UNIVERSITY OF SAN FRANCISCO

CS-686 Machine Learning - Case Study

GoodBelly: Using Statistics to Justify the Marketing Expense

Alec Hsu, Rong Liew

March 23rd, 2018

Contents

1. Executive Summary

GoodBelly was a new line of probiotic juice products produced by Colorado-based NextFoods Inc. With the aim of boosting sales, the company explored a series of in-store demonstration programs: Firstly, company representatives engaged with customers by handing out product samples, information and vouchers. Second, they positioned GoodBelly products at endcaps in store aisles - one of the store's most popular location. However, due to limited marketing resources, management was pressured to cut marketing expenses that did not contribute directly to GoodBelly's results, thus bringing into question the effectiveness of the promotional programs.

By leveraging data collected over sales and promotions within a 10-week period, we conclude that the promotional programs had a positive effect on sales results. Specifically, the estimates suggest endcap programs are most effective when used in combination with a regional sales representative, increasing weekly sales, on average, by 513.05 units whilst holding all other variables constant. Engagement in Demo programs on the corresponding week has a greater effect than other weeks, increasing weekly sales, on average, by 106.28 units whilst holding all other variables constant.

2. Introduction

2.1 Background

GoodBelly was a new line of probiotic juice products produced by Colorado-based NextFoods Inc. Since its first product launch in January of 2008, GoodBelly products were now on the shelves of nationwide retailers such as Whole Foods Market Inc. and Safeway Inc.

2.2 Product

GoodBelly products are organic juice-based drinks. While most probiotic products were dairy-based, GoodBelly's products were dairy-free, soy-free, and vegan. All GoodBelly products used live and active cultures of a proprietary probiotic strain, which had been used in Europe for more than 15 years and thoroughly tested in over 17 research trials.

2.3 Purpose

GoodBelly's marketing initiative focused on two promotional programs:

2.3.1 In-store demonstrations

GoodBelly representatives distributed product samples, informed consumers about the product and offered coupons to inspire purchase.

2.3.2 Endcap displays

Sales representatives competed for the highest number of stores they could convince to place GoodBelly's products at the endcap. The endcap is the hub at the end of an aisle — one of the store's most popular locations (Figure 2). They had also competed for the best decorated endcap.

Figure 2

GoodBelly Endcap Displays

Source: GoodBelly company documents, 2012.

2.3.3 Problem

Due to limited marketing resources, GoodBelly management was pressured to cut marketing expenses that did not directly contribute to GoodBelly's weekly sales results. Concerns were raised on the return on investment of the promotional programs, as such, in order to prevent a budget reduction, justification was needed for the continuation of in-store promotional programs.

3. Data Overview

The dataset collected from 126 Whole Foods stores over the 10 weeks, between May 4th and July 13th, includes 1386 observations. We focus on a subset of eleven key variables in our analysis:

- Date: The weekly period
- Region: The region of the given store
- Store: The area in which the store was located
- Units Sold: The number of units sold per week
- Average Retail Price: The average retail price for GoodBelly products per store per week
- Sales Rep: Defined as 1 if the store had a regional sales rep (face-to-face contact) and 0 if the store had only national sales rep (no face-to-face contact)
- Endcap: Defined as 1 if the store participated in an endcap promotion
- Demo: Defined as 1 if the store had a demo on the corresponding week
- Demo1-3: Defined as 1 if the store had a demo 1-3 weeks ago
- Demo 4-5: Defined as 1 if the store had a demo 4-5 weeks ago
- Natural: The number of other natural retailers within 5 miles of each store
- Fitness: The number of fitness centers within 5 miles of each store

4. Methods and Results

To examine the effect on Weekly Sales units (explained variable), we run several Multiple Linear Regression (MLR) models in which we regress our explained variable on subsets of selected explanatory variables. We take into consideration, also, the individual effects of certain explanatory variables by running Simple Linear Regression models and delve into the possible complementary effects between variables by testing combinations of interactive terms.

Our approach begins with a 'Best Subsets' approach to identify a set of 'good' models by selecting from a subset of the predictors that do the best at meeting some well-defined objective criteria at the model level. Subject to respectable results, we further conduct analysis at the parameter level.

We present only four key models. We find these results particularly helpful in yielding optimal conclusions, given our pursuit of justifying if the promotional programs were effective. Additional models tested can be found in the 'Appendix' section.

4.1 Full Model

We begin our analysis with what we term our 'Full Model' which includes eight explanatory variables, namely: Average Retail Price, Sales Rep, Endcap, Demo, Demo 1-3, Demo 4-5, Natural Retailers and Fitness Centers. Three numerical and five categorical variables.

Full Model				
Units_Sold = - 28.535365(Average_Retail_Price) + 77.436914(Sales_Rep) + 305.102123(Endcap) + 111.132849(Demo) + 73.517171(Demo1_3) + 67.569811(Demo4_5) - 1.594168(Natural) - 1.019671(Fitness) + 298.488131				
R-Squared	**Adj. R-squared**	**F-statistic**	**AIC**	**BIC**
0.673	0.671	353.7	1.546e+04	1.550e+04

Table 1: Summary of 'Full Model' key criteria

At the model level, we observe that the R-squared and Adjusted R-squared value suggest that 0.673 and 0.671 of the variation in Weekly Sales can be explained by variation in our explanatory variables. Although we find these results satisfactory, we find that the inclusion of all variables provide us little insight, relative to later models.

Further, the F-statistic at 353.7 is low relative to our later models, as we shall see, in which we have a large Mean Squared Error value and a small Mean Squared Residuals value (See Appendix A). Deeper analysis of the individual variables finds the coefficients for Natural Retailers and Fitness centers statistically insignificant at the 5% significance level which leads us to reject that there is a significant relationship with Weekly Sales (See Appendix A). As such, we omit them in our second model, Subset Model I.

4.2 Subset Model I

4.2.1 Omits ‘Natural Retailers’ and ‘Fitness Centers’ variables.

Subset Model I				
Units_Sold = - 28.609165(Average_Retail_Price) + 76.951206(Sales_Rep) + 304.959716(Endcap) + 111.260534(Demo) + 73.663094(Demo1_3) + 67.700203(Demo4_5) + 294.189036				
R-Squared	**Adj. R-squared**	**F-statistic**	**AIC**	**BIC**
0.672	0.671	471.4	1.545e+04	1.549e+04

Table 2: Summary of ‘Subset Model I’ key criteria

At the model level, we successfully validate our decision in omitting ‘Natural Retailers’ and ‘Fitness Centers’ from our model. As shown in the summary table, the F-statistic increases substantially to 471.4 and we obtain a reduction in both the AIC and BIC values. However, we observe no change to the Adjusted R-Squared value and only a negligible reduction in the R-Squared value by 0.001. Nonetheless, this model is a significant improvement in totality.

At the parameter level, all variables are statistically significant at the 5% significance level. We delve deeper, then, into each individual explanatory variables to verify our results by first plotting a correlation matrix (See Appendix C).

We make two key observations. Firstly, the correlation between Average Retail Price (ARP) and Weekly Sales is notably weak at -0.019, which is contrary to what we would expect. Prompting us to investigate the specific relationship independently. Second, we observe that the variables Sales_Rep and EndCap are most highly correlated with Weekly Sales units at 0.449 and 0.593 respectively. The correlation between Sales_Rep and EndCap is also low at 0.052, considering for any multicollinearity. We work on each of these findings below.

4.2.2 Analysis of Average Retail Price

Simple Linear Regression: Units_Sold ~ Average Retail Price				
Units_Sold = - 4.505868(Average_Retail_Price) + 272.326736				
R-Squared	**Adj. R-squared**	**F-statistic**	**AIC**	**BIC**
0.000	-0.000	0.4908	1.699e+04	1.700e+04

Table 3: Summary of Simple Linear Regression Model key criteria - Units Sold ~ ARP

We run a Simple Linear Regression model regressing Weekly Sales units on ARP which deduces the insignificance of ARP as an explanatory variable.

We observe that the coefficient is statistically insignificant at the 5% significance level, leading us unable to reject that there is no relationship with Weekly Sales. Additionally, the R-Squared and Adjusted R-Squared value is a dismal 0.000 in combination with an F-statistic at 0.4908 in which the MSR is lower than the MSE value (See Appendix D).

In sum, we omit ARP as an explanatory variable, bringing us to Subset Model II.

4.3 Subset Model II

Our analysis arrives at a subset selection of variables to explain the effectiveness of the promotional programs on Weekly Sales which coincides well with our initial focus - justifying the impact of the promotional programs. The subset includes only 5 categorical variables, inclusive of: Sales Rep, Endcap, Demo, Demo 1-3, Demo 4-5.

Referring to "Costs and Benefits of Including or Omitting Interaction Terms: A Monte Carlo Simulation" by Mikucka, Sarracino and Dubrow. Mikucka et al. finds that "wrongly including an interaction term has little effect on the bias of estimates or on the loss of predictive power. Whereas, wrongly omitting an interaction term creates larger biases and a loss of adjusted ". As such, we contrast two separate models, Subset Model II A and Subset Model II B, in which the latter includes an interaction term between Sales_Rep and Endcap.

4.3.1 Subset Model II A

Subset Model II A				
Units_Sold = 67.658944(Sales_Rep) + 310.331985(Endcap) + 110.751351(Demo) + 77.069866(Demo1_3) + 65.651948(Demo4_5) + 181.243383				
R-Squared	**Adj. R-squared**	**F-statistic**	**AIC**	**BIC**
0.660	0.659	535.2	1.550e+04	1.554e+04

Table 4: Summary of 'Subset Model II A' key criteria

4.3.2 Subset Model II B

Subset Model II B				
Units_Sold = 52.042236(Sales_Rep) - 0.255131(Endcap) + 106.287830(Demo) + 75.979833(Demo1_3) + 73.092376(Demo4_5) + 461.257332 (Endcap) * (Sales_Rep) + 189.593039				
R-Squared	**Adj. R-squared**	**F-statistic**	**AIC**	**BIC**
0.798	0.798	910.5	1.478e+04	1.482e+04

Table 5: Summary of 'Subset Model II B' key criteria

Subset Model II B delivers our optimal model. We also notice the significant improvement in the model when including an interaction term (See above).

R-Squared and Adjusted R-Squared values of 0.798 is greatly higher than our former models without the overhang of the additional variables. Further, we have leading F-statistic, AIC and BIC values across all models. Namely, the F-statistic at 910.5 represents the additional increase and decrease in MSR and MSE respectively (See Appendix F).

At the parameter level, all coefficients which represent "main effects" to our model are statistically significant at the 5% level, notwithstanding the insignificant conditional effect of categorical variable 'Endcap'.

Ultimately, we find that, in combination, having both a 'Regional Sales Rep' engaging in face-to-face contact and 'Endcap' competition programs increases Weekly Sales units on average by 513.05 whilst holding all other variables constant. This suggests that face-to-face contact does deliver increased store participations in endcap promotions which ultimately leads to the increased weekly sales units for GoodBelly products. Demo programs on the corresponding week increases Weekly Sales, on average, by 106.28 units, by 75.98 when demo programs were held 1-3 weeks ago and 73.03 when demo programs were held 4-5 weeks ago, whilst holding all other variables constant.

5. Conclusion

This paper exploits the data provided on GoodBelly sales and promotions spreadsheet to determine the effectiveness of the promotional programs on the Weekly Sales units of GoodBelly products. We find, after modelling several variable subsets and model modifications, that an optimal model includes five categorical variables with an interaction term involving 'Endcap' and 'Sales_Rep' in which we have leading results at both the model and parameter levels.

Based on our model estimates, we find that the interaction between ‘endcap’ and ‘Sales_Rep’ augments the singular effect of each variable individually. The estimates suggest an increased Weekly Sales, on average, by 513.05 units whilst holding all other variables constant when a regional Sales_Rep is used in combination with Endcap programs. Additionally, the engagement in Demo programs on the corresponding week has a greater effect than other weeks, increasing Weekly Sales, on average, by 106.28 units whilst holding all other variables constant. These estimates favour the use of such promotional programs and recommends continuation in order to maintain current Weekly Sales units, on average, with the assumption that all consumer preferences remain unchanged.

While recognizing the viability of the data provided, this paper also points to a second path forward: one in which better data are the key to deeper insights. We comment on the lack of information regarding quantity of consumers that were engaged by the sales representative, during the three hours of the demonstration. Moreover, data about how much products were used as samples. Ultimately aiding our analysis on the effectiveness of the promotional programs.

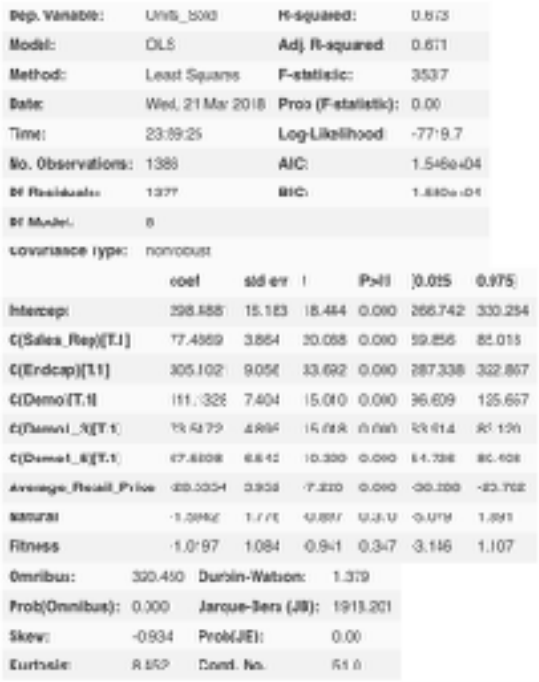
6. Appendix

Assumptions:

- α value = 0.05
- Consumer Income over the time period does not affect the demand for GoodBelly products
- Foot traffic remained constant throughout the data collected.
- All representatives for GoodBelly products were equally capable and received sufficient training such that the engagement with customers and stores were impactful.
- Sales representative had
- The three hours in which the demonstration was carried out continuously received sufficient foot traffic and at the same times each day.
- The reports provided by sales representatives were accurate and representative of the success of the demonstration programs.

A. Full Model

Full Model	
Units_Sold = - 28.535365(Average_Retail_Price) + 77.436914(Sales_Rep) + 305.102123(Endcap) + 111.132849(Demo) + 73.517171(Demo1_3) + 67.569811(Demo4_5) - 1.594168(Natural) - 1.019671(Fitness) + 298.488131	
Mean Squared Error	**Mean Squared Residuals**
4056.80146487	1434747.40613



Dep. Variable:	Units_Sold	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.871
Method:	Least Squares	F-statistic:	3537
Date:	Wed, 21 Mar 2018	Prob (F-statistic):	0.00
Time:	23:59:25	Log-Likelihood:	-7719.7
No. Observations:	1388	AIC:	1.546e+04
Df Residuals:	1377	BIC:	1.552e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	298.488	15.183	18.464	0.000	268.742	328.234
C(Sales_Rep)[T.1]	77.4369	3.864	20.038	0.000	69.856	85.018
C(Endcap)[T.1]	305.102	9.056	33.692	0.000	287.338	322.867
C(Demo)[T.1]	111.1328	7.404	15.010	0.000	96.609	125.657
C(Demo1_3)[T.1]	73.5172	4.895	15.018	0.000	63.914	83.120
C(Demo1_5)[T.1]	67.5698	6.643	10.170	0.000	54.736	80.403
Average_Retail_Price	-28.5354	3.958	-7.210	0.000	-36.300	-20.770
Natural	-1.5942	1.771	-0.900	0.368	-5.070	1.881
Fitness	-1.0197	1.084	-0.941	0.347	-3.146	1.107

Omnibus:	320.450	Durbin-Watson:	1.379
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1913.201
Skew:	-0.934	Prob(JB):	0.00
Kurtosis:	8.852	Cond. No.	51.0

OLS Summary for Full Model

B. Subset Model I

Subset Model I	
Units_Sold = - 28.609165(Average_Retail_Price) + 76.951206(Sales_Rep) + 304.959716(Endcap) + 111.260534(Demo) + 73.663094(Demo1_3) + 67.700203(Demo4_5) + 294.189036	
Mean Squared Error	**Mean Squared Residuals**
4055.49935895	1911943.5417

Dep. Variable:	Units_Sold	**R-squared:**	0.672
Model:	OLS	**Adj. F-squared:**	0.671
Method:	Least Squares	**F-statistic:**	471.4
Date:	Wed, 21 Mar 2018	**Prob (F-statistic):**	0.00
Time:	23:59:25	**Log-Likelihood:**	-7720.5
No. Observations:	1385	**AIC:**	1.545e+04
Df Residuals:	1379	**BIC:**	1.549e+04
Df Model	6		
Covariance Type:	nonrobust		

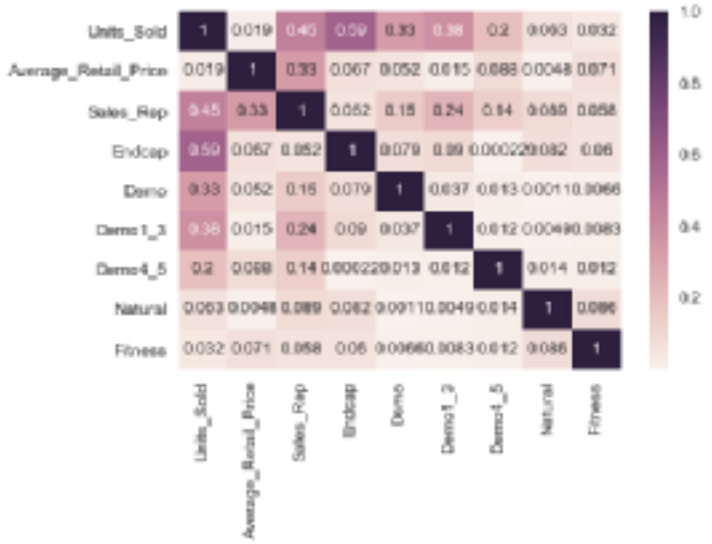
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	294.1890	15.387	18.635	0.000	263.220	325.158
C(Sales_Rep)[T.1]	76.9512	3.841	20.035	0.000	69.417	84.485
C(Endcap)[T.1]	304.9597	9.014	33.831	0.000	287.277	322.643
C(Demo)[T.1]	111.2605	7.401	15.033	0.000	96.742	125.779
C(Demo1_3)[T.1]	73.6631	4.851	15.060	0.000	64.068	83.258
C(Demo4_5)[T.1]	57.7002	6.539	10.353	0.000	54.872	80.528
Average_Retail_Price	-28.6092	3.945	-7.253	0.000	-36.347	-20.871

Omnibus:	324.016	**Durbin-Watson:**	1.378
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1964.703
Skew:	-0.943	**Prob(JB):**	0.00
Kurtosis:	8.530	**Cond. No.**	41.0

OLS Summary for Subset Model I

C. Correlation Between Explanatory Variables

	Units_Sold	Average_Retail_Price	Sales_Rep	Endcap	Demo	Demo1_3	Demo4_5	Natural	Fitness
Units_Sold	1.000000	-0.016829	0.449440	0.593219	0.329358	0.384602	0.196331	0.062784	-0.031772
Average_Retail_Price	-0.016829	1.000000	0.328900	-0.057294	0.051664	-0.015093	0.007836	-0.004812	0.071175
Sales_Rep	0.449440	0.328900	1.000000	0.051887	0.151264	0.242227	0.138508	0.068319	0.067511
Endcap	0.593219	-0.057294	0.051887	1.000000	0.078630	0.090093	-0.000215	0.061884	-0.059535
Demo	0.329358	0.051664	0.151264	0.078630	1.000000	0.036549	-0.013806	-0.001106	0.006623
Demo1_3	0.384602	-0.015093	0.242227	0.090093	0.036549	1.000000	0.011710	-0.004856	0.008313
Demo4_5	0.196331	0.007836	0.138508	-0.000215	-0.013806	0.011710	1.000000	0.013648	-0.011907
Natural	0.062784	-0.004812	0.068319	0.061884	-0.001106	-0.004856	0.013648	1.000000	-0.085628
Fitness	-0.031772	0.071175	0.067511	-0.059535	0.006623	0.008313	-0.011907	-0.085628	1.000000



HeatMap of Correlation - Absolute Values

D. SLR Model with Average_Retail_Price

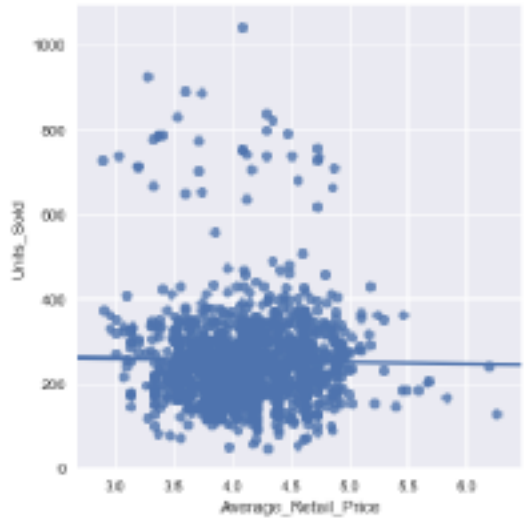
Simple Linear Regression: Units_Sold ~ Average Retail Price	
Units_Sold = - 4.505868(Average_Retail_Price) + 272.326736	
Mean Squared Error	**Mean Squared Residuals**
12325.2495238	6049.52517999

Dep. Variable:	Units_Sold	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	0.4908
Date:	Wed, 21 Mar 2018	**Prob (F-statistic):**	0.484
Time:	23:56:26	**Log-Likelihood:**	-8493.3
No. Observations:	1399	**AIC:**	1.699e+04
Df Residuals:	1384	**BIC:**	1.708e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	272.3267	26.088	10.244	0.000	220.180	324.474
Average_Retail_Price	-4.5059	6.432	-0.701	0.484	-17.123	8.111

Omnibus:	811.019	**Durbin-Watson:**	0.730
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8125.659
Skew:	2.585	**Prob(JB):**	0.00
Kurtosis:	13.570	**Cond. No.**	39.0

OLS Summary for SLR Model with Average_Retail_Price



Scatter Plot of Units_Sold V.S. Average_Retail_Price

E. Subset Model II A

Subset Model II A	
Units_Sold = 67.658944(Sales_Rep) + 310.331985(Endcap) + 110.751351(Demo) + 77.069866(Demo1_3) + 65.651948(Demo4_5) + 181.243383	
Mean Squared Error	**Mean Squared Residuals**
4207.13746488	2251669.03293

Dep. Variable:	Units_Sold	R-squared:	0.660
Model:	OLS	Adj. R-squared:	0.659
Method:	Least Squares	F-statistic:	535.2
Date:	Wed, 21 Mar 2018	Prob (F-statistic):	6.22e-320
Time:	22:59:27	Log-Likelihood:	-7745.4
No. Observations:	1395	AIC:	1.551e+04
Df Residuals:	1389	BIC:	1.554e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	181.2434	2.637	68.718	0.000	176.069	186.417
C(Sales_Rep)[T.1]	67.6589	3.688	18.348	0.000	60.425	74.893
C(Endcap)[T.1]	310.3320	9.150	33.915	0.000	292.382	328.282
C(Demo)[T.1]	110.7514	7.538	14.693	0.000	95.965	125.538
C(Demo1_3)[T.1]	77.0699	4.959	15.542	0.000	67.342	86.798
C(Demo4_5)[T.1]	65.6519	6.654	9.866	0.000	52.599	78.705

Omnibus:	336.166	Durbin-Watson:	1.375
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2050.781
Skew:	-0.990	Prob(JB):	0.00
Kurtosis:	8.021	Cond. No.	8.20

OLS Summary for Subset Model II A

F. Subset Model II B

Subset Model II B	
Units_Sold = 52.042236(Sales_Rep) - 0.255131(Endcap) + 106.287830(Demo) + 75.979833(Demo1_3) + 73.092376(Demo4_5) + 461.257332 (Endcap) * (Sales_Rep) + 189.593039	
Mean Squared Error	**Mean Squared Residuals**
2493.97335512	2270834.26825

Dep. Variable:	Units_Sold	R-squared:	0.798
Model:	OLS	Adj. R-squared:	0.798
Method:	Least Squares	F-statistic:	910.5
Date:	Fri, 23 Mar 2018	Prob (F-statistic):	0.00
Time:	16:58:12	Log-Likelihood:	-7383.5
No. Observations:	1386	AIC:	1.478e+04
Df Residuals:	1379	BIC:	1.482e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	189.5930	2.049	92.544	0.000	185.574	193.612
C(Sales_Rep)[T.1]	52.0422	2.884	18.043	0.000	46.384	57.700
C(Endcap)[T.1]	-0.2551	12.300	-0.021	0.983	-24.384	23.873
C(Demo)[T.1]	106.2878	5.805	18.309	0.000	94.900	117.676
C(Demo1_3)[T.1]	75.9798	3.818	19.900	0.000	68.490	83.470
C(Demo4_5)[T.1]	73.0924	6.139	14.261	0.000	63.031	83.154
C(Endcap)[T.1]:C(Sales_Rep)[T.1]	461.2573	14.973	30.805	0.000	431.804	490.710

Omnibus:	0.991	Durbin-Watson:	2.068
Prob(Omnibus):	0.609	Jarque-Bera (JB):	0.941
Skew:	-0.063	Prob(JB):	0.625
Kurtosis:	3.033	Cond. No.	16.4

OLS Summary for Subset Model II B

G. Subset Model with Demos Removed

Model	
Units_Sold = 93.623920(Sales_Rep) + 330.626533(Endcap) + 189.704820	
Mean Squared Error	**Mean Squared Residuals**
5828.0562491	4501996.53684

Dep. Variable:	Units_Sold	R-squared:	0.628
Model:	OLS	Adj. R-squared:	0.627
Method:	Least Squares	F-statistic:	772.5
Date:	Thu, 22 Mar 2018	Prob (F-statistic):	6.63e-295
Time:	21:19:53	Log-Likelihood:	-7973.6
No. Observations:	1386	AIC:	1.595e+04
Df Residuals:	1383	BIC:	1.597e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	188.7048	3.079	61.293	0.000	183.662	195.747
C(Sales_Rep)[T.1]	98.6239	4.127	22.894	0.000	85.528	101.720
C(Endcap)[T.1]	330.6255	10.707	30.873	0.000	309.522	351.631

Omnibus:	145.815	Durbin-Watson:	1.201
Prob(Omnibus):	0.000	Jarque-Bera (JB):	557.147
Skew:	-0.283	Prob(JB):	7.40e-122
Kurtosis:	5.611	Cond. No.	5.12

OLS Summary for Subset Model with Demos Removed

H. Subset Model with Endcap Removed

Subset Model with Endcap Removed	
Units_Sold = 70.225412(Sales_Rep) +129.103185(Demo) + 90.546768(Demo1_3) + 64.933301(Demo4_5) + 188.571246	
Mean Squared Error	**Mean Squared Residuals**
7708.25333599	1604774.2523

Dep. Variable:	Units_Sold	R-squared:	0.376
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	208.2
Date:	Thu, 22 Mar 2018	Prob (F-statistic):	8.09e-140
Time:	21:24:21	Log-Likelihood:	-8166.5
No. Observations:	1386	AIC:	1.634e+04
Df Residuals:	1381	BIC:	1.637e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	188.5712	3.558	52.999	0.000	181.591	195.551
C(Sales_Rep)[T.1]	70.2254	4.991	14.071	0.000	50.435	80.016
C(Demo)[T.1]	130.1033	10.177	12.786	0.000	110.140	140.066
C(Demo1_3)[T.1]	90.5468	6.691	13.533	0.000	77.422	103.672
C(Demo4_5)[T.1]	64.9333	9.007	7.209	0.000	47.265	82.602

Omnibus:	958.161	Durbin-Watson:	1.054
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13476.173
Skew:	3.103	Prob(JB):	0.00
Kurtosis:	16.959	Cond. No.	5.16

OLS Summary for Subset Model with Endcap Removed

I. Subset Model with Interaction Between Demo and Endcap Included

Subset Model with Interaction Between Demo and Endcap Included	
Units_Sold = 67.973535(Sales_Rep) + 291.497211(Endcap) + 98.790920(Demo) + 76.886283(Demo1_3) + 64.892823(Demo4_5) + 132.327043(Endcap) * (Demo) + 181.812105	
Mean Squared Error	**Mean Squared Residuals**
4131.72321206	1894424.75946

Dep. Variable:	Units_Sold	**R-squared:**	0.656
Model:	OLS	**Adj. R-squared:**	0.655
Method:	Least Squares	**F-statistic:**	458.5
Date:	Thu, 22 Mar 2018	**Prob (F-statistic):**	4.94e-324
Time:	21:23:34	**Log-Likelihood:**	-7733.4
No. Observations:	1388	**AIC:**	1.548e+04
Df Residuals:	1379	**BIC:**	1.552e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	181.8121	2.616	69.487	0.000	176.680	186.944
C(Sales_Rep)[T.1]	67.9735	3.655	18.596	0.000	60.803	75.144
C(Endcap)[T.1]	291.4972	9.786	29.786	0.000	272.300	310.695
C(Demo)[T.1]	98.7909	7.827	12.622	0.000	83.437	114.145
C(Demo1_3)[T.1]	76.8863	4.914	15.645	0.000	67.246	86.527
C(Demo4_5)[T.1]	64.8928	6.596	9.838	0.000	51.954	77.832
C(Endcap)[T.1]:C(Demo)[T.1]	132.3270	25.868	5.117	0.000	81.602	183.052

Omnibus:	313.636	**Durbin-Watson:**	1.403
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1926.036
Skew:	-0.904	**Prob(JB):**	0.00
Kurtosis:	8.486	**Cond. No.**	18.1

OLS Summary for Subset Model with Interaction Between Demo and Endcap Included

J. Additional Information

i. Probiotics

Probiotics are live bacteria that are beneficial to the host organism. Probiotics and other pathogens enter the bloodstream through the stomach lining. Over 100 trillion bacteria and other microorganisms live in the intestines, and the influx of good bacteria can aid digestion and support the immune system.

ii. The Market

By 2008, the global market for probiotic and prebiotic food and beverages was substantial, at $15.4 billion, and still growing. Probiotic products grew between 5% and 30% in 2008—exceptional growth considering that the overall food market grew only 1-2%. There were hundreds of probiotic products on the market, from yogurt to pizza to chocolate. Perhaps the most notable brand was Activia, a family of yogurts produced by The Dannon Company Inc. Activia contained a proprietary strain of Bi dobacterium called Bi dobacterium animalis DN 173 010, which sparked an increase in probiotic awareness in the US.

iii. Mallows Cp

Mallow's Cp is a technique for model selection in regression (Mallows 1973).

The Cp statistic is defined as a criteria to assess fits when models with different numbers of parameters are being compared.

iv. AIC

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data.

- AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters.
- The penalty discourages overfitting, since increasing the number of parameters in the model almost always improves the goodness of the fit.

v. Schwartz Bayesian Criterion

The formula for the Bayesian information criterion (BIC) is similar to the formula for AIC, but with a different penalty for the number of parameters.

- With AIC the penalty is 2k, whereas with BIC the penalty is ln(n) k.
- the model with the lowest BIC is preferred.

K. Reference

i. Costs and Benefits of Including or Omitting Interaction Terms: A Monte Carlo Simulation