Alla Topp
MSDS 660

Linear Regression Model

The Carseats data set tracks sales information for car seats. It has 400 observations (each at a different store) and 11 variables:

- Sales: unit sales in thousands
- CompPrice: price charged by competitor at each location
- Income: community income level in 1000s of dollars
- Advertising: local ad budget at each location in 1000s of dollars
- Population: regional pop in thousands
- Price: price for car seats at each site
- ShelveLoc: Bad, Good or Medium indicates quality of shelving location
- Age: age level of the population
- Education: ed level at location
- Urban: Yes/No
- US: Yes/No

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes
11	9.01	121	78	9	150	100	Bad	26	10	No	Yes
12	11.96	117	94	4	503	94	Good	50	13	Yes	Yes
13	3.98	122	35	2	393	136	Medium	62	18	Yes	No
14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes
15	11.17	107	117	11	148	118	Good	52	18	Yes	Yes
16	8.71	149	95	5	400	144	Medium	76	18	No	No

1. Pick 2-3 predictors (independent variables) and one response (dependent variable). List them. Perform appropriate data explorations. State your research questions.

Y = Sales, X = Price, X1= CompPrice, X2 = Income.

Alla Topp
MSDS 660

I would like to see how Sales of the car seats are affected by Prices, Competitors Prices and Income of Population, if sales and Prices, Income and Competitor's prices have positive or negative relationships.

```
> summary(Carseats) # get summary of your data set/data frame
     Sales          CompPrice       Income        Advertising       Population      Price        ShelveLoc        Age          Education    Urban       US
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0   No :118   No :142
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
> |
```

```
> str(Carseats)      # display structure
'data.frame':   400 obs. of  11 variables:
 $ Sales       : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice   : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income      : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price       : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc   : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
 $ Age         : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education   : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

2. Test the entire model (or significance of the model) using a global F-test. State both null and alternative hypothesis.

H0: All the regressors (Price, Comprice, Income) are not statistically different from zero (or do not contribute significantly to the model)

H1: At least one of the independent variable has a contribution to the model

```
> lm.model = lm(Sales ~ Price+CompPrice+Income, data= Carseats)
> summary(lm.model)

Call:
lm(formula = Sales ~ Price + CompPrice + Income, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1166 -1.5039 -0.2224  1.4806  6.1195

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.950236   0.981383   5.044 6.95e-07 ***
Price       -0.087197   0.005816 -14.991  < 2e-16 ***
CompPrice    0.092786   0.008996  10.315  < 2e-16 ***
Income       0.015251   0.004005   3.809 0.000162 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.231 on 396 degrees of freedom
Multiple R-squared:  0.3805,    Adjusted R-squared:  0.3758
F-statistic: 81.08 on 3 and 396 DF,  p-value: < 2.2e-16
```

What is the value of the F-statistic test or p-value? What is your conclusion?

From the results we can see that the value of p-value of the entire model is 2.2e-16 which is much less than 0.05.

We reject H0 and say that variables (or at least one of them) are significantly different from zero.

R-squared value which is 0.3805 tells us that 38% of variance in the measure of sales can be predicted by Price, CompPrice, Income.

3. Test significance of each explanatory variable (X). State both null and alternative hypothesis of each explanatory variable (list each in pairs). T-test?

1) Sales(y) and Price(x)

H0: There is no relationship between "Sales" and "Price".

H1: There is relationship between "Sales" and "Price".

```
> #3) Testing significance of each x variable. Gives t-statistic, p-value and 95% confidence interval.
> t.test(Sales, Price) # Performs a t-test of means between two variables X and Y for the hypothesis H0 : μX = μY .

        Welch Two Sample t-test

data:  Sales and Price
t = -90.837, df = 410.35, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -110.6423 -105.9550
sample estimates:
 mean of x  mean of y
  7.496325 115.795000
```

Here we can reject the null hypothesis. The linear regression suggests a relationship between price and sales given the low p-value of the t-statistic. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.

2) Sales(y) and CompPrice(x)

H0: there is no relationship between "Sales" and "CompPrice".

H1: there is relationship between "Sales" and "CompPrices".

```
> t.test(Sales, CompPrice) #t-test of means between Sales and CompPrices

        Welch Two Sample t-test

data:  Sales and CompPrice
t = -150.69, df = 426.04, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -119.0111 -115.9463
sample estimates:
 mean of x  mean of y
  7.496325 124.975000
```

According to results of the t-test, we can reject the null hypothesis as well and state that there is relationship between Sales and Competitor's prices. The coefficient (0.09279) states a positive relationship between two variables.

3) **Sales(y) and Income(x)**

H0: There is no relationship between Sales and Income

H1: There is relationship between Sales and Income

Based on the t-test we can reject the null hypothesis. There is relationship between Sales and income of the population, coefficient states that there is positive relationship between sales and Income.

Give a brief interpretation of each coefficient in the model. Which predictor(s) will cause you to reject the null hypothesis?

To see which predictor variables are significant, you can examine the coefficients table, which shows the estimate of regression beta coefficients and the associated t-statistic p-values:

```
> lm.model

Call:
lm(formula = Sales ~ Price + CompPrice + Income, data = Carseats)

Coefficients:
(Intercept)        Price    CompPrice       Income
    4.95024     -0.08720      0.09279      0.01525
```

There are all predictors in this model that cause us to reject the null hypothesis based on the t-test we ran.

For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero. For a given predictor variable, the coefficient can be interpreted as the average effect on y of a one unit increase in predictor, holding all other predictors fixed. Looking at the model above, we can see the coefficient for the price is -0.08720, it means that sales will decrease by 87 units on average if the prices go up. Then, Income coefficient suggests that if income will go up, then sales will grow by 15 units on average. The same with competitor`s prices, if they go up, then car seat sales will be positively affected.

4) Now, try a different model.

Alla Topp
MSDS 660

```
> lm2 = lm(Sales~Population+Age+Education)
> summary(lm2)

Call:
lm(formula = Sales ~ Population + Age + Education)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0136 -1.8285 -0.1146  1.8208  8.2170

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.152081   0.935079  10.857  < 2e-16 ***
Population   0.000684   0.000941   0.727    0.468
Age         -0.040093   0.008512  -4.710 3.43e-06 ***
Education   -0.050291   0.052875  -0.951    0.342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.752 on 396 degrees of freedom
Multiple R-squared:  0.05754,   Adjusted R-squared:  0.0504
F-statistic: 8.059 on 3 and 396 DF,  p-value: 3.18e-05
```

Which model fits the data better? How do you select the model? Explain your answers.

The first model fits the data better because all the predictors (Price, CompPrice, Income) show the t-test results less than 0.05. We also proved that all those predictors have negative or positive relationships with response (dependent variable). The first model`s p-values is much smaller than the second model`s p-value, it means that that model fits data better. According to the second model, we accepted the null hypothesis with predictors Population and Education because those two independent variables don`t have relationships with independent variable.

One of the ways to select the model is to run global F-test and look at all the coefficients and their p-values, it will most likely show us the relationships between the variables, also p-value and the coefficients, so we can identify the relationship between variables and make an assumption.

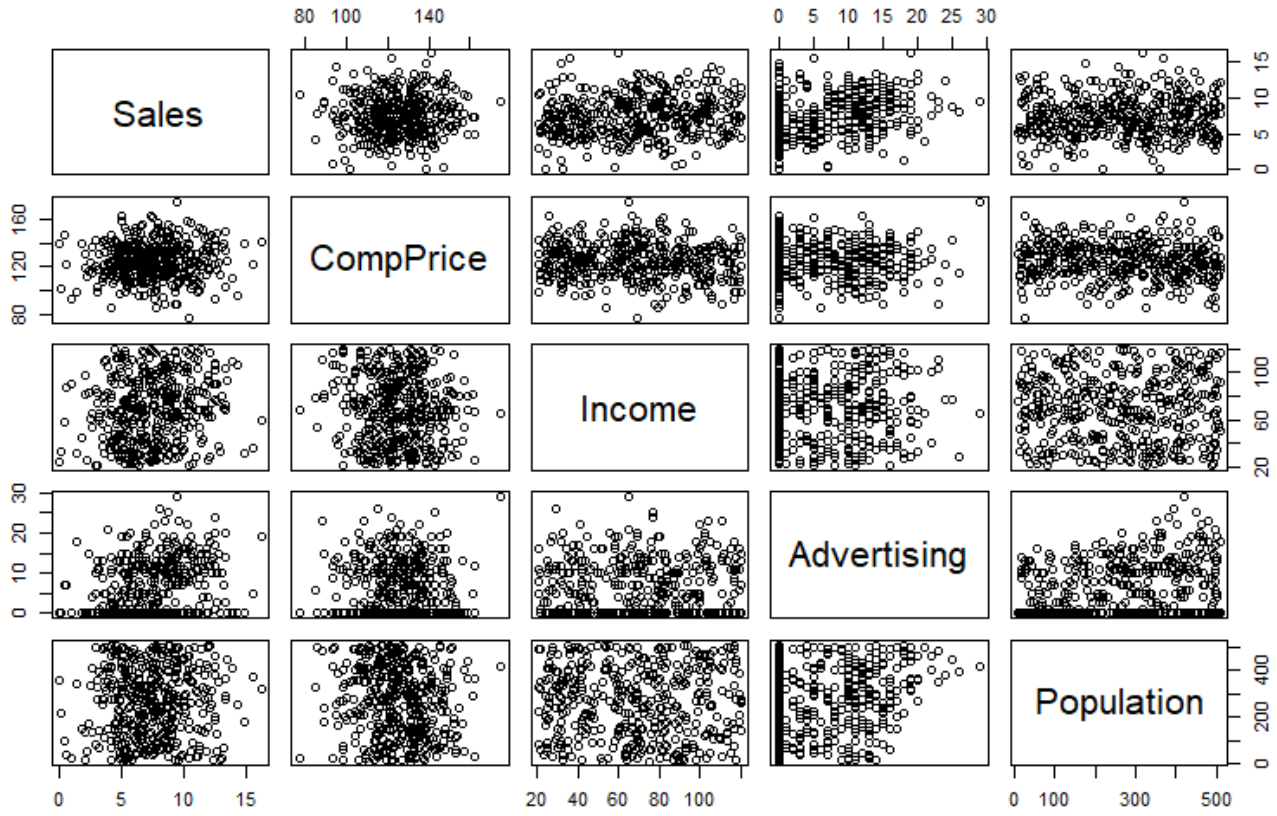
6) Multicolinarity

Correlations with quantitative data:

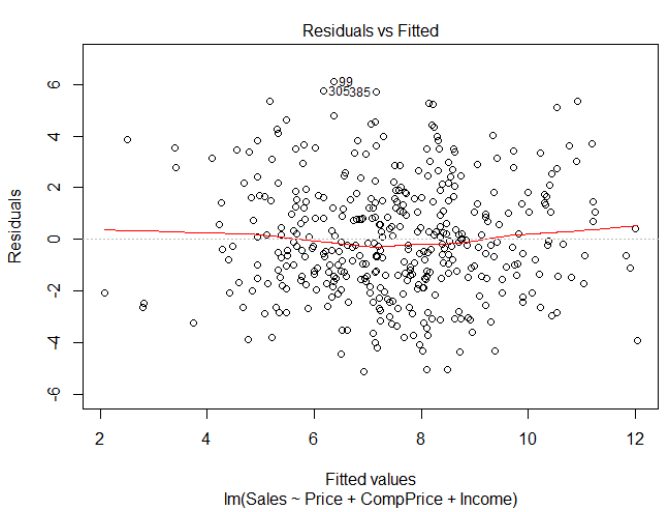
```
> cor(subset(Carseats, select=-c(ShelveLoc,Urban,US)))  # omit qualitative data
                Sales   CompPrice       Income  Advertising   Population       Price           Age     Education
Sales       1.00000000  0.06407873  0.151950979  0.269506781  0.050470984 -0.44495073 -0.231815440 -0.051955242
CompPrice   0.06407873  1.00000000 -0.080653423 -0.024198788 -0.094706516  0.58484777 -0.100238817  0.025197050
Income      0.15195098 -0.08065342  1.000000000  0.058994706 -0.007876994 -0.05669820 -0.004670094 -0.056855422
Advertising 0.26950678 -0.02419879  0.058994706  1.000000000  0.265652145  0.04453687 -0.004557497 -0.033594307
Population  0.05047098 -0.09470652 -0.007876994  0.265652145  1.000000000 -0.01214362 -0.042663355 -0.106378231
Price      -0.44495073  0.58484777 -0.056698202  0.044536874 -0.012143620  1.00000000 -0.102176839  0.011746599
Age        -0.23181544 -0.10023882 -0.004670094 -0.004557497 -0.042663355 -0.10217684  1.000000000  0.006488032
Education  -0.05195524  0.02519705 -0.056855422 -0.033594307 -0.106378231  0.01174660  0.006488032  1.000000000
```

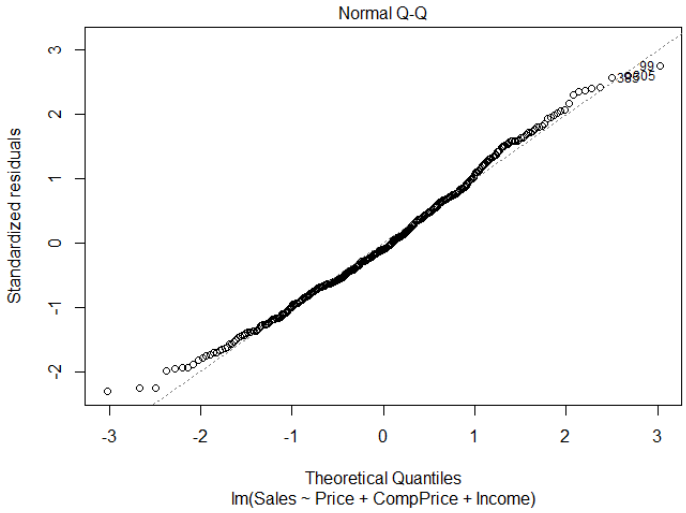
I chose to calculate correlation between y and all x variables we have and see the relationships between them. Here we can identify if relationships are negative or positive. Below I posted the graph that shows the correlation as well but I chose not all the variables.

Alla Topp
MSDS 660

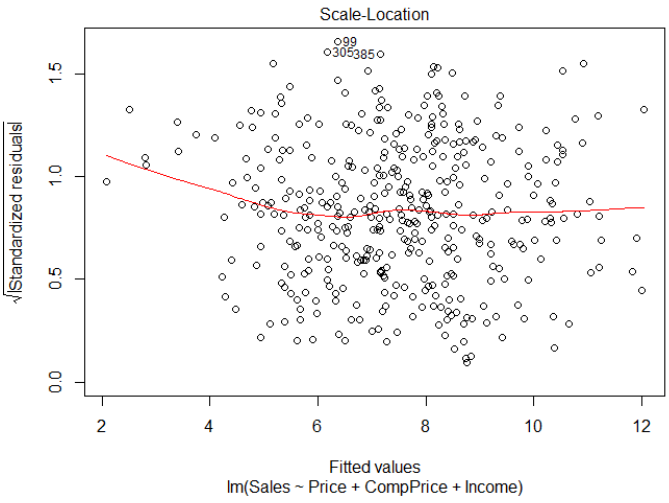


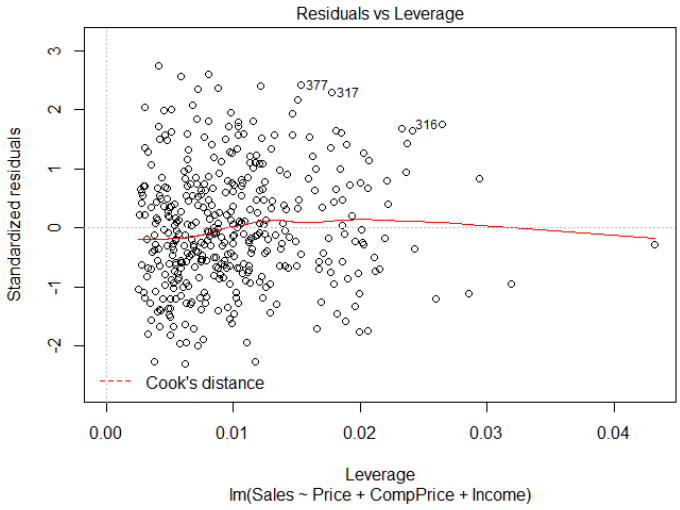
The model assumptions with residual analysis with plot(lm.model)





Alla Topp
MSDS 660





The assumptions for linear regression are:

1. Linearity: the relationship between x,x1,x2 and the mean of y is linear

2. Homoscedasticity: the variance of residual is the same for any value of x

3. Independence: observations are independent of each other

4. Normality: for any fixed value of x, y is normally distributed

The other comments I can make here is that residuals are a little bit spread out, which is not a problem here.

Variance Inflation Factor

The variance inflation factor quantifies the effect of collinearity on the variance of our regression estimates.

In practice it is common to say that any VIF greater than 5 is cause for concern.

```
install.packages("car")
library(car)
vif(lm.model) #calculates the VIFs for each of the predictors of a model.
```

Output:

```
> vif(lm.model)
    Price CompPrice    Income
 1.520076  1.525111  1.006687
> vif(lm2)
Population        Age  Education
  1.013251   1.001827   1.011450
```

We can see that all the results are much less than 5, so there are no multicollinearity issues. We could predict it on earlier steps when we compared two models and saw that the first model fits data better. So to see some different results we can explore the whole data and look at multicollinearity there.

Alla Topp
MSDS 660

```
> summary(carseats.lm)

Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8692 -0.6908  0.0211  0.6636  3.4115

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice         0.0928153  0.0041477  22.378  < 2e-16 ***
Income            0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising       0.1230951  0.0111237  11.066  < 2e-16 ***
Population        0.0002079  0.0003705   0.561    0.575
Price            -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood     4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium   1.9567148  0.1261056  15.516  < 2e-16 ***
Age              -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education        -0.0211018  0.0197205  -1.070    0.285
UrbanYes          0.1228864  0.1129761   1.088    0.277
USYes            -0.1840928  0.1498423  -1.229    0.220
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

```
> vif(carseats.lm)
                GVIF Df GVIF^(1/(2*Df))
CompPrice   1.554618  1        1.246843
Income      1.024731  1        1.012290
Advertising 2.103136  1        1.450219
Population  1.145534  1        1.070296
Price       1.537068  1        1.239785
ShelveLoc   1.033891  2        1.008367
Age         1.021051  1        1.010471
Education   1.026342  1        1.013086
Urban       1.022705  1        1.011289
US          1.980720  1        1.407380
```

I don`t see any huge numbers or issues of the VIF values in the whole model.