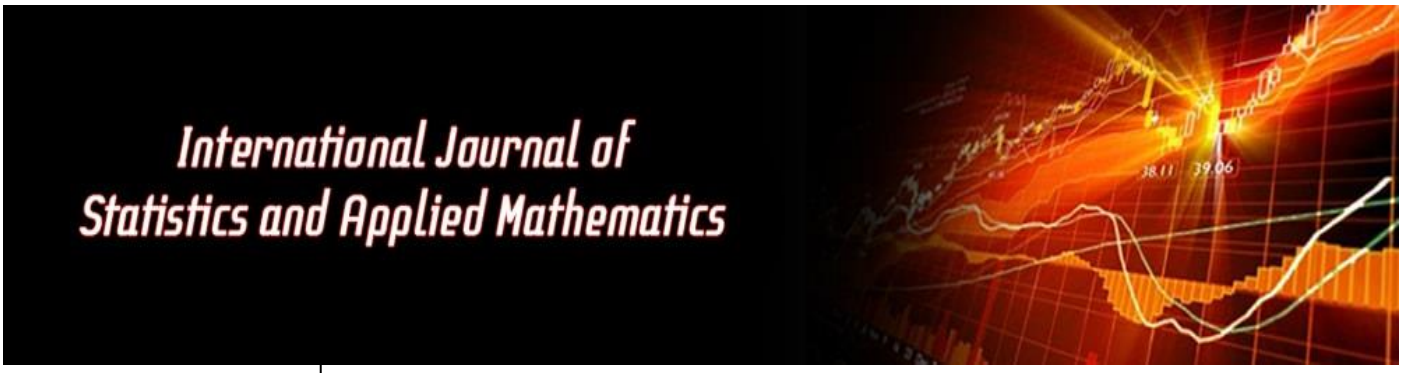
International Journal of Statistics and Applied Mathematics 2023; 8(6): 133-137



ISSN: 2456-1452
Maths 2023; 8(6): 133-137

https://www.mathsjournal.com
Received: 09-08-2023
Accepted: 16-09-2023

P Anandhi
Assistant Professor of Statistics, Department of Mathematics, Sona College of arts and science, Salem, Tamil Nadu, India

Dr. E Nathiya
Assistant Professor, Department of Statistics, Government Arts College for Women, Salem, Tamil Nadu, India

Corresponding Author:
P Anandhi
Assistant Professor of Statistics, Department of Mathematics, Sona College of arts and science, Salem, Tamil Nadu, India

Application of linear regression with their advantages, disadvantages, assumption and limitations

P Anandhi and Dr. E Nathiya

DOI: https://dx.doi.org/10.22271/maths.2023.v8.i6b.1463

Abstract
Regression analysis is one of the most commonly used strategies in statistics. The simple purpose of regression analysis is to match a version that finely describes the connection among one or more predictor variables and a reaction variable. Regression strategies are the most extensively used statistical strategies hired on a huge form of optimization troubles within the area of carried out studies. The fundamental forms of linear regression strategies could be reviewed along with their applications, advantages, and drawbacks to endorse a manner of choosing regression strategies for specific forms of optimization troubles.

Keywords: Direct retrogression, Simple direct retrogression, multiple direct retrogression

Introduction
In this paper, the application, advantages, assumptions and limitations and disadvantages of following linear regression strategies in studies are provided. Linear regression technique to are expecting rainfall in India turned into utilized by reference [1]. Reference [2], used the electricity regression technique to look at the impact of amassed oxygen deficit. According to the exponential electricity distribution, reference [3] advanced Bayesian evaluation for the linear regression version with random mistakes distribution. In reference [4], a contrast of estimating diffusive CH4 through closed chambers the use of linear and exponential regression turned into made. 12 ELEKTRON MAGAZINE Reference [5], provided the bearing Residual Useful Life (RUL) estimation through featuring a brand new method through combining data-pushed and version primarily based totally strategies. Estimation of bulk electricity structures the use of linear regression-primarily based totally disturbance value approach turned into provided through reference [6]. In reference [7], a more than one linear regression technique turned into used to forecast constructing strength performance. Reference [8], provided using more than one linear regression strategies with interactions to version and forecast hourly electric powered load. In reference [9], the strength performance of the economic homes turned into modelled the use of the bushy linear regression approach.

Direct retrogression
Linear regression is a method of analyzing data that predicts the value of unknown data using another associated and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a direct equation. Retrogression is a statistical fashion enforced in the fields of engineering, business, finance, clinical care, and other disciplines with the thing of discovering the correspondence between one dependent variable and a chain of different unprejudiced variables. There is numerous retrogression strategies described in the literature that are used for study purposes.
For illustration, one would conceivably need to determine a courting pattern among the burdens according to their heights through the use of the direct retrogression system. Before trying to fit a direct interpretation to the set up data, one needs to first test whether there's-- or no longer is-- a dating of pursuits among the variables. To estimate the robustness of the connection among variables, a matter plot may be a salutary tool. (1(♂ =). The canonical expression used for the direct retrogression system is proven through equation (1), in which y

is the established variable, x is the unprejudiced variable, a is the intercept value (while = 0), and b is the pitch of the line. Figure 1 suggests the direct retrogression wind. Linear retrogression is a machine literacy conception that's used to make or train models (Fine models or equations) for working supervised literacy problems related to prognosticating nonstop numerical values. Supervised literacy problems represent the class of problems where the value (Data) of the independent or predictor variable (Features) and the dependent or response variables are formerly known. The known values of the dependent and independent variables are used to come up with a fine model or formula, also called a direct retrogression equation, which is latterly used to prognosticate or estimate affair given the value of input features (the independent variable). In machine literacy tasks, direct retrogression is used to make the vaccination of numerical values from a set of input values. The following is an illustration of a uni variate direct retrogression analysis representing the relationship between height and weight in grown-ups using the retrogression line. The regression line is superimposed over the size-to-weight scatter diagram to illustrate the direct relationship.

Advantages of Direct Regression

1. Linear Regression plays properly while the dataset is linearly separable. We can use it to locate the character of the connection many of the variables.
2. Linear Regression is simpler to implement, interpret and really green to train.
3. Linear regression is vulnerable to over-becoming however it is able to be averted the use of a few dimensionality discount strategies, regularization (L1 and L2) strategies and cross-validation
4. After the linear regression method, the exponential regression method is an smooth one to recognize and observe due to the fact most effective 3 data portions are required for exponential regression.
5. It produces correct forecasts. The forecast is correct if the estimate of the difference among the real projections and what has occurred is lower.
6. When you realize the connection among the impartial and based variable have a linear relationship, this set of rules is the pleasant to apply due to it is much less complexity to as compared to different algorithms.

Disadvantages of Linear Regression

Main trouble of Linear Regression is the idea of linearity among the established variable and the impartial variables. In the actual world, the facts are hardly ever linearly separable. It assumes that there may be a straight-line courting among the established and impartial variables which is wrong many times.

1. On the alternative hand in linear regression approach outliers could have large consequences at the regression and obstacles are linear on this approach.
2. Diversely, linear regression assumes a linear dating among based and unbiased variables. That way it assumes that there may be a straight-line dating among them. It assumes independence among attributes.
3. But then linear regression additionally seems at a dating among the implies of the based variables and the unbiased variables. Just because the imply isn't a entire description of a unmarried variable, linear regression isn't a entire description of relationships amongst variables.

Implementation of linear regression

1. Epidemiology

Relating smoking to mortality and illnesses got here from the observational studies imposing linear regression evaluation. For example, we've got thirteen ELEKTRON MAGAZINE a linear regression model in which cigarette smoking is the explanatory variable, and the mounted variable is the lifespan of an character measured in years.

2. Finance

Linear regression and the beta concept are used for assessment and evaluation of the systematic chance of investment. This comes straight away from the beta coefficient of the linear regression model that relates the cross again on the investment to the cross again on all risky assets.

3. Econometrics

Linear regression is applied in economics as an optimization tool. In current econometrics. turning into the street through data elements reflecting paired values of the impartial and set up variables can be completed the use of linear regression estimating model.

4. Environmental Science

Environmental generation finds a large style of linear regression applications. Environmental effect monitoring on fish and benthic surveys to estimate the effect of metal mine or paper pulp at the aquatic environment uses linear regression techniques.

Assumption of Linear Regression

Linear regression is a beneficial statistical approach we are able to use to apprehend the connection among variables, x and y. However, earlier than we behaviour linear regression, we need to first make certain that 4 assumptions are met:

1. Linear dating

There exists a linear dating among the impartial variable, x, and the structured variable.

2. Independence: The residuals are impartial. In particular, there may be no correlation among consecutive residuals in time collection data.

3. Homoscedasticity: The residuals have regular variance at each degree of x.

4. Normality: The residuals of the version are typically distributed. If one or greater of those assumptions are violated, then the consequences of our linear regression can be unreliable or maybe misleading. In this post, we offer an cause of every assumption, the way to decide if the belief is met, and what to do if the belief is violated.

The line is continually a directly line- There isn't anyt any curve or grouping element for the duration of the conduction of a linear regression. There is a linear courting among the variables (based variable and unbiased variable). If the facts fails the assumptions of homoscedasticity or normality, a nonparametric take a look at is probably used. (For example, the Spearman rank take a look at). Example of facts that fails to fulfil the assumptions: One might imagine that cured meat intake and the occurrence of colorectal most cancers within side the U.S have a linear courting. But later on, it involves the information that there's a completely excessive variety distinction among the gathering of facts of each the variables. Since the homoscedasticity assumption is being violated here, there may be no linear regression take a look at. However, a

Spearman rank take a look at may be finished to realize approximately the connection among the given variables.

Limitations of the Linear Regression

We cannot take a look at linear regression blindly on any of the datasets. The records desires to be within side the constraint such that we're capable of take a look at a Linear Regression set of guidelines on it. There are a few boundaries that need to be satisfied. These are:

- **Linearity**
- **Constant Error Variance**
- **No autocorrelation of the**
- Residuals
- Normal Errors
- Multicollinearity 4% error
- Exogeneity or Omitted Variable Bias

Linearity

The relationship many of the aim variable and the independent variable want to be linear. Linear Relationship vs No Relationship amongst independent and established variables Sometimes the immediately line may not be the right in shape to facts and we can also need to choose the polynomial function like beneathneath root, square root, log, and so on to in shape the facts**.**

Constant Error Variance (Homoscedasticity or no Heteroskedasticity)

Homoscedasticity describes a scenario in which the error term is the same at some stage in all values of the independent variables. If we have got were given a dataset wherein the spread of the data or variance will growth as X will growth then there can be a problem. And it'd now no longer be the great idea to use linear regression in such scenarios. Or in unique words, the residuals of the elements want to now no longer observe any pattern. Let's plot a scatter plot amongst primarily based totally and independent variables: To check the heteroskedasticity of the data, we plot residual plot and the expected cease end result is that the plot want to be randomly spread out and there want to now no longer be any patterns.

Independent Error Terms or No autocorrelation of the residuals

The residual term need to now no longer depend on the previous residual term. Or in unique words, y(x) is relying on y(x+1). This assumption makes experience while we are dealing with time series related data. Consider an example of the stock rate, wherein the present-day rate is relying at the previous rate. This violates the concept of the Independent Error Terms.

Normal Errors

Residual want to conform with a bell-fashioned distribution with the advice of 0. In unique words, if we draw a histogram of the residual term it wants to be a bell shape curve having an average close to 0 with regular popular deviation. Residual Terms following normal distribution. The normality assumption of errors is important because of the truth on the identical time as predicting individual facts points, the self-guarantee interval spherical that prediction assumes that the residuals are generally distributed. We need to use 'Generalised Linear Models' if we want to lighten up the normality assumption.

Residual want to comply with a bell delivery with the recommendation of 0. In unique words, if we draw a histogram of the residual term it wants to be a bell shape curve having an average close to 0 with regular popular deviation. Residual Terms following normal distribution The normality assumption of errors is important because of the truth on the identical time as predicting individual facts points, the self -guarantee interval spherical that prediction assumes that the residuals are generally distributed. We need to use 'Generalised Linear Models' if we want to lighten up the normality.

Multicollinearity

Multicollinearity takes place while the independent variable X is predicated upon on the alternative independent variable. In a model with correlated variables, it's miles difficult to determine out the real relationship many of the independent and primarily based totally variable. In unique words, it turns into difficult to find out which independent variable is simply contributing to assume the primarily based totally variables. Additionally, with correlated variables, the coefficient of the independent variable is predicated upon on the alternative variables present within side the dataset. If this happens we will come to be with an incorrect cease of independent variables contributing to the prediction of the primarily based totally variable. The great way to check for the multicollinearity is with the useful resource of the use of plotting heat map. The variables having immoderate correlations are multicollinearity. Heat map of sklearn boston dataset.

Exogeneity or Omitted Variable Bias:

Before we understand Exogeneity it's far crucial to understand while we generate a linear regression line there can be an mistakess associated with it. (It is not like residual).

$$y = ax_1 + bx_2 + \ldots\ldots + nx_n +$$

wherein represents all of the factors that impact the aim variables and is not covered within side the model. Consider a feature A which is not covered with inside the model. So, it's far the part of the error term. And moreover, A has a immoderate correlation with the x2 and y variable. This will make coefficient b as biased and will now no longer be an actual coefficient. (i.e. Sample is not a reflected photo of population value) Or in unique words, A variable is correlated with an independent variable within side the model, and with the error term. And the real model to be anticipated is: but we by skip over z_i while we run our regression. Therefore, zi receives absorbed thru the error term and we will definitely estimate: (where in) If the correlation of and is not 0 and one after the alternative affects, then is correlated with the error term. Therefore, Exogeneity or Omitted Variable Bias takes place while a statistical model leaves out one or extra relevant variables. The exceptional way to deal with endogeneity issues is through instrumental variables (IV) techniques. And the most now no longer unusual place IV estimator is Two Stage Least Squares (TSLS).

Types of Linear Retrogression

Typically, direct retrogression is divided into two types multiple direct retrogression and simple direct retrogression. so, for better concurrence, We will bandy these types in detail.

Simple Linear Retrogression: Simple direct retrogression is a statistical system that allows us to epitomize and study connections between two nonstop
(Quantitative) variables. Using simple direct retrogression, it's possible to identify connections between two quantitative variables. One can use simple direct retrogression to establish,

1. How tightly are two variables related to one another (for case, how downfall and soil corrosion are related)?
2. The Quantum of the independent variable at a specific position that the dependent variable is at (e.g., the quantum of soil corrosion at a certain position of downfall).

Advantages of simple linear regression

The biggest advantage of linear regression models is their linearity – this means that the estimation procedure is easy to understand and follow on a modular level. Additionally, these equations are straightforward to interpret, making them easier to comprehend than nonlinear models.

Disadvantages of simple linear regression

The sup position of linearity between dependent and independent variables it is frequently relatively prone to noise and overfitting. Linear regression relatively sensitive to outlier. It's prone to multicollinearity.

Applications of simple linear regression

Marks scored by scholars grounded on number of hours studied (immaculately)- Then marks scored in examinations are independent and the number of hours studied is independent. Predicting crop yields grounded on the quantum of downfall- Yield is a dependent variable while the measure of rush is an independent variable. Predicting the Salary of a person grounded on times of experience- thus, Experience becomes the independent while Salary turns into the dependent variable.

Multiple Linear Regression

Multiple direct retrogression relate to a statistical trend that's used to prognosticate the consequence of a variable grounded on the values of two or further variables. It's occasionally known simply as multiple retrogression and it's an extension of direct retrogression. The variable that we want to predict forecast is known as the dependent variable, while the variables we use to prognosticate forecast the value of the dependent variable are known as independent or annotative variables.

Multiple Linear Regression Formula

$$y_i=\beta_0+\beta 1X_i1+\beta 2X_i2+\ldots+\beta pX_ip+Є$$

Where,

- yi is the dependent or predicted variable
- β_0 is the y-intercept, i.e., the value of y when both x_i and x_2 are 0.
- β_1 and β_2 are the regression coefficients representing the change in y relative to a one-unit change in x_{i1} and x_{i2}, respectively.
- p is the slant coefficient for each independent variable
- ϵ is the model's random error (residual) term.

Advantages of Multiple linear regressions

There are main blessings to reading statistics using a multiple regression model. The first is the capability to determine the relative have an effect on of one or more predictor variables to the criterion value. The real belongings agent may also need to find out that the dimensions of the homes and the extensive type of bedrooms have a sturdy correlation to the fee of a home, on the identical time because the proximity to schools has no correlation at all, or perhaps a horrible correlation if it`s far generally a retirement community. The second benefit is the capability to understand outliers, or anomalies.

Disadvantages of Multiple Regressions

Any downside of the usage of a a couple of regression model usually it comes right all the way down to the records being used. Two examples of this are the usage of incomplete records and falsely concluding that a correlation is a causation. Linear regression executes poorly when there are non-linear relationships.

Applications of Multiple Regressions

It can be used to prognosticate the relationship between reckless driving and the total number of road accidents caused by a motorist or, to use a business illustration, the effect on deals and spending a certain amount of money on advertising. Retrogression is one of the most common models of machine literacy.

Conclusion

Regression techniques are the types of predictive modelling techniques that investigate the correspondence among two variables in which one is dependent and the other is an independent variable. Many regression techniques have been developed and many more are in process of making. I have discussed about linear regression. Linear regression is simple to implement but does not give accurate results. Regression techniques are useful statistical methods that can be leveraged to estimate the degree to which independent variables are affecting dependent variables. These regression techniques should be implemented according to the limits defined on the given data set. One of the best ways to explore which regression technique should be implemented on the problem is to check the family of the variables involved in that problem**.**

References

1. Wi YM, Joo SK, Song KB. Holiday load forecasting using fuzzy polynomial regression with 17 Elektron Magazine weather feature selection and adjustment. IEEE Trans Power Syst. 2011;27(2):596–603.
2. Hu Z, Gao J. Uncertain Gompertz regression model with imprecise observations. Soft Comput. 2020;24(4):2543–2549.
3. Roush W, Dozier W, Branton S. Comparison of Gompertz and neural network models of broiler growth. Poult Sci. 2006;85(4):794–797.
4. Gabriel JM, Moris YG, Moral CP, Calvo CM, Beltrá RL. A Gompertz regression model for fern spore germination. Anales del Jardn Botánico de Madrid. 2015;72(1):1–8.
5. Wu L, You S, Dong J, Liu Y, Bilke T. Multiple direct retrogression-ground disturbance magnitude estimations for bulk power systems. In: 2018 IEEE Power & Energy Society General Meeting (PESGM). IEEE; c2018. p. 1–5.
6. Ciulla G, Amico AD. Structure energy performance soothsaying: A multiple direct retrogression approach. Appl Energy. 2019;253:113500.

7. Hong T, Gui M, Baran ME, Willis HL. Modelling and vaticinating hourly electric cargo by multiple direct retrogression with relations. In: IEEE PES General Meeting. IEEE; c2010. [Specify page range].
8. Chung W. Using the fuzzy direct retrogression system to standard the energy effectiveness of marketable structures. Appl Energy. 2012;95:45–49.