JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

Determination of Significant Variables of Medical Data Through Machine Learning Approach

Parthasarathy Sundararajan

SRM IST, Ramapuram,Department of Mathematics,Chennai-89

Email: parthass@srmist.edu.in

Received: 14 Feb 2020 Revised and Accepted: 25 March 2020

ABSTRACT: In this article we are using diabetes data to calculate p value separately for every independent variable present in the data. This calculation done through ordinary least square estimation method. According to our p value we have selected highly significant variable to achieve expected model accuracy.

KEYWORDS: Human resources policies, HRM, UCB's. HR, Performance.

I. INTRODUCTION

The industry of health care plays a vital role in offering valued based service to the entire society, which at the same time makes some countries, top revenue earners. The word health care always associates with some terms such as quality, value and outcome[2]. Nowadays healthcare promise, a lot and simultaneously innovations are required to keep up the promise[1]. But such innovations are just seconds apart today even though population burden plays very critical role. Some more critical parameters include billing, patient record and patient care etc, where in today's technology alternate staffing methods, IP capitalization techniques are applied enabling smart healthcare thus minimizing administrative and supply cost[6]. Slowly ML is taking up a major role in the healthcare industry where as it is partially playing a vital role in various other situations in the field of healthcare[4]. ML is majorly seen in this analysing n number of data thus providing suggestions, risk scores which includes various other applications. This article focusses on single application of ML which would change the basic aspect of healthcare in 2019 and beyond[3]. This will ensure hand in hand development of data analysis and innovation. This can ensure helping millions of people even without their knowledge, by integrating data of various patient of various cases belonging to different countries just to merely improve the present treatment facility[5].

II. Ordinary Lease Square - Regression

In statistics, **ordinary least squares** (**OLS**) is a type of linear least squares method for estimating the unknown parameters in a linear regression model[15]. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function[14].

2.1 Linear Model

Consider the data consists of n observations $\{y_i, x_i\} i = 1\ to\ n$.There exists a scalar response y_i and a column vector x_i of p parameters $x_{ij}$for j=1,…p, for each and every observation of i. the response variable,,$y_i$is a linear function of the regressors In linear regression model.

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots a_p x_{ip} + \varepsilon_i \quad (1)$$

The vector form of the above equation is

$$y_i = x_i^T A + \varepsilon_i \quad (2)$$

where xi is a column vector of the i^{th} observations of all the explanatory variables; A is a p×1 vector of unknown parameters; and the scalarsε_i represent unobserved random variables, which account for influences upon the responses y_i from sources apart from the explanators xi. This model also can be written in matrix notation as

$$y = XA + \varepsilon, \quad (3)$$

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

where y and ε are $n\times1$ vectors of the values of the response variable and the errors for the various observations, and X is an n×p matrix of regressors, also sometimes called the planning matrix, whose row i is x_i^{T} and contains the ith observations on all the explanatory variables[8].

As a rule, the constant term is usually included within the set of regressors X, say, by taking $x_{i1} = 1$ for all $i = 1, ..., n$. The coefficient A_1corresponding to this regressor is called the *intercept.*

Here relationship of desired kind can be availed. But it is to be clearly noted that linear regression cannot exist. In particular, the third regressor could be the square of the second regressor. In that case, the model would be quadratic in the second regressor[16], but since the model is linear in the parameter A, it will be considered as linear model.

2.2 Estimation

Suppose 'c' is a static value for the parameter vector 'a'. The quantity $y_i - x_i^{\mathrm{T}}c$, called the residual for the i-th observation, measures the vertical distance between the data point (x_i, y_i) and the hyperplane $y = x^{\mathrm{T}}c$, and thus assesses the degree of fit between the particular data and therefore the model. The sum of squared residuals (SSR) (also called the error sum of squares (ESS) or residual sum of squares (RSS)) may be a measure of the general model fit[10]:

$$S(c) = \sum_{i=1}^{n}(y_i - x_i^T c)^2 = (y - Xc)^T(y - Xc), \qquad (4)$$

where T denotes the matrix transpose and therefore the rows of X, denoting the values of all the independent variables related to a specific value of the variable, are $X_i = x_i^{\mathrm{T}}$. The value of 'b' which minimizes this sum is called the OLS estimator for 'A'. The function $S(c)$ is quadratic in 'c' with positive-definite Hessian, and therefore this function possesses a unique global minimum at $c = \hat{A}$, where $\hat{A} = argmin_{c\in R^p}S(c) = (X^TX)^{-1}X^Ty$.

The product $N=X^{\mathrm{T}}X$ is a normal matrixand its inverse, $Q=N^{-1}$, is the *cofactor matrix* of 'A', closely related to its covariance matrix, c_A. The matrix $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}=Q\,X^{\mathrm{T}}$ is called the Moore–Penrose pseudoinverse matrix of X. This above alignment clearly depicts that, when, or only when, imperfect multicollinearity prevails between explanatory variables, estimation can be carried out. After we have estimated 'A', the fitted values (or predicted values) from the regression will be

$\hat{y} = X\hat{A} = By$, where $B = X(X^TX)^{-1}X^T$is the projection matrix onto the space V spanned by the columns of X. This matrix P is additionally sometimes called the hat matrix because it "puts a hat" onto the variable y[18]. Another matrix, closely related to P is the annihilator matrix $N = K_n - B$:

this is a projection matrix onto the space orthogonal to vector. Both matrices B and N are symmetric and idempotent(meaning that $B^2 = B$ and $N^2 = N$), and relate to the data matrix X via identities $BX = X$ and $NX = 0$. Matrix M creates the residuals from the regression:

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{A} = Ny = N(XA + \varepsilon) = (NX)A + N\varepsilon = N\varepsilon \qquad (5)$$

Using these residuals we can estimate the value of σ^2 using the reduced chi-squared statistic:

$$s^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p} \qquad (6)$$

$$\hat{\sigma}^2 = \frac{n-p}{n}s^2, \qquad (7)$$

where n-p, is the statistical degrees of freedom. The first quantity, s^2, is the OLS estimate for σ^2, whereas the second $\hat{\sigma}^2$ is the MLE estimate for σ^2. The two estimators are quite similar in large samples; the primary estimator is usually unbiased, while the second estimator is biased but features a smaller mean squared error[17]. In practice s^2 is employed more often, since it's more convenient for the hypothesis testing. The square root of s^2 is called the regression standard error, standard error of the regression, or standard error of the equation[7].

It is common to assess the goodness-of-fit of the OLS regression by comparing what proportion the initial variation within the sample are often reduced by regressing onto X[13]. The coefficient of determination R^2 is defined as a ratio of "explained" variance to the "total" variance of the dependent variable y, in the cases where the regression sum

of squares equals the sum of squares of residuals

$$R^2 = \frac{\sum(\hat{y_i}-\bar{y})^2}{\sum(y_i-\bar{y})^2} = 1 - \frac{RSS}{TSS} \qquad (8)$$

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

where TSS is the total sum of squares for the dependent variable, $L = I_n - 11^T/n$, and 1 is an $n \times 1$ vector of ones. (L is a "centering matrix" which is equivalent to regression on a constant; it simply subtracts the mean from a variable.) In order for R^2 to be meaningful, the matrix X of knowledge on regressors must contain a column vector of ones to represent the constant whose coefficient is that the regression intercept[12]. In that case, R^2 will always be a number between 0 and 1, with values close to 1 indicating a good degree of fit

III. Diabetes Dataset

Reference	http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html
Source	These data are courtesy of Dr John Schorling, Department of Medicine, University of Virginia School of Medicine
References	Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* 90:814-820; 1997
Description of the Data	The data consist of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. According to Dr John Hong, Diabetes Mellitus Type II (adult onset diabetes) is associated most strongly with obesity. The waist/hip ratio may be a predictor in diabetes and heart disease. DM II is also agssociated with hypertension - they may both be part of "Syndrome X". The 403 subjects were the ones who were actually screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes.
Format	http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/Cdiabetes.html

Using Feature reduction technique based on p value and Output of OLS is as below

IV. Output for OLS

OLS Regression Results

Dep. Variable:	y	R-squared:	0.063
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	1.866
Date:	Tue, 28 Apr 2020	Prob (F-statistic):	0.0285
Time:	23:57:01	Log-Likelihood:	-2867.9
No. Observations:	403	AIC:	5766.
Df Residuals:	388	BIC:	5826.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	843.3350	385.877	2.186	0.029	84.663	1602.007
x1	-0.0033	0.001	-2.521	0.012	-0.006	-0.001
x2	0.1292	0.667	0.194	0.846	-1.182	1.440
x3	-1.0433	0.435	-2.396	0.017	-1.900	-0.187
x4	1.4072	2.071	0.680	0.497	-2.664	5.479
x5	-3.0585	23.274	-0.131	0.896	-48.818	42.701
x6	27.6141	10.760	2.566	0.011	6.459	48.769
x7	-0.7759	1.192	-0.651	0.515	-3.119	1.567
x8	0.8751	4.307	0.203	0.839	-7.592	9.343
x9	-0.3269	0.827	-0.395	0.693	-1.952	1.299
x10	-0.8805	1.144	-0.769	0.442	-3.130	1.369
x11	-0.0867	1.700	-0.051	0.959	-3.428	3.255
x12	0.3030	1.593	0.190	0.849	-2.830	3.436
x13	-5.2527	2.823	-1.861	0.064	-10.803	0.298
x14	0.7132	5.687	0.125	0.900	-10.468	11.895

Omnibus:	66.165	Durbin-Watson:	1.854
Prob(Omnibus):	0.000	Jarque-Bera (JB):	95.791
Skew:	1.093	Prob(JB):	1.58e-21
Kurtosis:	3.963	Cond. No.	5.08e+05

V. Discussion and Conclusion

Mathematical probabilities like **p-values range** from 0 (no chance) to 1 (absolute certainty). So 0.5 means a 50 per cent chance and 0.05 means a 5 per cent chance. ... If the **p-value** is under 0. 01, results are considered statistically significant and if it's below 0.005 they are considered highly statistically significant variable[9].

In this case, p-value for column variable x3 which is stab.glu is having a p-value of 0.017. this will become high statistically significant column and can be used for predicting the results.

We are using Feature Reduction Technique. What it means is that to identify and eliminate irrelevant information from the data set using p-value

It is also suggested to eliminate one variable at a time (which has high p-value) and rerun the OLS program with the newer set, will give us different output. Our aim is to eliminate highest p-value backward and find out most accurate result[11].

The American Statistical Association (ASA) published these recommendations on the proper use of the p value:

- p-values can indicate how incompatible the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

VI. REFERENCES

[1] Aiken L, Clarke S, Sloane D, Sochalski J, Busse R, Clarke H, Giovannetti P, Hunt J, Rafferty A, Shamian J. 2001. Nurses' report on hospital care in five countries. Health Affairs 20(3):43–53

[2] Aiken LH, Smith HL, Lake ET. 1994. Lower Medicare mortality among a set of hospitals known for good nursing care. Medical Care 32:771–787.

[3] Andrulis DP, Kellermann A, Hintz EA, Hackman BB, Weslowski VB. 1991. Emergency departments and crowding in United States teaching hospitals. Annals of Emergency Medicine 20(9):980–986

[4] Baker DW, Shapiro MF, Schur CL. 2000. Health insurance and access to care for symptomatic conditions. Archives of Internal Medicine 160(9):1269–1274.

[5] Burnham, Kenneth P.; David Anderson (2002). *Model Selection and Multi-Model Inference* (2nd ed.). Springer. ISBN 0-387-95364-7.

[6] Draper DA, Hurley RE, Lesser CC, Strunk BC. 2002. The changing face of managed care. Health Affairs 21(1):11–23.

[7] Davidson, Russell; MacKinnon, James G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford Universiry Press. p. 33. ISBN 0-19-506011-3.

[8] Goldberger, Arthur S. *(1964).* "Classical Linear Regression". Econometric Theory. *New York: John Wiley & Sons. pp.* 158. ISBN 0-471-31101-4.

[9] Giudici P (2003) Applied Data Mining. Wiley pp:76-77.

[10] Hayashi, Fumio (2000). *Econometics*. Princeton University Press. p. 15.

[11] Han J, Kamber M (2006) Data Mining Concepts and Techniques. Morgan Kaufmann pp: 229-230

[12] Kenney, J.; Keeping, E. S. (1963). *Mathematics of Statistics*. van Nostrand. p. 187.

[13] Mac Queen JB (1967) Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Statistics 1: 281-297

[14] Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (Second ed.). New York: J. Wiley & Sons. p. 319. ISBN 0-471-70823-2.

[15] Tan PN, Steinbach M, Kumar V (2006) Introduction to Data Mining. Addison-Wesley pp: 328

[16] Westhpal C, Blaxton T (1998) Data Mining Solutions: Methods and Tools for Solving Real-World Problems. Wiley pp: 186-189

[17] Williams, M. N; Grajales, C. A. G; Kurkiewicz, D (2013). "Assumptions of multiple regression: Correcting two misconceptions". *Practical Assessment, Research & Evaluation*. 18 (11).

[18] Zwillinger, D. (1995). *Standard Mathematical Tables and Formulae*. Chapman&Hall/CRC. p. 626. ISBN 0-8493-2479-3.