European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

Assessing the Relative Importance of Predictors in Linear Regression

Srinivasa Rao. D[1], S Jyothi Kannipamula[2]

[1]*Professor, MBA, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India*

[2]*Research Scholar, MBA, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India*

Abstract

Regression is the most extensively used statistical technique for explaining theoretical relationships and for prediction. This method can be viewed as a mapping from input or response variables space to an outcome variable space. If the assumption of the model is met, metrics like R2 F statistic and significance of t-values of the regression coefficients are used to judge the goodness of fit of the regression model. Similarly Mean Square Error (MSE) is used to judge the predictive power of the regression model. For judging the relative importance of the response variables in an estimated regression model, the magnitude and signs of the regression coefficients are considered. However, this approach is quite arbitrary and many a times inconclusive. In this context the present paper demonstrates the use of some of the relative importance metrics (lmg (Lindemann, Merenda and Gold,1980, pmvd (Feldman,2005)) which provides the decomposition of variance explained by a regression model into nonnegative components. It is shown that these relative measures are comparatively better than the magnitude and sign of regression parameters for assessing the relative importance of individual predictors in regression.
Key Words: Relative importance, variance decomposition, R2, regression model, LMG, PRATT

1. Introduction

Regression models are a set of statistical techniques that allow us to track the relationship between one outcome variable and a set of predictors and these models are popular across several disciplines. To start with regression analysis assesses how strong is the relationship between the outcome variable and the predictor variables and then with some ambiguity assesses the relative importance of each of the predictors to the relationship both in terms of explanation and for prediction. The relative contribution of each of the predictors in a multiple regression model is de facto judged by the t-statistics associated with the predictors. However, empirical evidence suggest that statistical significance measures are incomplete measures of relative contribution (Feldman ,2005). For example, when two predictors of a regression model are correlated their joint marginal contribution to the model increases but their marginal contributions to the explained variance decreases leading to low statistical significance of the regression coefficients. The full contribution of each of the predictors to the model can only be gauged by Relative importance measure. The statistical measure of relative importance can be helpful us in reducing the time, effort and skill required to identify the joint correlations present among the predictors. Thus, it is important to use various metrics of Relative importance to assess the individual contribution each of the predictors and the confidence interval associated with them.

The present paper aims at demonstrating the computation and visualisation of various Relative importance metrics in multiple regression model. Section 2 of the paper presents an overview of the previous studies on relative importance. Section 3 of the paper explains the basic features of the example dataset

'mtcars' that is used for computation. Section 4 presents the specification of multiple linear Regression Model and the relative importance metrics. Section 5 describes the estimation of multiple regression model and computation of relative importance metrics from the estimated model using R programming and the library 'relaimpo' package (Gromping, 2018).

2. Review of Literature

The concept of relative importance was first proposed by Hooker and Yule (1908). Achen (1982) tried to differentiate between theoretical importance and dispersion importance in the context of relative importance. However the need for relative importance metrics was expressed in medicinal sciences by Healy (1990) and Schemper(1993) , in management science by Soofi and Retzer(2000), and in Social sciences Kruskal and Majors(1989). With regard to the actual theoretical measures of relative importance, King (1987) provides a critique of the standard regression coefficients. Goldberger (1995) and Heckman (1995) severely critical of the adhoc use of relative importance measures. Johnson and Leberton(2004) defined relative importance as the proportional contribution of each of the predictors to the overall R2 of the regression model .The most frequently used method for relative importance is known as averaging method due to Lindeman, Merenda and Gold(1980 (LMD) which consists of variance decomposition by averaging the marginal contribution of each of the predictors over the ordering of all predictors Another similar averaging method was due to Chevan and Sutherland(1987). Feldman (2005) proposed an alternative measure of relative importance known as Proportional Marginal Decomposition (PMD). A novel method of R2 decomposition as a metric for relative importance was introduced by Zuber and Strimmer (2010) known as CAR. Hoffman's (1960) Natural Decomposition of R2 (PRAT) is somewhat controversial metric of relative importance. Both LMD and PMD methods are computer intensive and implemented in R Language by Gromping (2006) The present paper implements the R package 'relaimpo' for illustrating the relative importance metrics like LMD, prat and car in linear regression model.

Section 3 of the paper provides an overview of the example dataset mtcars that is used to demonstrate the computation of relative importance metrics and visulisation of these in multiple linear regression model.

3. The Example Dataset

We consider here the R built in data set 'mtcars' for building the Multiple Linear Regression Model and from it the various relative importance metrics: lmd, prat and car. The mtcars dataset consists of 32 observations on 12 variables comprising of fuel consumption and 10 aspects of automobile design for 32 varieties of Cars. The outcome variable is mpg (miles per gallon) and the response variables are:

Cyl: no of cylinders, disp: Displacement, hp: Horse Power, drat: Rear Axle Ratio, wt=Weight, vs: Engine type, am: Transmission type, qsec:1/4-mile time, gear: Number of Gears and carb: No of Carburettors.

This dataset is useful for comparing the various metrics of relative importance of response variable in

In a regression model in the context of multicollinearity problem. For this first we examine the correlation matrix of the dataset.

Correlation Matrix of 'mtcars'

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

The above correlation structure clearly shows a significant negative correlation between mpg and cyl, disp, hp and wt. From this structure there is strong multicollinearity among the outcome variables. First we shall find the best regression model with mpg as the outcome variable using step wise regression methodology. This enables us allocating the relative importance across the best predictors.

4. The Multiple Linear Regression Model and Measures of Relative Importance

A multiple linear regression model can be specified as follows:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \text{--------} + \beta_n x_{in} + \varepsilon_i \qquad (1)$$

where y is the outcome variable, x's are response variables and ε is the error term. And the total variance explained by the response variables is indicated by the coefficient of determination called as R2 which is defined as:

$$R^2 = \frac{Model\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} \qquad (2)$$

The model to be estimated is given by:

$$mpg \sim cyl+disp+hp+drat+wt+sec+vs+am+gear+carb \qquad (3)$$

Observations: 32
Dependent Variable: mpg
Type: OLS linear regression
The best model is determined by the following R code
> library(c('MASS',"jtools'))
> step AIC (lm(mpg ~., mtcars))
The following output shows the best model selection based on Akaike Information Criterion(AIC)
Step: AIC=61.31

Model: mpg ~ wt + qsec + am (4)

The following R code results in the estimation of the model (4)

```
> model = lm(mpg ~ wt+ qsec+am,data=mtcars); summ(model)
```

Output

MODEL FIT:

F(3,28) = 52.75, p = 0.00
R^2 = 0.85
Adj. R^2 = 0.83

	Est.	S.E.	t val.	p
(Intercept)	9.62	6.96	1.38	0.18
wt	-3.92	0.71	-5.51	0.00
qsec	1.23	0.29	4.25	0.00
am	2.94	1.41	2.08	0.05

From the above output we observe that R2 is 0.83 and all the response variables except the intercept are statisticall y significant at 5% level. We find the weight is negatively influeincing the mpg whereas am(Transmission type) and qsec(¼ mile time) are positively related which are on expected lines

4.1. The Relative Importance Metrics

The relative importance metrics used in this paper are as follows:

- LMG : is the contribution of ordered predictors to R2 (Lindeman, Merenda and Gold(1980)
- CAR: is the Decomposition of R2 proposed by Zimmer and Strmmer(2010)
- PRATT: is the product of correlation and standardised coefficient.

We shall use the reliampo package to calculate the above metrics for the multiple regression model (4) with the following R code.

```
> library('reliampo')
 > calc. relimp(lm(mpg~am+qsec+wt, mtcars),type=list('lmg','pratt','car'), rela=T)
```

Output

Response variable: mpg
Total response variance: 36
Analysis based on 32 observations

3 Regressors:
am qsec wt
Proportion of variance explained by model: 85%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

	lmg	pratt	car
am	0.25	0.17	0.20
qsec	0.19	0.18	0.19
wt	0.56	0.65	0.61

Average coefficients for different model sizes:

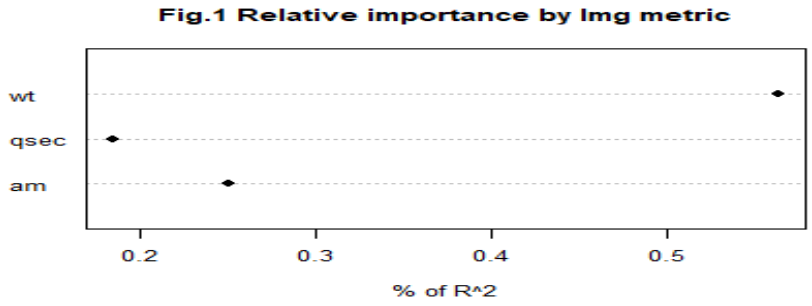
	1X	2Xs	3Xs
am	7.2	4.4	2.9
qsec	1.4	1.5	1.2
wt	-5.3	-5.2	-3.9

From the above output we observe that all the relative importance metrics of each of the predictor variables show slight variations depending upon the metric considered. All the three metrics considered the predictor, wt as the most significant contributor for the variance in the outcome variable mpg. Whereas the predictor qsec's relative importance is around 19%. But there seems to be some difference in the relative contribution of am across the three methods under consideration.

5. Visualising the Relative importance metrics

Now we shall use the dot chart to visualise the relative importance of the three predictors of our multiple linear regression model (4). The following R code is used for this purpose

```
> mod= calc. relimp(lm(mpg~am+qsec+wt, mtcars),type=list('lmg','pratt','car'),rela=)
> dotchart(mod@lmg,pch=19,main='Fig.1 Relative importance by lmg metric')
>dotchart(mod@car,pch=19,main='Fig.1 Relative importance by car metric')
>dotchart(mod@lpratt,pch=19,main='Fig.1 Relative importance by pratt metric')
```



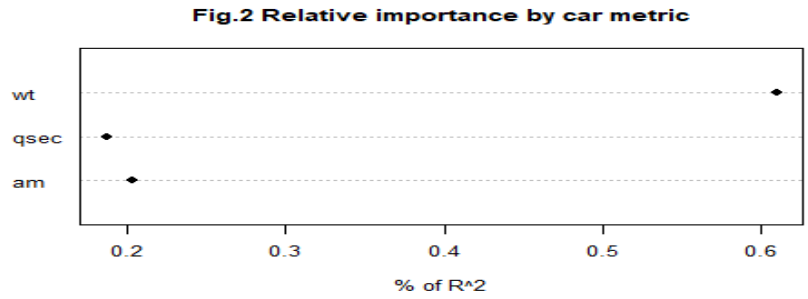
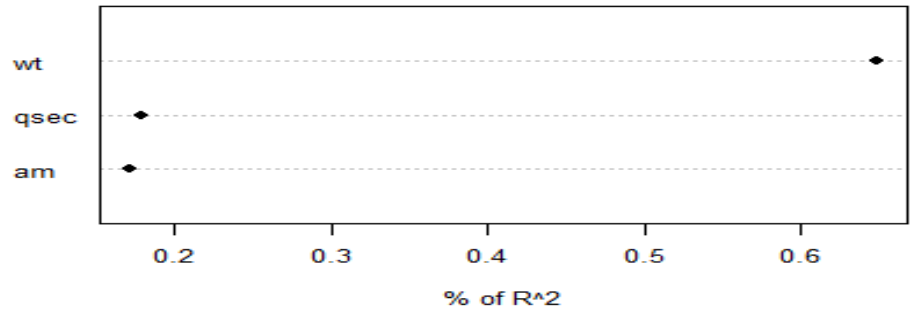


Fig.3 Relative importance by pratt metric



6. Summary and Conclusions

In this paper an attempt was made to measure the relative importance of predictors in a multiple regression model. In many managerial applications such as marketing a large group of regressors are

Supposed to be influencing an outcome variable and there is no reliable measure of the relative importance of all the regressors that are considered. In such a scenario the present demonstrates the use of the relative importance metrics like lmg,car and pratt. These metrics can be used to identify the most significant explanatory variables that may be influencing the variable of interest. How to measure the relative importance of the predictors are demonstrated by using an example dataset built into R software and also R package 'reliampo' was used for calculations. Future research in this may consider the hierarchical structure in the predictors rather than individual contributions as an extension of the present study.

References

1. Barbara Tabachnick and Linda Fidel (2018) : Using Multivariate Statistics, Pearson.
2. Chevan, A. and Sutherland, M. (1991) Hierarchical Partitioning. The American Statistician 45, 90–96.
3. Darlington, R.B. (1968) Multiple regression in psychological research and practice. Psychological Bulletin 69, 161–182.
4. Feldman, B. (2005) Relative Importance and Value. Manuscript (Version 1.1, March 19 2005), downloadable at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2255827
5. Genizi, A. (1993) Decomposition of R2 in multiple regression with correlated regressors. Statistica Sinica 3, 407–420.
6. Groemping, U. (2006) Relative Importance for Linear Regression in R: The Package relaimpo Journal of Statistical Software 17, Issue 1.
7. Lindeman, R.H., Merenda, P.F. and Gold, R.Z. (1980) Introduction to Bivariate and Multivariate Analysis.
8. Glenview IL: Scott, Foresman. Zuber, V. and Strimmer, K. (2010) Variable importance and model selection by decorrelation.
9. Dr. B. Kishore Babu, N. Rajeswari and Naidu Mounika, An Empirical Study on Consumer Green Buying Behaviour, Vijayawada, Andhra Pradesh, International Journal of Civil Engineering and Technology,9(3), 2018, pp. 648–655
10. Kishore Babu, P. Pavani, Engineering Students Perception Towards Social Media Advertising For Social Causes, Vijayawada: An Empirical Study, International Journal of Recent Technology and Engineering (IJRTE), Volume-7, Issue-6, March 2019, pp. 1901-07
11. Sailaja, V.N., A study on investors awareness towards mutual funds investment, International Journal of Civil Engineering and Technology, Volume -9, issue 3,2018, pp.376-382
12. Kannipamula, S. J., & Srinivasa Rao, D. (2017, December 1). Impact of human resource information system (HRIS) on firms performance: A conceptual framework. Journal of Advanced

Research in Dynamical and Control Systems. Institute of Advanced Scientific Research, Inc.Volume: 9 | Issue: 1,Pages: 462-472.
13. Jyothi Kannipamula, S., & Srinivasa Rao, D. (2019). HRIS driving health care institutes. International Journal of Innovative Technology and Exploring Engineering, 8(11), 3917–3920. https://doi.org/10.35940/ijitee.K1571.0981119
14. Jyothi Kannipamula, S., & Srinivasa Rao, D. (2019). An empirical study on organizational learning capability in it industry. International Journal of Recent Technology and Engineering, 7 (6), pp. 916-919.
15. Srinivasa Rao, S Jyothi Kannipamula, (2018) .HRIS impact on organizational efficiency, International Journal of Creative Research Thoughts.Volume 6, Issue 2 April 2018, pp.585-587.
16. Venkata Ramana J., Hanuma Reddy D., Venkateswara Kumar K.S., Sirisha K. (2019), 'An empirical study on marketing of handloom fabrics in Andhra Pradesh (A case study with reference to Guntur district)', International Journal of Innovative Technology and Exploring Engineering, 8(8), PP.1071-1075.
17. Ramana J.V., Sridhar P. (2019), 'The movement of industrially applicable yellow metal and its impact on global currencies', International Journal of Recent Technology and Engineering, 8(3), PP.7066-7070.
18. Srinivasa Rao, D..,Y.Meduri (2019). Humanitarian Efficiency & Role of Relief Workers: Testing a competency based approach, International Journal of Business Science and Applied Management, 1753-0296.
19. D.Shri Jyothi, Srinivasa Rao,D,(2019). An Analysis on Perspectives of Investors on Commodity Trading and Risk Management in India, International Journal of Innovative Technology and Exploring Engineering, pp.2278-3075
20. Srinivasa Rao,D, Shri Jyothi D., (2019). Financial Risk Quantification of Indian Agro-Commodities using Value At Risk, International Journal of Engineering and Advanced Technology ,pp-2249-8958