Qiutong Shi

BU.450.740 Retail Analytics

Homework 2 (Individual): Linear regressions in R

Prof. Mitsukuni Nishida

In this exercise you will be using market-level cross-sectional sales data: “minivan_hw2.csv” from an actual car manufacturer and its retail dealers. The data are private; Please do not share with anyone outside the class. The data set is a csv file containing a major car maker’s annual sales in quantity of a particular model of their minivans. For a given zipcode, this manufacturer’s dealers in that zip code sells this model and report sales in units. Assume for now that this model is identical in year, specs, options, and MSRP for all cars sold.

Variable description:

- Location: City name of the geographic market
- Zip: US zip code for the geographic market
- q_sold: Quantity of minivan sold in # of cars
- ave_p: Average price sold (in thousand USD)
- comp_p: Average Chrysler’s minivan price (in thousand USD)
- adv: Advertising expenditure (in thousand USD)

Questions: 20 points (+ bonus 5 points) in total

1. [4 points] Descriptive analysis
    1. Describe the summary statistics of all variables except zip

```
> summary(minivan$Location,minivan$q_sold)
   Length     Class      Mode 
       44 character character 
> summary(minivan$Location)
   Length     Class      Mode 
       44 character character 
> summary(minivan$q_sold)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   25.75   33.00   32.84   41.25   55.00 
> summary(minivan$Location)
   Length     Class      Mode 
       44 character character 
> summary(minivan$q_sold)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   25.75   33.00   32.84   41.25   55.00 
> summary(minivan$ave_p)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.39   19.25   20.09   20.05   20.91   23.06 
> summary(minivan$comp_p)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.57   19.56   20.18   20.36   20.95   23.61 
> summary(minivan$adv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.1200  0.6775  1.6700  2.2993  3.8800  5.1900 
```

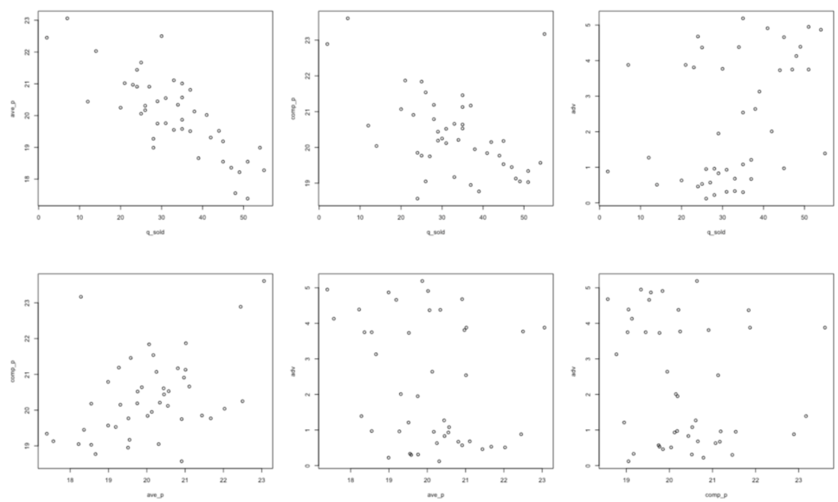
2. Generate two-way plots (i.e., plot of two variables) for all variables except location and zip. Do you observe any patterns?

I observe a linear relationship between ave_p and q_sold, comp_p and ave_p, comp_p and q_sold.

```
> attach(minivan)
The following objects are masked from minivan (pos = 3):

    adv, ave_p, comp_p, Location, q_sold, Zip

> par(mfrow=c(2,3))
> plot(q_sold, ave_p)
> plot(q_sold, comp_p)
> plot(q_sold, adv)
> plot(ave_p, comp_p)
> plot(ave_p, adv)
> plot(comp_p,adv)
```



2. [4 points] Simple linear regression
   1. Perform a regression of q_sold on ave_p. Is there any relationship between q_sold and ave_p?

```
> simple_lr <- lm(q_sold ~ ave_p, data = minivan)
> summary(simple_lr)

Call:
lm(formula = q_sold ~ ave_p, data = minivan)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.8636  -4.0869   0.3968   5.7093  15.7176 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 184.4653    17.4677  10.560 2.15e-13 ***
ave_p        -7.5637     0.8696  -8.698 6.08e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.311 on 42 degrees of freedom
Multiple R-squared:  0.643,	Adjusted R-squared:  0.6345 
F-statistic: 75.65 on 1 and 42 DF,  p-value: 6.085e-11
```

   2. Discuss the fit of the model using R_squared

According to the R_squared, 64.3% of variation of a q_sold is explained by the independent ave_p.

```
> summary(simple_lr)$r.sq
[1] 0.6430036
```

3. [6 points] Multiple linear regression
    1. Perform a multiple regression of q_sold on ave_p, adv, and comp_p. Is there any relationship between q_sold and controls: ave_p, adv, and comp_p? Are the signs of the estimated parameters reasonable?

The signs of the parameters are reasonable. According to the law of demand, when price increases, demand decreases. In this case, as average price sold (in thousand USD) and average Chrysler's minivan price (in thousand USD) increase, quantity of minivan sold will decrease, reflected by the negative signs of ave_p and comp_p. It is reasonable for increasing advertising to generate increasing quantity sold, reflected by the positive sign of adv in the model.

```
> multiple_lr <- lm(q_sold ~ ave_p+adv+comp_p, data = minivan)
> summary(multiple_lr)

Call:
lm(formula = q_sold ~ ave_p + adv + comp_p, data = minivan)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.6249  -4.8934  -0.0172   4.3332  15.1604 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 190.9227    22.4622   8.500 1.68e-10 ***
ave_p        -6.6867     0.9277  -7.208 9.65e-09 ***
adv           1.2152     0.6330   1.920    0.062 .  
comp_p       -1.3181     1.0269  -1.283    0.207    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.996 on 40 degrees of freedom
Multiple R-squared:  0.6886,	Adjusted R-squared:  0.6653 
F-statistic: 29.49 on 3 and 40 DF,  p-value: 3.163e-10
```

    2. Discuss the model fit using R_squared

According to the R_squared, 68.9% of variation of a q_sold is explained by the independent ave_p.

```
> summary(multiple_lr)$r.sq
[1] 0.6886314
```

4. [6 points] If you are a data analyst for this company, which variables would you

   suggest to add in the last estimation equation (i.e., multiple regression) to make

more precise argument regarding the effect of price on unit sold? These variables you will propose to add may not currently be in your data. Which variables would you suggest to add and why? You can think of this question as proposing a “wish list.” We may not always have access to the variables we wish, but we can always keep a lookout for them.

I want to have a list of minivan’s substitutes and their prices to learn about the price of minivan relative to its substitutes. This is because substitutes can impact the demand curve therefore impacting sales. If substitutes do affect sales of minivan, perhaps the marketing team can put more effort into points of differentiation of minivans. I would also want to have variable of gas prices because gas is a complementary good to cars that could impact the demand curve of minivan. If increasing gas price decreases minivan sales, perhaps the marketing team can target the minivan as “eco.” I also want to know if mortgage quote has an impact on sales, because some consumers may be willing to pay for minivan if the interest is low or they get a quote generous enough relative to their credit score or financial circumstance when they cannot afford paying at full price. I also want to know if the specific car manufacturer or retailer has a discount for consumers. It is possible that discounted price looks more attractive to consumers therefore increasing sales.