UCDAVIS
UNIVERSITY OF CALIFORNIA

Trend Model - Advertising vs Sales – HW 2

Anna Gaudette, Jiaqi Zhou, Prithvi Bisht, Prathyusha Malapaka, Ayan Ghosh
Master of Science in Business Analytics
University of California Davis

UCDAVIS
UNIVERSITY OF CALIFORNIA

Table of Contents

1. Executive Summary:

Advertising is one of major factors driving sales for any firm, and more so for a Business to Consumer, product firm in cosmetics industry. In our study we were given to examine the effect of 13 advertising variables on Sales for a product, launched by a firm 4 years ago.

We formulated our problem by first conceptualizing a framework of our own called 3C-DIUM methodology for solving any advertising analytics problem. This gave a structure to our problem-solving approach, post which we moved on to our next step of exploratory data analysis. Here, we first analyzed the data in excel table, through different forms of selective filtering. We then plotted scatter and line graphs, and followed this up by finding correlation coefficient between all the variables. Ultimately, we were able to trim down to 9 variables, using these techniques.

It was also stated in the problem statement, that the variables have a diminishing return in the form of a square root function. Similarly lagged sales variable was also to be considered as an extra independent variable. Incorporating both, we ran 14 multiple regression models. In two of these we used Stepwise Reduction technique of variable selection. After observing adjusted R squared, AIC, BIC and p values for each, we selected our final focal model which consisted of 7 variables, including lagged sales.

After running our model in R, we observed that all but one variable for offline advertising has a negative coefficient. Similarly for online advertising, the variable has a highly positive coefficient and elasticity values, pointing towards greater impact on sales. The overall model, especially, due to the negative coefficients doesn't have a high adjusted R squared, even though it has lowest AIC.

The point to note however, is that the selected model is still doing better than all other similar models and could be the best possible option for any business recommendations. Our analysis concludes by mentioning that the effect of confounding variables like seasonality, should be considered. Also, the management should give a push towards increasing the online medium of advertising and start leveraging social media platforms as well, the spend for which as per the data is either is either zero or null.

2. Problem

The study aims to examine the effect of dollars spent towards different platforms of advertising, on total sales of a month over a period of 42 months.

Objective

The objective of the study is to check the effectiveness of different mediums of advertising spend, and recommend the right advertising mix of activities, that might lead to increased sales, substantiating the same through model development. We will do it with two pre-defined constraints on the diminishing returns and lagged sales over a given time period.

3. Problem Formulation

With constantly evolving mediums of advertising it is important that managers must evaluate advertising data, holistically and put incremental sales into management objectives, with a single goal of effective customer outreach. Hence, to solve advertising analytics problems such as these our team has cconceptualized a framework called *3C-DIUM*, for problem structure and formulation.

3C – ***C****onsumer*, ***C****onsumer and* ***C****onsumer*. Implying the value, we are putting into the actual end user of our study, in this case the profile of an average cosmetics brand customer

DI – *Data* ***D****iscovery and Data* ***I****ngestion*. This is explained more on the next section and will incorporate variable selection on the basis of Exploratory Data Analysis

U – *Fit for* ***U****SE*. To ensure that our final model is fit for use, we performed the following 2 steps:

- Our data incorporates the effect of diminishing returns through a square root function on advertising variables, as well taking the impact of lagged sales by just 1 unit i.e., a month.
- Subsequently, we zeroed down on our model selection by running 14 regression models - including Stepwise Selection Method of variable selection – and select the model which fits our data best. We leveraged adjusted R^2 values as well as AIC, BIC & p values for the same.

M – *Model Development.* Finally based on above, we selected our focal model and examined the results to possibly further optimize it, and gauging the use of independent advertising variables based on statistical values. With our final focal model known, we further deep-dived into analysis through below steps:

- Usage of multiple functional form like – Square root function, log function, log-lin function, etc,
- We also checked for synergies between variables, by calculating the interaction effect
- Lastly, we calculated elasticities of all the variable for managerial recommendations

4. Data Description

The data provided for the study consist of sales data for 42 months, along with 13 other variables, all of which consist of marketing activities like Offline Advertising, Online Advertising, Newsletters, search, etc. We performed exploratory data analysis using techniques, including but not limited to:

- Line and scatter plots, between all the variable and finding patterns in the data
- Finding correlation between different independent variables
- Check for possible statistically significant collinearities

Based on all of the above, we can conclude that *Adv_Online* is the sum of 6 other variables pointing out to online advertising mediums and *Adv_Offline* is the sum of another 4 variable, consisting of offline medium with 3 catalog advertising and one mailing variable. Similarly, Adv_Total is the sum of Adv_Offline and Adv_Online. Hence, these 3 variables could be removed in straightforward manner.

The other check we performed was, removing attributes based on zero or null values. We used a metric of 75% as the threshold of missing percentage of values, beyond which we will remove a variable from our regression model. We have kept the threshold value high, considering small value of maximum sample size i.e., 42.

We also noticed, based on correlation matrix, that there is a statistically significant high correlation of *0.88* between search and portal, *0.83* between search and retargeting, and lastly *0.70* between retargeting and portal. After running the model and taking out possible combination of 2 at a time, we examined the

AIC values for each, removing *search* and *retargeting* in the process, and adding only *portals*, out of the 3, ensuring our model has least AIC.

5. Model Development

The study aims to find the best possible model for business recommendation based on AIC, BIC, adjusted R squared and f statistic values. With an aim to find out the best fit model for the study, the different combination of multiple regression models (14), were tested. All these possible models followed the over-arching equation of

$$Y_t = \lambda Y_{t-1} + \beta_1 Z_{1t} + \beta_2 Z_{2t} + \beta_3 Z_{3t} + \beta_4 Z_{4t} + \beta_5 Z_{5t} + \cdots + intercept + \epsilon_t$$

Where λ measures the carry-forward effect from the past Sales outcomes, *t* denotes the unit of time period i.e., month, and β_i captures the efficacy of a particular advertising medium. The point to note is that none of the other 13 model is deemed fit for our final recommendation, due to following factors:

- Unusually large number of negative coefficients making the model unreasonable
- Not statistically significant
- Unusually high adjusted R squared (for the ones with no intercept)
- Functional forms or lagged sales time period is different from what is directed for analysis
- And most importantly high AIC and BIC values.

To counter the above effects, we selected the below as our *focal model*, taking the 7 variables, along with an intercept. The logic of selecting these variables, is already explained in *Data Description* above

$$Y_t = \lambda Y_{t-1} + \beta_1 sqrt(Catalogs\_Exist_t) + \beta_2 sqrt(Catalogs\_Winback)_t + \beta_3 sqrt(Catalogs\_NewCust_t) + \beta_4 sqrt(Mailings_t) + \beta_5 sqrt(Newsletter)_t + sqrt(Portals)_t + intercept + \epsilon_t$$

Addition to what is already explained in data description, the selection of this model is based on the following criteria below:

a) p value of 0.0154 which is statistically significant at 5% significance level

b) AIC of 660.96, which is least considering that we factor in the study constraints, as well as make sound business decision, of not taking out variables, which might skew our analysis towards either offline or online mediums.

6. Results

The selected focal model gives out the following output in R.

```
Call:
lm(formula = Sales ~ Stm1 + SCatlg.Exist + SCatlg.Winback + SCatlg.NewCust +
    SMailings + SNewsl + SPortal)

Residuals:
     Min       1Q   Median       3Q      Max
-1154.79  -422.49    79.34   366.83  1757.57

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2122.2149  1100.0980   1.929   0.0624 .
Stm1              0.1433     0.1946   0.736   0.4667
SCatlg.Exist    -24.4168    16.5790  -1.473   0.1503
SCatlg.Winback   53.7332    25.2653   2.127   0.0410 *
SCatlg.NewCust  -26.7490    14.3242  -1.867   0.0708 .
SMailings        -9.6636    42.8265  -0.226   0.8229
SNewsl          168.6910   131.9458   1.278   0.2100
SPortal         819.1793   300.6776   2.724   0.0102 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 685.8 on 33 degrees of freedom
Multiple R-squared:  0.3874,    Adjusted R-squared:  0.2574
F-statistic: 2.981 on 7 and 33 DF,  p-value: 0.01545
```

AIC - 660.96 ; BIC – 676.3;

Elasticity Values

Catalog Exist	Catalog Winback	Catalog New Cust	Mailings	Newsletter	Portal
(0.0289)	0.0282	(0.0221)	(0.0033)	0.0718	0.1837

Below are the key observations based on the above results:

a) Lagged sales with a coefficient value of 0.14 has minimal positive impact on sales

b) Offline mediums like Catalog for exiting customers, new customers and Mailing has a negative impact as per our model, which is an unlikely scenario and points out towards a possibility of lurking variables, which are not present in the model

c) Online mediums – news and Portal - have a highly positive coefficient, pointing towards greater impact for not just these 2 mediums, but also for Search and Retargeting, as both these are correlated with Search, like mentioned earlier.

7. Recommendations and Managerial Implications

Below are the key decisions that we recommend to the advertising manager/head of the firm:

Shift towards online advertising: Online advertising mediums, based on elasticity results have a significantly greater impact on sales. Hence the firm should invest more on mediums like social media, which uptill now has now spend. Multiple studies have proven and it is a common knowledge now, that social media advertising with its targeted marketing, is the way to move forward for firms across the global

Possible Lurking Variables: After running different combinations of advertising variables, we can conclude that none of the model fits in well and is a good predictor of sales. We can attribute this to lurking variables like seasonality, discounts on certain brands, packaging design, product quality index, results, etc.

Interaction Effects: From the interaction summary of focal model, in the Appendix, we could see that there is significant interaction between Catalogs for Winback Customer and Newsletter, as well as Catalogs sent to Existing Customer vs Winback customers. These interactions should be taken into account.

Other possible functional forms: Changing to other function forms like log-log and lin-log is giving a better regression model with lower AIC values and higher Adjusted R squares (refer appendix code)

8. Conclusions

The study clearly points out that Sales and advertising, doesn't have highly significant linear relationship with low values of adjusted R squared across models, leaving us with three possible reasons. First is the possibility of other missing advertising variables in our study, which the data excluded. Second is the null or zero entries across some advertising mediums, which can lead to unusual results. And third is the presence of lurking variables, like seasonality – evident from the trend graph in appendix – as well as variables like discounts.

9. Appendix

a) All the models and statistical values

Modelling_Results_Table_final

b) Interaction effect/Synergy on focal model variables

```
Residuals:
    Min      1Q  Median      3Q     Max
-875.43 -282.84  -37.96  291.47  988.91

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                6709.738   1308.739   5.127 1.37e-05 ***
SCatlg.Exist               -159.329     51.928  -3.068 0.004361 **
SCatlg.Winback             -260.729     65.853  -3.959 0.000393 ***
SCatlg.NewCust              -22.829      9.800  -2.329 0.026308 *
SNewsl                       56.873    100.081   0.568 0.573817
SPortal                    -345.645    546.429  -0.633 0.531520
SCatlg.Exist:SCatlg.Winback    5.728      1.803   3.177 0.003291 **
SCatlg.Winback:SNewsl         40.055     16.749   2.391 0.022832 *
SCatlg.Exist:SPortal          44.143     21.706   2.034 0.050342 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 483.8 on 32 degrees of freedom
Multiple R-squared:  0.7044,    Adjusted R-squared:  0.6304
F-statistic:  9.53 on 8 and 32 DF,  p-value: 1.243e-06

> AIC(step.model.focal.synergy)
[1] 633.0911
>
```

c) RMD file with inference and code for all the models, including functional forms – Refer attachment with code