Svedka Case Analysis

James Machado, Justin Kaplan, Jonny Shakerchi

4/13/2017

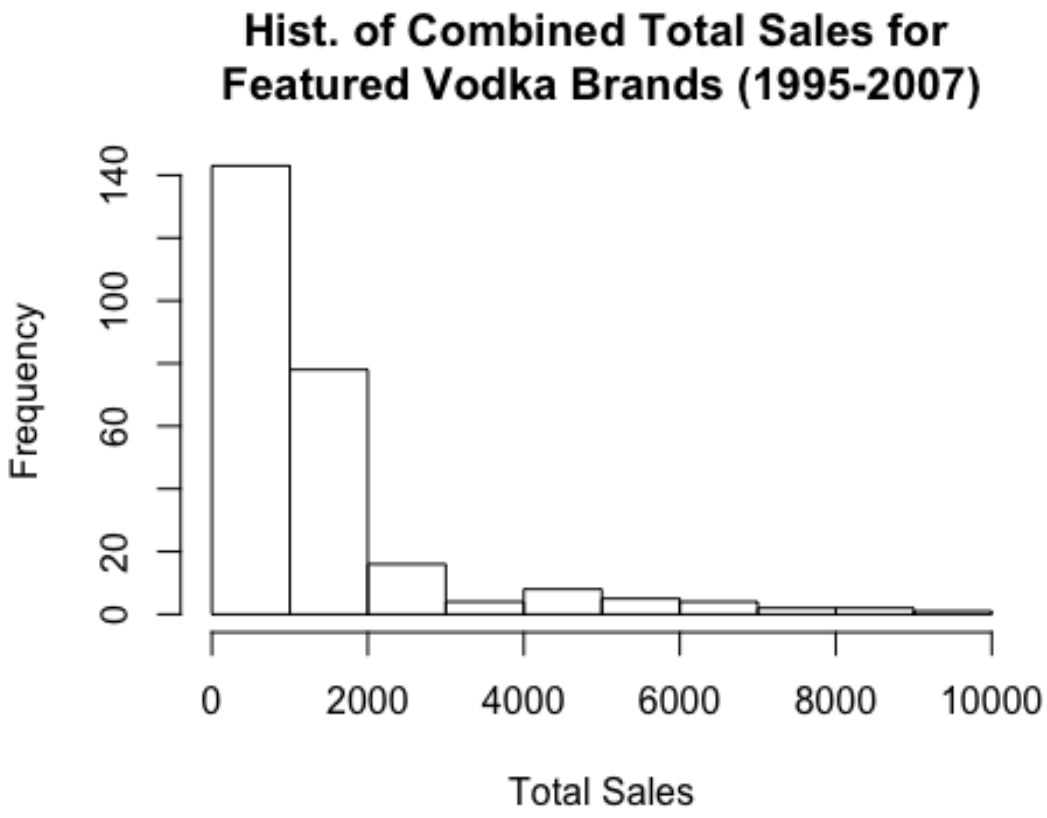
Code to get started

```
setwd("~/Desktop")
library(readr)
Svedka_data <- read_csv("Svedka data.csv")

## Warning: Missing column names filled in: 'X68' [68], 'X69' [69],
## 'X70' [70], 'X71' [71], 'X72' [72], 'X73' [73], 'X74' [74], 'X75' [75]

## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   BrandName = col_character(),
##   LnSales = col_double(),
##   LnLSales = col_double(),
##   Ln2Lsales = col_double(),
##   LnDiff = col_double(),
##   diff = col_character(),
##   DollarSales = col_double(),
##   PriceRerUnit = col_double(),
##   LagPrice = col_double(),
##   LnPrice = col_double(),
##   LnLPrice = col_double(),
##   Mag = col_double(),
##   News = col_double(),
##   Outdoor = col_double(),
##   Broad = col_double(),
##   Print = col_double(),
##   LnMag = col_double(),
##   LnNews = col_double(),
##   LnOut = col_double(),
##   LnBroad = col_double()
##   # ... with 12 more columns
## )

## See spec(...) for full column specifications.

View(Svedka_data)
Svedka_data = subset(Svedka_data, select = -c(68:75) )
library(plyr)
Svedka_data <- rename(Svedka_data, c("PriceRerUnit"="PricePerUnit"))
```

Question 1. (5) Create a simple histogram of our target variable, TotalSales.

Comment on the shape of the distribution that you see.

```
hist(Svedka_data$TotalSales,
     main = 'Hist. of Combined Total Sales for \nFeatured Vodka Brands (1995-
2007)',
      xlab = 'Total Sales')
```



Answer: The shape of the distribution is extremely downward trending with the majority of Total Sales at $2,000 or less. This means that most brands had Total Sales of $2,000 or less.

Question 2. (10) Run a regression of the natural logarithm of total sales on the the following variables: price, print marketing expenditure, outdoor marketing expenditure, broadcast marketing expenditure, and previous year's sales. Keeping in mind your answer to #1, explain why it makes sense to use ln(TotalSales) for the dependent variable. Comment on the relative influence of the five variables on sales.

```
RegQ2 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales,
data = Svedka_data)
summary(RegQ2)
```

```
##
## Call:
## lm(formula = LnSales ~ PricePerUnit + Print + Outdoor + Broad +
##     LagTotalSales, data = Svedka_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79999 -0.25496  0.07956  0.35411  1.23370
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.470e+00  8.615e-02  75.100  < 2e-16 ***
## PricePerUnit  -7.026e-03  7.840e-04  -8.962  < 2e-16 ***
## Print          5.149e-05  7.572e-06   6.800 7.27e-11 ***
## Outdoor       -3.663e-04  9.815e-05  -3.732 0.000234 ***
## Broad         -4.529e-05  4.802e-05  -0.943 0.346558
## LagTotalSales  5.334e-04  3.884e-05  13.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5475 on 257 degrees of freedom
## Multiple R-squared:  0.761,  Adjusted R-squared:  0.7564
## F-statistic: 163.7 on 5 and 257 DF,  p-value: < 2.2e-16
```

Answer: It makes sense to use ln(Sales), because according to the histogram we made, most of the data was skewed to the left. Using a natural log un-skews the data, making it easier to interpret. Based on this regression, it is apparent that price, print marketing expenditure, outdoor marketing expenditure, and previous year's sales are significant variables affecting vodka sales by very marginal percentages less than 1%. However, broadcast marketing expenditure is not a significant variable in determining sales.

Question 3. (15) Sometimes we can improve model fit by taking logs on independent variables. Run a second regression of the natural logarithm of change in sales on the natural logarithm of previous period's prices, and the natural log of marketing expenditures on print, outdoor, and broadcasting. Comment on the comparison of your two models at this point in the analysis.

```
RegQ3 <- lm((LnSales-LnLSales) ~ LnPrice + LnPrint + LnOut + LnBroad, data =
Svedka_data)
summary(RegQ3)
```

```
##
## Call:
## lm(formula = (LnSales - LnLSales) ~ LnPrice + LnPrint + LnOut +
##     LnBroad, data = Svedka_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62884 -0.06978 -0.00528  0.04199  1.10794
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.117365   0.098396  -1.193  0.23405
## LnPrice      0.036774   0.025379   1.449  0.14855
## LnPrint      0.014776   0.004600   3.212  0.00149 **
## LnOut       -0.012622   0.005786  -2.182  0.03004 *
## LnBroad     -0.005764   0.005158  -1.117  0.26489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1672 on 258 degrees of freedom
## Multiple R-squared:  0.1252, Adjusted R-squared:  0.1116
## F-statistic: 9.228 on 4 and 258 DF,  p-value: 5.532e-07
```

Answer: We resize the independent variables by taking the natural log of them to better understand the effects on the natural log of the independent variable. Unlike in the last question, when all variables are resized by the natural log, we can more clearly see the percent affects of each independent variable on the dependent variable. However, now we can see that only the natural log of print marketing expenditure and outdoor marketing expenditure are the only significant variables on sales, respectively adding or subtracting an additional percentage point for each additional dollar spent on print or outdoor ads.

Question 4. (15 pts) To understand the influence of vodka quality, expand your regression model from question 2 by adding the tier 1 and tier 2 dummy variables (that indicate whether a vodka brand belongs to the first or second quality tiers) to the set of independent variables named in question 2. How does quality influence sales?

```
RegQ4 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales
+ Tier1 + Tier2, data = Svedka_data)
summary(RegQ4)
```

```
##
## Call:
## lm(formula = LnSales ~ PricePerUnit + Print + Outdoor + Broad +
##     LagTotalSales + Tier1 + Tier2, data = Svedka_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6851 -0.2262  0.0440  0.2705  1.0332
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.552e+00  8.807e-02  74.401  < 2e-16 ***
## PricePerUnit  -7.382e-03  1.387e-03  -5.324 2.23e-07 ***
## Print          4.002e-05  7.385e-06   5.418 1.39e-07 ***
## Outdoor       -3.177e-04  8.878e-05  -3.578 0.000414 ***
## Broad         -4.362e-05  4.301e-05  -1.014 0.311443
## LagTotalSales  5.586e-04  3.510e-05  15.914  < 2e-16 ***
## Tier1          2.591e-01  2.105e-01   1.231 0.219445
## Tier2         -4.645e-01  1.197e-01  -3.879 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4903 on 255 degrees of freedom
## Multiple R-squared:  0.8099, Adjusted R-squared:  0.8046
## F-statistic: 155.2 on 7 and 255 DF,  p-value: < 2.2e-16
```

Interpreation: According to our regression, a tier 1 vodka is not a significant variable to determining sales. However, a tier 2 vodka is significant. If the vodka is tier 2, sales can be expected to decrease by 46%.

Question 5. (15) To understand the influence of competition and brand power, expand your model again and run a regression by adding the sum of sales of all the competing brands in the previous year ("lagtotalminussales") to the independent variables in question 3. What additional insight does this model provide?

```
RegQ5 <- lm((LnSales-LnLSales) ~ LnPrice + LnPrint + LnOut + LnBroad + LagTot
alMinusSales, data = Svedka_data)
summary(RegQ5)
```

```
##
## Call:
## lm(formula = (LnSales - LnLSales) ~ LnPrice + LnPrint + LnOut +
##     LnBroad + LagTotalMinusSales, data = Svedka_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64492 -0.06413 -0.01537  0.04681  1.09548
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.970e+00  5.453e-01  -3.613 0.000364 ***
## LnPrice            -5.790e-03  2.775e-02  -0.209 0.834864
## LnPrint             1.827e-02  4.618e-03   3.956 9.86e-05 ***
## LnOut              -6.701e-03  5.921e-03  -1.132 0.258800
## LnBroad             3.567e-03  5.730e-03   0.623 0.534138
## LagTotalMinusSales  3.197e-05  9.261e-06   3.452 0.000650 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1637 on 257 degrees of freedom
## Multiple R-squared:  0.1639, Adjusted R-squared:  0.1477
## F-statistic: 10.08 on 5 and 257 DF,  p-value: 7.969e-09
```

Answer: The LagTotalMinusSales variable is a significant variable, but it has an extremely small effect on total sales. Instead, this variable provides some explanation on the variance in the regression and gives us a stronger R^2.

Question 6. (10) To measure the sales growth of new brands compared to the existent ones, include the variable "firstintro" to the independent variable set in question 4. "Firstintro" is equal to 1 in the first three years after a brand is introduced, and equals 0 elsewhere. How does it help to include this variable in the model?

```
RegQ6 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales
+ Tier1 + Tier2 + Firstintro, data = Svedka_data)
summary(RegQ6)
```

```
##
## Call:
## lm(formula = LnSales ~ PricePerUnit + Print + Outdoor + Broad +
##     LagTotalSales + Tier1 + Tier2 + Firstintro, data = Svedka_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73621 -0.23657  0.04249  0.27209  1.01342
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.583e+00  8.643e-02  76.170  < 2e-16 ***
## PricePerUnit  -7.692e-03  1.357e-03  -5.669 3.90e-08 ***
## Print          3.527e-05  7.329e-06   4.812 2.56e-06 ***
## Outdoor       -2.282e-04  9.011e-05  -2.532 0.011943 *
## Broad         -4.392e-05  4.200e-05  -1.046 0.296774
## LagTotalSales  5.419e-04  3.459e-05  15.667  < 2e-16 ***
## Tier1          3.654e-01  2.076e-01   1.760 0.079657 .
## Tier2         -4.264e-01  1.174e-01  -3.632 0.000340 ***
## Firstintro    -9.411e-01  2.579e-01  -3.649 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4788 on 254 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.8136
## F-statistic:   144 on 8 and 254 DF,  p-value: < 2.2e-16
```

Answer: Again, adding this significant variable explains more variance in the regression. The effect of this variable on sales is also extremely significant; for the first 3 years of sales of a new vodka brand, we can expect sales to decrease by 94% in each of the first three years. This implies that branding and competition are extremely strong barriers to entry.

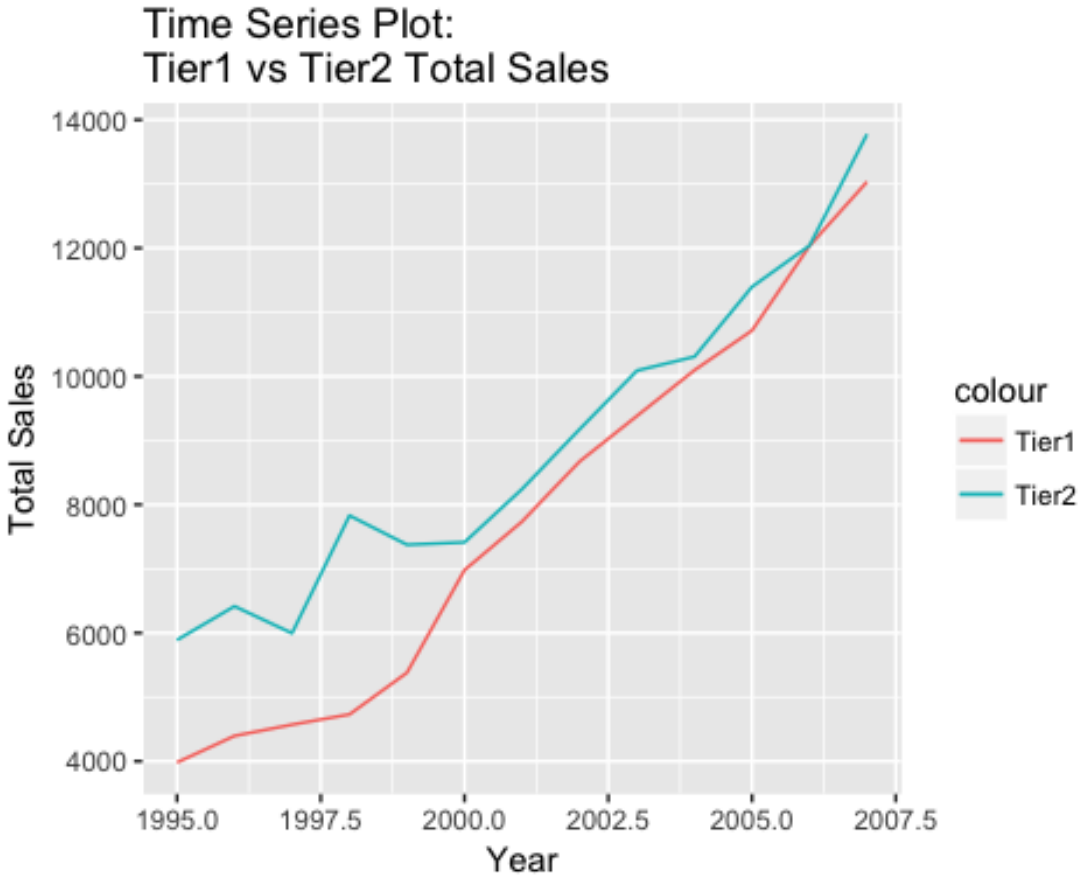
Question 7. (10) Examine the coefficients of the price and advertising variables in your last four regressions. Why do the coefficients of price and advertising change in the above regressions?

Answer: The coefficients keep changing as we add more variables, because as we add more variables, we better explain the variance in each regression. Before, the price and advertising variable coefficients were accounting for the effects of the added variables that were not originally included. When included, the price and advertising coefficients changed to more accurately reflect their effects on total sales.

Question 8. (10) Create a time-series plot with two lines on it: total industry sales units for Tier 1 brands and total industry sales units for Tier 2 brands. NOTE: This will require some aggregation and pre-processing of the data, and is more of a challenge than it might appear.

```
Tier1_Agg <- aggregate(Svedka_data$TotalSales ~ Svedka_data$Year + Svedka_dat
a$Tier1, FUN = sum)
View(Tier1_Agg)
Tier1_Agg <- rename(Tier1_Agg,c("Svedka_data$Tier1" = "Tier1","Svedka_data$Ye
ar" ="Year",
                                "Svedka_data$TotalSales" = "Total Sales"))

Tier2_Agg <- aggregate(Svedka_data$TotalSales ~ Svedka_data$Year + Svedka_dat
a$Tier2, FUN = sum)
View(Tier2_Agg)
Tier2_Agg <- rename(Tier2_Agg,c("Svedka_data$Tier2" = "Tier2","Svedka_data$Ye
ar" ="Year",
                                 "Svedka_data$TotalSales" = "Total Sales"))
```

```r
Tier1 <- Tier1_Agg[which(Tier1_Agg$Tier1==1 & Tier1_Agg$Year),]
View(Tier1)
Tier1$Tier1 <- NULL
View(Tier1)

Tier2 <- Tier2_Agg[which(Tier2_Agg$Tier2==1 & Tier2_Agg$Year),]
View(Tier2)
Tier2$Tier2 <- NULL
View(Tier2)

library(ggplot2)
ggplot(Tier1,aes(Year,`Total Sales`)) +
  geom_line(aes(color="Tier1")) +
  geom_line(data=Tier2,aes(color="Tier2")) +
  ylab("Total Sales") +
  xlab("Year") +
  ggtitle("Time Series Plot: \nTier1 vs Tier2 Total Sales")
```



Question 9. (10) Conclude with a short summary of your findings. How do the 4 elements of the marketing mix influence unit sales in this industry? What insights should we communicate with M. Cuvelier?

Answer:

Product - over time, tier 2 vodka brands have more sales, and it is extremely hard to introduce a new brand due to high barriers to entry.

Price - significant variable to sales; consumers are price sensitive.

Promotion - it seems outdoor and print ads are the most effective; broadcast is insignificant.

Place - we did not analyze distribution.