social sciences. In the applied sciences themes of forecasting and prediction, regression acts as an indicator. Regression is currently regarded to be a strategy that connects one variable, known as the dependent variable, to at least one other variable, known as the independent variable, with the purpose of estimating and/or projecting the former's mean value in relation to the latter's known (fixed) value(s).

On the basis of the independent variables, a regression model may be used to define, predict, and regulate the dependent variable. Both quantitative and qualitative independent variables may be used in a regression model. An independent variable with a numerical value that is along the real line is said to be quantitative. Non-numerical variables are qualitative independent variables. The location of the conditional expectation of the dependent variable versus fixed values of the explanatory variable is represented by the geometric regression curve (s).

Regression analysis is a statistics subject. It is useful in scientific endeavors that frequently require statistical testing and estimation of behavioral hypotheses. Given that the name "regression" alludes to a dependence relationship, an econometric model with only one equation can be called a regression model. Regression models are a type of statistical model, much like econometric models are a type of economic model. A regression model is a statistically dependent relationship. Regression models may clearly be employed in a wide range of physical and behavioural fields.

Regression's dependence relation definition is far too general to be of any assistance in understanding the term's more technical meaning in today's context. The model is often expressed as an equation with all the variables having clearly defined statistical features. There are several methods for the analysis, but the least squares approach is the one that is most usually employed.

2. PRELIMINARIES

Regression analysis is a statistics subject. It is useful in scientific endeavors that frequently require statistical testing and estimation of behavioral hypotheses. Given that the name "regression" alludes to a dependence relationship, an econometric model with only one equation can be called a regression model. Regression models are a type of statistical model, much like econometric models are a type of economic model. A regression model is a statistically dependent relationship. Regression models may clearly be employed in a wide range of physical and behavioural fields.

The definition of regression's dependency connection is considerably too broad to be useful in comprehending the term's more technical meaning in today's context. The model is frequently stated as an equation, with all variables having well-defined statistical properties. There are other ways for analyzing data, but the least squares method is the most commonly employed.

Relationship research involves two components. The first is how connections are made, and the second is the strength of those relationships. In most situations, correlation for quantitative elements

and association for qualitative variables decide the strength. The connection pattern reveals how much one variable will change quantitatively for every unit change in the other. When numerous independent variables are present, the regression becomes multiple. The structure and expression of the regression may alter depending on the number of independent variables and the type of the dependent variable.

To answer the question "How much does one variable change for a certain change in one or more of the other variables?" you must first understand the nature of the connection. Regressions can be linear or non-linear, and both are widely used. The utility and meanings of linear connections are investigated under two key categories. Both the first and second contain methods for a single dependent variable, one of which is quantitative and the other qualitative.

A simple linear regression is the most common type of regression. This is only supported if the dependent variable grows or decreases at a steady rate as the regression result increases. Although the "optimal" value of the dependent variable predicted by the regression model may be far from reality. The usefulness of the model is determined by the successful selection of the regression. All persons who have an impact on the dependent variable should be included. However, there should be no substantial association between the explanatory elements.

A regression model can only reliably predict each of the answer variables if the explanatory variable values are within the range utilized to build the regression equation. If one is unable to judge the statistical significance of the model, which may occur if the assumptions are violated, its practical relevance is greatly decreased.

The steps for fitting regression for numerical variables are as follows:

(1) Identifying the dependent variable and the independent factors

(2) Specifying the type of regression (linear or non-linear) to be studied

(3) Estimating the regression coefficients

(4) Testing the model's goodness of fit

(5) The importance of each regression coefficient

(6) Estimating the residuals to ensure the assumptions' validity.

A model is considered more accurate when it includes fewer explanatory variables yet a sufficient R2 value. To pick the best explanatory variables, many approaches such as stepwise regression, backward estimation, and forward selection can be utilised. In the forward selection, the coefficient of a variable with the highest statistical significance (i.e., the lowest p-value) is put first, followed by the coefficients with progressively decreasing statistical significance.

Backward estimating entails include all explanatory factors in the first step and then deleting them one by one, beginning with the least significant. The third process, known as step-wise regression, reassesses the effectiveness of previously input elements and removes those that lose significance with the addition of additional explanatory variables. For models containing qualitative dependent variables, the techniques of logistic regression and the log-linear model are applied.

2.1 Variable selection

During the regression analysis, we assumed that the model's regressor variables were significant. We focused on approaches for ensuring that the model's functional form was correct and

that the underlying assumptions were followed. In certain circumstances, theoretical considerations or past information may help in selecting the regressor to include in the model. Even if the analyst frequently has a pool of prospective regressors that should comprise all of the key aspects, the precise subset of regressors that should be included in the model must be decided. The "Variable Selection Problem" refers to the difficulty of picking an appropriate selection of regressors for the model.

2.2 Variance inflation factor

The wording is the fault of Marquardt. Every term in the model has a variance inflation factor (VIF) that computes the total influence of the regressors' dependencies on the term's variance. When explanatory factors are considered to be uncorrelated with the remaining explanatory variables, the variance of predicted regression coefficients is defined as the ratio of the actual variance to the variance that would have occurred in the absence of the explanatory variables.

The VIF implicitly compares the current situation to an ideal condition, with an ideal situation being one in which all explanatory variables are uncorrelated with the remaining explanatory variables and each other. The VIF of each explanatory variable is calculated independently.

Only in extremely extreme situations, such as those where R2 = 1.0 or the lowest eigen value is very near to zero, are the VIFs effective for removing certain variables and imposing parameter limitations. Indicators of the degree of multicollinearity include the greatest VIF among the independent variables and the mean of the independent variables' VIF. Multicollinearity may significantly affect the least squares point estimates if the greatest VIF is bigger than 1.0 or if the mean of the VIF is much higher than 1. For instance, if the VIF for the three-variable models is higher than that for the two-variable model, the multicollinearity of the three-variable model will be higher.

2.3 Path analysis

The route coefficient analysis technique is essentially a tool for analysing the breakdown of a correlation coefficient into direct and indirect effects within a framework of causal interactions among linearly related variables. The path coefficient analysis has observed the direct impacts of independent variables and their indirect effects through other independent variables with which the former is related to evaluate the relative relevance of causative factors and estimate the correlation coefficient. We provide a path diagram to illustrate the path that the causal components traverse.

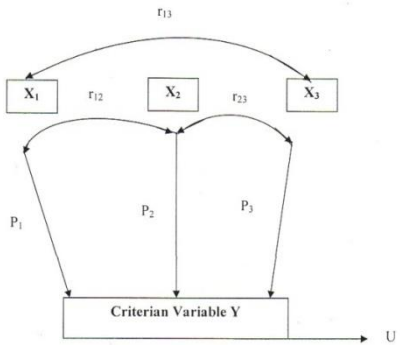


Fig 1: Path co-efficient model

The techniques of path coefficients may be facilitated as productive search in determining relative importance by using a few principles for tracing, connecting the path between two or more variables in a causal connection.

3. METHODOLOGY

3.1 Standardized technique

Let us consider the following linear multiple regression of Y on $X_1, X_2, \ldots, X_k$ as

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K \quad \ldots (1)$

The model is anticipated to be properly defined and fulfill the classical linear regression (CLR) criteria when Xi is changed by one unit while the other variables stay constant. The regression coefficient is also referred to as partial regression coefficients to emphasize the fact that the projected effect of changing a variable is reliant on the other variables in the regression equation being constant. The R^2 is commonly used to evaluate how well variables predict Y when the independent variables are uncorrelated. R^2can be divided between the independent variables, using the definition as follows:

$R^2 = \frac{Explained\ sum\ of\ squares}{Total\ sum\ of\ squares}$

3.2 Presentation and Discussion of Empirical Results

This section goes into great detail on standards and partial standardization procedures. As stated in section, the issues given by standardized techniques are recommended to be handled by a partial standardization strategy. We show this numerically by utilizing data from sixteen patients at a community health centre within a private hospital in Tamilnadu, where blood pressure is regressed on weight and height. For this problem, patients who are considered for our model to execute the study, have been divided into four groups with height held constant. The proposed model is,

Blood Pressure $= \alpha + \beta_1$ height $+ \beta_2$ weight $\quad \ldots(2)$

3.3 Path Analysis

The path analysis can be demonstrated as follows:

Let $X_1, X_2, \ldots X_k$ be measurements of K(clinical) factors supposed to influence the effect measured by Y. Let the effect Y be the sum total of these causal factors $X_1, X_2, \ldots X_k$ and a residual U. That is

$Y = X_1 + X_2 + \ldots + X_k + U \quad \ldots (3)$

Table 1 : Blood Pressure, Height and Weight for 16 Patients

Group	Blood Pressure(B.P.)	Height	Weight	Standard deviation (S^*) of Weight within thegroup
1	106	140	51	
1	120	150	44	
1	126	160	65	4.1205
1	124	140	61	
2	126	145	56	
2	104	155	56	
2	120	145	66	6.2655
2	120	135	67	
3	122	160	56	
3	120	150	61	
3	122	140	66	5.1183
3	126	150	64	

4	124	165	60	
4	120	165	62	
4	124	155	62	3.1515
4	120	155	60	

Table 2: Blood Pressure, Height and Weight for 16 Patients

Group	Blood Pressure(B.P.)	Height	Weight	Standard deviation (S^*) of Weight within thegroup
1	126	160	56	
1	130	160	62	
1	136	160	66	4.3205
1	134	160	64	
2	136	165	66	
2	124	165	55	
2	130	165	60	7.3655
2	140	165	72	
3	130	170	60	
3	130	170	64	
3	132	170	70	6.2183
3	136	170	74	
4	134	175	72	
4	140	175	78	
4	134	175	74	3.6515
4	140	175	80	

3.4 Machine Learning Techniques

Statistical model fitting, an older topic, is frequently closely connected to machine learning. Building good probabilistic models is critical to machine learning's goal of extracting useful information from a corpus of data. Machine learning is the process of programming computers to maximize a performance criterion using sample data or existing knowledge.

The anticipated model accuracy can be the optimized criterion in a modeling problem, but the fitness or evaluation function value can be the optimized criterion in an optimization problem.

In a modeling problem, the term "learning" refers to running a computer programme to generate a model using training data or prior information. Because the purpose of machine learning is to derive inferences from samples, statistical theory is applied while developing computational models.

The two main parts in this approach are inducing the model by analyzing a large amount of data and efficiently describing the model. In general, machine learning strategies fall into two categories:

Supervised Learning, where the output has been given a priorilabeled or the learner has some prior knowledge of the data; and Unsupervised Learning, where no prior information is given tothe learner regarding the data or the output.

3.5 Supervised Machine Learning

A classification problem involves classifying a collection of items into groups. Upon submission of an element from the collection, a class is established based on the attributes of the element and a set of classification rules. The only data provided is a list of labeled samples when this set of criteria is unknown in the real world (i.e., a set of instances associated with a class). Using supervised classification paradigm approaches, the classification rules are derived from the data.

We are seeking for second cancer identification in our second cancer prediction experiential study, and the presence or absence of second cancer may provide instructional value for the class. Because this is a supervised classification problem, a set of labelled instances, or sequences of whether a second malignancy occurs or not, together with their label, is required. After gathering the training data, a classifier may be built. After being trained, the classifier may be used to classify new sequences by inputting the nucleotide present at each location into the classifier and receiving the supplied label as an output.

3.6 Classification and Prediction Methodologies

The existence or absence of a second cancer may have educational value for the class in our experiential research on second cancer prediction. We are seeking for second cancer detection. It is necessary to have a set of labeled instances, or sequences of whether a second malignancy occurs or not, together with their labels, as this is a problem involving supervised classification. After gathering the training data, a classifier may be produced. The classifier may be used to label new sequences after being trained by providing the nucleotide present at each location and receiving the supplied label as an output.

There are a number of learning algorithms that might be discussed in this section despite the fact that machine learning is a relatively young field of research; nevertheless, just five strategies that are frequently used to solve data analysis issues (usually classification) are mentioned.

3.7 Support RD-RANSAC classifier

Support RD-RANSAC (Robust Distance- Random Sample Consensus) technology was created by Ravi.J (2017). The RD-RANSAC approach organizes data by creating a multidimensional hyper plane that maximizes the margin between the two data clusters and provides the best discrimination between two classes. By converting the input space into a multidimensional space using certain nonlinear functions, or kernels, this method offers great discriminative capability.

To prevent over fitting, this method is based on locating linear hyper planes in kernel space and input space. The SVM classifier determines the hyper plane (P0) bisecting the nearest points of the data that can be linearly separated if the training sample data consists of n pairs (x_1, y_1), (x_2, y_1),..., (x_n, y_n) with xi R_p and y_{i-1}, $_1$. The definition of the P_0 is

$(P_0) = \{x: f(x) = x'\beta + \beta_0 = 0\}$.....(3)

Classifier creates a parallel hyper plane P_1 which is defined as

$(P_1) = \{x': f(x') = x' \beta + \beta_0 = -1\}$ … (4)

On a point in the class -1 closest to P_0 and second hyper plane P_2 isdefined as

$(P_2) = \{x: f(x) = x\beta + \beta_0 = 1\}$ on a point in class closet to P_0.

If the data set cannot be split, this approach puts it into a higher dimensional space where the training set may be separated using a transformation or extended feature space. Linear or Gaussian kernels are typically utilized in this context.

3.8 k- Nearest neighbor classifier

The k-Nearest Neighbor (kNN) algorithm employs distance measurement techniques. kNN assigns the most common class label among the training examples to the test sample after selecting k instances from the training data that are most similar to the test case. When categorising a new sample, the distance to each characteristic in the training data must be calculated. Only the k closest examples from the training set are evaluated further. The phrase "closest" is defined using a distance metric, such as Euclidean distance. The unknown example is assigned the most frequent class among its k nearest neighbours for k-nearest neighbour categorization.

3.9 Random forest classifier

Breiman recently created the classification or regression tool "Random Forest." It consists of many tree predictors that were each built separately. The strategy is clear to understand and has worked well as a nonlinear tool. The decision tree ensemble learning framework underlies the RF method. By merging many learning machines, ensemble learning is useful for increasing model accuracy.

h(X, k), k = 1, K, where k are independent identically distributed random vectors, is a classifier built up of tree-structured classifiers. Each tree casts one unit of vote for the input X's most popular class. The method enables users to replicate in the initial phase of RF and construct k- data sets from the original data. For each set of bootstrapped data, a decision tree is constructed in the second step, and the solution is selected by a majority vote. The impurity index, which depicts the fluctuation of data at the node, is decreased by RF. Each piece of data becomes identical as the index gets closer to 0. RF uses the Gini variance index as the impurity index.

Where, *i(t)* is Gini variable index at node t, P(j/t) is the ratio of j to allthe samples at node t, c is the number of clusters.

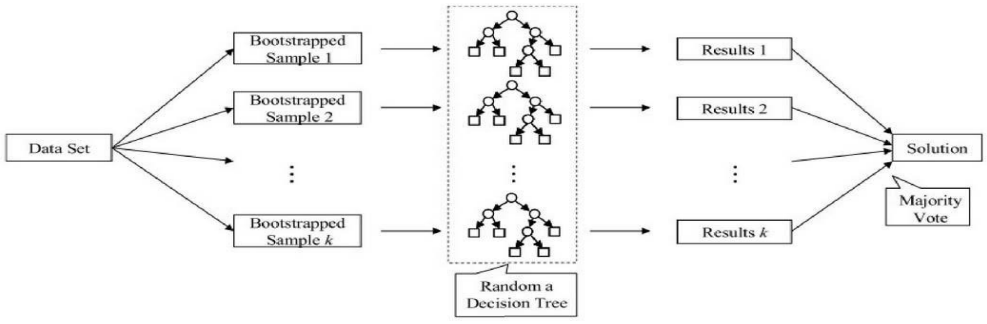


Fig. 2 The structure of Random Forest

The decrease of the impurity is defined as the difference of impurity between the parent and children nodes. It may be written as: $\Delta i(t)=i(t)-P_L i(t_L)-P_R i(t_R)$

Where PL i(tL) and PR i(tR) represent the proportion of data at the left- and right-side children nodes, respectively, to all the data. The relevance of the variable is a key factor in determining which variable is more crucial. It concentrates on each variable's impurity reduction. You may write it like this:

3.10 McNemar's Test:

This test is used to determine the model's relevance. It will be calculated using a 2x2 confusion matrix. Two binary variables' marginal homogeneity is studied.

3.11. Kappa Statistics:

Another statistic utilised in this study is the Kappa coefficient (K). Kappa is a number between 0 and 1, with 0 representing agreement based completely on chance. The total agreement between the two data sets is given as 1, or 1. Although negative values are feasible, they are untrue. It is commonly represented as a percentage (%). The Kappa statistic is a more sophisticated measure of classifier agreement that provides better interclass discrimination than overall accuracy. The Kappa statistic value is shown below.

4. Research Results

Using the same variables as the logistic regression formulation, the five approaches listed above are applied to the Second Cancer Data. Gender, Undergo Radiation Treatment, Family Ever Had Cancer, Smoke, Diet & Exercise, Employment, Marital Status, Education, Income, Stress, Area, Age (in intervals), Age at First Cancer (in intervals), and Obesity are the explanatory factors revealed from 836 samples.

Table 3: Variable importance /weight ages given by each supervised machinelearning methods

Explanatory Variable	Machine Learning Methods				
	Support Vector Machine	k Nearest Neighbor	NaïveBayes	Random Forest	Neural Network
Smoke	100.00	100.00	65.6	38.70	23.11
Age Interval	72.67	72.67	64.07	24.94	28.47
Family Ever Had Cancer	28.05	28.05	57.25	8.21	8.46
Undergo Radiation Treatment	23.57	23.57	42.42	10.15	10.14
Income	6.16	6.16	46.2	0.53	2.08
Diet & Exercise	5.05	5.05	46.48	1.65	3.24
Employment	3.35	3.35	51.75	14.79	1.15
Area	2.14	2.14	52.33	0.09	1.63
Stress	0.89	0.89	51.49	2.51	3.00
Education	0.68	0.68	51.32	3.78	1.87
Obesity	0.40	0.40	51.01	-1.58	0.35
Gender	0.04	0.04	49.62	19.74	14.52
Marital Status	0.00	0.00	50.26	7.50	1.99

Table 3, which uses supervised machine learning techniques, shows that smoking, becoming older, and family history have the greatest effects on the risk of getting a second cancer.

Table 4: Measure of Model performance metrics

Methods	Accuracy	Kappa	Mcnemar's test p-Value	Sensitivity	Specificity
Support VectorMachine	0.689	0.3587	0.7143	0.7143	0.6763
k Nearest Neighbor	0.7368	0.4609	0.0000	0.8	0.705
Naïve Bayes	0.7177	0.3332	0.0003	0.4786	0.8381
Random Forest	0.8313	0.6456	0.0000	0.8857	0.804
Neural Network	0.7273	0.4255	0.0000	0.725	0.7284

The Random Forest technique beat the other approaches with 83% accuracy, followed by the k-Nearest Neighbor, Neural Network, and Nave Bayes approaches. Random Forest has a good Kappa score (0.6456) in comparative statistics when compared to other learning systems. According to Mcnemar's test p-Value, Random Forest, k-Nearest Neighbor, and Neural Network outperform other approaches on the supplied dataset.

Table 2 reveals that, when compared to the other supervised algorithms, Support RD-RANSACand Nave Bayes perform poorly on the second cancer data.

Conclusion

Using supervised machine learning approaches, we investigated the influence of demographic characteristics on second cancer patients in this study. In this study, social and demographic risk factors for having a second cancer were investigated. Support Vector Machine, k-Nearest Neighbor, Nave Bayes, Random Forest, and Neural Network were all investigated. Gender, Undergo Radiation Treatment, Family Ever Had Cancer, Smoke, Diet & Exercise, Employment, Marital Status, Education, Income, Stress, Area, Age (in intervals), Age at First Cancer (in intervals), and Obesity are the socio - demographic risk variables obtained from the 836 samples utilised.

We discovered that smoking, age, and family history are the most important factors of developing the risk of second cancer in our study utilising supervised machine learning approaches. Our research shows that Random Forest outperforms other strategies for the given dataset among all Supervised Machine Learning Methods. Furthermore, Support RD-RANSACand Nave Bayes do not outperform the other learning algorithms.

References

[1]. Agarwal, S. et al., 2002, Birth weight patterns in rural undernourished pregnant women, *Indian Pediatrics,* Vol. 39, March, p. 17.

[2]. Barone S., Lombardo A. and Tarantino P., 2007, A weighted logistic regression for conjoint analysis and Kansei engineering, *Quality and Reliability Engineering International*, Vol. 23 Issue 6, pp. 689 – 706.

[3]. Biswas, S.C., et al., 2000, Impact of some biosocial variables on infant and child mortality, Demography India. Vol. 29, No. 2, pp. 211-221.

[4]. Chernick, M., R., 2008, Bootstrap Methods, A Guide for Practitioners and Researchers. Second Edition, Wiley, Inc., New York.

[5]. Ferri C, Flach P., Hernandez-Orallo J, 2002, "Learning Decision Trees Using the Area under the ROC Curve", Nineteenth International Conference on Machine Learning (ICML 2002); Morgan Kaufmann; pp. 46-139.

[6]. Hazarika et al., 2000, Differences in blood pressure level and hypertension in three ethnic groups of northeastern India, *Asia-pacific journal of public health*, Vol. 12, No. 2, pp. 71-78.

[7]. Ravi.J & et. al., 2017 , A robust measure of pair-wise distance estimation approach: RD-RANSAC, *International Journal of Statistics and Applied Mathematics*, 2(2): 31-34.

[8]. Ravi.J & et. al., 2018, A New Robust Distance Machine Leaning Approach to Edge and Corner Detection, RD- FAST, *Journal of Statistics and Mathematical Engineering*, 4(3), 31-44.

[9]. Weber, S., O.; and Michalik G., W., 2008, Incorporating sustainability criteria into credit risk management, *Business Strategy and the Environment*, Vol. 19 Issue 1, pp. 39 - 50.