ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 3, March 2024

A Short Review on Correlation Analysis and Linear Regression for Business Analytics

[1] Nitish Kumar Jha

[1] Department of Mathematics, University Institute of Sciences, Chandigarh University, Mohali, Punjab, India
Corresponding Author Email: [1] nitish.jha97@outlook.com

Abstract*— This paper discusses how to enhance the business using the statistical terms "Correlation Analysis" and a machine learning algorithm "Linear Regression". In many research projects, correlation and regression analysis are the most often utilized statistical tools. The purpose of correlation analysis is essentially the same in quantitative analytical investigations, making it advantageous to look into the link between independent and dependent variables. In order to show how to use a widely common statistical tool called correlation and regression analysis for beginning researchers, this study used secondary data. Regression analysis comes after correlation. It begins with the idea of a simple correlation coefficient, which indicates how linearly related two variables are to one another. A scatter plot should be created to check for a linear relationship between the two variables. If explanatory variables change by one unit, regression analysis technique exposes the relevance of variables and the degree of change in exogenous variables. The results of this study have ramifications for how to analyze data and how to do correlation and regression analyses.

Keywords: Correlation, Linear Regression, Mean Absolute error, Root Mean Square error.

I. INTRODUCTION

Correlation Analysis shows the linear relationship between two variables. It measures both the magnitude of the linear relation between independent variables and also shows the direction of their relation [1].

Regression [2] is a technique used for two things. First, it establishes the link of dependent variable with one or more independent variables. Second, it is also used for forecasting and prediction.

The format of this paper is as follows: The specifics of the approaches utilized are illustrated in Section 1; the methods' implementation and outcomes are covered in Section 2.

Correlation Analysis

A correlation coefficient (r) is a metric that expresses a relationship, or a statistical link between two variables. A positive correlation occurs when the dependent variable increases as the independent variable does. A negative correlation indicates that one variable falls as another rises.

A +1 of correlation coefficient shows a perfect correlation. This means the 1st variable is in exact relation with the second positive variable. While -1 correlation coefficient shows a perfect negative correlation. This means 1st variable is in negatively linear relation with 2nd variable. If the correlation coefficient is 0 then there is no linear relation between both the variables. If the value of r lies in between 0.5 to .75 it is considered as moderately correlated variables.

Linear Regression

By fitting a linear equation to the observed data, linear regression makes an attempt to predict the relationship between two variables. One of the variables is regarded as an explanatory variable and the other as a dependent variable. Linear regressions are of two types. I) Simple Linear Regression and (ii) Multiple Linear Regression.

Simple Linear Regression: The easiest model to forecast the value of one variable in relation to another is simple linear regression. [3]. Finding a relationship between two continuous variables can be done using this regression where independent variable is the predictor and the others are dependent variable [4].

This regression is a parametric test, which mean it has some pre-assumptions about the data. These assumptions are:

1. Homogeneity of variance: The amount of our forecast's error is rather constant in size across the independent variable's range of values.
2. Independence of observations: Since the dataset's observations were compiled using statistically acceptable sampling methods, there are no undetected correlations between the variables.
3. Normality: The observed data are thought to exhibit normal distribution.
4. The relationship between the explanatory and response variable is linear: based on the straight line that most closely matches the data points.

The formula of this Regression is:

$$У = \beta_0 + \beta_1 X + \varepsilon$$

- У is the response variable in light of the explanatory variable's observed values (x).
- When the value of X is at origin, the constant value β_0 is called the intercept.
- β_1 is the coefficient of independent variable.
- ε is our estimation's inaccuracy or the extent of its change.

The optimal line to match the data is found using linear regression by identifying the regression coefficient (β_1) with minimum error (ε).

Multiple Linear Regressions: Using a variety of explanatory factors, it is a statistical technique that predicts the outcome of a responsive variable. Modelling the link between the independent variable and the dependent variables is the goal of MLR. [5].

Equation of Multiple Linear Regressions:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

Where, i = for n observations:
Y_i = outcome variable
X_i = Predictor variables
β_0 = y – intercept (constant term)
β_p = each explanatory variable's slope coefficient
ε = the model's inaccuracy.

The multiple regression models are based on the following assumptions:

- The independent and dependent variables exhibit linear correlation.
- The independent variables barely have any correlation with one another.
- A normal distribution with a mean of 0 and variance of 1 is ideal for residuals.

Residual Sum of Squares: Residual analysis plays a critical role in regression analysis. It gauges the degree of variation in a regression model's error term or residuals. The formula to measure this is given below [5,6]:

$\boldsymbol{RSS = \Sigma (y_i – \acute{y}_i)^2}$ [Taking i = 1 to n]. where y_i is the actual values of the raw data and $\acute{y}_i$ is the predicted values.

Mean Squared Error: It is used to figure out the gap between values predicted by the model and actual values. [7].

It is the mean of squared residuals (e^2) and is calculated by dividing RSS by the number of data values.

$$MSE = RSS/n$$

$$MSE = (1/n)\, \Sigma\, (y_i – \acute{y}_i)^2$$

The root means square deviation (RMSD) of an estimator is the square root of its mean square error [8,9].

$$RMSD = \sqrt{MSE}$$

Example of Linear Regressions:

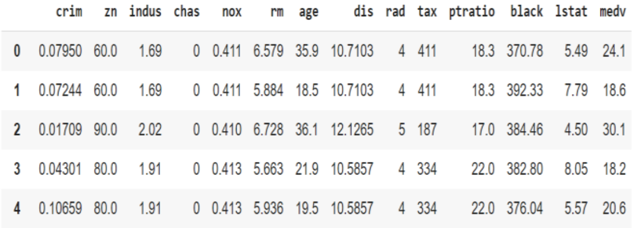
To illustrate the above-discussed concepts effectively, this research employed a dataset (the Boston Housing Dataset). This housing dataset is derived from the data gathered by the U.S. Census Agency about housing in the Boston, Massachusetts, area.

The polynomial regression model has been used in this study due to the significance of model combination.

II. METHODOLOGY

A. Data pre-processing.

'Boston House Price' dataset containing 506 rows and 14 columns. These variables serve as the characteristics of the dataset. These characteristics aid in predicting the typical housing price in Boston. The below figure shows the little introduction about the data.



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0	0.07950	60.0	1.69	0	0.411	6.579	35.9	10.7103	4	411	18.3	370.78	5.49	24.1
1	0.07244	60.0	1.69	0	0.411	5.884	18.5	10.7103	4	411	18.3	392.33	7.79	18.6
2	0.01709	90.0	2.02	0	0.410	6.728	36.1	12.1265	5	187	17.0	384.46	4.50	30.1
3	0.04301	80.0	1.91	0	0.413	5.663	21.9	10.5857	4	334	22.0	382.80	8.05	18.2
4	0.10659	80.0	1.91	0	0.413	5.936	19.5	10.5857	4	334	22.0	376.04	5.57	20.6

Fig. 1 First five rows and all the columns of the data.

Finding the null values and features that have a strong correlation with the output variables came next. Details on each are provided in the table below:

Attribute	Detail
CRIM	Per capita crime rate by town
ZN	Proportion of Residential land zoned for lots over 25000 sq. ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River Dummy variable (1 if tract bounds, 0 otherwise)
NOX	Nitric oxide concentration
RM	Average number of rooms per dwelling
AGE	Proportions of owner-occupied units built prior to 1940
DIS	Weighted distance to five Boston Employment centers
RAD	Accessibility score for radial highways.
TAX	Entire value of property tax for every $10,000
PTRATIO	Pupil teacher ration by town.
LSTAT	% Lower level of population status
MEDV	Owner-occupied residences' median values in the $1,000 range

B. Data Analysis

Before creating a regression model, exploratory analysis is an essential step. It helps researchers to discover the pattern of the data which ultimately help to choose the right machine learning algorithm [10].

To do this, first import all the required libraries into the software (python e.g.). After analysis being processed, we analyze the pattern of target variable ('medv'). The graph (Fig.1) given below help us to analyze the distribution of the target variable;

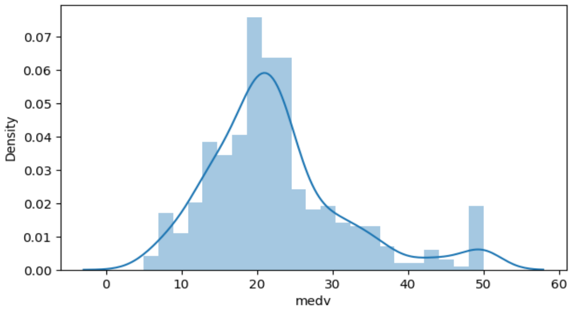


Fig.2 This shows the distribution followed by the target variable.

This graph shows that the average price of houses is centered around 20K. Apart from the target variable, other variables also play a significant role in the model's performance. We analyze the correlation between two variables. It helps to choose the correct set of attributes while developing the model [11].

Upon checking the correlation values, there were two variables which were highly correlated with the target variables. The correlation of both the variables have shown below:

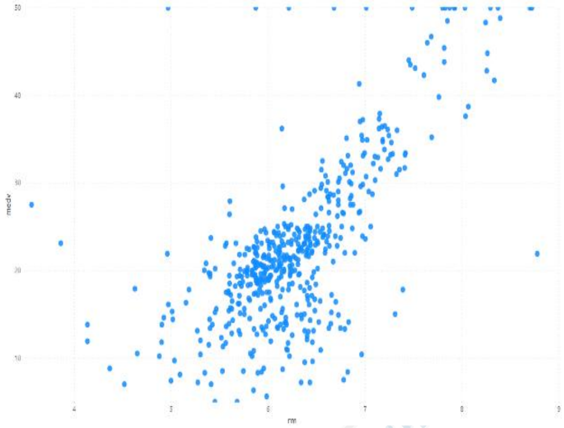


Fig. 3 Positive Correlation of 'medv' and 'rm'.

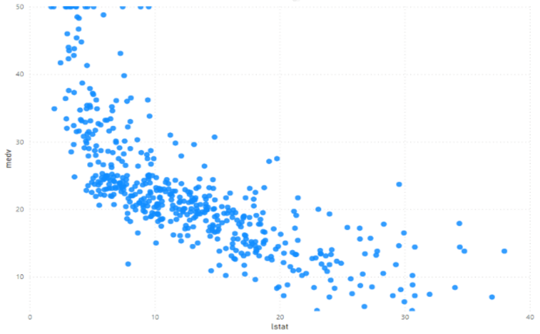


Fig. 4 Negative Correlation of 'lstat' and 'medv'.

C. Model Selection

The data should be appropriately pre-processed before creating models. Next, Using the scikit-learn package, A train set and a test set of data were created [12]. In the software, the required libraries were imported for the train-test split and Linear Regression model.

D. Regression Model

Regression models show that the independent factors or variables forecast the values of the dependent variables. [13].

Machine learning methods frequently divide the data into train and test datasets. This enables us to evaluate the model's performance using both known and unknown data. Here, the multi-variable polynomial regression formula is used internally by the linear regression model.

After training the model, the unobserved or unseen data were predicted.

It became important to visualize how the model's predicted behavior would behave given the observed data. This makes it easier to see how the model behaves in relation to the initial values.

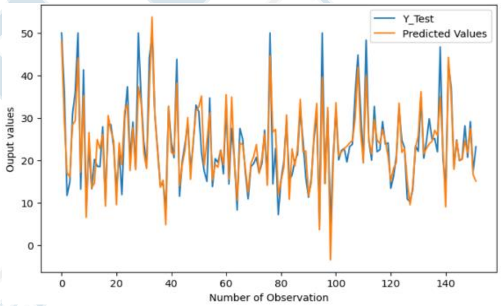


Fig. 5 Comparison of Regression's predicted and original values.

Figure 5 above demonstrates how well the model has learned the data and can reasonably anticipate prices. Measuring the performance of a linear regression model involves using the "Mean Absolute Error," "Mean Squared Error," and "r2" score. It is necessary to compute the mistakes in order to comprehend the model's suitability for making predictions in the business world. The errors were measured for different degree of polynomials and it was visualised also. The 2nd order polynomial regression model was found to have the least error. Finally, a graphic explanation of the selection of second-degree polynomial regression is provided below.

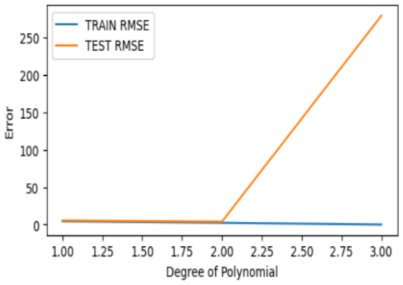


Fig. 6 Selecting the degree of polynomial

Figure 6 displays the inaccuracy as measured by various model of different polynomial degrees. The test error is lowest at the second order and rapidly increases after the second order polynomial. Therefore, choosing a polynomial regression model of second order is optimal for this problem.

III. CONCLUSION

Regression modelling and correlation are statistical methods commonly used for observational and experimental studies. This review paper has focused on understanding the concepts of correlation, data analysis and developing a regression model to predict future unseen data. After data pre-processing and data cleaning, to achieve the target of the problem different algorithm had been experimented. For example, at first simple regression was used to experiment but the errors were not satisfactory.

With multi linear regression which is used for more than two variables having linear relationship with the output variable, the error little improved but still it was not good enough. At last polynomial regression has been used. This regression is used when output variables are in polynomial relation with all the independent variables. Finally, the accuracy of the models was tested to assess the model's goodness of fit after attempting every sort of suitable algorithm. The model is picked with the least amount of tolerance for error. For example, this paper has shown the 2nd degree polynomial regression model has the least error and hence this model could be used for predicting the price of the city having the least error.

REFERENCES

[1] Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences by Patricia Cohen, Stephen G. West, Leona S. Aiken.

[2] J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142

[3] “Statistical Sampling and Regression: Simple Linear Regression”. Columbia University. Retrieved 2016-10-17. When one independent variable is used in a regression, it is called a simple regression.

[4] Altman, Naomi; Krzywinski, Martin (2015). "Simple linear regression". Nature Methods.

[5] Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley.

[6] Archdeacon, Thomas J. (1994). Correlation and regression analysis: a historian's guide. University of Wisconsin Press. pp. 161–162.

[7] Hyndman, Rob J.; Koehler, Anne B. (2006). "Another look at measures of forecast accuracy". International Journal of Forecasting.

[8] Ponitus, Robert; Thontteh, Olufunmilayo; Chen, hao (2008). “Components of information for multiple resolution comparison between maps that share a real variable”. “Environmental Ecological Statistics”.

[9] Willmott, Cort; Matsuura, Kenji (2006). "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators". International Journal of Geographical Information Science.

[10] Diniz-Filho, J. A., T. N., Lima, J.S., Dobrovolski, R, Landerio, V.L., de Campos Telles, M.P., Rangel, T.F., & Bini, L.M.(2013).

[11] Gogtay NJ, Deshpande S, Thatte UM. Measures of Association. J Assoc Phy Ind 2016; [in progress]

[12] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825–30.

[13] H. Roopa and T. Asha, "A linear model based on principal component analysis for disease prediction," IEEE Access, vol. 7, pp. 105314-105318, 2019.