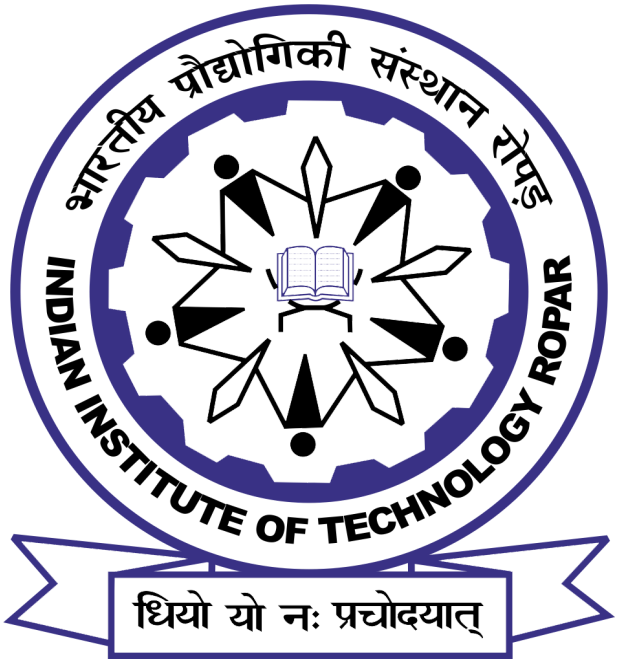
MA515 Project Report



Indian Institute of Technology, Ropar

Submitted To - Dr. Arun Kumar

Date - 30 November 2021

Submitted By -

Yogesh Vaidhya - 2018EEB1277

1. Project Details

Do exploratory data analysis on the data. Use multiple linear regression to predict the Sales. Further, use lasso technique to see if some coeeficients are 0. Use different lambda for comparision purpose.

2. Dataset Used

Carseats Data is used. The data sets are from the ISLR book. Below Fig. [1] shows is preview of the carseats dataset.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
0	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
1	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
2	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
3	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
4	4.15	141	64	3	340	128	Bad	38	13	Yes	No

Fig1. Carseats Data Set

3. Exploratory Data Analysis

Predictors datatype in Dataset:

```
Sales          float64
CompPrice        int64
Income           int64
Advertising      int64
Population       int64
Price            int64
ShelveLoc       object
Age              int64
Education        int64
Urban           object
US              object
```

So, in the dataset Sales have decimal value. CompPrice, Income, Advertising, Population, Price, Age, Educatoin have integer value. ShelveLoc has string as Good, Medium or Bad. Urban and US has string as Yes or No.

NaN values in Dataset:

Sales	0
CompPrice	0
Income	0
Advertising	0
Population	0
Price	0
ShelveLoc	0
Age	0
Education	0
Urban	0
US	0

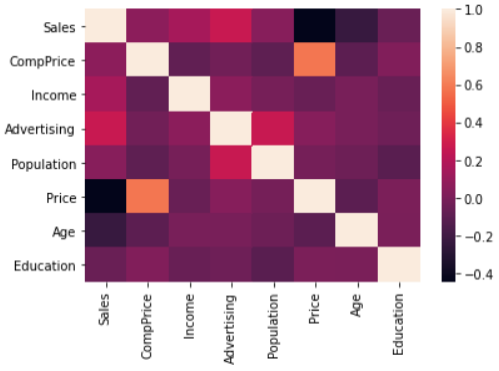
No Nan values are present in given dataset.

Summary of the data:

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	7.496325	124.975000	68.657500	6.635000	264.840000	115.795000	53.322500	13.900000
std	2.824115	15.334512	27.986037	6.650364	147.376436	23.676664	16.200297	2.620528
min	0.000000	77.000000	21.000000	0.000000	10.000000	24.000000	25.000000	10.000000
25%	5.390000	115.000000	42.750000	0.000000	139.000000	100.000000	39.750000	12.000000
50%	7.490000	125.000000	69.000000	5.000000	272.000000	117.000000	54.500000	14.000000
75%	9.320000	135.000000	91.000000	12.000000	398.500000	131.000000	66.000000	16.000000
max	16.270000	175.000000	120.000000	29.000000	509.000000	191.000000	80.000000	18.000000

Population have very large standart deviation of 147.37 .

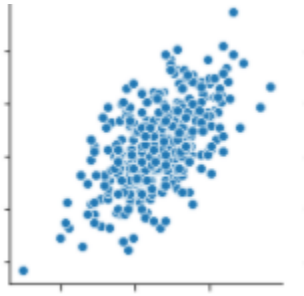
Heat Map Plots:



	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
Sales	1.00	0.06	0.15	0.27	0.05	-0.44	-0.23	-0.05
CompPrice	0.06	1.00	-0.08	-0.02	-0.09	0.58	-0.10	0.03
Income	0.15	-0.08	1.00	0.06	-0.01	-0.06	-0.00	-0.06
Advertising	0.27	-0.02	0.06	1.00	0.27	0.04	-0.00	-0.03
Population	0.05	-0.09	-0.01	0.27	1.00	-0.01	-0.04	-0.11
Price	-0.44	0.58	-0.06	0.04	-0.01	1.00	-0.10	0.01
Age	-0.23	-0.10	-0.00	-0.00	-0.04	-0.10	1.00	0.01
Education	-0.05	0.03	-0.06	-0.03	-0.11	0.01	0.01	1.00

Price and ComPrice have positive correlation with value 0.58 . Others have very low correlation between them.

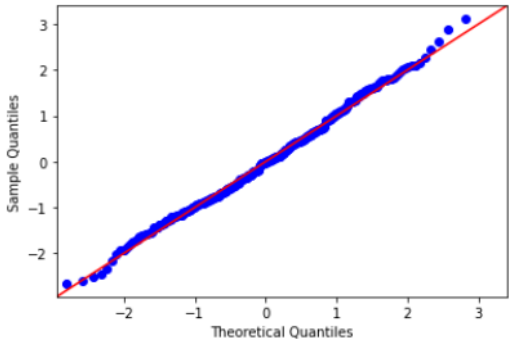
Scatter Plot:



This is the scatter plot between Price and ComPrice. From this plot, positive correlation can be seen between them.

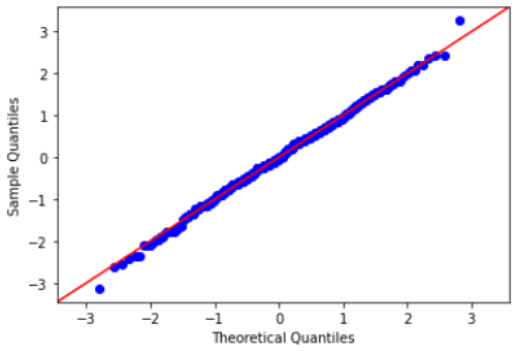
Q–Q (quantile-quantile) plot:

Sales:



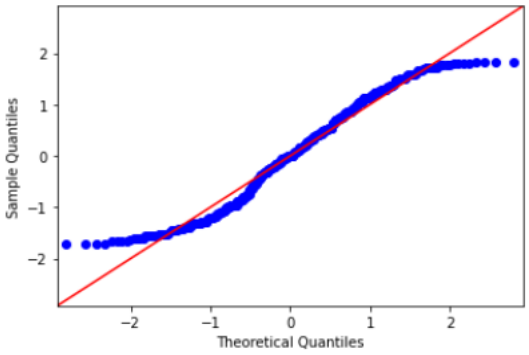
Sales have normal distribution.

ComPrice:



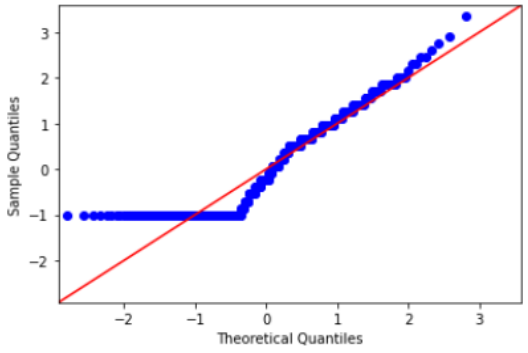
ComPrice have normal distribution

Income:



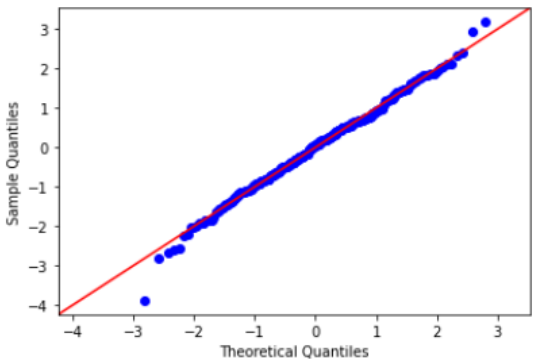
Income have both left and right skewness.

Advertising:



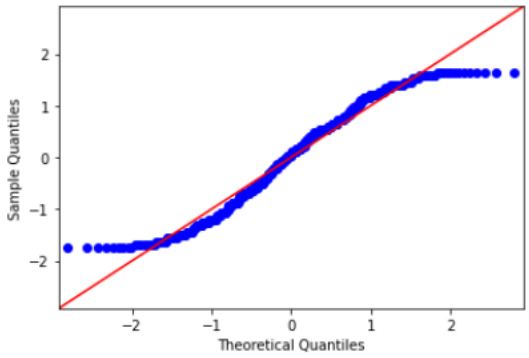
Advertising have left skewness.

Price:



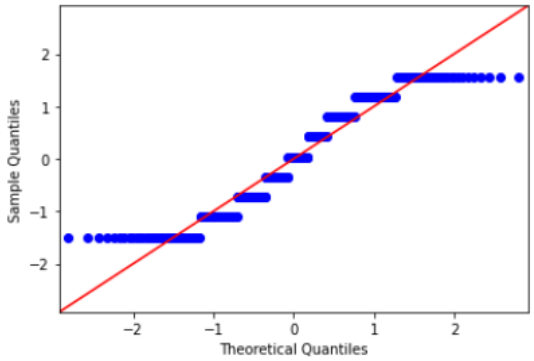
Price have normal distribution.

Age:



Age have both left and right skewness

Education:



Attribute makes more sense as discrete

4. Multiple Linear Regression

Preprocessing on Dataset:

- Separate the X and y from dataset
- For X, Converting Object string of ShelveLoc, Urban, US to numbers
- Then do Column Transformation on ShelveLoc, Urban, US transformed attributes

```
[[0.0 1.0 0.0 1.0 1.0 0.0 0.0 138 73 11 276 120 42 17]
 [0.0 1.0 0.0 1.0 0.0 1.0 0.0 111 48 16 260 83 65 10]
 [0.0 1.0 0.0 1.0 0.0 0.0 1.0 113 35 10 269 80 59 12]
 [0.0 1.0 0.0 1.0 0.0 0.0 1.0 117 100 4 466 97 55 14]
 [1.0 0.0 0.0 1.0 1.0 0.0 0.0 141 64 3 340 128 38 13]
 [0.0 1.0 1.0 0.0 1.0 0.0 0.0 124 113 13 501 72 78 16]
 [1.0 0.0 0.0 1.0 0.0 0.0 1.0 115 105 0 45 108 71 15]
 [0.0 1.0 0.0 1.0 0.0 1.0 0.0 136 81 15 425 120 67 10]
 [1.0 0.0 1.0 0.0 0.0 0.0 1.0 132 110 0 108 124 76 10]
 [0.0 1.0 1.0 0.0 0.0 0.0 1.0 132 113 0 131 124 76 17]
```

Transformed X

Predict Sales:

- Divided the dataset into training and testing in the ratio 70:30
- Used sklearn model selection library for splitting training and testing data
- Taken LinearRegression model from sklearn linear model library
- Use OLS model from statsmodels.api library to calculate different parameters.

Regression coefficient:

```
array([ 1.21764962e-01, -1.21764962e-01, -5.33085785e-02,  5.33085785e-02,
       -2.25186439e+00,  2.53136111e+00, -2.79496720e-01,  9.39027222e-02,
        1.42404765e-02,  1.30157013e-01,  1.30621176e-04, -9.57611854e-02,
       -4.36518640e-02, -1.97451934e-02])
```

Intercept value: 7.783335966340507
Regression score: 0.8834699579060779
R-squared: 0.873
Adj. R-squared: 0.870

5. Lasso Regression

Predict Sales:

- Divided the dataset into training and testing in the ratio 70:30
- Used sklearn model selection library for splitting training and testing data
- Taken Lasso model from sklearn linear model library.
- Calculated Regression coefficient and Regression score for different value of lambdas

Lambda	Regression Score	No. of Coefficient turned zero
0.001	0.8834597923337515	2
0.01	0.8830582981259949	3
0.1	0.8649118612794333	5
0.5	0.5112242171513168	7
1	0.505564877893838	8
2	0.4864646227418405	8
10	0.19487788070729173	9

λ = 0: Same coefficients as simple linear regression
λ = inf : All coefficients zero

6. Conclusion

With the above method, I am successfully able to perform exploratory data analysis on dataset which gives many insights about the data. The using Linear Regression predicted sales. The regession score is 0.88. The Adjusted R-squared value of 0.87. This is also a good result. Lasso regression help us in feature selection to reduce the predictor along with good regression score. For lamda = 0.01, regression score is 0.88 and 3 predictor have zero coefficient value. Also lamda = 0.1 is a good choice. It reduces parameters from 14 to 9 and have regression score of 0.86.