Dean&Francis
ISSN 2959-6157

Unveiling the Influencing Factors of Advertisement and Study by Linear Regression Approach

Xirui Zhao

Chongqing DEPU Foreign Language School, Chongqing, China

*Corresponding author: zhaoxirui@163offer.com

Abstract:

The linear regression method is important in natural science. This method is widely used in economics, data science, management, medicine, biology, sociology and so on. This statistic method helps people to find the relationship between inputs and outputs and classify massive data. In the following article, some basic theories, formulas and real problems with solutions will be mentioned. This essay firstly introduces some methods about linear regression, and then use the formula and basic theories to calculate and analyze the relationship between independent and dependent variables of some real massive data. Finally, it can be found that all of data have relationship and people can use these conclusions to predict or give a suitable strategy, so it is really useful for people to prepare solutions ahead of schedule. It is one of the most fundamental methods in dealing with data, because the model of linear regression is quite easy, and provide understandable mathematical formulas, then create some scientific results.

Keywords: Simple linear regression, Multiple linear regression, Advertisement.

1. Introduction

Regression line is a type of useful statistic method to organize and analyze data. Normally, It can be used to calculate some results by linear regression formula, and predict a reliable and scientific changing trend in the future, and helps people to prepare solution ahead of time. People need to make the model better fit to some unknown parameters, mainly use least square method to fit, this method can not only deal with linear model, but also suitable for non-linear model [1]. Linear regression is quite important in many fields, it can create predictions quickly, many enterprises use linear regression as a reliable and accurate way to convert original data to workable suggestions to their own business. Set a suitable model for each database, work out unknown coefficients, plug them into formulas to give an analyzable result.

ISSN 2959-6157

For example, use this month's expenses to predict next month's expenses [2].

In this essay, firstly, the author introduces basic knowledge background: the fields that mainly use linear regression, process of calculating correlation of data, the formula of setting a regression line model. The most important two parts of linear regression are simple linear regression and multiple linear regression. For simple linear regression, one can use Maximum Likelihood Estimate and Least Squares Estimate to calculate. For multiple linear regression, it is specially adapted for problems with more than one variable. Both of them need to set hypothesis and test the significance of each model, it also represents the validity of model. Then, the author will find two real examples from online database, the first one is about how advertising and promotion costs have influence on sales, and the other one is about whether sleeping and studying time influences test scores. Finally, both of their results show that there exists a positive relationship between independent variables and dependent variable. Thanks to these predictions, people can prepare strategies that solve problems faster and better, gain benefits as much as possible. This method not only improve the efficiency but also provide maximum advantages to people in the limited time and resources.

2. Basic information about line regression

2.1 Correlation of Data

This method is widely used in Statistics, especially when dealing with linear relationships between variables. For example, when discuss the relationship between the level of salary and life happiness. Consider the level of salary as x, and life happiness as y. Normally, y will increase with the increase in x, and then y and x have correlation [3].

There is a system to calculate the correlation coefficient, for a test with sample size n: (x_1,x_1), (x_2,y_2), …, (x_n,y_n). For a fixed j, x_j and y_j come from a same experiment. Next, calculate means of $\{x_j\}$ and $\{y_j\}$, represented as $\bar{x}$ and $\bar{y}$ respectively. Use s^2x to represent the sample variance of $\{x_j\}$ and s^2y to represent the sample variance of $\{y_j\}$, finally, introduce sample covariance

$$s_{xy}=\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(y_j-\bar{y}\right) \tag{1}$$

When $s_x s_y \neq 0$, then use $\hat{\rho}=\frac{s_{xy}}{s_x s_y}$ as the sample correlation coefficient of $\{x_j\}$ and $\{y_j\}$. If $\hat{\rho}$ is larger than 0, it can be seen that exist a positive correlation, if $\hat{\rho}$ is smaller than 0, it can be seen that exist a negative correlation, if $\hat{\rho}$ is equal to 0, means no correlation. The closer $\hat{\rho}$ is to 1 or -1, data more likely to lie on a straight line.

2.2 Regression Line

When $\{x_j\}$ and $\{y_j\}$ are highly correlated, data is scattered around a straight line, and this straight line is called regression line.

Generally, this method is used for observational data with sample size n, (x_1,y_1), (x_2,y_2), …, (x_n,y_n), then establish regression line $y=a+bx$. Since the distance between point (x_j,y_j) and the intersection point with regression line is $\left|yj-\left(a+bx_j\right)\right|$, so use these distances to calculate their sum of squares [4]

$$Q(a,b)=(y_1-a-bx_1)2+(y_2-a-bx_2)2+\ldots +(y_n-a-bx_n)2 \tag{2}$$

measuring how far the data is from a straight line. When constants a and b make the smallest value of $Q(a,b)$, then call this straight line a regression line.

The point with the smallest value of $Q(a,b)$ is $\hat{b}=\frac{s_{xy}}{s_x s_y}$, $\hat{a}=\bar{y}-\hat{b}\bar{x}$ And call $\hat{a}$ and $\hat{b}$ are the least squares estimate of the regression linear coefficients a and b. The core of this method is finding the optimal linear relation of independent variables and dependent variables of data. Make sure that regression line is close to practical data points by minimizing the sum of squares of error, so prediction can be more accurate.

2.3 Simple Linear Regression

In this method, one can use

$$\hat{\epsilon}_j=y_j-\hat{y}_j=y_j-\hat{a}-\hat{b}x_j \tag{3}$$

to represent prediction error or residual error when y_j is predicted value, and the sum of squares of residuals equals to $Q=\sum_{j=1}^{n}\hat{\epsilon}_j{}^2$. When Q is low, regression line shows the linear relationship between x and y: $y_j=\hat{a}+\hat{b}x_j+\hat{\epsilon}_j$, a

and b are unknown constants, $\{\epsilon_j\}$ are independent equally distributed random variables, while follows normal distribution $N(0,\delta^2)$, this model is called linear regression model [5].

Maximum Likelihood Estimate is widely used to estimate unknown parameters, find estimate of parameter by using known data to maximize likelihood function value. Likelihood function is defined as

$$L(\theta)=\prod_{i=1}^{n} f(x_i\mid\theta) \tag{4}$$

$f(x_i\mid\theta)$ is the probability density function of the function x. People want to simplify calculation, usually take the logarithm of the likelihood function, and then get log-likelihood function $\phi(\theta)=log_{L(\theta)}=\sum_{i=1}^{n} f(x_i\mid\theta)$, both of these two functions are monotonically increasing. Another commonly used method is Least Squares Estimate, the main aim of this method is minimizing the square difference between the observed value and the predicted value of the model. The whole calculating process involves four key steps, first of all, work out the residual error $\in_i=y_i-(\beta_0+\beta_1 x_i)$, finally calculate the sum of squares of residual errors, then derivative β_0, β_1 and let the derivative be 0, so the parameters that makes the smallest SSR. There are formulas to calculate least squares estimation, $\widehat{\beta}1=\dfrac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2}$ and $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$.

If and only if $b\neq 0$, regression result makes sense. And test b. State the null hypothesis $H_0:b=0$ vs. the alternative hypothesis $H_1:b\neq 0$. If H_0 is valid, $T=\dfrac{\hat{b}}{\dfrac{\hat{\sigma}}{\sqrt{l_{xx}}}}t(n-2)$, so the rejection filed of level α is $W=\left\{|T|>\dfrac{t_\alpha}{2(n-2)}\right\}$.

Predicted Confidence Interval estimates range between new prediction and prediction, also affect the uncertainty of predicted outcomes. And the formula of this is

$$SE(\hat{y}*)=\sqrt{\delta^2(1+\frac{1}{n}+\frac{\left(x-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}} \tag{5}$$

and therein $\delta^2=\dfrac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}{n-2}$.

2.4 Multiple Linear Regression

This method is typically used in finding the relationship between several independent variables and one dependent variable, in order to minimize the residual squares of model. The main expression of this model is $y=x\beta+?$, y represents independent variable of $n\times 1$, x represents independent variables of $n\times p$, β represents regression coefficient vector of $p\times 1$, ? represents error vector of $n\times 1$, therein n is observed quantity and p is the number of independent variables [6].

Least squares estimation is to find regression coefficient $\hat{\beta}$, then minimize the sum of residual squares. Use the formula

$$\hat{\beta}=\frac{1}{(X^TX)}x^Ty\#(6)$$

The regression significance can be tested, also called F test. Use the formula $F=\dfrac{(\hat{y}^T\hat{y}-\frac{(\sum y_i)^2}{n}/p}{(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-p-1)}$, if F is out of critical value, then reject null hypothesis, and regression model is significant. If only want to test single regression coefficient, use the formula $t_j=\dfrac{\hat{\beta}_j}{\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}\dfrac{1}{(X^TX)}j^2}}$, if the absolute value of t is out of critical value, then reject null hypothesis, and regression model is significant.

3. Appliacations and Examples

3.1 Applications

These methods are used for discussing the relationship between one or more independent variables and one dependent variable. In addition, linear regression can be divided into two types depend on the number of independent variables, involves simple linear regression and multiple line regression. For simple linear regression, it only can be used for the relationship between one dependent variable and one independent variable, so show the direct impaction of one certain dependent variable to the independent

ISSN 2959-6157

variable. For multiple line regression, it can be used for the relationship between more than one variable and one certain independent. This method can solve more difficult problems, and the result is more accurate and more reliable. These methods are widely used in many different aspects in daily life. Such as medicine and public health, social science, engineering. Especially in economics and finance, use regression line, economists can predict the changing trend in the future, useful strategies can be given to merchants. For example, the housing price will be affected by facilities, positions, transportation and so on.

After setting a suitable regression model, the next necessary step is assessment. Common assessment indictors are R^2 and error of mean square. For example, if the value of R^2 is closer to 1, that means the model is more suitable to the data. In addition, residual analysis is also essential, the data need to be tested to decide whether the hypothesis is established. Another important issue is the stability of model will be reduced when dependent variables are highly similar to each other. Then need to use Ridge Regression or Lasso regression to solve this problem.

3.2 An example

In this subsection, the paper mentions an example using methods of simple linear regression and multiple line regression. This example discusses the relationship of three key indicators. Some data shown in Table 1 is needed to use.

Table 1. Three variables comparing. x_1 and x_2 are independent variables, while y is dependent variable.

Advertising costs x_1	Promotion costs x_2	Sales y
1.5	2.0	20
2.0	2.5	23
2.5	3.0	26
3.0	2.5	30
3,5	4.0	34

For simple linear regression, one will use the formula of $y = \beta_0 + \beta_1 x + ?$. Therein y represents sales with unit of 10000 yuan, x represents shopping advertising expenses with unit of 10000 yuan, β_0 and β_1 are model's intercept and slope respectively, ? is the random error term. Firstly, one can calculate means of x and y. $\bar{x} = \frac{1.5+2.0+2.5+3.0+3.5}{5} = 2.5$ and $\bar{y} = \frac{20+23+26+30+34}{5} = 26.6$. Then, one will calculate the value of β_0 and β_1,

$$\beta_0 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{17.5}{2.5} = 7.0$$ and

$\beta_0 = \bar{y} - \beta_1 \bar{x} = 26.6 - 7.0 \times 2.5 = 9.1$. Finally, the linear equation is obtained, $y = 9.1 + 7.0x$. Through this calculating process, some conclusions can be gotten. First of all, there exists a positive relationship between shopping advertising expenses x and sales y, that means y will increase by 70000 yuan with each 10000 yuan increase in x. Secondly, sales will be 91000 yuan without any shopping advertising expense.

For multiple line regression, one will use the formula of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in$. Therein x_1 and x_2 represent advertising costs and promotional costs individually, while β_1 and β_2 are regression coefficients of these two variables individually. And through many calculating steps, the linear equation is obtained, $y = 6.53 + 1.87x_1 + 5.13x_2$. More conclusions can be gotten compared to simple linear regression. First of all, advertising costs and promotional costs are directly proportional to sales, that means y will increase by 18700 yuan and 51300 individually with per 10000 yuan increase in advertising costs and promotional costs. Moreover, sales will still be 65300 yuan without any advertising costs and promotional costs.

In a nutshell, these two methods have similarities, shops still have significant sales without any marketing methods, like advertisements or promotion mentioned before. That reflects other factors that can improve sales, like brand awareness and shopping services. However, many differences between them. The accuracy of using simple linear regression is less than that of using multiple line regression, since the former only focus on a single factor, so applicability and explanatory power are limited. Therefore, although the sample size of this research is not enough, but it can show the truth that the prediction from data can give useful strategies and solutions to merchants, helping them to enlarge their benefits of goods.

3.3 Another example

Here is another example using methods of sample linear regression and multiple line regression. This example discusses the relationship of another three key indicators. Some data needed to use are shown in Table 2.

Table 2. Three variables comparing. x_1 and x_2 are independent variables, while y is dependent variable.

Studying time x_1	Sleeping time x_2	Test scores y
2	6	75
3	5	78
4	7	82
5	6	85
6	8	88

For simple linear regression, one will use the formula of $y = \beta_0 + \beta_1 x + ?$, therein y represents test scores with unit of points, x represents studying time with unit of hour, β_0 and β_1 are model's intercept and slope respectively, ? is the random error term. Firstly, one will work out the mean of x and y, $\bar{x} = \frac{2+3+4+5+6}{5} = 4$ and $\bar{y} = \frac{75+78+82+85+88}{5} = 81.6$. Next, one can calculate the value of β_0 and β_1 and get the regression equation $y = 70 + 3x$, and then the relationship between studying time and test scores can be seen positively, test score will increase 3 points with per hour increase in studying time. And even pay no attention on studying, students can still get certain points.

For multiple line regression, one will use the formula of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in$. Therein x_1 and x_2 represent studying time and sleeping time respectively, while β_1 and β_2 are regression coefficients of these two variables individually. After complicated calculations, additional conclusions can be gotten. Both of studying time and sleeping time are directly proportional to test scores. Specifically, test scores will increase 4 and 2 points respectively with per hour increase in studying time and sleeping time. Additionally, the test score will be kept unchanged without studying time and sleeping time.

In a nutshell, both factors impact test score. But the study duration has a deeper effect compared to sleep duration. By contrast, if a student spends no time on studying or sleeping, still can get a basic score. Therefore, this example shows some suggestions to manage time suitably for students. And the truth that the testing result can be predicted from some living habits.

4. Conclusions

This essay can be divided into three main parts. The first part is basic introduction about linear regression, common fields of application, importance of linear regression, and general aims of linear regression. The second one is about some main theoretical introductions involve correlation of data, regression line, simple linear regression and multiple linear regression, and give formulas of them. The last one is real examples. Two practical applications about linear regression are the relationship between advertising, promotion costs and sales, and the relationship between studying, sleeping time and test scores. Use formula to calculate and organize data, and finally find some conclusions. From the first example, improve advertising and promotion costs can increase sales. This conclusion gives helpful suggestion for merchants to market, get a maximum benefit from goods. From the second example, longer studying and sleeping time can make test scores better. This conclusion gives useful advice for students, parents and teachers to improve learning efficiency, give them solutions that how to manage time better to get a better result in the limited time. In a nutshell, this method is too effective, maybe people can create a precise system to organize and calculate massive data by linear regression, then people can find the relationship between several variables directly without much calculation.

References

[1] Xiaogang Su, Xin Yan, Chih-Ling Tsai. Wiley Interdisciplinary Reviews: Computational Statistical. 2012, 4 (3): 275-294.

[2] Douglas C Montgomery, Elizabeth A Peck, G Geoffery Vining. John Wiley & Sons, 2021.

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor. An introduction to statistical learning: With applications in python, Springer, 2023.

ISSN 2959-6157

[4] Nikolaos Pandis. Multiple linear regression analysis. American journal of orthodontics and dentofacial orthopedics 149 (3), 431-434, 2016.

[5] Weixin Yao, Longhai Li. A New Regression Model: Modal Linear Regression. Scandinavian Journal of Statistics, 2014, 41(3), 656-671.

[6] Jain, Kirti, et al. Applications of Multitarget Regression Models in Healthcare. Machine Learning in Healthcare and Security, CRC Press, 2024.