Advances in Economics, Business and Management Research, volume 211
Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)

Linear Regression Model for Business Strategy
A Case Study of *SMARTFOOD Company*

Keyi Chen[1,] Liyao Dong[2,] Linxinyi Wang[3,] *

[1]*Management School, Lancaster University, Lancaster University, Lancaster, UK.*
[2]*SEM, Beijing University of Technology, Beijing, China.*
[3]*Mathematics Department, University of Nottingham, Nottingham, UK.*
**Corresponding author Email: shylw3@nottingham.ac.uk*
[†]*These authors contributed equally.*

ABSTRACT
This article is a case study that illustrates how a linear regression model can be implied in business strategy to find the relationship between each variable and advise price setting and revenue prediction. With the fast development of big data, a business can get a more thorough data set, so experiment in the real world is more reliable. The challenge is to clean, organize and detect wrong data. This paper aims to use the existing statistic modeling methods to make a business strategy for a company based on a relatively small data frame. R studio is used for all statistical calculations. This paper also examines the interaction effects for the centered and uncentered data set. The result of this study shows that price is negatively correlated to sales and advertising fees, and stores have a positive correlation. It also shows that centering makes a great difference in illuminating the collinearity of interaction terms. Finally, this article analyzes the estimation made by the marketing director and gives suggestions on business strategy in the end.

Keywords: *sale forecast, price, linear regression, interaction, business strategy.*

1. INTRODUCTION

1.1. Research Background and Motivation

The business environment is growing more complex—not just economically, but socially, politically, and legally—and firms must manage this complexity with strategies that match the amount and variety of complexity in their environment [1]. The product development department of SMARTFOOD planning to launch a new product in the diet food category, K-Pack, a new low-carbohydrate candy bar. The CEO of SMARTFOOD, Mr. Donovan, is pinning his hopes on the K-Pack. In the marketing director's opinion, most snack products have short life cycles, around 2 years. Therefore, if they want to seize the market, they should launch their product as soon as possible. Although SMARTFOOD can survive within 2 years, it is hard to guarantee K-Pack can still remain in the market after that. Marketing has been an effective tool that people can apply to increase sales (Jager 2007), so finding the right marketing strategy is a priority for the firm [2]. Therefore, Mary, the marketing director, carried out a small pre-test market study to see if they could maximize the financial interests, particularly during the first two years with little competition. In Mary's prediction, the annual sales of K-Pack (24 packages in a case) are 750,000 cases using a mix of 50 cents per packages, and revenue is 70 per cent of the retail price. For the manufacturing expenses, there were 1 million dollars fixed manufacturing cost and $1 variable cost per case, which in total was $1.75 million. Besides, SMARTFOOD spent 3 million dollars in advertising per year. So under Mary's prediction, K-pack will bring 1.55 million dollars to SMARTFOOD. In addition, the company plans to outsource some manufacturing processes of K-Pack, which means the 1 million dollars manufacturing cost is fixed in the following years. The marketing mix is commonly known as the 4Ps. McCarthy's (2000) consists of product, price, place and promotion [3]. Therefore, to get the optimal marketing mix, Mary plans to change the variables: month, price, advertising expenses, K-pack's display area, store volume, and city index. After collecting and analyzing the data, the business can evaluate different marketing mixes based on the pre-test result to optimize the sale tactic strategies.

1.2. Literature Review

The regular process for regression model fitting is model selection, assumption analysis and outlier detection. After fitting the model, it is important to test multicollinearity after generating a regression model. This is because if the explanatory variables are correlated, it is impossible to determine which variable is associated with the response variable. An extreme example of this is two variables, which are exactly the same measurements but different in the unit. They are both correlated or not correlated to the dependent variable. They have perfect collinearity, and hence only one of them is necessary for a model. Two common ways to detect multicollinearity are using covariance matrix and variance inflation factor (VIF). According to Freund et al. (2003), variance inflation factor can detect multicollinearity without centering the data [4]. However, according to Robinson and Schumacker (2009), if the model contains interaction terms, only centering would guarantee the consistent result from VIF because centered variables have low intercorrelation [5]. According to Blatna (2006), the outliers related to predictors are defined as leverage points. If a leverage point is far from the other points but in the vicinity of the regression line, then it is a good leverage point. On the other hand, if a leverage point is far from the fitted line, then it is seen as a bad leverage point [6]. The tools to identify outliers and leverage points are Standardized residuals, the Mahalanobis distance, the Cook's distance and DFITS. These methods may all fail if there is a masking effect. The masking effect occurs when outliers group together and make it hard to identify. In this case, robust regression is preferred because it is more sensitive to unusual points. Blatna (2006) also recommended that apart from those numerical diagnostic methods [6]. We should also draw residual diagnostics plots to analyze visually.

1.3. Research Contents and Framework

The pre-test result has taken the considerations for the average sales per month by incorporating the impact of the month, price, advertising, location of the supermarket, the size for the stores, and the city. This provides the comprehensive considerations to drive the quantitative analysis on maximizing the financial interest. This study uses R to select a multiple linear regression model based on the previous test. The process starts from a model with all variables and deletes one variable every time manually. This is followed by assumption tests, residual analysis and outlier detection.

The paper is organized as follows. Section 2 is about model selection and all the hypothesis tests. In Section 3, the paper predicts the profit of the following 2 years based on the equation of the final regression model. Section 4 discusses the result and propose the optimal marketing mix. Finally, the section shows the pros and cons of this study and gives further investigation direction.

2. METHODOLOGY

2.1. Exploratory Analysis

Studies can contribute to theory and practice by bringing together the existing literature on seller and consumer information about the subjects to develop and test practical strategy concepts [7]. The data frame has 96 observations with 7 variables. In Table 1, Sales is chosen to be the response variable, and the potential independent variables are Month, Price, Advertising, Location, Store, and City. For advertising, 0 and 1 represent advertisement investments of \$3 million and \$3.5 million, respectively. For the Location variable, 0 and 1 represent the product is put in the breakfast section and bakery section in the supermarket. The only store is considered as a numerical variable. Others are all categorical variables.

TABLE 1. SMARTFOOD DATA FRAME

Sales	Month	Price	Advertising	Location	Store	City
170	S1	50	0	0	45	1
213	S1	50	0	0	45	2
195	S1	50	1	0	43	3
241	S1	50	1	0	45	4
179	S1	50	0	1	49	1
115	S1	50	0	1	44	2
190	S1	50	1	1	49	3
317	S1	50	1	1	54	4
95	S1	60	0	0	54	1
157	S1	60	0	0	50	2
205	S1	60	1	0	44	3

A Scatter plot matrix (Figure1) is also made to see if there is an obvious correlation between each pair of variables for Sales, Price, Advertising, Store and city. This is a symmetric matrix, so only the upper triangle matrix or the lower one needs to be considered.

The matrix the worth notice is for Advertising and City. It seems that Advertising and City have a perfect correlation. City1 and City2 only used 3 million advertisement fees, and City3 and City4 only used 3.5 million advertisements. This means among cities and advertisements, which one will influence sales cannot be determined. Also, there might be some correlation between price and store volume as well, but it needs further evaluation. Looking at the Q-Q plot of sales in Figure 2, since all the points fall approximately along the reference line, and these points almost follow a straight line, data can be regarded as a normal distribution. From the Q-Q plot of store volume, although most of the points are within the confident interval, some extreme values exit the confident interval. Also, the histogram for the store does not suggest a normal distribution. However, additional information for the store in all the stores in that country is provided. From figure 3, Stores looks like the normal distribution, so no transformation is needed. For categorical variables, a boxplot (Figure 4) is used to briefly understand them. Based on this box plot, the combination of the lower price, higher advertising investment, location on bakery section, the 4th month and city 4, might lead to higher sales.

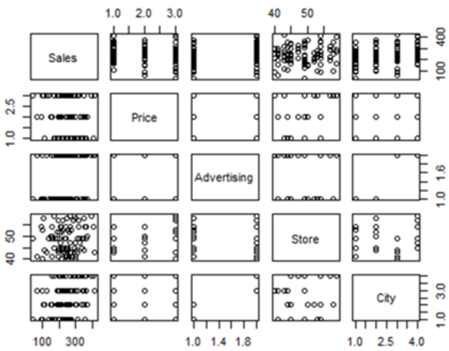

Figure 1. Scatter Matrix Plot

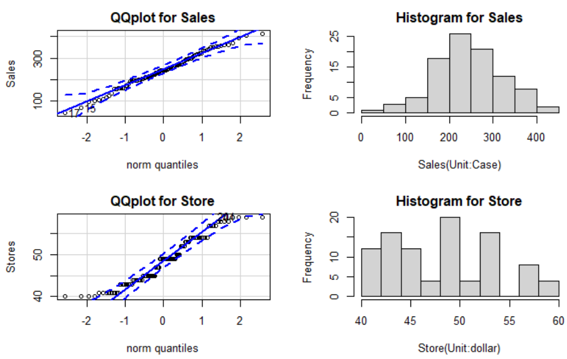

Figure 2. Q-Q plots and Histograms for Sales and Stores

Figure 3. Histogram for Stores Nationwide

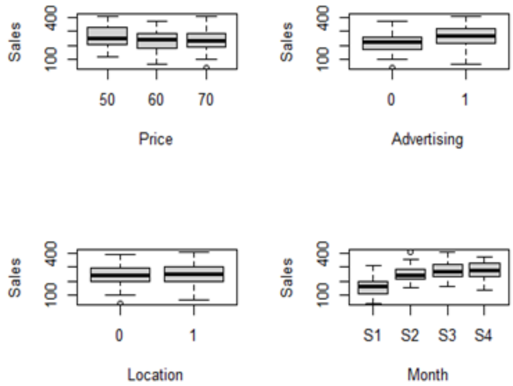

Figure 4. Boxplot for categorical variables

2.2. Assumptions

This paper assumes that the error terms are random and independent, which means no examine functional forms can be found. Secondly, the error residuals are assumed to be normally distributed with a mean of 0. This guarantees that the hypothesis tests are valid. Thirdly, errors are homoscedastic, which means they have constant variance and are independent of independent variables. If this does not hold, then a linear model may not be good enough.

2.3. Model Selection

This article uses the p-value approach and selects the model from the full model backwards. The article set the significant level to 0.1. The P-value approach means the paper starts with the full model and deletes the variable with a P-value bigger than 0.1. This process is repeated until all variables are significant. The model selection process starts by fitting the full model (called: full)，

$$full: Sales = \beta_0 + \beta_1 * Price + \beta_2 * Advertising + \beta_3 * Location + \beta_4 * Store + \beta_5 * City + \beta_6 * Month \quad (1)$$

β_i are arbitrary constant that needs to be determined using programing.

The estimated coefficient, error, t-value and p-value for each variable in the full model are shown in the following table:

TABLE 2. SUMMARY FOR FULL MODEL

full	Estimate	Error	t value	Pr(>|t|)
(Intercept)	-24.526	66.085	-0.371	0.711465
Price60	-31.742	13.293	-2.388	0.019166
Price70	-49.059	14.691	-3.339	0.001248
Advertising1	56.911	15.971	3.563	0.000603
Location1	12.430	10.827	1.148	0.254172
Store	3.614	1.370	2.638	0.009929
City2	8.397	15.547	0.540	0.590561
City3	2.512	19.232	0.131	0.896388
City4	—	—	—	—
MonthS2	93.583	15.304	6.115	2.84e-08
MonthS3	117.000	15.304	7.645	2.92e-11
MonthS4	118.583	15.304	7.749	1.81e-11

Overall, City is not significant. Since City is a categorical variable, F-test is used. The reduced model excluded City term is called fit_1 below

$$fit_1: Sales = \beta_0 + \beta_1 * Price + \beta_2 * Advertising + \beta_3 * Location + \beta_4 * Store + \beta_5 * Month \quad (2)$$

The hypothesis for the F test is:

H_0: the coefficients for each level of the city are all 0

H_1: at least one coefficient of the city is not 0

This article used R to compute and the p-value for the ANOVA test of full and fit_1 is 0.8492. Hence, we reject H_0. City is not related to Sales. This is no surprise because city have a perfect correlation with advertising, as we previously observed.

The estimated coefficient, error, t-value and p-value for each variable in fit_1 are shown in the following table:

TABLE 3. SUMMARY FOR fit_1

fit_1	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	-19.249	53.080	-0.363	0.71776
Price60	-31.725	13.150	-2.413	0.01794
Price70	-48.954	14.026	-3.490	0.00076
Advertising1	53.925	10.911	4.942	3.71e-06
Location1	12.435	10.720	1.160	0.24923
Store	3.591	1.069	3.359	0.00116
MonthS2	93.583	15.156	6.175	2.06e-08
MonthS3	117.000	15.156	7.720	1.83e-11
MonthS4	118.583	15.156	7.824	1.13e-11

From the p-value in the last column, we see that the intercept and Location are greater than 0.1. Hence are not significant. Since Location is also a categorical variable, we prove it again by conducting an F test. We firstly create fit_2 with Location deleted (Reduced model):

$$fit_2: Sales = \beta_0 + \beta_1 * Price + \beta_2 * Advertisin + \beta_3 * Store + \beta_4 * Month \quad (3)$$

The hypothesis Test is: H_0: Placing K_Pack in a different location in the supermarket will not lead to different Sales (Location is not correlated to Sales). H_1: Different locations will lead to different sales (Location is correlated to Sales)

ANOVA of fit_1 and fit_2 is used, and the p-value for this test is calculated as 0.2492 (bigger than 0.1), so we do not reject H_0 and conclude that location is not correlated to Sales. After deleting location, the model only has 4 variables: Price, Store, advertising and Month. Interaction between Price and Store is considered since it is plausible that factories store fewer products if the price is set to be high. The model with interaction term is called fit_{inter},

$$fit_{inter}: Sales = \beta_0 + \beta_1 * Price + \beta_2 * Advertisin + \beta_3 * Store + \beta_4 * Month + \beta_5 * Price * Store \quad (4)$$

Summary for fit_{inter} is shown below

TABLE 4. SUMMARY FOR fit_{inter}

fit_{inter}	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	264.117	127.069	2.079	0.04064
Price60	291.400	161.646	1.803	0.07494
Price70	249.506	148.897	1.676	0.09743
Store	9.034	2.723	3.318	0.00133
Advertising1	47.200	11.277	4.185	6.84e-05
MonthS2	93.583	14.959	6.256	1.48e-08
MonthS3	117.000	14.959	7.821	1.21e-11
MonthS4	118.583	14.959	7.927	7.43e-12
Price60*Store	-6.889	3.427	2.010	0.04758
Price70*Store	-6.299	3.103	2.030	0.04544

Here, we see that the p-values for interaction terms are all less than 0.1, so does the terms for price and store. The variance inflation factors for Price, Store, Advertising, Month, and interaction terms are 10.7, 2.8, 1.1, 1.0, 11.1, respectively. The conventional rule is that VIF should not be more than 10 [8]. Variance inflation factor does not show a good fit since VIF for the interaction term and price is higher than 10. Following Blatna (2006) method, we try to center the numerical variable Store [6]. After delete by the mean for Stores (48.54167), the data for

Store becomes -3.5416667, -3.5416667, -5.5416667, -3.5416667, 0.4583333, -4.5416667, 0.4583333, 5.4583333, 5.4583333, 1.4583333…. The model after centering is called fit_{center}, and VIF for each variable is shown below:

TABLE 5. VARIANCE INFLATION FACTOR

fit_{center}	VIF^(1/(2*Df))
Price	1.083792
Store	2.827342
Advertising	1.066107
Month	1.000000
Price*Store	1.707324

All the VIF values are less than 10. Hence, fit_{inter} is valid. If we neglect the interaction term, the created model is insignificant and have coefficients with a P-value of nearly 1. Therefore, fit_{center} is the final model.

2.4. Residual Analysis

All the test statistics are based on our original assumption for residuals. We use a Q-Q plot and histogram to check that the residual is normally distributed (Figure 5). For the Q-Q plot, the points mostly fall on or near the diagonal line and are almost in the confident interval. The histogram of the residuals is approximately normally distributed. For example, no. 82 and No.47 are potential outliers, but it is within the confidence interval.

To check the assumption of homoscedasticity and no function forms, we use the Pearson residuals plot (Figure 6). We display the residual plot for Store and the boxplot of residuals for Advertising and Price. The residual plot with store shows that residual scattered nicely around 0 with no obvious pattern. For the boxplot of advertising and price, the mean residual of each level is 0. The boxplots also indicate that No.47 might be an outlier. This is also true for the model with Price. The plot of fitted value versus Residual shows no pattern, so there are no functional forms. Points are evenly scattered around 0, so the assumption of homoscedasticity is satisfied. Besides, Since the data is independent, each residual are independent, there is no obvious pattern, so errors are independent. Hence, all the assumptions are satisfied. The model is acceptable.

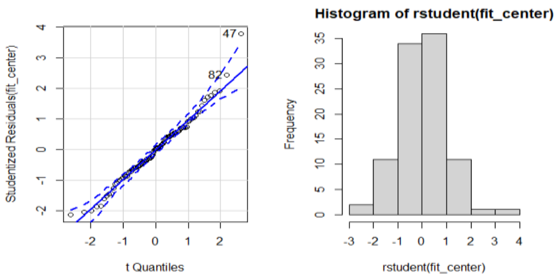

Figure 5. Q-Q Plot and Histogram for Residual

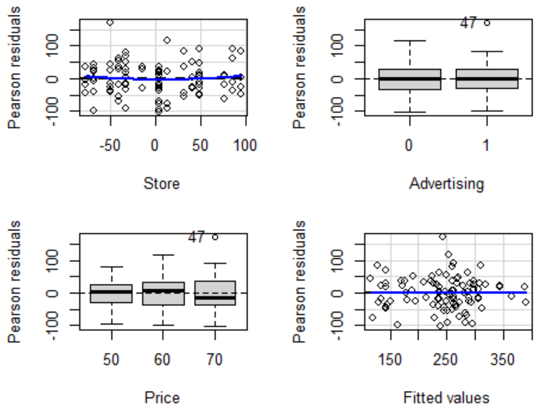

Figure 6. Q-Q Plot and Histogram for Residual

The plot of the true values versus predicted values is shown in Figure 7. The black diagonal line in the plot above is that the fitted value (prediction) equals the true value. Therefore, if most of the points are around the line and fit the trend of the line, the model performs generally well. In this case, the plot above indicates that fit_{inter} is generally good though some points are far away from the diagonal line. There is a large deviation when the value is smaller than 100 or bigger than 400.

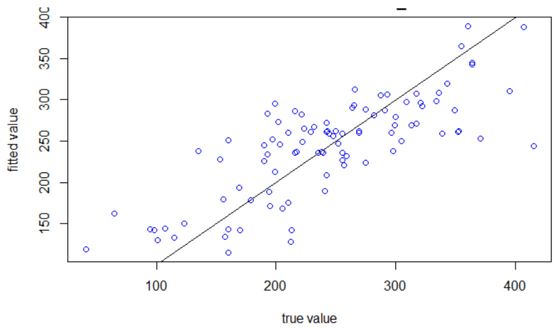

Figure 7. Fitted-value VS True-value for fit_center

2.5. Unusual and Influential Point Analysis

The fact that we have a relatively explicit solution means we can consider the comparative statics of the problem [9]. Since the data set is relatively small, outliers would have a huge impact on the linear regression model, and we try to detect it using Studentized Residuals, Hat value and Cook Distance. Figure 8 shows the most influential points in the data set. A big circle of data shows the greater influence. So, No.47 and No.82 are the two most influential points. The table below shows that No.15, No.47, and 82 data points have absolute studentized residuals bigger than 2. According to Blatna (2006), they are seen as outliers [6]. If the hat value is bigger than 2p/n=2*10/96=0.20833, no point is considered as a leverage point. The rule of thumb is to set the cut of point of Cook's distance to be 4/n [6]. In this case, it is 4/96=0.04167. The Cook distances for No.15, No.82, No47 are greater than 0.04167, so they are considered outliers.

TABLE 6. VARIANCE INFLATION FACTOR

Data Point	StudRes	Hat	CookDistance
8	0.9968624	0.20474817	0.025586886
15	-2.04115	0.11984	0.054713
32	-0.20475	0.204748	0.001091
47	3.783796	0.11929	0.16792
82	2.434872	0.077619	0.047185

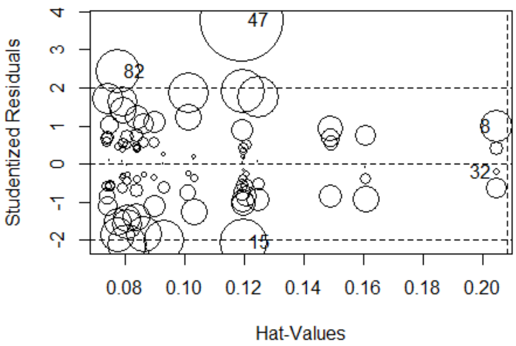

Figure 8. Studentized Residuals

Mahalanobis distance can also be used to detect leverage points. However, since Mahalanobis distance is only used for continuous variables, and we have only 1 continuous variable and many categorical variables in this study, it is not suitable. No.47 has the largest Studentized Residual and the highest Cook's Distance, which is the most influential point. We try to delete No.47 first. The new model is called fit_3. The adjusted R squared for fit_{center} and fit_3 are 0.56 and 0.60 respectively. So, deleting No.47 increases the accuracy of the model.

We try to delete No.15, No.47 and No.82 altogether to create a new model called fit_4 and compare the coefficients with fit_{center} in the table below. The new model is called fit_4.

TABLE 7. COMPARISON OF COEFFICIENT

	fit_{center}	fit_4
(Intercept)	174.4	181.2
Price60	-43.0	-44.8
Price70	-56.3	-65.7
Store	9.03	9.03
Advertising1	47.20	47.37
MonthS2	93.6	80.8
MonthS3	117.0	112.3
MonthS4	118.6	108.7
Price60*Store	-6.89	-8.25
Price70*Store	- 6.30	-5.11

Model fit_{center} uses the original data set. fit_4 uses 94 sample data with outliers deleted. We see there is a significant difference in terms of coefficient. The adjusted r squared for fit_4 is 0.61, which is slightly larger than fit_3. We can also draw a graph comparing fitted value and true value for fit_{center} and fit_4. In figure 9, the blue points are the predicted value from fit_{center}, the red dots are the predicted value from fit4. We can see that red dots are generally closer to the fitted line, especially for large and small values. The original model seems to have a bigger error for extreme value.

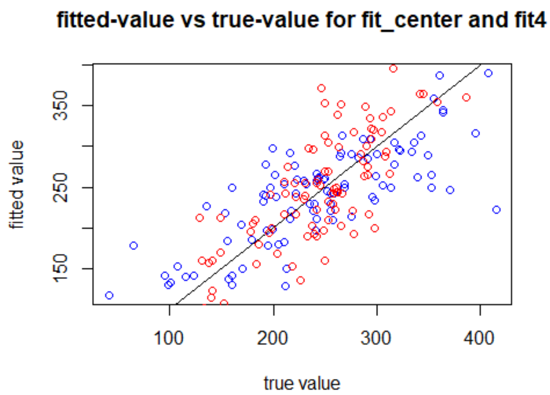

Figure 9. Fitted Value Comparison

Therefore, the final model is fit_4: (Calculated by R Studio)

$$fit_4\colon\ Sales = 181 - 45 * Price60 - 66 * Price70 + 9 * Store + 47 * Advertising1 + 80 * MonthS2 + 112 * MonthS3 + 109 * MonthS4 - 8 * Price60 * Store - 5 * Price70 * Store \tag{5}$$

3. PREDICTION

The marketing director's initial prediction for the new product is shown in Table VIII. Based on this, the general calculation formula for this specific product could be generated.

TABLE 8. INITIAL ESTIMATION BY MARKETING DIRECTOR

Sales	750,000 cases	
Revenue	\$6.3 million	(assume 70 percent of the retail price is revenue to the manufacturer)

Costs:	manufacturing	$ 1.75 million	($1 million fixed manufacturing cost plus $1 variable cost per case)
	advertising	$ 3 million	
Net Margin		$1.55 million	

Since this paper only has data for 4 months, assume the sales is not changing for the next 8 month, and the estimates of the first-year sales (unit: Cases) can be calculated by:

$$Annual\ Sales = 3 * [Sales(monthS1) + Sales(monthS2) + Sales(monthS3) + Sales(monthS4)] \tag{6}$$

The revenue of sales is calculated by: (unit: million dollars)

$$Revenue = Annual\ Sales * 24 * Price * 10^{-8} * 70\% \tag{7}$$

The cost of the sales is calculated by: (unit: million dollars)

$$Cost = 1 + Sales * 10^{-6} + Advertising \tag{8}$$

The net margin is calculated by: (unit: million dollars)

$$Net\ Margin = Revenue - Cost \tag{9}$$

4. RESULT AND DISCUSSION

Based on the final model fit_4 in (5), the relation can be interpreted by looking at the coefficients. Only Price have a negative relationship with Sales. Higher storage volume and higher advertising fee lead to higher Sales. However, there is no clear relationship between month and Sales. In fact, this variable is unclear because we don't know which specific months they are. As for the interaction term, there is a negative relationship between price and store, as we suspected before. Therefore, to generate higher sales, a combination of higher advertisement fees, lower price and higher storage should be applied. Since price has 3 levels and advertising has 2 levels, which means for each month, there are 6 possible combinations:

TABLE 9. BUSINESS STRATEGIES

Scenario	Price (cents)	Advertisement (million)
1	50	3
2	50	3.5
3	60	3
4	60	3.5
5	70	3
6	70	3.5

The following figures illustrate the prediction of sales and revenue. We see that even though lower price leads to higher profit, higher prices have a higher net margin. The highest sales we predict by fit_4 model is 0.94 million (cases) and the highest profit is 2.8 million dollars. The average sales and profit are 0.705 million cases and 2.01 million dollars respectively. These figures are all higher that the market director's initial prediction. Therefore, we briefly conclude that the initial estimation is quite conservative.

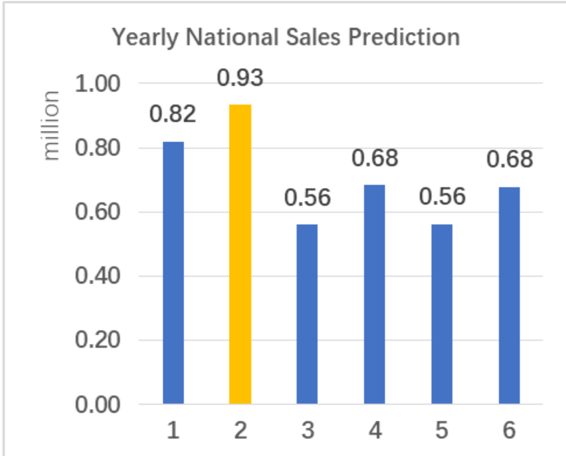

Figure 10. Sales Prediction

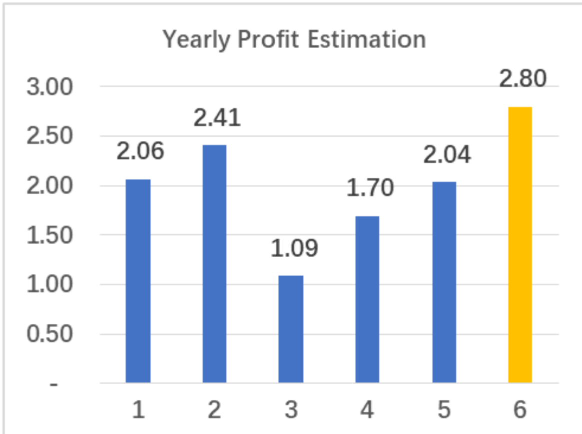

Figure 11. Profit Prediction

5. CONCLUSION

The result of this study accords market strategy rules, which proves that the final linear regression model is valid to some extent. Also, adjusted R squared value of 0.61 shows accuracy. Furthermore, we proved that centering is crucial to illuminate the collinearity of interaction term, so that the variance inflation factor can be lowered down. Also, we used the methods from previous research to find outliers, and it indeed increased adjusted R square.

However, we only use one criterion to select the model, which is the P-value. Mean square error can also be a criterion, but it may show a different result. The result is not complete for this case study because there is a perfect correlation between variables City and Advertising. The final model shows that the combination of the highest price and highest advertisement has the highest profit. However, it may not be true if we keep raising the selling price. The reason for this inaccurate conclusion might be that we only have three different prices (50,60,70 cents), and they are very close to each other. In addition, we are not sure whether price follows normal distribution because we only have three distinct values. Also, we assumed that the Sales between each season are the same. However, it might not always be true.

Further research should pay attention to the city's influence, and different advertising strategies should be used in each city. It is also suggested that other methods to select models can be used. For example, data modeler creates the model from the null model. Automated variable selection such as best subsets regression and stepwise regression can also be used. These methods are especially convenient if the data set is large. The different month should be examined so that we can see if there is any seasonal difference. Also, it is recommended to use bigger data sets and a wider range of prices, for example, set price to 40, 60, 80, 100 cents.

REFERENCES

[1] Tiller, S. (1999). Integrating market, technology, and policy opportunities in e-business strategy. The Journal of Strategic Information Systems.

[2] Jager, W. (2007). The four p's in social simulation, a perspective on how marketing could benefit from the use of social simulation. Journal of Business Research, 60(8), 868-875.

[3] McCarthy, J. E., & Perreault, W. D. (2000). Basic Marketing, A Managerial Approach. United States: McGrow Hill.

[4] Freund, R.J., Littell, R.C., and Creighton, L. (2003). Regression Using JMP. Cary, NC: SAS Institute, Inc.

[5] Robinson, C., & Schumacker, R. E. (2009). Interaction effects: centering, variance inflation factor, and interpretation issues. Multiple linear regression viewpoints, 35(1), 6-11.

[6] Blatná, D. (2006). Outliers in regression. Trutnov, 30, 1-6.

[7] Meredith Lawley; Dawn Birch (2016). Exploring Point of Sale Strategies for Improving Seafood Retailing: The Case of the Australian Oyster Industry. Journal of Food Products MarketingVolume 22, Issue 7. 2016. PP 792-808

[8] Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley.

[9] David Hobson; Yeqi Zhu (2016). Optimal Consumption and Sale Strategies for a Risk Averse Agent. SIAM Journal on Financial MathematicsVolume 7, Issue 1.PP 674-719