Sales Prediction Carseats

Lana Korošec

2023-08-09

Dataset Carseats frpm ISLR package is a study case of sales prediction.

We will forecast sales based on variables: price, advertising, ShelveLoc (3-factor variable - positioning quality)

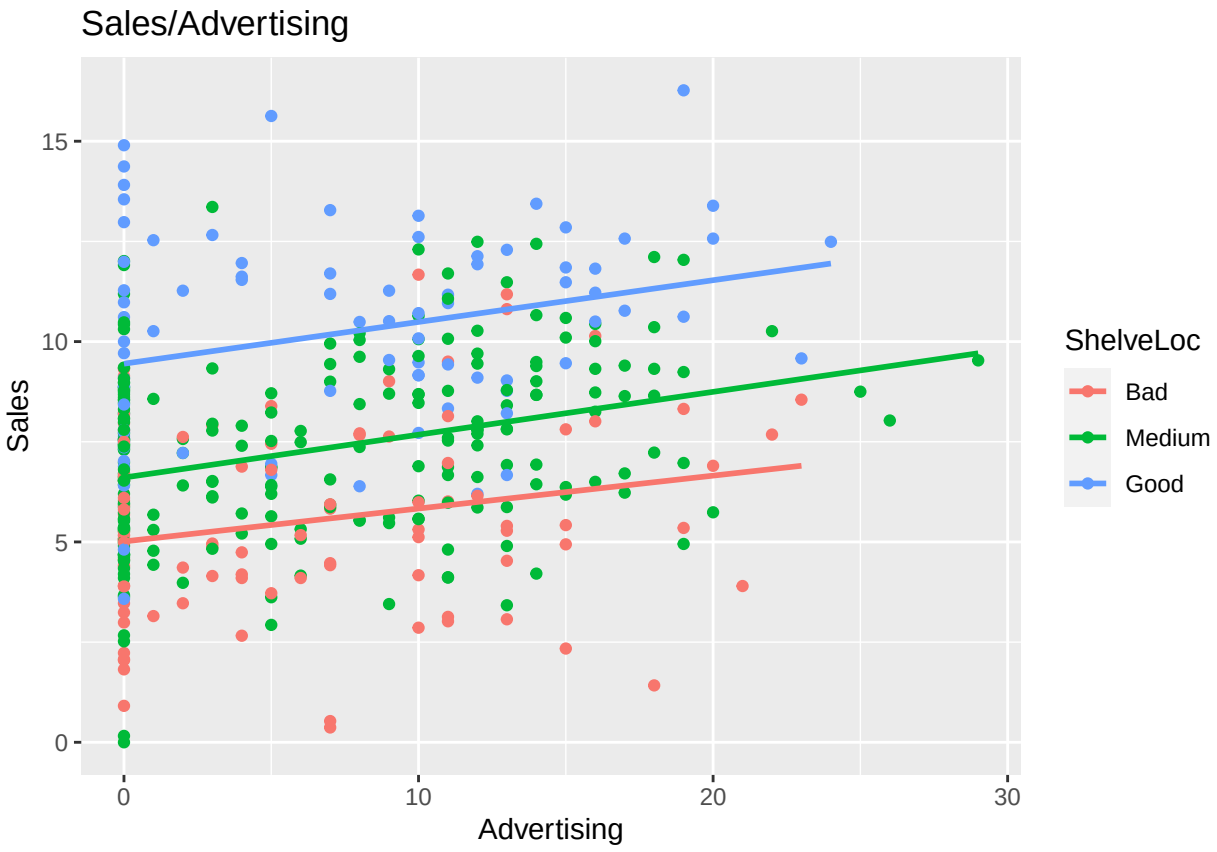
Regarding given data, we change factos in variable ShelveLoc to have "natural" order from "bad to good"

Data presentation

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

```
## `geom_smooth()` using formula = 'y ~ x'
```


```
## `geom_smooth()` using formula = 'y ~ x'
```



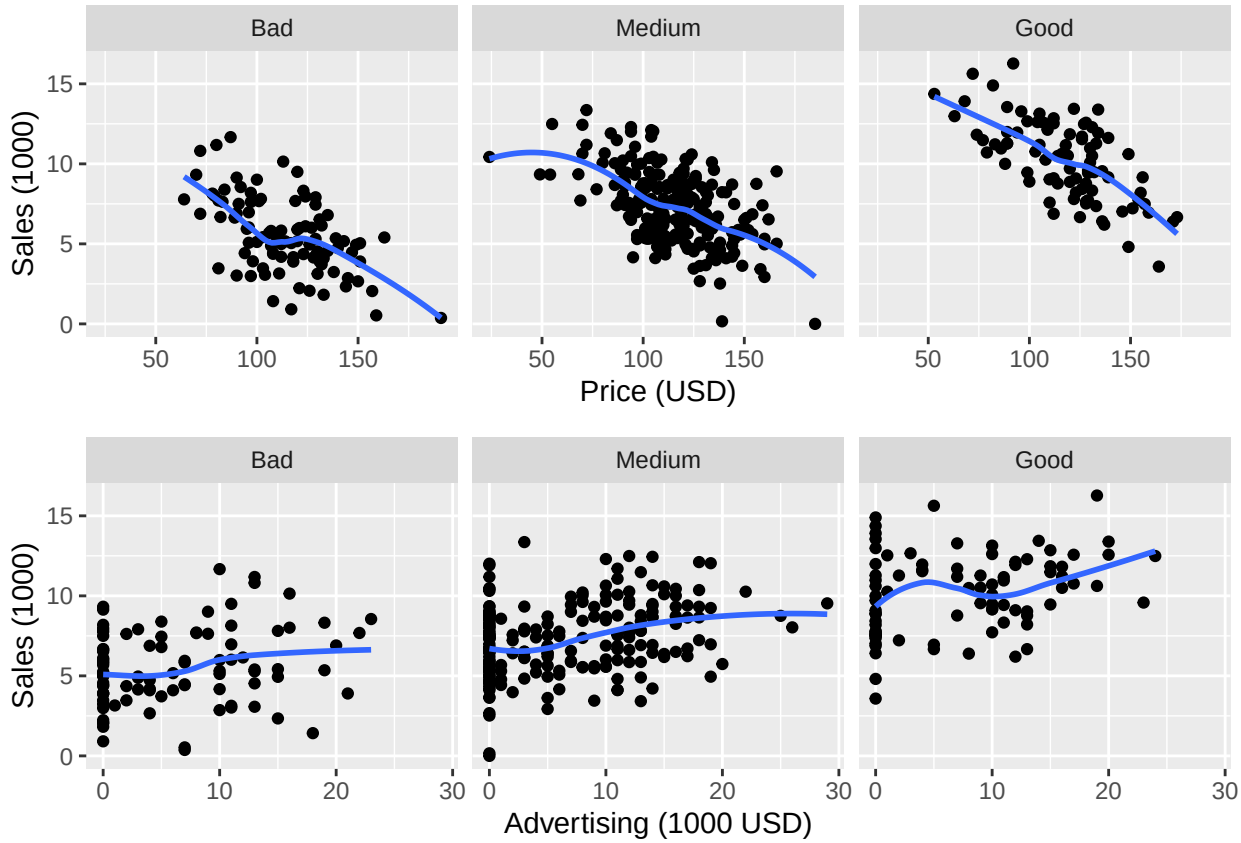
The influence of ShelveLoc, that is shelve location, is clear from both graph. Whereas we see that Price in comparison with Advertising has greater impact on sales.

These are first impressions, but lets dive in the research itself.

Analysis

First, we would like to see if there is interaction between pairs of predictor variables: - ShelveLoc and Price - ShelveLoc and Advertising

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Using smoother we discover that it might be reasonable to take into account interaction between advertising and shelveLoc when creating linear model.

Further we use statistical test to get correct answer.

We create one model without interaction and one model with interaction and check ANOVA. It turns out the interaction is unsignificant.

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7542 -1.1455 -0.0064  1.1768  4.1628
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.468018   0.470930  24.352  < 2e-16 ***
## Price           -0.057975   0.003764 -15.404  < 2e-16 ***
## Advertising      0.109305   0.013405   8.154 4.72e-15 ***
## ShelveLocMedium  1.828803   0.217492   8.409 7.64e-16 ***
## ShelveLocGood    4.776488   0.265261  18.007  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.776 on 395 degrees of freedom
```

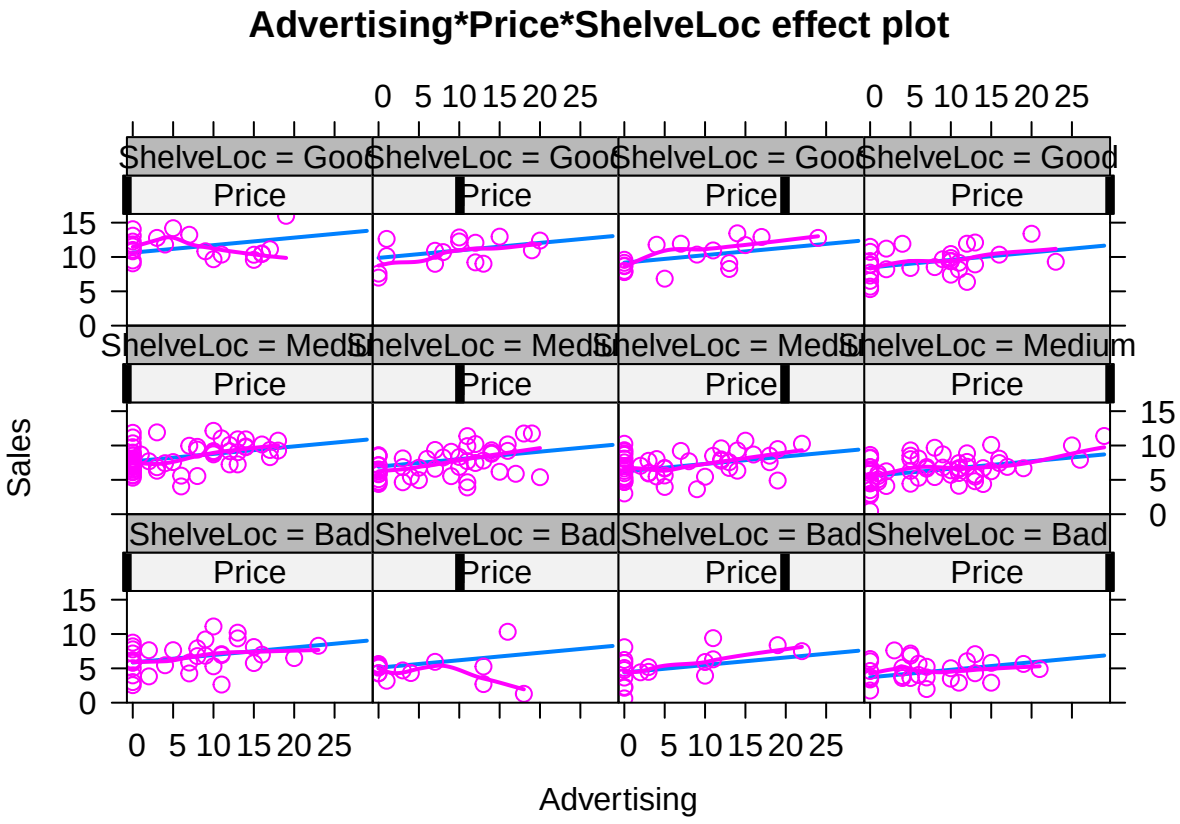
```
## Multiple R-squared:  0.6085, Adjusted R-squared:  0.6045
## F-statistic: 153.5 on 4 and 395 DF,  p-value: < 2.2e-16
```

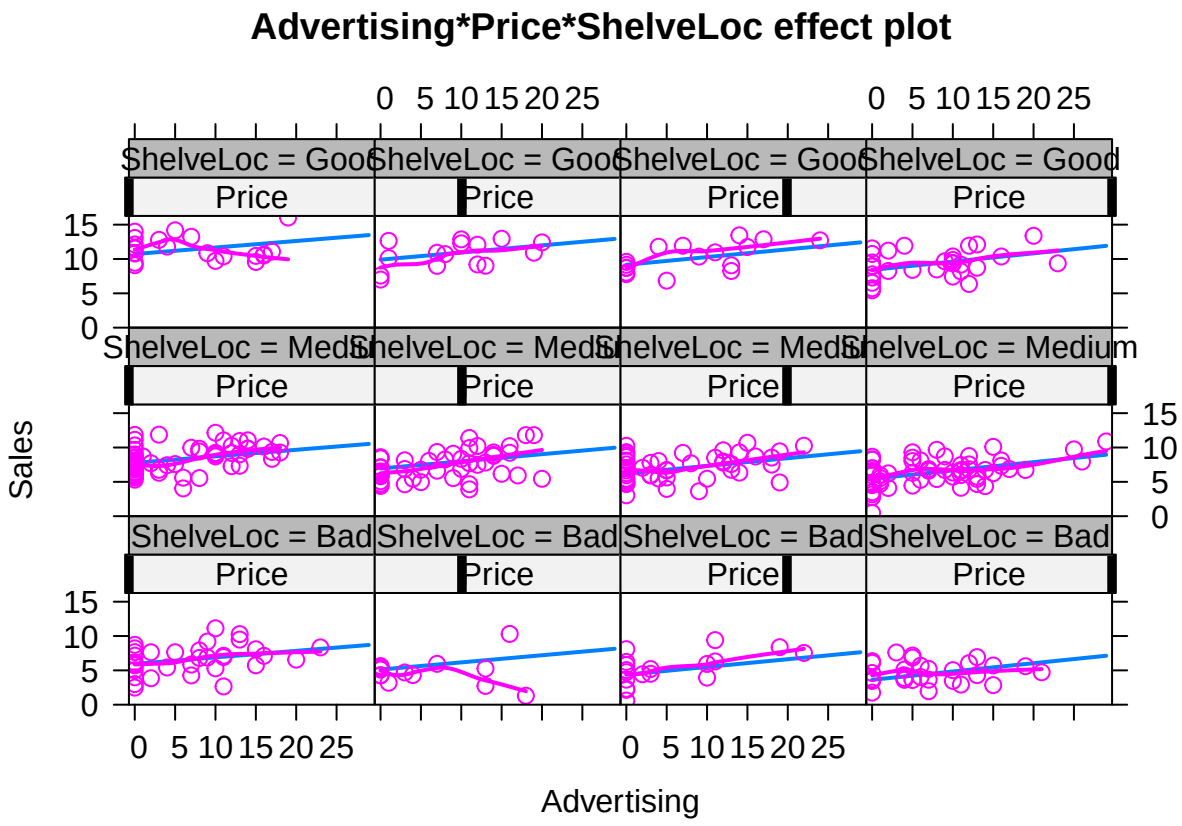
```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + ShelveLoc + Price:ShelveLoc +
##     Advertising:ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5120 -1.0930 -0.0013  1.1684  4.2863
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 11.3779153  0.9076968  12.535  < 2e-16 ***
## Price                       -0.0561211  0.0076860  -7.302 1.60e-12 ***
## Advertising                  0.0897301  0.0282785   3.173  0.00163 **
## ShelveLocMedium              1.5880017  1.0992568   1.445  0.14937
## ShelveLocGood                5.7900588  1.3213434   4.382 1.51e-05 ***
## Price:ShelveLocMedium        0.0003524  0.0092986   0.038  0.96979
## Price:ShelveLocGood         -0.0089812  0.0109023  -0.824  0.41056
## Advertising:ShelveLocMedium  0.0311600  0.0335730   0.928  0.35392
## Advertising:ShelveLocGood    0.0082494  0.0401925   0.205  0.83749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.78 on 391 degrees of freedom
## Multiple R-squared:  0.6106, Adjusted R-squared:  0.6027
## F-statistic: 76.64 on 8 and 391 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Advertising + ShelveLoc
## Model 2: Sales ~ Price + Advertising + ShelveLoc + Price:ShelveLoc + Advertising:ShelveLoc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    395 1245.9
## 2    391 1239.1  4     6.796 0.5361 0.7093
```

Therefore we continue with model.seat.

At this point we check whether there is interaction between Price and Advertising


Advertising*Price*ShelveLoc effect plot
ShelveLoc = Good
ShelveLoc = Medium
ShelveLoc = Bad
Price
Sales
Advertising
0 5 10 15 20 25
0 5 10 15



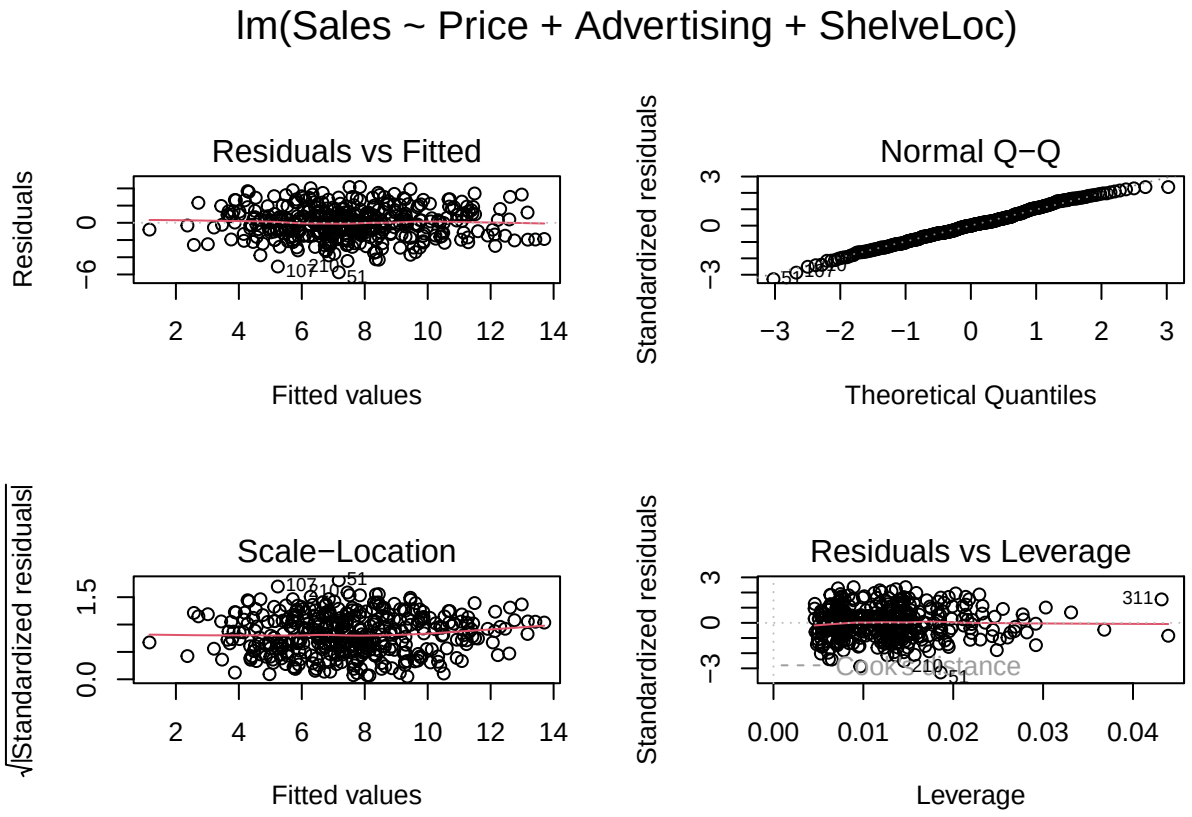

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Advertising + ShelveLoc
## Model 2: Sales ~ Price * Advertising + ShelveLoc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    395 1245.9
## 2    394 1241.3  1    4.6247 1.4679 0.2264
```

We see that interaction between numeric predictors is as well insignificant.

Model diagnosis model.seat

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7542 -1.1455 -0.0064  1.1768  4.1628
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.468018   0.470930  24.352  < 2e-16 ***
## Price           -0.057975   0.003764 -15.404  < 2e-16 ***
## Advertising      0.109305   0.013405   8.154 4.72e-15 ***
```

```
## ShelveLocMedium  1.828803   0.217492   8.409 7.64e-16 ***
## ShelveLocGood    4.776488   0.265261  18.007  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.776 on 395 degrees of freedom
## Multiple R-squared:  0.6085, Adjusted R-squared:  0.6045
## F-statistic: 153.5 on 4 and 395 DF,  p-value: < 2.2e-16
```

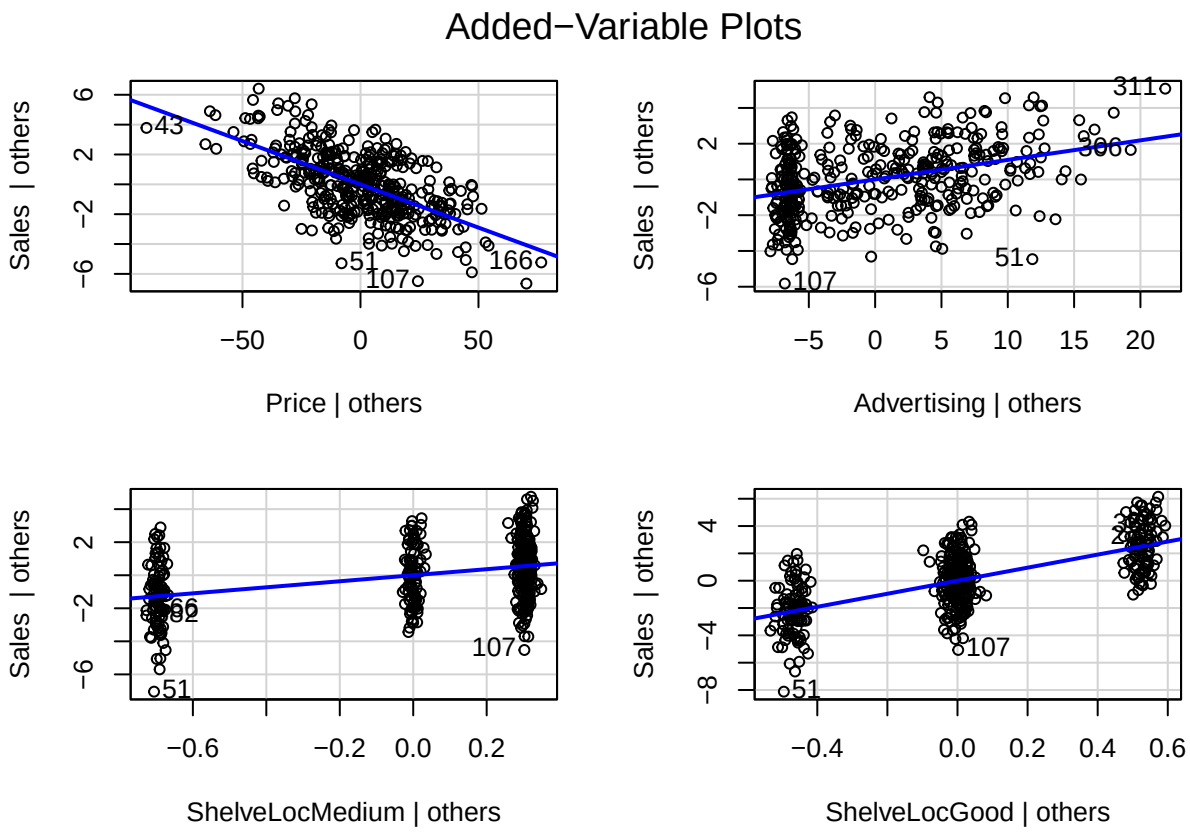



```
##                     2.5 %      97.5 %
## (Intercept)     10.54217589 12.39386048
## Price           -0.06537400 -0.05057587
## Advertising      0.08295169  0.13565903
## ShelveLocMedium  1.40121654  2.25639000
## ShelveLocGood    4.25498762  5.29798794
```

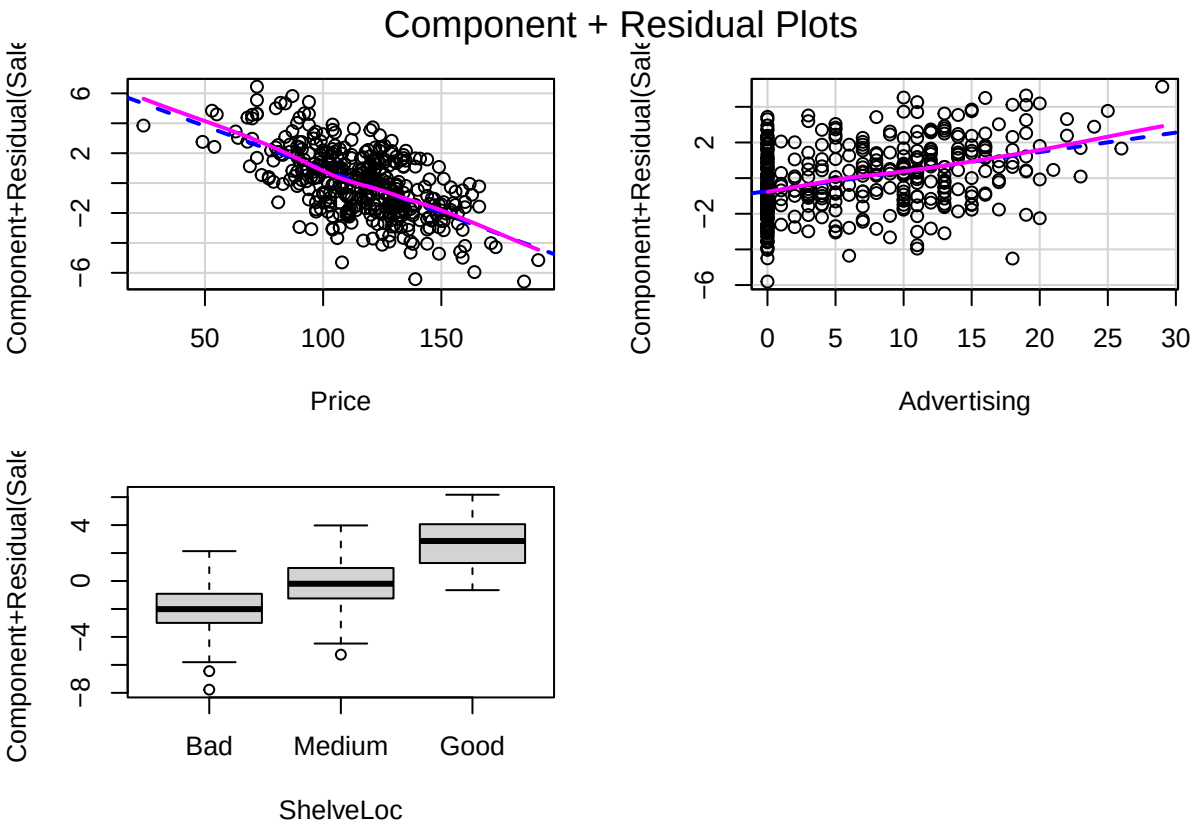
The coefficient of determination R2 has a value of 0.609.

The model explains 60.9 % of variability of Sales.

The expected Sales is 10.54. 1 m/s increase in wind speed results in almost 8.5 hundreths decrease in 110 metres hurdles run result.



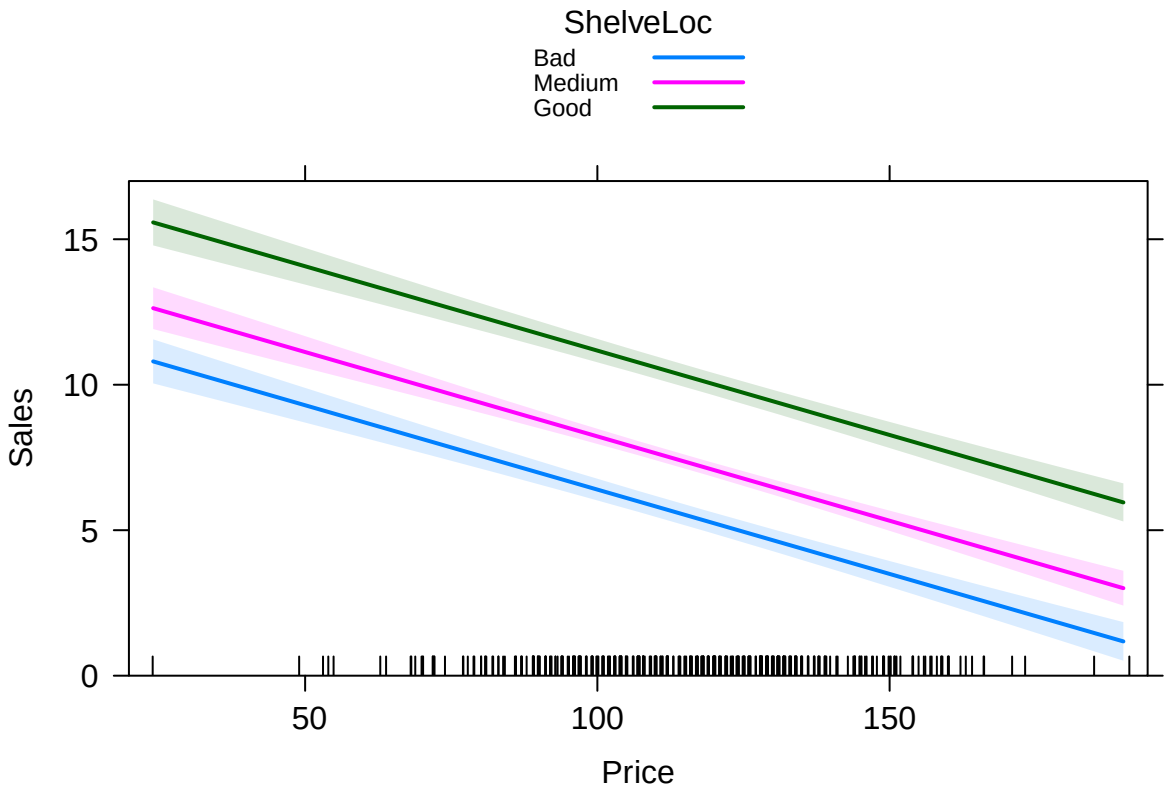
Price has the most significant impact on Sales, which represents the slope of the line.

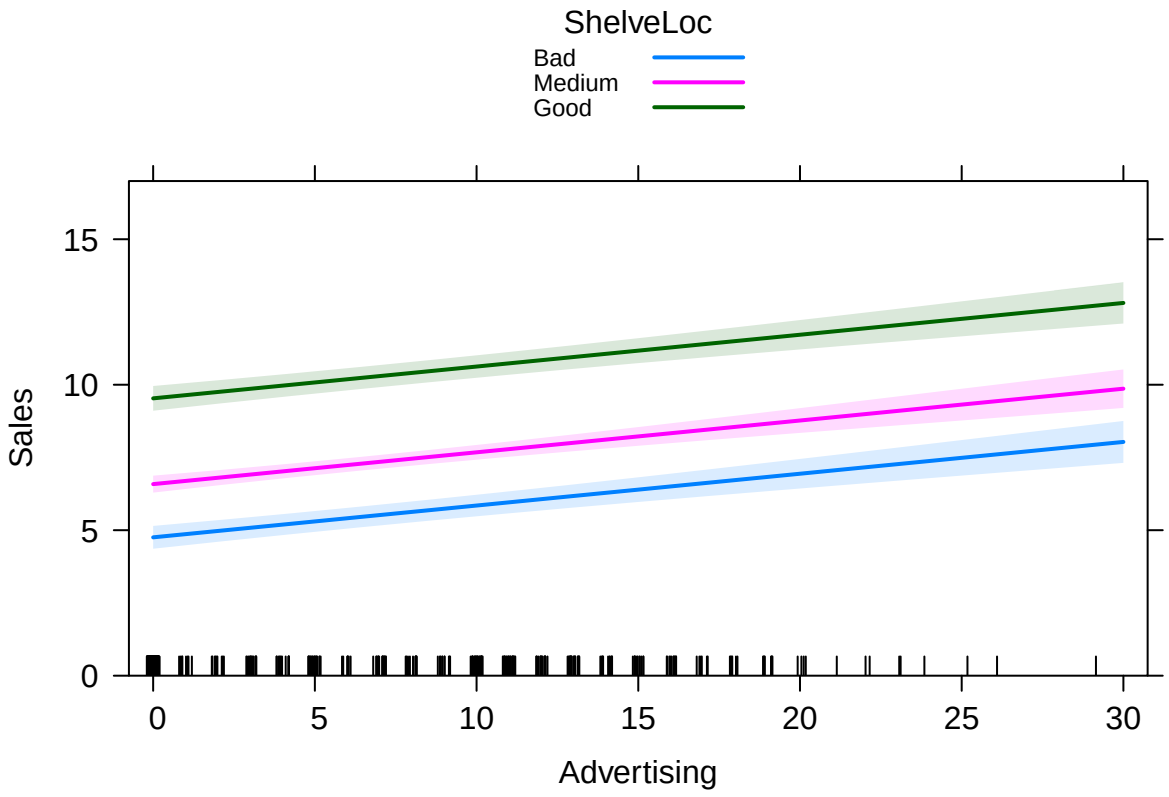



```
## Analysis of Variance Table
##
## Response: Sales
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Price         1  630.03  630.03 199.743 < 2.2e-16 ***
## Advertising   1  266.91  266.91  84.621 < 2.2e-16 ***
## ShelveLoc     2 1039.42  519.71 164.767 < 2.2e-16 ***
## Residuals   395 1245.91    3.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results

For "official results" to make sure that confidence intervals are correct, though we are testing multiple hypothesis, we use correction with the use of multcomp glth


ShelveLoc
Bad
Medium
Good
Sales
15
10
5
0
50
100
150
Price

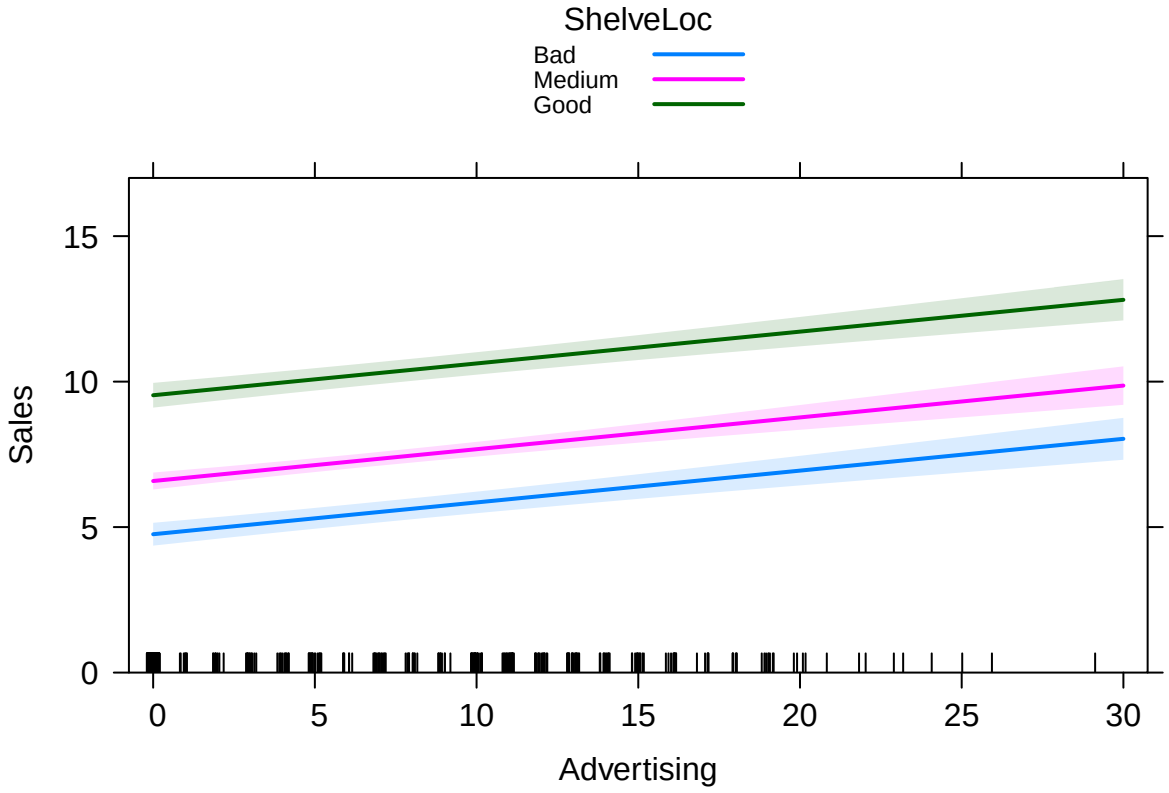


The upper graphics represents results assuming an average value of advertising.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)
##
## Linear Hypotheses:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0     11.468018   0.470930  24.352   <1e-10 ***
## Price == 0           -0.057975   0.003764 -15.404   <1e-10 ***
## Advertising == 0      0.109305   0.013405   8.154   <1e-10 ***
## ShelveLocMedium == 0  1.828803   0.217492   8.409   <1e-10 ***
## ShelveLocGood == 0    4.776488   0.265261  18.007   <1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)
##
## Quantile = 2.5267
## 95% family-wise confidence level
##
```

```
##
## Linear Hypotheses:
##                       Estimate lwr      upr
## (Intercept) == 0     11.46802 10.27812 12.65792
## Price == 0           -0.05797 -0.06748 -0.04847
## Advertising == 0      0.10931  0.07544  0.14318
## ShelveLocMedium == 0  1.82880  1.27927  2.37834
## ShelveLocGood == 0    4.77649  4.10625  5.44672
```



The upper graphics represents results assuming an average price (reduction).

Remark

The case study was done and has been presented in the course Linear Models by Damjana Kastelec UNI LJ 22/23