Class Project: Multiple Linear Regression on Market Share data

Ashwin Rathore

07/09/2021

Importing data from market_share.xlsx file and creating a dataframe.

```
library("readxl")
ms <- read_excel("market_share.xlsx")
ms =within(ms, rm(idnum))
ms$month <- as.numeric(as.factor(ms$month))
ms
```

```
## # A tibble: 36 x 7
##    marketshare price gnrpoints discount promotion month  year
##          <dbl> <dbl>     <dbl>    <dbl>     <dbl> <dbl> <dbl>
##  1        3.15  2.20       498        1         1    12  1999
##  2        2.52  2.19       510        0         0    11  1999
##  3        2.64  2.29       422        1         1    10  1999
##  4        2.55  2.42       858        0         1     3  1999
##  5        2.69  2.18       566        1         0     5  2000
##  6        2.38  2.21       536        0         0     4  2000
##  7        3.02  2.13       585        1         1     8  2000
##  8        2.52  2.21       310        1         0     1  2000
##  9        2.45  2.31       211        0         0     9  2000
## 10        2.42  2.26       504        0         1     7  2000
## # ... with 26 more rows
```

Applying linear regression between market share and the rest of the covariates one by one to find the relationship.

```
lm.price <- lm(ms$marketshare ~ ms$price, data = ms)
summ.price <- summary(lm.price)
summ.price
```

```
##
## Call:
## lm(formula = ms$marketshare ~ ms$price, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3976 -0.2063 -0.0463  0.2237  0.4596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3746     0.6365   5.302 6.97e-06 ***
## ms$price     -0.3058     0.2732  -1.119    0.271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2634 on 34 degrees of freedom
## Multiple R-squared:  0.03553,    Adjusted R-squared:  0.007168
## F-statistic: 1.253 on 1 and 34 DF,  p-value: 0.2709
```

```
lm.gnrpoints <- lm(ms$marketshare ~ ms$gnrpoints, data = ms)
summ.gnrpoints <- summary(lm.gnrpoints)
summ.gnrpoints
```

```
##
## Call:
## lm(formula = ms$marketshare ~ ms$gnrpoints, data = ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44903 -0.19130 -0.02349  0.21748  0.51365
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6197063  0.1132896  23.124   <2e-16 ***
## ms$gnrpoints 0.0001139  0.0002684   0.424    0.674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2675 on 34 degrees of freedom
## Multiple R-squared:  0.005266,   Adjusted R-squared:  -0.02399
## F-statistic:  0.18 on 1 and 34 DF,  p-value: 0.6741
```

```
lm.discount <- lm(ms$marketshare ~ ms$discount, data = ms)
summ.discount <- summary(lm.discount)
summ.discount
```

```
##
## Call:
## lm(formula = ms$marketshare ~ ms$discount, data = ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31810 -0.12702  0.02095  0.10893  0.32190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.42000    0.04240  57.080  < 2e-16 ***
## ms$discount  0.41810    0.05551   7.532 9.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1642 on 34 degrees of freedom
## Multiple R-squared:  0.6253, Adjusted R-squared:  0.6142
## F-statistic: 56.73 on 1 and 34 DF,  p-value: 9.584e-09
```

```
lm.promotion <- lm(ms$marketshare ~ ms$promotion, data = ms)
summ.promotion <- summary(lm.promotion)
summ.promotion
```

```
##
```

```
## Call:
## lm(formula = ms$marketshare ~ ms$promotion, data = ms)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.465 -0.205 -0.035  0.160  0.425
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.57500    0.06386  40.321   <2e-16 ***
## ms$promotion  0.16000    0.08568   1.867   0.0705 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2555 on 34 degrees of freedom
## Multiple R-squared:  0.09302,    Adjusted R-squared:  0.06635
## F-statistic: 3.487 on 1 and 34 DF,  p-value: 0.07049
```

```
lm.month <- lm(ms$marketshare ~ ms$month, data = ms)
summ.month <- summary(lm.month)
summ.month
```

```
##
## Call:
## lm(formula = ms$marketshare ~ ms$month, data = ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46171 -0.22318 -0.02402  0.22541  0.50009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.612222   0.094783  27.560   <2e-16 ***
## ms$month    0.007949   0.012878   0.617    0.541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2667 on 34 degrees of freedom
## Multiple R-squared:  0.01108,    Adjusted R-squared:  -0.01801
## F-statistic: 0.3809 on 1 and 34 DF,  p-value: 0.5412
```

```
lm.year <- lm(ms$marketshare ~ ms$year, data = ms)
summ.year <- summary(lm.year)
summ.year
```

```
##
## Call:
## lm(formula = ms$marketshare ~ ms$year, data = ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42906 -0.19018 -0.03354  0.22456  0.48646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.63188   94.73597   0.334    0.741
```

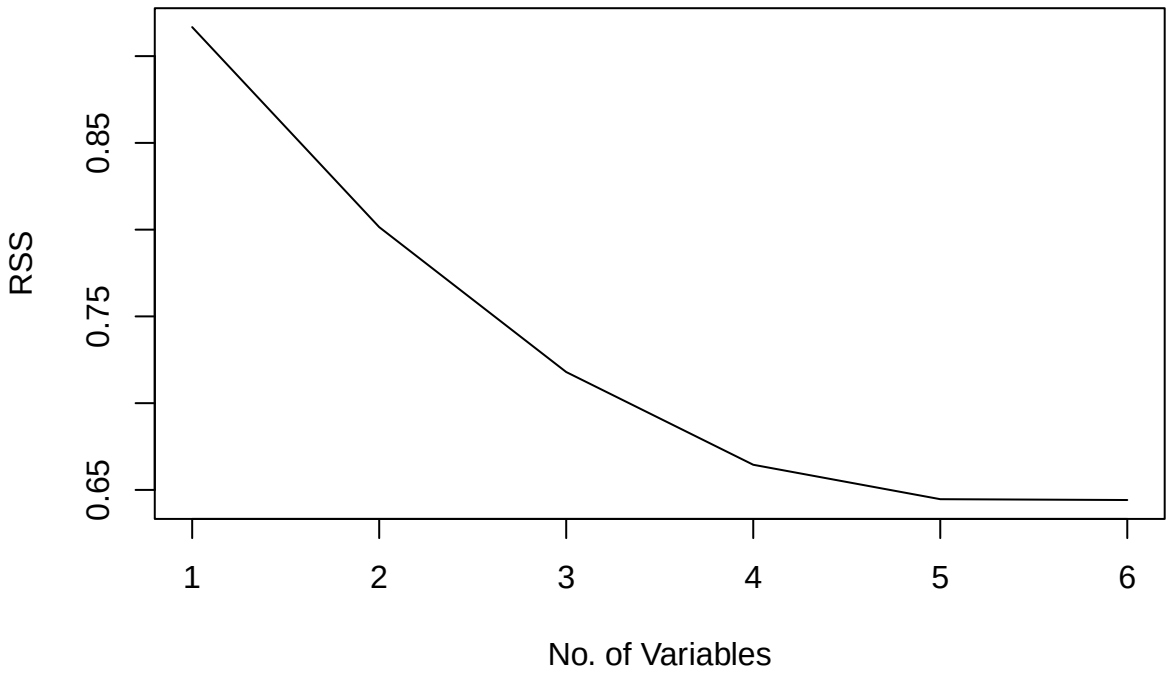
```
## ms$year     -0.01448    0.04735  -0.306    0.762
##
## Residual standard error: 0.2679 on 34 degrees of freedom
## Multiple R-squared:  0.002742,   Adjusted R-squared:  -0.02659
## F-statistic: 0.0935 on 1 and 34 DF,  p-value: 0.7616
```

Building the "best" regression model for Y

```
library(leaps)
b <- regsubsets(ms$marketshare ~., data = ms, nvmax=6)
rs <- summary(b)
rs$rsq
```

```
## [1] 0.6252543 0.6723710 0.7065091 0.7283602 0.7364704 0.7366661
```

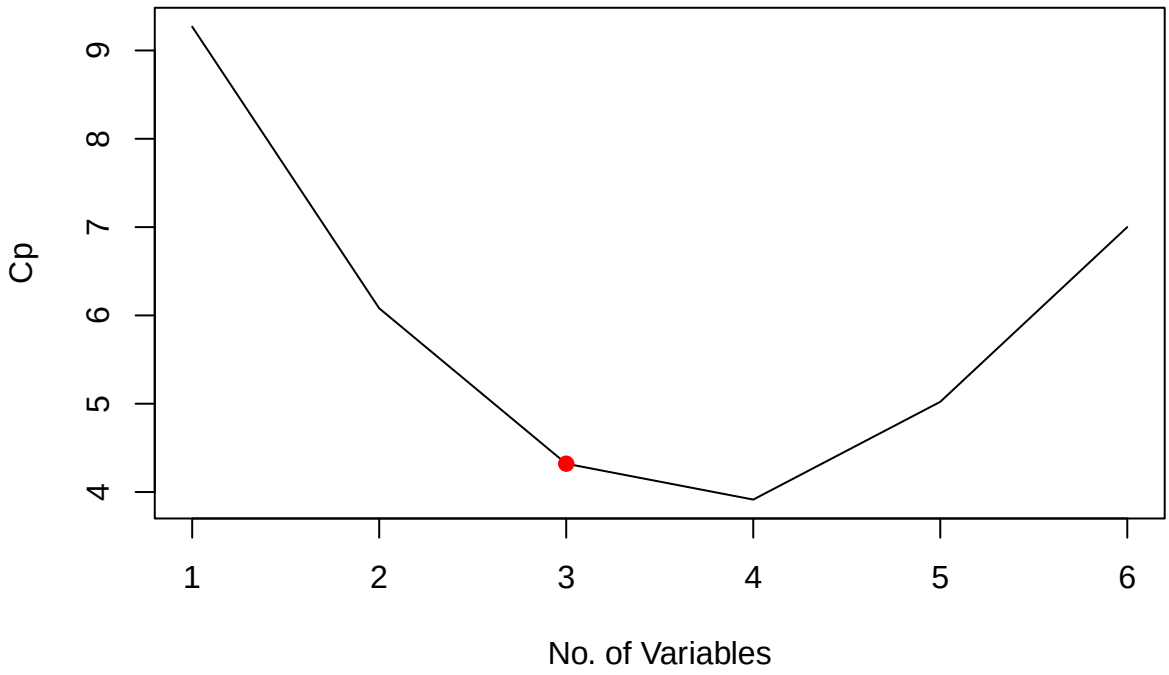
```
plot(rs$rss,xlab='No. of Variables',ylab='RSS',type='l')
```



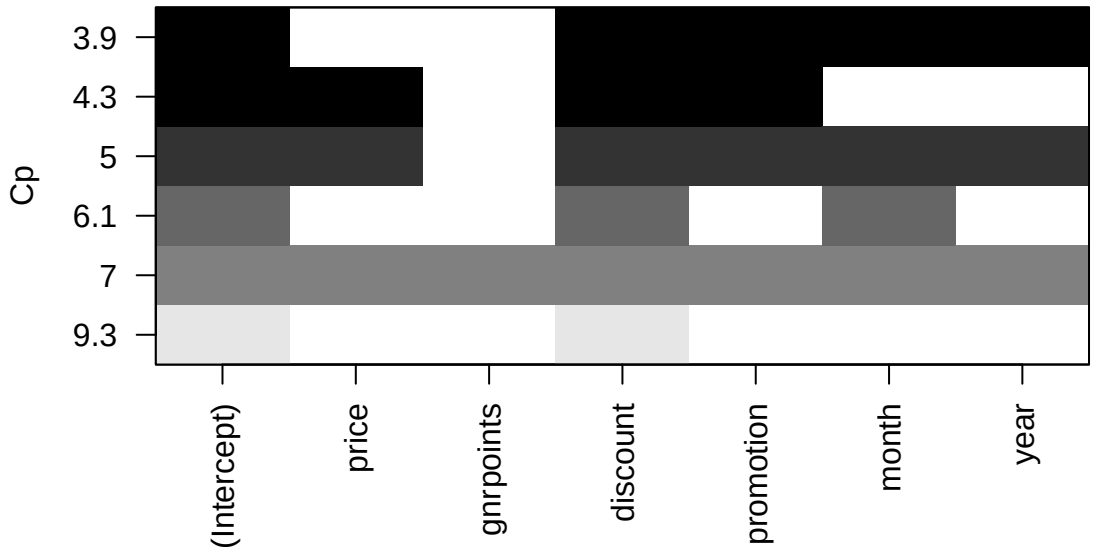

```
plot(rs$cp,xlab='No. of Variables',ylab='Cp',type='l')
which.min(rs$cp)
```

```
## [1] 4
```

```
points(3,rs$cp[3],pch=19,col='red')
```




```
plot(b,scale='Cp')
```



The best model according to both R^2 and Cp is the one that uses price, discount, promotion, month and year. We can remove teh gnrpoints as it does not relate much with the market share.

Fitting the model.

```
ms_data <- data.frame(ms$price, ms$discount, ms$promotion, ms$month, ms$year)
library(VIF)
vif(ms_data)
```

```
## Warning in mean.default(y): argument is not numeric or logical: returning NA
```

```
## [1] "m should be less than or equal to n"
```

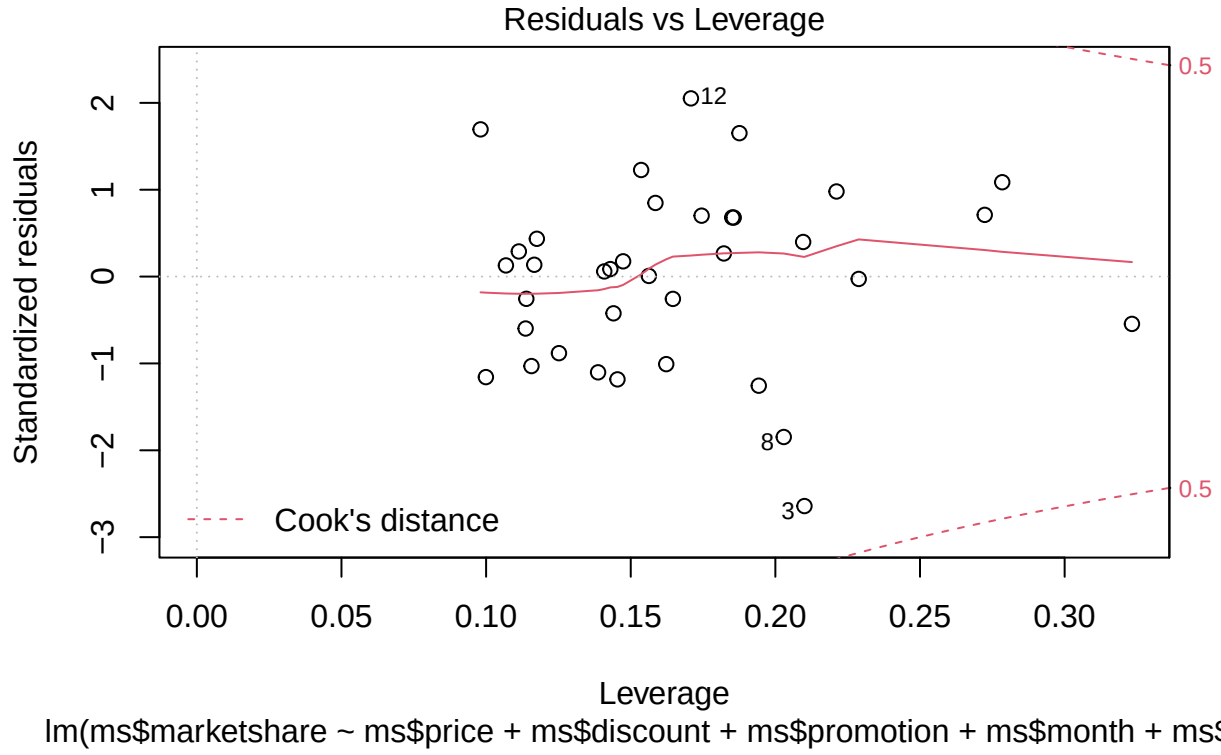
```
## [1] 0
```

A commonly used practice is if a VIF is > 10, you have high multicollinearity. In our case, with values around

0, we are in good shape, and can proceed with our regression.

– Outliers and influential points

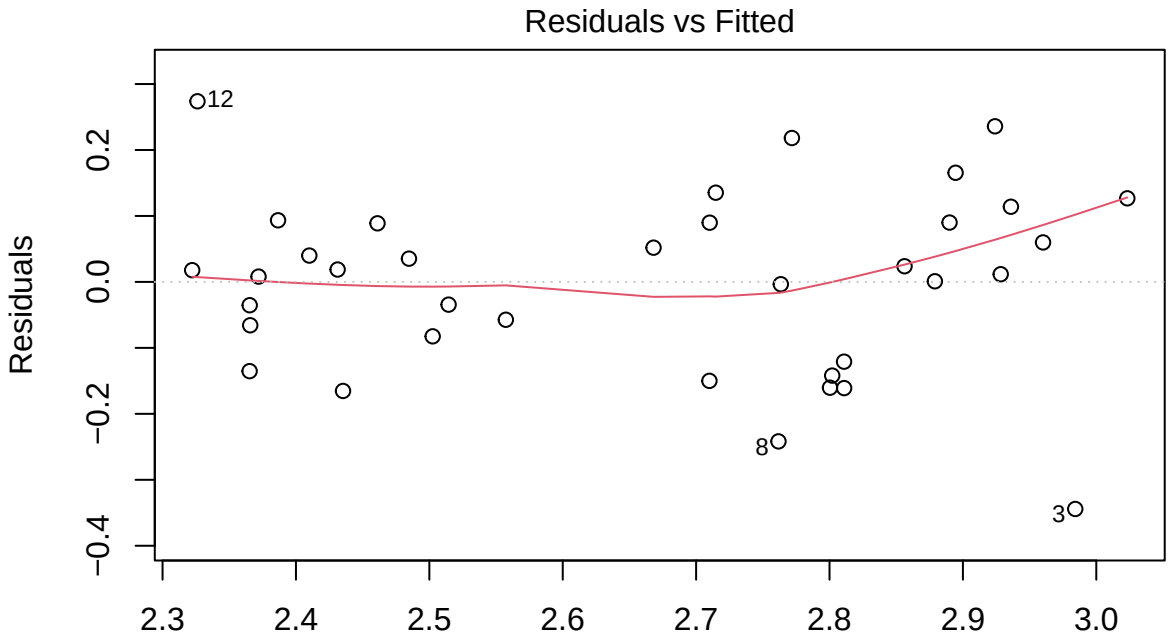
```
lmms <- lm(ms$marketshare ~ ms$price+ ms$discount+ ms$promotion+ ms$month+ ms$year, data = ms)
plot(lmms, 5)
```



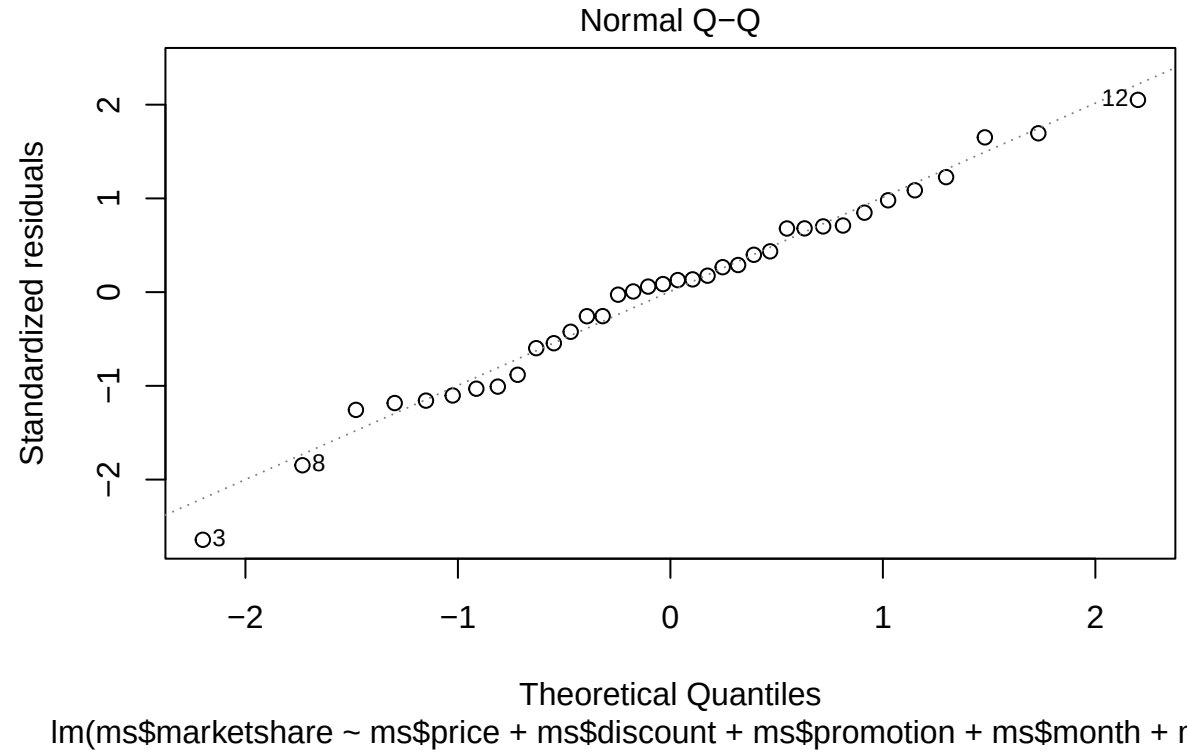
In the Residuals vs Leverage plot, no influential points are outside the Cook's distance lines (a red dashed line). Thus we can assume there are no outliers in the data.

– Appropriateness of predictors (i.e., is any transformation of predictors necessary?)

```
plot(lmms, 1)
```

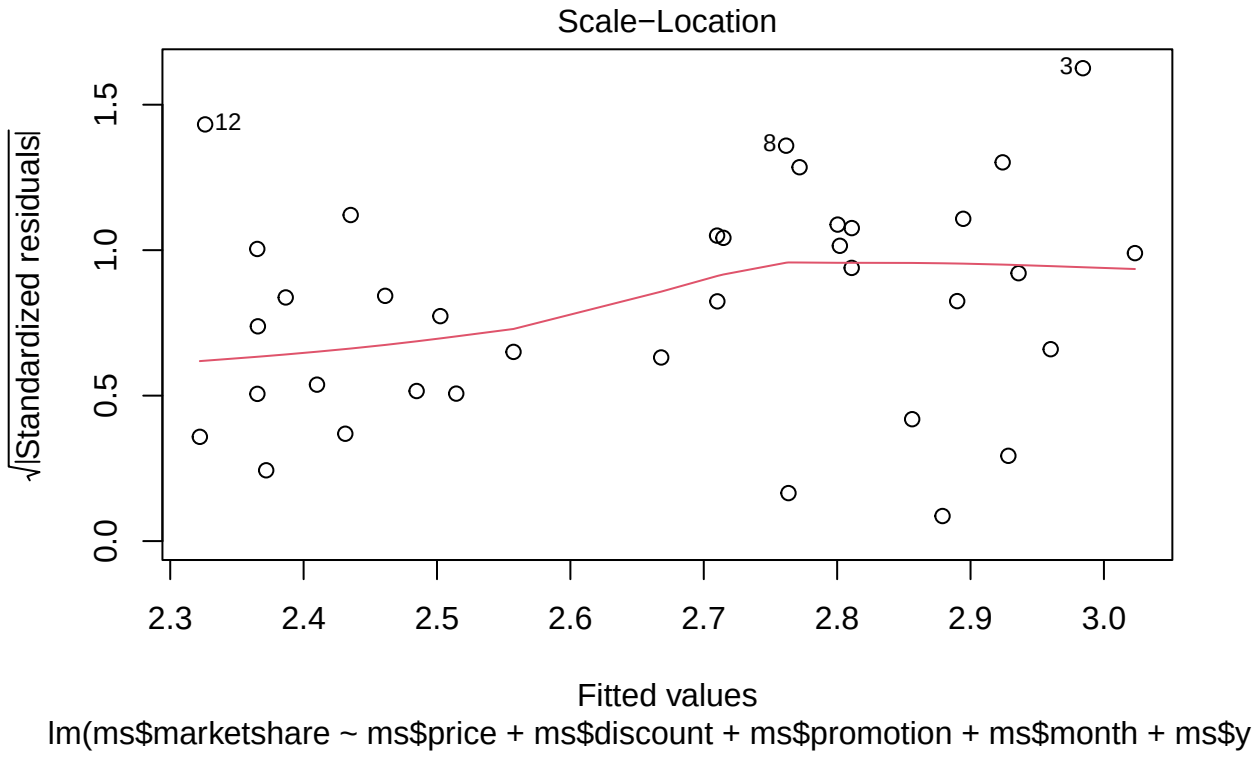



```
plot(lmms, 2)
```



– Constant variance of residuals

```
plot(lmms, 3)
```


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
12
8
3
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
Fitted values
lm(ms$marketshare ~ ms$price + ms$discount + ms$promotion + ms$month + ms$y