Application of Multiple Linear Regression on Sales Prediction

Chen Dong

Woodland Christian High School, Ontario, N0B 1M0 Canada

jing@ito-sh.com.cn

Abstract. Hitherto, companies still facing issues of pricing and allocation of promotional funds. Some evidence already shows the relationship between the sales of the product and three elements which include price, in-store spending, and online advertisement spending. This project aims to predict sales to determine the equation between these factors to help companies maximize the efficiency of promotional funds and balance the product's price and its sales. With the use of multiple linear regression and the least square method, predicted sales can be indicated. This is done through using both the R-square and graph the R-square to evaluate the equation between sales and the three factors mentioned earlier, discovering the fitness of equation is 0.8 (the upper limit is 1) according to the R-square and the trend of predicted values meet the observed values according to the graph. This finding can be used to predict the sales trend for the product and help the enterprise manage the allocation of promotional funds.

Keywords: Sales Prediction, Multiple Linear Regression, least square method.

1. Introduction

Sales prediction and pricing strategy are two of the most important factors for achieving higher efficiency in balancing and optimizing the cost and the profit. In-store spending, online advertisement spending and price can influence the sales. Depending on the promotional funds' allocation and pricing strategy, they will affect the sales to different degrees. For instance, according to the magnitude of influence created by a certain quantitative of promotional funds, putting most of the expenditure into the area with a more significant impact is more effective for the company, and at the same time, it is also necessary for the company to measure the impact of product prices on total sales. This research aims to determine the relationship between sales and each factor.

In-store spending mainly involves the indoor environment. Store capacity, store decoration, and in-store promotions can affect sales to different degrees by attracting store traffic, among these, a narrow range of a store can negatively affect store traffic, while new product promotions can better encourage customers to purchase products [1]. Besides, there is evidence indicating that the atmosphere of the store has an impact on arousing customers' emotions of excitement or pleasure. These emotions will cause customers more likely to stay in the store for a long time and cost more than their budgets [2]. In further research, through overall consideration of environmental psychology and strategies of sales promotion, stores can create a sustained impact by combining short-term and long-term promotional strategies due to the feature of consumption is more related to short-term promotion, while the increasing attraction is produced by long-term promotion [3]. Online advertisements will positively affect the sales [4]. Especially, when advertisements try to use strong emotion to influence the audience [5], at the same time, though the promotional funds spent on repeated advertisements have little influence on customers' decision-making they have already spent on the specific product, it continues to remind the other the brand of the advertisement, which can promote the increase of the number of customers for their acceptability towards the product has been enhanced [6]. The last factor is price, most product sales will decline when the price increases according to research [7], however, there still exists some exceptions, the high consumption of products representing reliable quality [8]. According to the trend of product price, it is possible to predict sales [7].

Overall, this research aims to predict the sales through the promotional funds the enterprise spends on each element and price using linear regression. The organization of this paper is the followings. Section 2 will introduce the method and theory for the prediction, Section 3 will illustrate the result

of the multiple linear regression application by using the approach mentioned in section 2, and the last Section is devoted to the conclusion.

2. Method and Theory

2.1. Multiple Linear Regression

Through predicting sales and drawing up a pricing strategy, companies can dynamically adjust commodity prices, which will boost profits during periods of peak demand as well as sales during periods of low demand. Furthermore, after a prediction of sales, companies can better manage their inventory to reduce the waste of storage space in order to lower operating costs. So that companies can optimize their expenditures to enhance their competitiveness, and the manufacturers can also benefit from an accurate prediction [9]. There are some researches using linear regression to solve cosmic distance scale problems in various ways [10], medical treatment field to analysis lifetimes [11], coping plans for epidemic disease [12] and Face Recognition application [13]. While these researches are more focused on social, or as above mentioned, focusing on how variables lead to changes in total sales, this research faces the prediction of the sales. Inspired by these applications of linear regression, through these main elements, sales can be predicted through linear regression and the least square method.

Besides, there are still some researchers learning the possible sales based on machine learning techniques [14] or using customer demographics to make the prediction [15]. These projects forecast the sales using in other ways. With the increasing number of methods that exist, it has been found that when a model contains price information, the result will be more accurate [16]. Therefore, except for the publicity both online and offline, this project selects the product price as a factor for the prediction. This research includes two methods, the least square method and multiple linear regression. The least square method is to select the least residual sum of squares for β estimating. This aims to reduce the differences between the predicted value and the observed value. The result can reflect the effect of independent value on influencing the dependent value

$$\hat{\beta} = \arg\min_{\beta \in R^P} \|Y - X\beta\|^2 = \arg\min_{\beta \in R^P} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{p-1} \beta_k x_{ik} \right)^2 \quad (1)$$

with $x_{i0} = 1$. When X is full rank ($p < n$), the β that minimizes the sum of squared residuals can be obtained as $\hat{\beta}$ = (X'X)-1X'Y. In this case, $\hat{\beta}$ is referred to as the least squares estimate of β, and the fitted value is

$$\hat{Y} = X\hat{\beta} = (X'X)^{-1}X'Y. \quad (2)$$

Linear regression is a tool that use two or more variables to explain the predicted value. When there is only one variable $y = \beta_0 + \beta x$. and while the variables increase, the formula becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots. \quad (3)$$

Through put the independent value into the equation in section 2.2.1, β can be derived, after putting them into Eq. 3. The relationship between the independent variables (in-store spending, online advertisement spending and price) and dependent variable (y) can be obtained.

2.2. Dataset

To study the project, the author applies a dataset from Kaggle to the research [17]. The dataset contains 992 data with three variables that can affect sales, in-store spending, online advertisement spending and price, which are all proven to be related to the overall sales. Its basic unit is one dollar, the expenditures spent on each aspect affect the Product impact of one specific product, and ultimately affect total sales. The differences in promotional funds of each aspect have a relatively prominent

difference, allowing comprehensive research in a wide range. Building on this dataset, linear regression is effective in calculating the relationship between sales and in-store spending, online advertisement spending and price.

3. Result

The research uses data from Kaggle with 992 numbers of data in it. The dataset shows the sales change and lists three independent variables in it (in-store spending, online advertisement spending, and price) collected from a specific product. As Table 1 shows, there is a Summary Statistics indicates the basic features of the dataset applied in the research. The differences between the maximum and minimum values for each allow more comprehensive research.

Table 1. Summary Statistics of Sales Data

cases	Sales	In-store spending	Price	Online advertisement spending
std	81397.84	17.49	8.72	927.47
mean	171327.12	30.59	14.60	1569.50
min	1992.00	0.19	0.14	1569.50
max	393914.00	59.96	29.99	3198.27

Through the least square method, it provides coefficients for the three variables (in-store spending, online advertisement spending and price) and the intercept, 2,833.61306, 0.52944, 6,543.00074, 179,319.74170 respectively. According to the formula of linear regression (as shown in Eq. 3). Combining the formula and the coefficients, the equation between in-store spending, online advertisement spending and price and price can be found:

$$y = 179{,}319.74170 + 2{,}833.61306 \times x_1 - 6{,}543.00074 \times x_2 + 0.52944 \times x_3 \quad (4)$$

In this case, y represents the sales, and x represents the in-store spending, price, and online advertisement spending. The intercept β_0 refers to the sales that products can most likely achieve regardless of the variables change.

Table 2. Inaccuracies between Sales and Prediction

observation	Sale	Predicted	Residual	Standardized
Minimum	1992.000	-3860.653	-66847.148	-1.953
Maximum	393914.000	344305.784	71723.908	2.096
Mean	171327.118	171327.118	-1.833e-13	-6.379e-18

As the Eq. 4 shows, comparing the publicity, in-store spending (2833.61306) is much more efficient than online advertisement spending (0.52944). According to this information, allocating more promotional funds on in-store spending will be beneficial for product popularity and its sales. The coefficient of price stands at -6543.00074. While the price per unit increases, it will reduce the overall sales. It is important to evaluate these coefficients. By determining the variables that bring varying degrees of change to product sales, companies can find the optimal scheme that will maximize their profits. After the calculation, there were some inaccuracies between the observed value and the predicted value as Table 2. The sale and predicted value for both maximum and minimum have a significant difference in it, whereas the mean value fits the reality with a slight error. R^2 can evaluate the goodness of fit of the equation. It tells the capacity of the equation to explain the whole trend. Its formula shows below:

$$R^2 = 1 - \frac{SCE_p}{SCE_{tot}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}, \quad (5)$$

Where SCE_p denotes the sum of squares relates to regression and SCE_{tot} represents the total sum of square [16]. By measuring the value of R^2 using programming. The R^2 equals 0.82381 as the

result shows, When R^2 more approaches 1, meaning that the result is more fit the observed situation, otherwise, when the values are close to zero, the prediction is unreliable. In this case, the R^2 value is closer to 1. Another way to calculate the R^2 is to find the R value. It is the relationship between the independent variables and dependent variables with the range between zero to one. After determining the value, square the value. For example, the R-value for the equation in this research equals 0.90764. $0.90764^2 = 0.8238103696$.

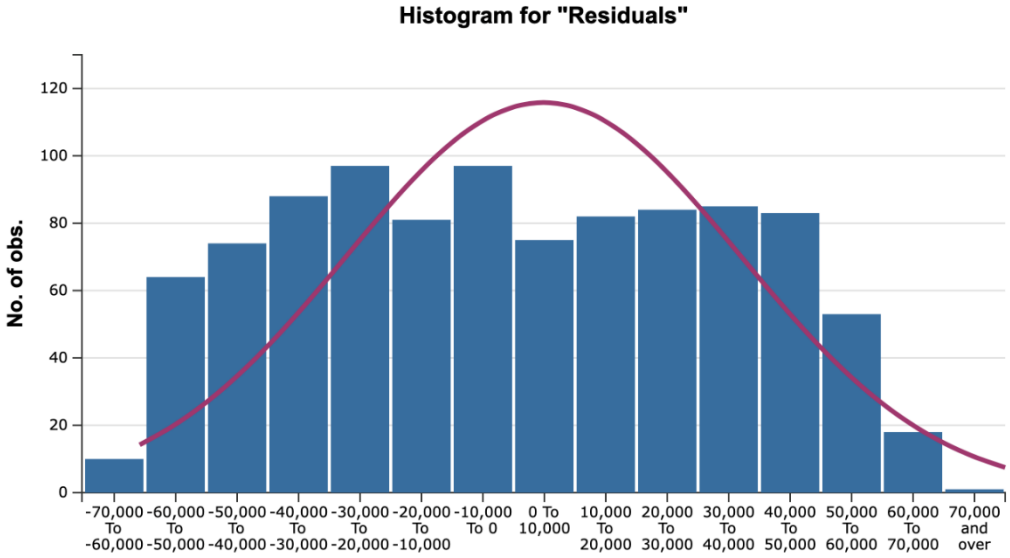

Fig. 1 Residual relationship between actual and forecast sales shown in histogram

Fitness can be tested by a residual graph, as shown in Fig. 1 and Fig. 2. In the Fig. 1, the vertical axis represents the number of data be observed, the horizontal axis is the sales ranges and the red dashed line is the trend of number of observed values for each range. According to the red line, the residuals are clearly divided. All the values are evenly distributed. On the right and left side of the red curve, the predicted value is higher than the observed one though the predicted value gradually increases or decreases with the trend of the red curve. In the middle of the curve, the predicted value is lower than the reality sales. Due to these features, residuals can be indicated by the graph. For Fig. 2, according to the horizontal line when residuals equal zero. Based on this line comes the analysis, most of the values have the same tendency with the line, but differences and outliers still exist.

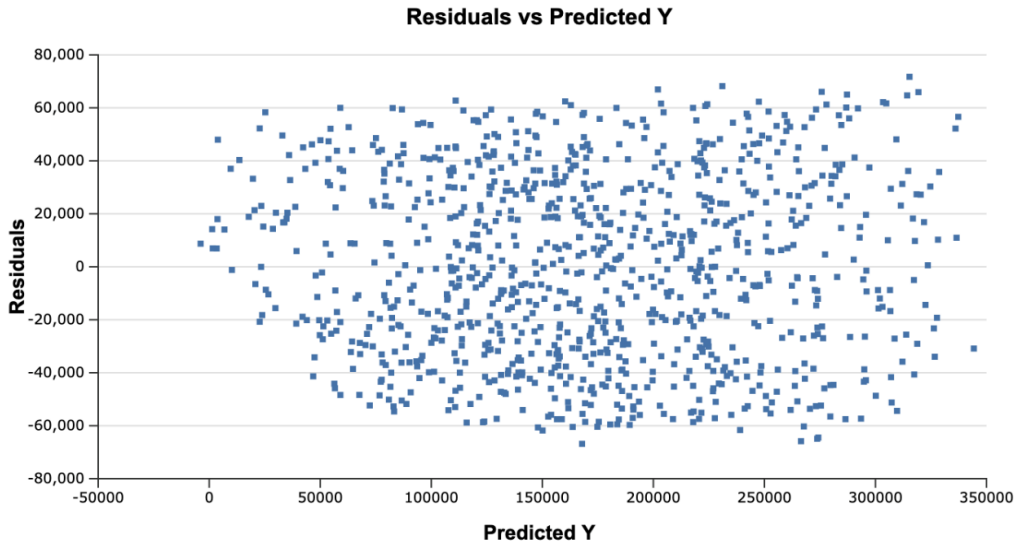

Fig. 2 Residual relationship between actual and forecast sales shown in scatter

4. Conclusion

By using linear regression analysis, this research predicts that the possible sales of a product can achieve through the promotional funds spent on in-store spending, online advertisement spending, and the price of the product. Through finding the equation between sales and the three independent variables, the relationship is derived. This equation shows the more efficient way to advertise the product and the relationship between price and overall sales, allowing companies to allocate promotional funds and exert their maximum utility and decide a proper sales price. Although some errors are still indicated when calculating the R square and graphing the residuals, the values already reach a close approximation with the observed value.

Because of the reason that the data were collected from one specific product sales from a company, the equation has a limited area to be applied. It can only be applied to the same series of products or products with high similarity. Besides, as the third section mentioned, the equation cannot explain all the situations that happened in the sales, indicating that there are still some defects for the model to be improved. For instance, the residual value is shown, instead of the differences between the observed value and the predicted value, some outliers exist. To address the problems, future research can be more focused on the scope of data collection to reach a wider range of data to promote the comprehensiveness of the relationship. The existence of the residuals means that there is still room for improvement, in further research, combining with other methods to have a comprehensive consideration will be helpful to the accuracy of the equation. Due to this research only focuses on utilizing linear regression to predict the overall sales and find out the relationship between sales and the three variables (in-store spending, online advertisement spending and price), however, it does not consider the optimize the allocation of the promotional funds and decided the sales price. Utility maximization problem can be considered in additional study.

References

[1] Shun Yin Lam, Mark Vandenbosch, John Hulland, Michael Pearce. Evaluating Promotions in Shopping Environments: Decomposing Sales Response into Attraction, Conversion, and Spending Effects. Marketing Science 2001, 20(2):194-215.

[2] Donovan, Robert, Rossiter, J. Store Atmosphere: An Environmental Psychology Approach. Journal of Retailing, 1982, 58: 34-57.

[3] Donovan R. Store atmosphere and purchasing behavior. Journal of Retailing, 1994, 70(3): 283-94.

[4] Ge, Jiaoju, et al. Effect of Short Video Ads on Sales through social media: The Role of Advertisement Content Generators. International Journal of Advertising, 2021, 40(6): 870-96.

[5] Karim, Y. Impact of emotional ads, online ads and repetition ads on customer buying behavior. Journal of Marketing and Consumer Research, 2017, 31: 22–27.

[6] Deighton, John, Et al. The Effects of Advertising on Brand Switching and Repeat Purchasing. Journal of Marketing Research, 1994, 31(1): 28-43.

[7] Gaur, Vishal, and Marshall L. Fisher. In-Store Experiments to Determine the Impact of Price on Sales. Production and Operations Management, 2005, 14(4): 377-87.

[8] Indah Safitri, Albari. The Influence of Product Price on Consumers' Purchasing Decisions. Review of Integrative Business and Economics Research, 2018, 7(2): 328-337.

[9] Taylor, Terry A., and Wenqiang Xiao. Does a Manufacturer Benefit from Selling to a Better-Forecasting Retailer? Management Science, 2010, 56(9): 1584-98.

[10] Isobe, T., Feigelson, E. D., Akritas, M. G., Babu, G. J. Linear regression in astronomy. Astrophysical Journal, 1990, 364: 104-113.

[11] Aalen, Odd O. A Linear Regression Model for the Analysis of Life Times. Statistics in Medicine, 1989, 8: 907-25.

[12] Anita Mfuh Y. Lukong and Yahaya Jafaru. Covid-19 Pandemic Challenges, Coping Strategies And Resilience Among Healthcare Workers: A Multiple Linear Regression Analysis. African Journal of Health, Nursing and Midwifery, 2021, 4(1): 1-18.

[13] Naseem, I., Togneri, R., Bennamoun, M. Linear regression for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 2106–2112.

[14] Tsoumakas, G. A survey of machine learning techniques for food sales prediction. Artif Intell Rev 2019, 52: 441–447.

[15] Giering, M. Retail sales prediction and item recommendations using customer demographics at store level. SIGKDD Explorations, 2008, 10(2): 84–89.

[16] Akossou, A. Y. J., Palm, R. Impact of Data Structure on the Estimators R-Square and adjusted R-Square in Linear Regression. International Journal of Mathematics and Computation, 2013, 20(3): 84–93.

[17] AI0909 Kaggle. (2024) Sales Data for Company Product. https://www.kaggle.com/datasets/ai0909/sales-data-for-company-product.