BABES-BOLYAI UNIVERSITY CLUJ-NAPOCA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

MASTER OF APPLIED COMPUTATIONAL INTELLIGENCE

Multiple Linear Regression

Analysis

Supervisor

Assoc. Prof. PhD Pop F. Horia

Author

Grigore Mihai Alin

2021

**1.Abstract**

Perhaps one of the most common and comprehensive statistical and machine learning are linear regression. The goal of this work is to present the theoretical ideas of multiple linear regression. Multiple linear regression is used to find a linear relationship between one or more predictors. In this paper, we will introduce the simple linear regression and we will present the multiple linear regression that is an extension of the simple linear regression. We will present the method of calculating the multiple regression and a small algorithm, interpretation of regression coefficients and other output, correlation coefficient (Pearson Correlation Coefficient) and the test for significance of correlation coefficient (R). We will also show how to use multiple linear regression for prediction. In this paper, our aim is to provide a clearest possible analysis of multiple linear regression, of its coefficients and output.

**2. Introduction**

We will focus on how two or more variables are interrelated. We shall understand the two concepts in statistics: regression analysis and correlational analysis and the relationship between the concepts. We shall discuss what to look for in the output of regression analysis and how the output can be interpreted. We shall elucidate on how to use regression analysis for forecasting.

Linear regression analysis is a technique used to predict the value of one quantitative variable by using its relationship with one or more additional quantitative variables. For example, if we know the relationship between height, weight and gender, we can use regression analysis to predict weight given a particular value for height and gender.

The relationship between height, weight and gender is familiar to us; generally, the taller a person is, the more he weighs. Another example of a familiar relationship is that of crop yield and the amount of fertilized applied to the land; the more fertilized applied to the land the greater the yield-to a point. If too much fertilized is applied, the crop will be killed off by fertilized chemicals-the land will be “burned”. An important relationship in business is the relationship between the allocation of dollars to advertising effort and the level of sales of a product; the more money expended in advertising, the greater the level of sales.

In this paper, we will emphasize the development of regression analysis when more predictors variables are used to predict the variable of interest and where the relationship between the variables is linear. In this context, the variables which is used to predict the variable of interest is called independent variables, and the variable we are trying to predict is called the dependent variable. The analysis used is called multiple linear regression analysis – multiple because there is more than one predictor or independent variable, and linear because of the assumed linear relationship between the dependent and the independent variables.

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know how strong is the relationship between two or more independent variables and one dependent variable (temperature, rainfall) and when you want to know the value of the dependent variable at a certain value of the independent variables.

**3. Linear Regression Analysis**

**3.1. Construction of line fit plots**

In showing the relationship between two variables, we can draw a line across the variables after plotting the scatter plot and ensuring the line passes through as many points as possible. The straight line that gives the best approximation in a given set of data is referred to as the line of best fit. Least squares method is the most accurate method of finding the line of best fit of a given dataset.

**3.2. Types of regressions analysis**

There are many regression analyses which are based on different assumptions. I mention a few, and I will dwell on the first and second mentioned below:

1.simple linear regression;

2.multiple regression;

3.ridge regression;

4.quantile regression;

5.Bayesian regression.

**3.3. Uses of regression analysis**

Regression analysis can be used for the following:

i. Causal analysis: to establish the relationship between two or more variables.

ii. Forecasting an effect: it is used to predict a response variable fully knowing the independent variables.

iii. Forecasting a trend: regression analysis can be used to predict trend in a dataset.

**3.4. Simple linear regression**

Before starting the presentation of the multiple regression, I will present the simple linear regression because these two are in a close connection.

**3.4.1. Theoretical**

Simple linear regression is a statistical technique to show the relationship between one dependent variable and one independent variable. The dependent variable is denoted by Y while the independent variable is denoted by X. The variables X and Y are linearly related. Simple linear regression can be used for the following:

i. Description of the linear dependence of one variable on another variable;

ii. Predinction of one variable from the values of another variable;

iii. Correction for the linear dependence of one variable on another variable.

**3.4.2. Methods**

The most common method used to find a regression line is the least squares method. The equation of a line is:

y = bx + a [4]

where b equals the slope of the line and a is where the line crosses the y axis.

The least squares method minimizes the vertical distances between our data points(the observed values) and our line(the predicted values).

The equation of the line we are trying to find is:

= bx + a [4]

Where is the predicted value of y for some x.

We have to calculate b and a.

[4]

Where X is the sum of products:

X)) [4]

And is the sum of squares for X:

[4]

is the average of the points and .

The regression line will always pass through the point () so we can plug this point into our equation to get a, where the line passes through the y-axis.

a = - b [4]

**3.4.3. Algorithm for Simple Linear Regression**

The linear regression works on the following algorithm [4]:

Step 1: Take the values of variable

Step 2: Calculate the average for variable

Step 3: Calculate the average for variable

Step 4: In a ‘for’ will be calculated X and value of regression coefficient b

Step 5: Calculate the value of another regression coefficients a by substituting the values of b , average of in equation: a = - b.

Step 6: Substitute the value of regression coefficients a and b in the equation: y = bx + a

**4. Multiple Linear Regression**

**4.1 Theoretical and Methods**

In simple linear regression, we considered two variables where one is the response and the other one is explanatory variable. In the case of multiple linear regression, it is an extension of simple linear regression whereby we have two or more explanatory variables is associated with a value of the response variable Y. Tue multiple linear regression model is of the form:

Where is a constant term, are regression coefficients, is error term, and The notation means normally distributed with mean and variance [4]

To simplify the computation, the multiple regression model in terms of the observations can be written using matrix notation.

The model will be written in the form:

Y = X + [4]

Where:

Y = is an n × 1 dimensional random vector that consisting of the observations,

X = is an n × (k + 1) matrix determined by the predictors,

β = is an n × 1 vector of unknown parameters,

= is an n × 1 vector of random errors

The first step in multiple linear regression analysis is to the vector (is the vector of least squares estimators), which gives the linear combination that minimizes the length of the error vector. An important property in multiple regression analysis is that the variables , are linearly independent. The correlation between each . Since the objective of multiple regression is to minimize the sum of the squared errors, the regression coefficients that meet this condition are determined by solving the least squares normal equation:

[4]

How , are linearly independent than the inverse of , will exist.

To find we will multiply in both equation with and we will obtain:

[4]

The mathematical software packages such as Mathematica, Stata, and MATLAB provide matrix commands to determine the solution to the normal equation.

**4.2. Interpretation of regression coefficients and other output**

a. *Regression coefficients*: The magnitude of the coefficient of each independent variable gives the size of the effect that the independent variable has on dependent variable and the sign (+/-) on the regression coefficient tells the direction of the effect. In general, the coefficient gives how much the dependent variable would change when the independent variable change by 1 unit, keeping all other independent variables constant. [4]

For example, if we fit a mode, Y = 2.5 + 0.5 – 0.8, this can be interpreted as: the Y-intercept can be interpreted as the predicted value for Y when both and are zero. Therefore, you would expect 2.5 unit when and . For every unit increase in Y would lead to 0.5 unit increase in while is held constant, also for every unit in Y would lead to 0.8 unit decrease in holding constant. [4]

b. *t-value*: This is the ratio of the coefficient and the standard error of the coefficient. The rule of thumb is that the absolute value of t must be 2 or more to show the significance of the coefficient. T-value is used to determine the p-value corresponding to Student t-distribution. [4]

c. *p-value*: It indicates that the probability that the estimated coefficient is not reliable. The less the p-value the more it is reliable under the significance level. For example, if the level of significance is 5% or 10% it means than the value of p is less than 5% or 10% indicates that the estimated coefficient is reliable, otherwise it is unreliable and it should be discarded from the model. [4]

d. *Multiple R-squared*: This shows the fraction of the variation in a response variable that is accounted for by independent variables in the model. It indicates how well the terms fit the data. In addition, the adjusted R-squared is used to adjust for the number of terms in a model. As long as you add more independent variables to a model, the R-squared continue to increase in value, even when the variable is useless in the model. However, he adjusted R-squared will increase if you add useful independent variable in the model, otherwise the value of adjusted R-squared will decrease. The R-squared rages from 0 to 1 but adjusted R-squared can dip down to the negative value. [4]

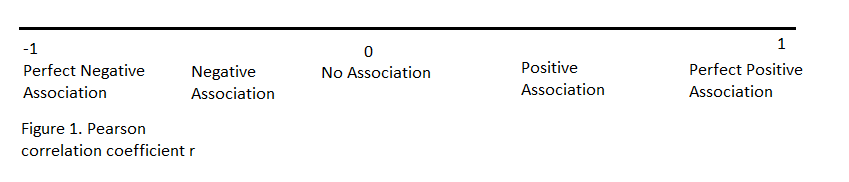
e. *F-statistics*: This test for the significance of the overall coefficients whether the regression

model provides a better fit to the data than a model with no independent variables. [4]

f. *Durbin-Watson*: It is used to test for the autocorrelation assumption of the error terms. That is, to make sure that no correlation between the error terms . Autocorrelation may be caused by omission of important explanatory variable, misspecification of the model, and systematic error in measurement. The consequences of autocorrelation include the least square estimators will be inefficient and the estimated variances of the regression coefficients will be biased and inconsistent, thus hypothesis testing is no longer valid. Furthermore, Durbin-Watson range from 0-4. The value of 2 indicates no autocorrelation between the error terms, between 0 to < 2 is a positive autocorrelation and between 2 and 4 is negative autocorrelation. A rule of thumb for Durbin-Watson is that for a relatively normal data, the test statistic should fall within 1.5-2.5. [4]

**4.3. Pearson correlation coefficient**

The multiple linear regression analysis shows the relationship between two or more variables that are linearly related. The correlation analysis is a measure of strength or level of association between variables. The correlation is measured by the Pearson correlation coefficient and it is denoted by r. The statistic r ranges from -1 to +1. If the values of r are zero, it implies that there is no linear association between the variables. As the values of correlation coefficient r move close to zero, the linear association between the variables becomes weaker. Conversely, as the value of the correlation coefficient r far away from zero and approaches -/+1, the linear association between the variables becomes stronger.[4] When the correlation coefficient is exactly 1, it is called a perfect positive correlation and when the correlation coefficient is exactly -1, then it is known as a perfect negative correlation.



The significance of a relationship is determined whether the Pearson’s correlation coefficient r is a meaningful reflection of the linear relationship between the two variables or whether the relationship occurred by chance. For a given significant value , the probability that Pearson’s correlation coefficient r value comes by chance is 5% or less.

Pearson’s correlation coefficient r between variables X and Y can be defined as:

. [4]

Alternatively,

= . [4]

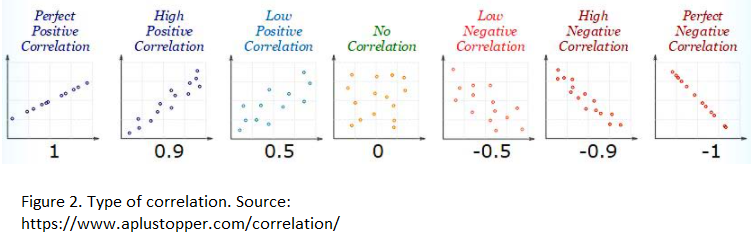
Where are the means of X and Y,

Cov(X, Y) is the covariance between X and Y, Cov(X,Y) = [4]

are standard deviations of X and Y. [4]

**4.4. Test for significance of correlation coefficient (R)**

This significance test is used to test the null hypothesis that . We should bear in mind that the sampling distribution of r is approximately normal when sample size is large (n ≥ 30), and distributed t when the sample size is small (n ≤ 30). [4]



The t-test statistics for the significance of the correlation coefficient r is defined as:

~ , [4]

However, when n is sufficiently large (n ≥ 30), we use the standardized score for the r using Fisher z-transformation (z’) to test for the significance of the correlation coefficient r:

[4]

**4.5. Algorithm for Multiple Linear Regression**

The multiple linear regression works on the following algorithm: [4]

Step 1: Take the values of variable in two vectors X and Y.

Step 2: Calculate the value for variable

Step 3: Calculate the value for variable .

Step 4: Calculate the value for variable

**4.6. Using Multiple Regression for Prediction**

A standard example for using multiple regression for prediction is: we would like to specify the height of a person knowing age, gender and weight. We would use standard multiple regression in which age, gender and wight would be the independent variables and height would be the dependent variable. The resulting output would tell us how much of the variance in height is accounted for by the joint predictive power of knowing a person’s wight, age and gender. Generally, this value is denoted by Another thing that would tell us the output is if the model allows the prediction of a person’s height at a rate better than change. Generally, this is denoted by the significance level of the model. Significance level of the model means a measure of how likely it is to draw a false conclusion in a statistical test, when the results are really just random variations. A significance level of 0.05 is considered the standard for what model is acceptable.

In our example if the significance level is less or equal to 0.05, then the model is considered significant. There is only a 5 in 100 or less chance that there really is not a relationship between height, weight, age and gender. If the significance level is between 0.05 and 0.10, then the model is considered marginal. The model is fairly good at predicting a person’s height, but there is between a 5-10% probability that there really is not a relationship between height, weight, age and gender.

Standard multiple regression show us how well each independent variable predicts the dependent variable, controlling for each of the other independent variables. The regression analysis would tell us how well weight predicts a person’s height, controlling for gender, as well as how well gender predicts a person’s height, controlling for weight or how well age predicts a person’s height.

**4.7. Hypotheses**

There are several hypotheses about the data that must be met to perform a linear regression analysis (multiple linear regression):Linearity: The relationship between independent and dependent variables is assumed to be linear. Although this assumption can never be fully confirmed, examining a scatter plot of the variables can help with this determination. If there is a curve in the relationship, you can consider transforming variables or explicitly allowing nonlinear components; Normal: The residuals of your variables are assumed to be normally distributed. That is, the errors in predicting the value of Y (the dependent variable) are distributed in a way that is close to the normal curve. You can look at normal histograms or probability charts to inspect the distribution of your variables and their residual values; Independence: It is assumed that the error in the prediction of the value of Y are independent of each other (they are not correlated); Homoscedasticity: It is assumed that the variance around the regression line is the same for all values of the independent variables.

**4.8. Influential Observations**

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of influential observations and to take them into consideration when interpreting the results.

**Influence**

The influence of an observation can be thought of in terms of how much the predicted scores for other observations would differ if the observation in question were not included. Cook’s D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

A common rule of thumb is that an observation with a value of Cook’s D over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

An observation’s influence is a function of tow factors: how much the observation’s value on the predictor variable differs from the mean of the predictor variable and the difference between the predicted score for the observation and its actual score. The former factor is called the observation’s leverage. The latter factor is called the observation’s distance. [13]

**Calculation of Cook’s D**

The first step in calculating the value of Cook’s D for an observation is to predict all the scores in the data once using a regression equation based on all the observations and once using all the observations except the observation in question. The second step is to compute the sum of the squared differences between these two sets of predictions. The final step is to divide this result by 2 time the MSE (The mean squared error tells you how close a regression line is to a set of points.) [13]

**Leverage**

The leverage of an observation is based on how much the observation’s value on the predictor variable differs from the mean of the predictor variable. The greater an observation’s leverage, the more potential it has to be an influential observation. For example, an observation with the mean on the predictor variable has no influence on the slope of the regression line regardless of its value on the criterion variable. On the other hand, an observation that is extreme on the predictor variable has, depending on its distance, the potential to affect the slope greatly. [13]

**Calculation of Leverage (h)**

The first step is to standardize the predictor variable so that it has a mean of 0 and a standard deviation of 1. Then, the leverage (h) is computed by squaring the observation’s value on the standardized predictor variable, adding 1, and dividing by the number of observations. [13]

**Distance**

The distance of an observation is based on the error of prediction for the observation: The greater the error of prediction, the greater the distance. The most commonly used measure of distance is the studentized residual. The studentized residual for an observation is closely related to the error of prediction for that observation divided by the standard deviation of the errors of prediction. However, the predicted score is derived from a regression equation in which the observation in question is not counted. The details of the computation of a studentized residual are a bit complex. An observation with a large distance will not have that much influence if its leverage is low. It is the combination of an observation’s leverage and distance that determines its influence. [13]

**4.9. Multiple linear regression models in outlier detection.**

In 2012, three researchers wrote a paper using linear regression in outlier detection.

Outlier is an observation that is numerically distant from the rest of the data. Outliers are most extreme observations with maximum or minimum sample. A more general definition of an outlier is given in: an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Outlier detection methods have been used to detect and remove anomalous values from data. There are three approaches in the outlier detection process.

Determine the outliers through learning approach. No prior knowledge of the data is similar to unsupervised clustering. This approach processes the data as a static distribution, pinpoints the most remote points and identify them as potential outliers.

Pre-labeled data are marked as both normal and abnormal. This approach is similar to supervised classification. This is a semi-supervised recognition or detection task. This algorithm learns to recognize. [12]

**4.10. Applications of multiple linear regression**

Luminto worked on a multiple linear regression model to predict rice cultivation time and the result showed highest farmer’s exchange rate. The data are obtained using NSA (National Statistical Authority’s weather) and will be used to get a multiple linear regression model to detect the weather FR association. The factors are: “Radiation from Solar”, “Rainfall”, “Average Temperature” and “Average Moisture”. Prediction can be achieved by checking all the combinations of variables that caused a low value of RMSE. Evaluations show a cumulative RMSE of 0.39 - 1.34 in a suggested study in two different regions. [1]

Sousa, Martins, Alvim-Ferraz, Pereira uses multiple linear regression (MLR) to predict ozone concentration. This multivariate statistical technique transforms the original data set into a set of linear combinations of the original variables. They set out to predict next day hourly ozone concentrations. The experiment takes place in Oporto situated in North Portugal. The study considered as predictor variables the hourly concentrations of (ozone), NO (nitrogen monoxide), (nitrogen dioxide) and hourly means of temperature (T), wind velocity (WV) and relative humidity (RH) and this data were collected from an urban site with traffic influence (in Oporto) and integrated in the Air Quality Monitoring Network of Oporto Metropolitan Area.[2]

Thanh, Degener, Kappas uses multiple linear regression, cubist regression and random forest algorithms to estimate daily air surface temperature. The study area is located in northwest Vietnam. The study examines the mentioned combinations of four MODIS-LST datasets and shows that different combinations and differently applied algorithms produce various Ta (air surface temperature) estimation accuracies. [3]

An overview of most recent multiple linear regression techniques [11]

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Pros and Cons | Results and Accuracy | References |
| Pima Indian Diabetes | PCA-LRM achieves higher precision than other approaches. | 82.1% | [5] 2019 |
| Aero-Material | The very strong fit effect of the MLR can be shown to be smaller than 0.8 per cent for all relative prediction errors such that good prediction outcomes are produced by the MLR prediction. | 99.89% | [6] 2018 |
| 3D coordinate Medical Data | Provides a reliable basis for a detailed study of Chinese bone-setting manipulation and the utility of creation essence in Chinese traditional medicine | X: 92%  Y: 90%  Z: 80% | [7] 2018 |
| Marketing | Level of prediction really strong. | 84% | [8] 2018 |
| Rice, Wheat and Maize | MARS production for rice and wheat is better than MLRM and RFR and MLRM is better than maize RFR n and MARS. | Rice: 97% Wheat: 92%  Maize: 80%, | [9] 2018 |
| Weather | The fit of the model is very similar to 1, which means that the fit of the model is very good and the value of F statistics is very high. There is a strong linear correlation between variables, which satisfies the criteria for MLR. | 30 min 100%,  1 hour 99%,  3 hours 95% | [10] 2017 |

**5. Conclusion**

This paper presents theoretical characteristics about multiple linear regression as well as their application method.

In conclusion we will list two advantages of using multiple linear regression. The first advantage would be the ability to determine the relative influence of more predictor variables to the criterion value. The second advantage is the ability to identify outliers or anomalies. But the disadvantage is that there is no feedback of the message by the receiver.

We consider that multiple linear regression is useful because we can use it together with techniques such as variable recoding, transformation, or segmentation.

**6. BIBLIOGRAPHY**

1.H. Luminto, "Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate," in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications, Denpasar, Indonesia, 2017, pp. 16-18.

2. S.I.V Sousa, F.G. Martins, M.C.M. Alvim-Ferraz M.C. Pereira, “Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations”, in 2006 Available online at www.sciencedirect.com .

3. Phan Thanh Noi, Jan Degener and Martin Kappas, “Comparison of Multiple Linear Regression, Cubist Regression and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data”, in 2017 Available at remote sensing articles.

4. A. Mustapha, “Introduction to Statistics Using R”. Morgan Calypool Publishers, Washington University, St. Louis, 2019

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

6. Y. Yang, "Prediction and analysis of aero-material consumption based on multivariate linear regression model," in 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 628-632.

7. D. Wei, M. Xing, J. Zhang, C. Zhang, and H. Cao, "Applied Research of Multiple Linear Regression in the Information Quantification of Chinese Medicine Bone-setting Manipulation," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 1912-1916.

8. T. Gopalakrishnan, R. Choudhary, and S. Prasad, "Prediction of Sales Value in Online shopping using Linear Regression," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-6.

9. Najat, N., & Abdulazeez, A. M. (2017, November). Gene clustering with partition around mediods algorithm based on weighted and normalized Mahalanobis distance. In 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS) (pp. 140-145). IEEE.

10. X. Feng, Y. Zhou, T. Hua, Y. Zou, and J. Xiao, "Contact temperature prediction of high voltage switchgear based on multiple linear regression model," in 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2017, pp. 277-280.

11. Dastan Hussen Maulud, Adnan Mohsin Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning”, Journal of Applied Science and Technology Trends Vol. 01, No. 04, pp. 140 –147, (2020)

12. S.M.A. Khaleelur Rahman, M. Mohamed Sathik, K. Senthamarai Kannan, “Multiple linear regression models in outlier detection” International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 2 (2012) pp.23-28

13. David M. Lane, David Scott, Mikki Hebl, Ruby Guerra, Dan Osherson “Introduction to Statistics” Rice University and University of Houston

14. “boundless-statistics” Simple Book Production, Lumen Learning , “https://courses.lumenlearning.com/boundless-statistics/chapter/multiple-regression/”