Smith SCHOOL OF BUSINESS Queen's University

Global Master of Management Analytics

GMMA 831
Marketing Analytics

Dr. Ceren Kolsarici

Data On The Spot - Assignment #1
Saturday, January 23, 2021

Team New York

Order of files:

Filename	Pages	Comments and/or Instructions

Additional Comments:

Team Members:
William Chan
Nicolas Gonthier
Jessee Ho
Nicole Johnson
Shveta Kanetkar

Linear Regression Model Analysis Highlights

Data Investigation

Upon initial data investigation and looking at summary statistics we observe that there are a total of 1,000 records in the data set with 7 variables (See Figure 1). Looking at the minimum, median, mean, and maximum for each variable, we notice that the mean and median are close to each other and approximately in between the minimum and the maximum, indicating to us that the variables are approximately normally distributed. We can confirm this by looking at the 'gpairs' plot and visually see the variable distributions. As a note, we decided not to transform any of the variables as they all appeared to have either linear or random relationships to the sales parameter (See Figure 2). We can also look at our correlation plot and observe the following notable correlations (See Figure 3):

- In-store advertisement and sales are moderately correlated at 0.51,
- Billboard advertisement and sales have the strongest correlation at 0.75,
- Customer satisfaction and sales are weakly correlated at 0.26, and
- Competitor's advertisement spending and sales are weakly correlated at -0.24

Violations of Assumptions

Looking at our final model's diagnostic plots (Figure 4), we observe that there are no major violations of linear regression assumptions. Specifically, our Residuals vs. Fitted plot does not exhibit any non-linear trends. The data points also appear to be well spread out and evenly distributed across both sides of the line. Our Q-Q plot also shows that our residuals are normally distributed and follow a straight line with very little to no deviations when compared to our diagonal line. Looking at our Scale-Location plot, our data points show an equal spread along the fitted value range with a linear pattern indicating to us that our data is homoscedastic. Finally, our Residuals vs. Leverage plot indicates to us that there is no evidence of major outliers in our dataset. There is one data point that is further out, which could warrant a quick investigation, but the Cook's Distance lines don't even appear in the plot itself meaning that none of our data points are close to having high residual or high leverage. Thus, our model won't be biased or skewed because of outlier effects.

Model Comparison and Holdout Prediction (Overfitting Test)

We ran nine models in total starting with price as a single variable in our model all the way up to nine variables including our interaction variables (See Figure 5) to predict our dependent variable, sales. We started with an adjusted r-squared of 0.059 (train) and 0.044 (test) and finished with 0.925 (train) and 0.928 (test), indicating that our model is not overfitted on our in-sample data set and is robust enough to generalize with out-of-sample data (See Figure 6). At each model building phase (i.e., adding/removing variables to the model), we performed ANOVA tests by checking the F-statistic to determine statistically significant differences between our models. We noticed that our early models improved in adjusted r-squared with inclusions of 'store', 'billboard', and 'satisfaction'. The intercepts and standard errors of those models shrank indicating to us that our variables were capturing and explaining more effects

(See Figure 7 and 8). In the end, despite having three models with the same performance on our training and testing adjusted r-squared, we opted with model #9 as it was the most parsimonious (See Figure 9). Our conclusion, based on the significance of our variables and the adjusted r-squared (0.928), is that our model has strong ability to predict our firm's total sales.

Results Interpretation and Implications for Managers

The final model (M9) to predict sales was a function of the product price (price), in-store advertisement spending (store), billboard advertisement spending (billboard), customer satisfaction level (sat), competitors advertisement spending (comp), and a pairwise channel interaction between in-store and billboard advertisement (store:billboard). The final R2 for both train and testing data sets was 0.925 and 0.928, respectively. There were in fact two other models (M8 and M7) that had very similar results to our final model (M9), however based on the results of our ANOVA test, we selected the more parsimonious model.

Using our model, we concluded that advertising does lead to higher sales volume as there are three advertising-related features in the model, in-store advertisement, billboard advertisement, and the combined effect of in-store and billboard advertisement. Both in-store and the in-store:billboard interaction variables are significant and have a positive effect on total sales. Specifically, for every dollar of in-store advertisement, total sales go up by $1.39 keeping all other variables constant. It should be noted that on average billboard advertising alone doesn't impact total sales given the insignificance of the variable in our model. In our interaction variable, for every $1.00 of in-store and billboard advertisement, total sales go up by $0.00544. Therefore, we conclude that the most effective channel for advertising is in-store, followed by billboard advertising but only combined with in-store. Billboard advertising alone without in-store doesn't impact sales, on average. We also conclude that our customers are price sensitive. In our model we observed that for every $1.00 increase in price, total sales declines by $199.60.

Business Relevance

Our linear model can give us important insights on the best allocation of marketing spend. As business managers we are looking for sales return on our marketing investments and understanding which levers drive sales is important to maintaining and growing a healthy business. For example, in our coefficient plot, we see that customer satisfaction and competitor advertisement spending has huge effects on our store sales (See Figure 10). This gives us an indication that customer service and the quality of service could aid as a differentiator between our business and competitors. Implementing a customer-centric strategy could help increase sales for our business.

Specifically, our linear model allowed us to conclude the following:

- Advertising does lead to higher sales volume as evidenced by the strong positive correlation and significant coefficients
- Billboards + In-store advertisement or In-store advertisement alone are the most effective ad channels as these were channel parameters that came out most significant

- Our customers are price sensitive as observed by the fact that the ‘price’ variable was significant and had a large negative coefficient
- Our competitor’s advertising does impact our sales, as their advertising spend goes up our sales go down. This is evidenced by a negative coefficient
- Analytics, specifically linear models can be used to predict sales as our model has a high adjusted r-squared on both training and testing data, indicating strong ability to predict and generalize with out-of-sample data

Based on these insights we would recommend that the marketing department:

- Continue to monitor these parameters by collecting data and refreshing the model on a regular basis
- Limit price increases, due to the highly sensitive nature of these customers
- Monitor competitor advertising spend and make further efforts to collect more granular data on where they are advertising
- Allocate advertising spend to in-store alone or billboards with in-store advertisements

Appendix

Figure 1 - Summary Statistics

```
> # basic descriptive statistics
> summary(df)
       X             store        billboard       printout          sat             comp           price      
 Min.   :   1.0   Min.   :1150   Min.   : 219   Min.   :  26.0   Min.   :54.00   Min.   : 230.0   Min.   : 85.00  
 1st Qu.: 250.8   1st Qu.:1831   1st Qu.: 845   1st Qu.: 626.0   1st Qu.:66.00   1st Qu.: 656.0   1st Qu.: 96.00  
 Median : 500.5   Median :1990   Median :1003   Median : 814.0   Median :70.00   Median : 788.0   Median :100.00  
 Mean   : 500.5   Mean   :1993   Mean   :1003   Mean   : 806.7   Mean   :69.45   Mean   : 791.7   Mean   : 99.75  
 3rd Qu.: 750.2   3rd Qu.:2153   3rd Qu.:1168   3rd Qu.: 979.2   3rd Qu.:73.00   3rd Qu.: 931.0   3rd Qu.:103.00  
 Max.   :1000.0   Max.   :2798   Max.   :1791   Max.   :1555.0   Max.   :85.00   Max.   :1339.0   Max.   :117.00  
     sales       
 Min.   : 5819  
 1st Qu.:15062  
 Median :17385  
 Mean   :17586  
 3rd Qu.:20038  
 Max.   :33767  
```

Figure 2 - 'gpairs' plot

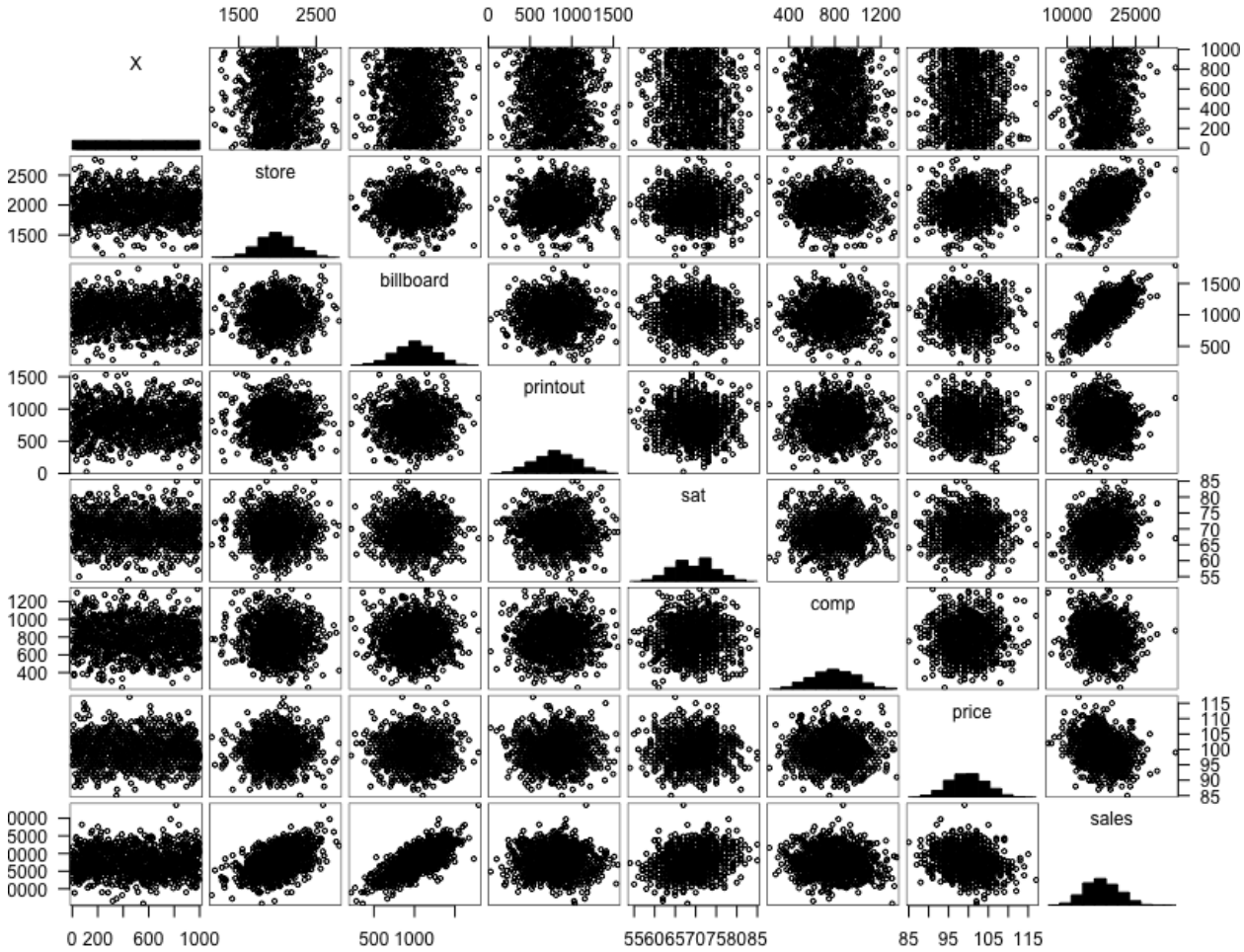

Figure 3 - corr plot

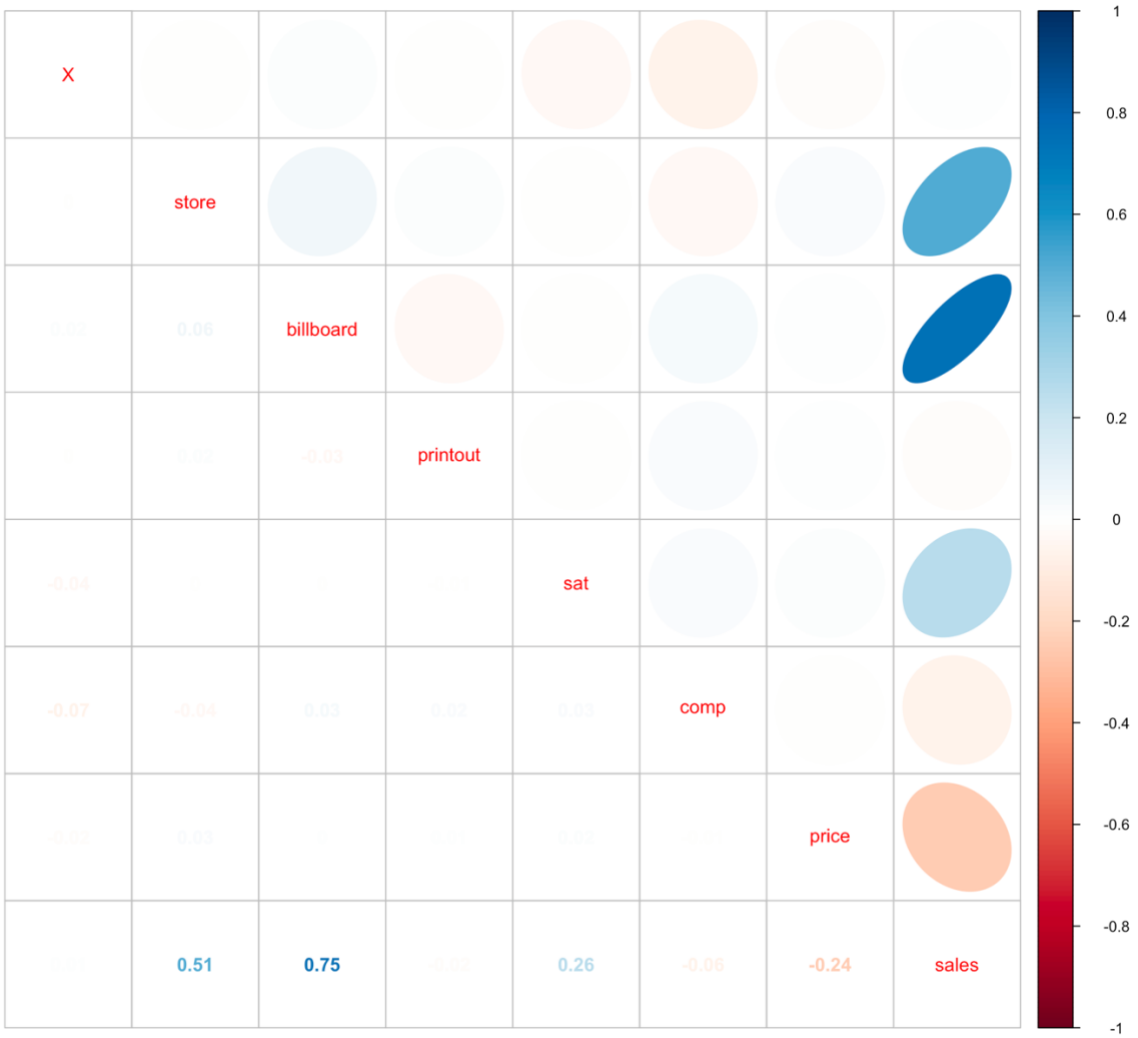

Figure 4 - Final Model Diagnostic Plot

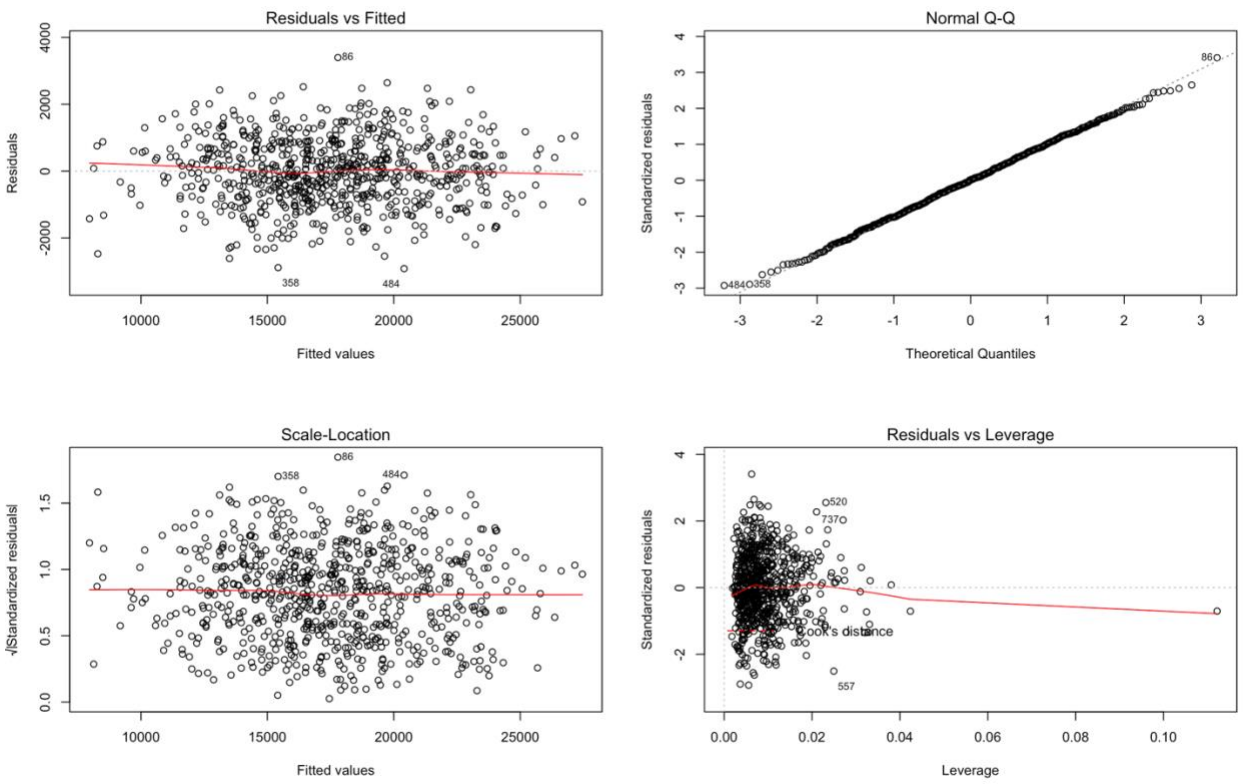

Figure 5 - Model Comparison

Model	Variables Used	Train R^2	Test R^2
M1	Price	0.059	0.044
M2	Price, Store	0.324	0.312
M3	Price, Store, Billboard	0.839	0.854
M4	Price, Store, Billboard, Printout	0.839	0.853
M5	Price, Store, Billboard, Printout, Satisfaction	0.913	0.916
M6	Price, Store, Billboard, Printout, Satisfaction, Competitors	0.912	0.920
M7	Price, Store, Billboard, Printout, Satisfaction, Competitors, Store:Billboard, Store:Printout, Billboard:Printout	0.925	0.928
M8	Price, Store, Billboard, Printout, Satisfaction, Competitors, Store:Billboard	0.925	0.928
M9 - Final Model	**Price, Store, Billboard, Satisfaction, Competitors, Store:Billboard**	**0.925**	**0.928**

Figure 6 - Model 1 vs. Model 9 Comparison

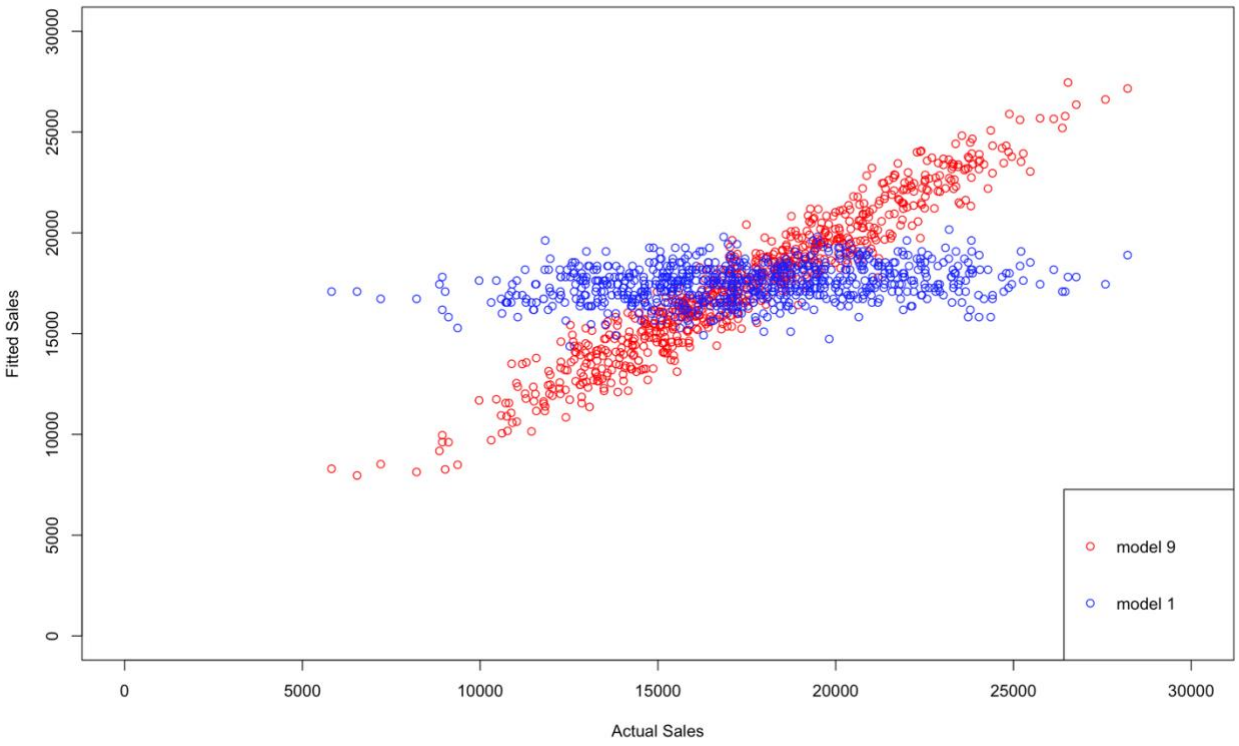

Figure 7 - Model #1

```
Call:
lm(formula = sales ~ price, data = train.df)

Residuals:
   Min     1Q Median     3Q    Max
-11268  -2482     14   2305  10139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35538.63    2595.40  13.693  < 2e-16 ***
price        -180.90      25.99  -6.959 7.48e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3541 on 748 degrees of freedom
Multiple R-squared:  0.06081,   Adjusted R-squared:  0.05956
F-statistic: 48.43 on 1 and 748 DF,  p-value: 7.475e-12
```

Figure 8 - Final Model

```
> #model 9 - train - final model
> m9.train <- lm(sales ~ price + billboard + store +  sat + comp + store:billboard, data=train.df)
> summary(m9.train)

Call:
lm(formula = sales ~ price + billboard + store + sat + comp +
    store:billboard, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-2919.2  -694.2     9.0   688.6  3397.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.180e+03  1.551e+03   5.918 4.99e-09 ***
price           -1.992e+02  7.355e+00 -27.084  < 2e-16 ***
billboard        1.967e+00  1.206e+00   1.631 0.103419
store            2.365e+00  6.230e-01   3.796 0.000159 ***
sat              1.994e+02  7.356e+00  27.102  < 2e-16 ***
comp            -1.530e+00  1.871e-01  -8.176 1.27e-15 ***
billboard:store  4.463e-03  6.038e-04   7.392 3.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1000 on 743 degrees of freedom
Multiple R-squared:  0.9256,    Adjusted R-squared:  0.925
F-statistic:  1540 on 6 and 743 DF,  p-value: < 2.2e-16
```

Figure 9 - ANOVA Test (Model 8 vs. Model 9)

```
> par(mfrow=c(2,2))
> plot(m9.train)
> anova(m8.train, m9.train)
Analysis of Variance Table

Model 1: sales ~ price + store + billboard + printout + sat + comp + store:billboard
Model 2: sales ~ price + billboard + store + sat + comp + store:billboard
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    742 742305449
2    743 743253665 -1   -948216 0.9478 0.3306
```

Figure 10 - Coefficient Plot