Exploring Factors Affecting Store Operating Profit Using Multiple Regression

Jiaqian Sun

International Department, Jiangxi University of Finance and Economics, 330013, Nanchang, China

2202101914@stu.jxufe.edu.cn

Abstract. The aim of this paper is to increase the profit of the company more efficiently. With the continuous development of science and technology and big data at this stage, the degree of influence of Marketing Spend, Administration, Transport and Area factors on Profit is quantified through the use of multiple linear regression models. This article is a linear regression analysis using the operational data collected from Kaggle for a company's 50 shops in 3 regions. Through the analysis, the model has a good fit and significant linear relationship, which can predict the future profit of the enterprise. At the same time, the model can also guide the enterprise how to improve the operating profit more efficiently, which has a certain guiding effect on the development of business operations. Based on the results, it can be seen that Marketing Spend has the greatest impact on the enterprise's profit.

Keywords: Multiple Linear Regression, Marketing Spend, Profit.

1. Introduction

Profit is the result of an enterprise's operations in a given accounting period. Under normal conditions, if an enterprise realizes a profit, it indicates that the equity of the owner of the enterprise will grow as well as the business performance will improve; on the contrary, if the enterprise incurs a loss, it indicates that the owner's equity of the enterprise will decrease and the performance will decline. Profit is one of the indicators for evaluating the performance of an enterprise's management, and investors can serve it as an important reference to guide them how to develop well and other users of financial reports when making decisions.

With the continuous development and progress of the Internet and big data, companies record the data generated in the course of their operations. Moreover, by analyzing big data, they can analyze the company's profits and thus make reasonable predictions about them. By collecting the data searched from *Kaggle*, it is found that the relationship between the interaction between various expenses and profits in the business process of enterprises is increasingly significant. Furthermore, it is crucial to analyze the relationship between profit and expenses in order to increase profitability, improve the ability to cope with the competitive market, and promote the long-term development of the company.

This paper investigates 50 stores of a company in *Ctg, Dhaka and Rangpur* districts respectively based on the data available on *Kaggle*. Profit is taken as the dependent variable and marketing spend, administration, transport and area are taken as the independent variables, and a multiple regression linear model is built to analyze the five factors of marketing spend, administration, transport, *Ctg* and *Dhaka* as independent variables to analyze the impact on corporate profits and improve the model by multiple covariance test.

The use of multiple linear regression model to predict profit serves two purposes.

(1) It can measure the degree of influence of different factors on profit so that the enterprise focuses on the variables that have the greatest influence on profit. In this case, the company can improve the amount of profit of the enterprise more scientifically.

(2) Through the establishment of the model, the future profit amount of the enterprise can be predicted so as to judge the future development of the enterprise.

2. Method

2.1. Data analysis methods

The data in this paper uses 50 stores in three districts of *Ctg*, *Dhaka*, and *Rangpur* of an enterprise in *Kaggle*, and by recording the data of marketing spend, administration, transportation, and profit of these 50 stores, we can observe the general distribution of the data through the MIN, MAX, and AVERAGE functions in Excel. functions in Excel to observe the approximate distribution of the data.

The specific distribution of the data is further reflected through charts and graphs. The distribution of stores in different regions is shown through pie charts, density charts show the number of stores included in each cost range, and box plots show the average profit in different regions and the degree of dispersion of the data.

Multiple linear regression models are usually used when one variable is subject by more than one variable. For example, the profits explored in this paper may be affected by multiple factors of Marketing Spend, Administration, Transport, and Area.

2.2. Multiple Linear Regression Modeling Approach

The multiple linear regression model expression is

$$P = \beta 1 \times MS + \beta 2 \times AD + \beta 3 \times TR + \beta 4 \times CT + \beta 5 \times DA \quad (1)$$

Where βj (j = 1, 2, 3, 4, 5) are the regression coefficients, MS is the variable marketing spend, AD is the variable administration, TR is the variable transport, CT is the district variable *Ctg* and DA is the district variable Dhaka. Since the district variables were taken to be brought into the model with Python coding process to avoid multicollinearity, *Rangpur* was not put as a variable in the expression.

2.3. Relevance Analysis

There is a statistical analysis method named correlation analysis that examines the correlation between two or more random variables that are on equal terms. To test the correlation between the variables, the correlation coefficient formula is applied:

$$\beta(x, y) = \frac{Cov(x,y)}{\sigma x \sigma y} \quad (2)$$

2.4. Stepwise Regression Analysis

Stepwise regression is in the process of screening variables in regression analysis, builing a stepwise regression model that uses stepwise regression from a series of variables that can be selected, allowing the system to automatically identify influential variables. In order to avoid interference between independent variables affecting the accuracy of regression modeling operations, stepwise regression analysis is applied to test the data.

3. Data Description

3.1. Research ideas

Table 1. Explanation of variables

Variables	Description
Area	Take the same company's stores in three regions
Marketing Spend	Costs incurred by the company to promote its products and services; including advertising, promotions, market research and etc.
Administration	Various expenses incurred by the administrative departments of enterprises for the organization and management of production and operations
transport	Costs incurred by enterprises due to transportation
Profit	Profitability of the enterprise

Table 2. Distribution of data

The case of the values of the independent variables	minimum value	maximum values	average value	upper quartile
Marketing Spend	0	165349.2	73721.62	73051.08
Administration	51283.14	182645.6	121344.6	122699.8
Transport	0	471784.1	211025.1	212716.2

By summarizing the previous research, it is found that the profit of enterprises is most obviously affected by the four factors of market overhead, transportation, management and region, and Python is used to build a multiple regression model to analyze the data empirically. In the process of analysis, the data were analyzed by analyzing the data of independent variables and establishing multiple linear regression model, after which the data were analyzed and tested by using correlation coefficient analysis, stepwise analysis method.

As can be seen from the data in Table 2, from the values of the three variables marketing spend, administration and transportation, the average overhead marketing spend<administration<transport; of these three variables, marketing spend and transport have the smallest value of 0. Among the 50 stores surveyed, there exist stores with no overhead in marketing and transport.

3.2. Analysis of variable data

The distribution of data from the five independent and dependent variables of Area, Marketing Spend, Administration, Transport, and Profit are analyzed separately.

3.2.1 Distribution of area data

From the above data, it can be seen that the data of this research includes 50 stores of the same company in three regions, namely *Ctg*, *Rangpur* and *Dhaka*. a pie chart is used to represent the distribution of stores in the three regions, and the angle of the pie represents the number of stores: the larger the angle of the fan, the more the number of stores in the corresponding region.

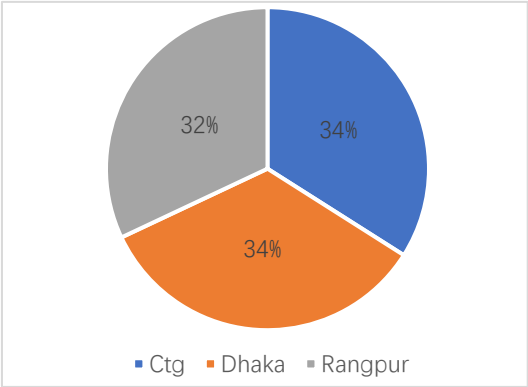

Figure 1. The proportion of stores of different companies

As can be seen from the pie diagram, *Ctg*, *Dhaka* and *Rangpur* account for 34%, 34% and 32% of the total number of stores respectively. Thus, it can be seen that the number of selected stores is comparable in these three regions.

3.2.2Distribution of marketing spend data

By selecting$[0,5000)$, $[5000,10000)$, $[10000,15000)$, $[15000,20000)$ these four intervals, calculate the number of stores corresponding to each of these four intervals, and connect them with a smooth curve to make a density map about marketing spend.

Figure 2. Density map on marketing spend

From the graph, it can be seen that the number of stores with marketing spend around $100000 is the most among all the stores surveyed. Most of the stores have marketing spend in the range of $50,000-$150,000. The image approximates a left-skewed normal distribution graph.

3.2.3 Distribution of administration data

By selecting $[0, 5000)$, $[5000, 10000)$, $[10000, 15000)$ and $[15000, 20000)$ these four intervals, calculate the number of stores corresponding to each of these four intervals and make a density map about administration.

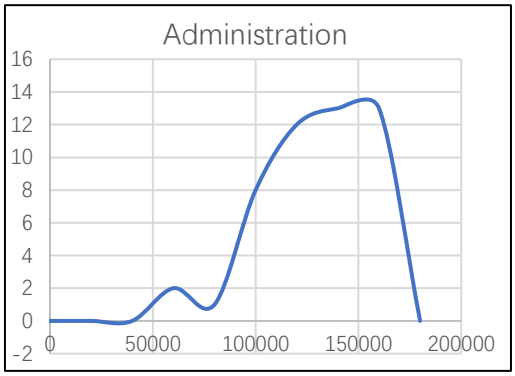

Figure 3. Density map on administration

As can be seen from the above chart, the costs of management in these three regions are relatively even, with most stores spending $100,000-$180,000 on management, which is relatively centralized, and there is not much difference between stores in different regions.

3.2.4Transport data distribution

By selecting $[0, 10000)$, $[10000, 20000)$, $[20000, 30000)$, $[30000, 40000)$, $[40000, 50000)$ and $[50000, 60000)$ these six intervals, calculate the number of stores corresponding to each of these six intervals, connect the corresponding scatter points, and make a density map about transport.

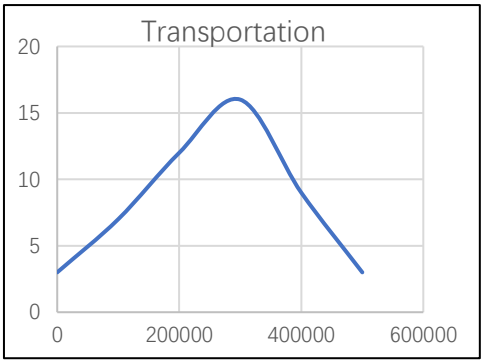

Figure 4. Density map on transport

As can be seen from the figure above, the overall distribution of the cost of transportation for stores in these three regions is $0-$500,000, with most stores having a distribution of $100,000-$400,000 for the transportation component. The distribution of overhead costs is similar to a normal distribution for all the stores in the sample.

3.2.5 Profit data distribution

Using box plots, abnormal data can be ruled out by figuring out the 25/% quartile, 75/% quartile, and mean, while observing the distribution of normal data.

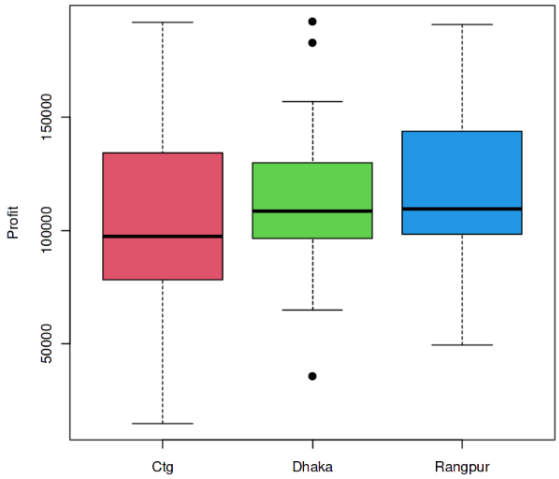

Figure 5. Box plot on PROFIT data

From the above data, it can be seen that the average value of store profit in *Ctg* region is around $100000, while the average value of store profit in *Dhaka* and *Rangpur* is above $100000, which is slightly higher than that of *Ctg*. Based on the shape of the box plots, it can be seen that the distribution of the data below the average in these three regions is more concentrated than that of the data above the average. Moreover, the distribution of the data in the *Dhaka* region is the most concentrated and the *Ctg* region has the most dispersed data, indicating that *Dhaka* stores have the most concentrated profit values, while *Ctg* stores have the largest profit variance.

4. Multiple regression analysis

Based on the generalized form of multiple linear regression the equation can be set as follows.

$$Y = \beta 1 \times X1 + \beta 2 \times X2 + \cdots + \beta n \times Xn \quad (3)$$

Where β1, β2, β3, β4 are unknown regression constants and ε is the random error. y is the explanatory variable and $X1, X2, Xn$ is the explanatory variable.

4.1. Modeling

The districts were coded using Python to create the following multiple regression model:

$$P = \beta 1 \times MS + \beta 2 \times AD + \beta 3 \times TR + \beta 4 \times CT + \beta 5 \times DA + \varepsilon \quad (4)$$

In the above model, P is the explanatory variable and MS, AD, TR, CT, and DA are marketing spend, administration, transportation, *Ctg*, and *Dhaka* explanatory variables, respectively. *Ctg* and *Dhaka* are coded into the multiple regression model using solo heat coding in Python.

4.2. Initial multiple regression analysis

The data were analyzed by multiple regression using Python and the results are shown in Tables 1 and 2. In order for the data to be brought into Equation (2), the data for the independent variables Marketing Spend, Administration, and Transport need to be normalized. Using Min-Max standardization, the raw data is linearly transformed so that all variables take values between $[0, 1]$.

Table 3. Summary of model results

Model	R-square	Adjusted R-square	Standard estimation error	(number of) degrees of freedom (physics)
	0.9012	0.8899	13370	44

Notes: a. Predictor variables: M, T, AD, AR; b. Dependent variable: P

Table 4. Least squares regression results

Variable	coefficient	Std. Error	t-Statistic	Prob.
Intercept	43143.1	7214.8	5.98	3.61E-07
MS	114584.6	10119	11.324	1.26E-14
AD	4852.1	9561.9	0.507	0.6144
TR	32594.2	10696.1	3.047	0.0039
Dhakal	1029.4	4723.4	0.218	0.8285
Ctgl	778.3	4773.9	0.163	0.8712

The results of the least squares regression after standardization, are presented in the table 5 below:

4.3. Analysis of regression results

The regression equation was established from Table 5:

$$P = 0.6453MS + 0.0273AD + 0.1835TR + 0.0044CT + 0.0058DA \quad (5)$$

4.4. Test the model

As can be seen in Table 3 and Table 4, economic significance: the estimated regression coefficients are β1=114584.6>0, β2=4852.1>0, β3=32594.2>0, β4>0, which indicates that when other conditions remain unchanged, Profit changes in the same direction as Marketing Spend, Administration, Transport and Area. Area in the same direction; when Marketing Spend, Administration, Transportation and Area increase at the same time, it has the greatest impact on profit.

Goodness-of-fit test: adjusted R^2=0.8899, which indicates that marketing spend, administration, transportation and area can explain 88.99% of the variation in profit and only 11.01% of the variables are not explained, which is a good fit.

Overall significance test of the regression model: f-statistic>F0.05(5,44), indicates that the joint effect of PROFIT is significant, which is quite significant as shown by the F-value of 0.00000.

Individual model test: in terms of the effect of individual factors, the t-statistic of β 1 is 11.324, at 5% level of significance. Checking the t-distribution table, with 44 degrees of freedom, t0.05(44) <11.324, indicating that final marketing spend has a significant effect on profit. β2 has a t-statistic of 0.507<t0.05(44), showing that administration does not have a significant effect on profit. β3 has a t-statistic of 3.047>t0.05(44), indicating that administration does not have a significant effect on profit. amounted to 3.047>t0.05 (44), indicating that TRANSPORT has a significant effect on PROFIT. This can be seen from their p-values of 0.00000, 0.6144, 0.0039, 0.8285 and 0.8712.

Table 5. Standardized least squares regression results

Variable	coefficient	Std. Error	t-Statistic	Prob.
Intercept	0.160275	0.040629	3.944882	0.000283
Marketing Spend	0.645255	0.056982	11.32374	1.26E-14
Administration	0.027323	0.053846	0.507439	0.614381
Transport	0.183546	0.060233	3.047284	0.003896
Dhakal	0.004383	0.026883	0.163033	0.871239
Ctgl	0.005797	0.026599	0.217926	0.828494

4.5. Correlation analysis

According to equation (5), the correlation between the independent variables is calculated to explain the degree of interaction between the respective variables.

Table 6. Correlation coefficient results

correlation	MS	AD	TR
MR	1	0.23	0.69
AD	0.23	1	-0.03
TR	0.69	-0.03	1

The two-by-two correlation coefficients of the three variables MS, AD, and TR are calculated, as shown in Table 6, which shows that marketing spend and administration present a weak positive correlation, and a strong positive correlation with transportation; administration and transportation present a weak negative correlation, and due to the r = -0.03 can be be ignored.

4.6. Stepwise regression analysis

Since each independent variable is positively correlated with the dependent variable, in order to avoid the linear relationship between each independent variable presenting a significant, stepwise regression analysis was performed on the above data to adjust the optimization formula (3), and the results are shown in Table 6.

Based on the ANOVA results in Table 7, the adjusted F=351.16, significance is less than 0.001 and the regression equation is significant. Based on Table 8, the adjusted regression equation was established as

Table 7. ANOVA Results

Model		square sum (e.g. equation of squares)	(number of) degrees of freedom (physics)	mean square	F	significance
1	**regression (statistics)**	2.220785	1	2.220785	351.1551	.000b
	residual	0.303563	48	0.006324	--	--
	(grand) total	2.524348	49	--	--	--
2	**regression (statistics)**	2.273136	2	1.136568	212.6441	.000c
	residual	0.251212	47	0.005345	--	--
	(grand) total	2.524348	49	--	--	--

Notes: a) Dependent variable: p

b) Predictor variables: (constants), MR

c) Predictor variables: (constants), MR, TR

$$P = 0.767MS + 0.0273AD + 0.1732TR + 0.0044CT + 0.0058DA \quad (5)$$

The regression equation model indicates that for every unit increase in marketing spend (MS), Profit (P) increases by 0.767 units; for every unit increase in administration (AD), Profit (P) increases by 0.0273 units; for every unit increase in transportation (TR), Profit (P) increases by 0.1732 units; for each unit increase in *Ctg* (CT), Profit (P) increases by 0.0044 units; and for each unit increase in *Dhaka* (DA), Profit (P) increases by 0.0058 units.

Table 8. Adjusted model regression results

Model		Unstandardized coefficient		Standardized coefficient	t	significance
		B	**standard error**	**Beta**		
1	**(Constant)**	0.206183	0.021434	--	9.619632	8.88E-13
	MR	0.766873	0.040924	0.937948	18.73913	1.02E-23
2	**(Constant)**	0.178138	0.021646	--	8.229441	1.16E-10
	MR	0.656033	0.051669	0.802381	12.69671	8.44E-17
	TR	0.173185	0.055337	0.19778	3.129623	0.003006

a: Dependent variable: P

5. Conclusion

5.1. Conclusion

According to the adjusted linear regression equation, it can be seen that store profit is affected by marketing spend (MS) and transportation (TR) factors, and both of them are positively correlated; it is not significantly affected by administration (AD) and area factors.

From the final model, it can be seen that profit is most significantly affected by the marketing spend factor. By applying stepwise regression analysis, the multiple linear regression was adjusted so that the new regression model fitted better than the original model. Moreover, the model allows companies to understand more the impact of different factors intuitively on operating profit. Finally, the model is a guide for future business programm and can also be used to predict future operating income.

5.2. Recommendation

From the results of the survey of this enterprise, the best way to increase the profitability of the stores is to increase marketing spend, because the independent variable increases by the same unit, marketing spend (MS) has the most obvious effect on profit; followed by transport (TR) which has a significant effect on profit. Therefore, firms' investment in stores should focus on both marketing spend and transportation.

From the findings of the business, it is clear that the region in which the business's stores are located does not have a significant impact on PROFIT. Therefore, there is no need to be overly concerned that the profitability of the stores will be affected by the difference in regions.

References

[1] Dawei Feng. An empirical analysis of the logistics industry and primary industry in Hainan Province based on multiple regression. Logistics Technology 94-98, 19 (2023).

[2] Weiwei Wang. (2016). Analysis of Influencing Factors of China's GDP Growth Based on Multiple Regression Model. China's Collective Economy 56-57, 9 (2016).

[3] Sunxin Qi, Gege Fu & Xinyu Zhang. Analysis of GDP Influencing Factors in Shenyang City Based on Multiple Regression Model. Modernisation of Shopping Malls 170-171, 9 (2017).

[4] Landini Fernando & Conti Santiago.Factors contributing to rural extension agents' support for a transfer of technology (ToT) approach: a multiple linear regression analysis. The Journal of Agricultural Education and Extension 5 (2023).

[5] Zhu Xuanxuan. (2023). Retraction notice to "Multivariate linear regression analysis on online image study for IoT" 312–316, 52 (2018).

[6] Zhengjiang Chen & Xi'an Pu. A comparative study of multiple linear regression analysis and stepwise regression analysis. Journal of Mudanjiang Institution of Education 131-133, 5. (2016).

[7] Yewei Fu, Wenkan Zhang, Shihao Xu & Kai Su. Research on waterfront pit seepage control and precipitation measures based on multiple linear regression analysis method. Hydropower and New Energy 13-17, 8 (2023).

[8] Yao Wang. Multiple Linear Regression Based on Multiple Covariance Correction (Master's Thesis, Yili Normal University). (2023).

[9] Zimei Cai. Sales Forecast Analysis of Building Material Consumables (Anchor Bolt) Based on Multi-Model Approach (Master's Thesis, Shanghai University of Finance and Economics). (2022).

[10] Lingling Liu & Xin Dai. The Impact of Financial Support on Rural Industrial Integration Development under Multiple Linear Regression Models--The Case of Lianyungang City, Jiangsu Province. Times Economy and Trade 61-63, 9 (2022).