APAN4335 — Homework 2

Ronak Shah

Feb 8th, 2018

---

1 Bias-Variance Tradeoff

Identifications for the nine labels of the provided figure include …

1.1 Identify lables

- A. Test MSE curve
- B. Variance curve
- C. Squared Bias curve
- D. The values corresponding to smallest test MSE
- E. Axis (components of MSE)
- F. Expected test MSE
- G. Squared bias of estimator function
- H. Variance of estimator function
- I. Variance of error terms

1.2

False: Because MSE also includes variance of error terms

1.3

True: Because it is squared

1.4

False: Variance increase with flexibility, as the model tends to over-fit

1.5

True: More flexible models tend to reduce the bias

1.6

False: there is insufficient evidence that more data should reduce the bias
Reference:
https://www.researchgate.net/publication/2456576_On_the_Effect_of_Data_Set_Size_on_Bias_and_Variance_in_Classification_Learning
(https://www.researchgate.net/publication/2456576_On_the_Effect_of_Data_Set_Size_on_Bias_and_Variance_in_Classification_Learning)

1.7

True: Variance can decrease when training set size increases

---

2 K Nearest Neighbors

First, let's load the data:

```
x     <- c(1, 1, 2, 2, 2, 6, 6)
y     <- c(4, 2, 6, 5, 1, 5, 1)
color <- c('black', 'red', 'red', 'blue', 'blue', 'red', 'red')
```

2.1 Euclidean distance

Then we use dist function to get a matrix of distances between every point

```
z <- as.matrix(dist(cbind(x,y), method = "euclidean")) # the function dist returns euclidean distance of each pair
 of points
z
```

```
##          1        2        3        4        5        6        7
## 1 0.000000 2.000000 2.236068 1.414214 3.162278 5.099020 5.830952
## 2 2.000000 0.000000 4.123106 3.162278 1.414214 5.830952 5.099020
## 3 2.236068 4.123106 0.000000 1.000000 5.000000 4.123106 6.403124
## 4 1.414214 3.162278 1.000000 0.000000 4.000000 4.000000 5.656854
## 5 3.162278 1.414214 5.000000 4.000000 0.000000 5.656854 4.000000
## 6 5.099020 5.830952 4.123106 4.000000 5.656854 0.000000 4.000000
## 7 5.830952 5.099020 6.403124 5.656854 4.000000 4.000000 0.000000
```

The first column on the matrix corresponds to the distances of black point with each of the other points.
This can be extracted using following

```
d <- (z[-1,1])
```

Next, we reorder the points in ascending order of distance and get the sorted vector of color

```
d.order <- order(d)
d.order
```

```
## [1] 3 1 2 4 5 6
```

```
color[-1][d.order]
```

```
## [1] "blue" "red"  "red"  "blue" "red"  "red"
```

KNN when K=1 will be Blue, The first point in the ordered list (the color corresponding to point number 3)
KNN when K=3 will be Red. (by 2/3 ratio) The first three points on the ordered list OR The maximum of the colors corresponding to point number 3,1,2 i.e. blue,red,red.

2.2 Inverse Square and Gaussian distance

Note, since K=6, we use all the available data points

First we create a numeric vector of inverse squared distances. Then, get sum of the weights grouped by color

```
d.invSq <- 1/(d)^2
aggregate(x = d.invSq, by = list(color[-1]), sum)
```

```
##   Group.1         x
## 1    blue 0.6000000
## 2     red 0.5178733
```

As we can see, Blue wins as it has higher sum

Repeat the same for Gaussian distance

```
d.gaus1 <- exp(-0.2*d^2)
d.gaus2 <- exp(-0.4*d^2)
aggregate(x = d.gaus1, by = list(color[-1]), sum)
```

```
##   Group.1         x
## 1    blue 0.8056553
## 2     red 0.8238387
```

```
aggregate(x = d.gaus2, by = list(color[-1]), sum)
```

```
##   Group.1         x
## 1    blue 0.4676446
## 2     red 0.3372635
```

For Alpha = -0.2, Red wins
For Alpha = -0.4, Blue wins
We observe that higher negative values of alpha gives preference to proximity or penalizes more for distance

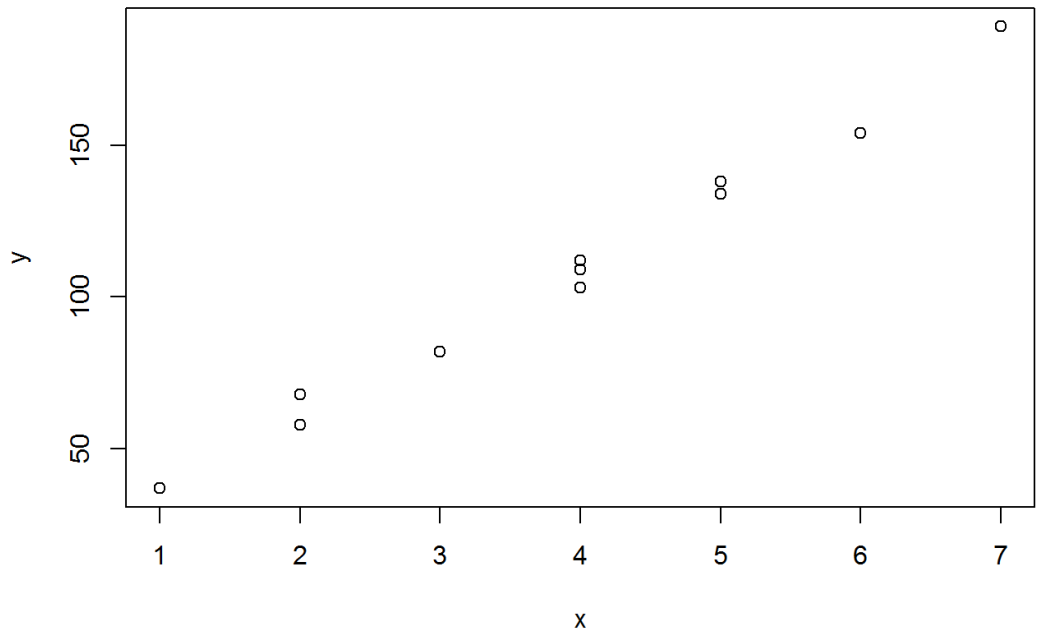
---

3 Simple Linear Regression

```
data <- read.csv("compsys.csv")
str(data)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ x: int  4 2 5 7 1 3 4 5 2 4 ...
##  $ y: int  109 58 138 189 37 82 103 134 68 112 ...
```

```
plot(data)
```




```
lm.model <- lm(y~x, data = data)
```

3.1 Estimate the coefficients

```
summary(lm.model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8729 -2.9696 -0.4751  2.8260  7.3315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4641     3.4390   3.334  0.00875 **
## x            24.6022     0.8045  30.580 2.09e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.615 on 9 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9894
## F-statistic: 935.1 on 1 and 9 DF,  p-value: 2.094e-10
```

As shown above, intercept (β_0) is 11.464 and β_1 is 24.602.
It can be interpreted that a basic call, without any microcomputer will take 11.464 minutes
And each additional single microcomputer adds 24.602 minutes to the call

3.2 Confidence Interval for β_1

```
confint(lm.model)
```

```
##                 2.5 %   97.5 %
## (Intercept)  3.684472 19.24371
## x           22.782272 26.42215
```

We are 95% confident that β_1 lies between 22.78 and 26.42 minutes

3.3 Call time for 6 microcomputers

```
predict(object=lm.model, newdata = data.frame(x = c(6)), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 159.0773 154.1388 164.0159
```

We are 95% confident that on average the time to serve 6 microcomputers will be between 154.1 to 164.01 minutes

3.4 Prediction time for 6 microcomputers

```
predict(object=lm.model, newdata = data.frame(x = c(6)), interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 159.0773 147.5279 170.6268
```

We are 95% sure that he prediction interval for 6 microcomputers lies between 147.5 and 170.6 minutes

3.5 Hypothesis test

```
summary(lm.model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8729 -2.9696 -0.4751  2.8260  7.3315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4641     3.4390   3.334  0.00875 **
## x            24.6022     0.8045  30.580 2.09e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.615 on 9 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9894
## F-statistic: 935.1 on 1 and 9 DF,  p-value: 2.094e-10
```

```
t = (24.6022-0)/0.8045 # Note that t statisitc for testing against mean = 0 is (b-0)/SE
qt(c(.025, .975), df=9) # df = n-2 = 11-2 = 9
```

```
## [1] -2.262157  2.262157
```

```
t
```

```
## [1] 30.58073
```

Since the value of t is much higher than the 95% confidence range, we can safely reject the Null hypothesis

---

4 Multiple Linear Regression

4.1 Load the data and view summary

```
cigsales <- read.table("cig_sales.txt", stringsAsFactors = F, header = T)
str(cigsales)
```

```
## 'data.frame':    51 obs. of  8 variables:
##  $ State : chr  "AL" "AK" "AZ" "AR" ...
##  $ Age   : num  27 22.9 26.3 29.1 28.1 26.2 29.1 26.8 28.4 32.3 ...
##  $ HS    : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 55.2 52.6 ...
##  $ Income: num  2948 4644 3665 2878 4493 ...
##  $ Black : num  26.2 3 3 18.3 7 3 6 14.3 71.1 15.3 ...
##  $ Female: num  51.7 45.7 50.8 51.5 50.8 50.7 51.5 51.3 53.5 51.8 ...
##  $ Price : num  42.7 41.8 38.5 38.8 39.7 31.1 45.5 41.3 32.6 43.8 ...
##  $ Sales : num  89.8 121.3 115.2 100.3 123 ...
```

```
summary(cigsales)
```

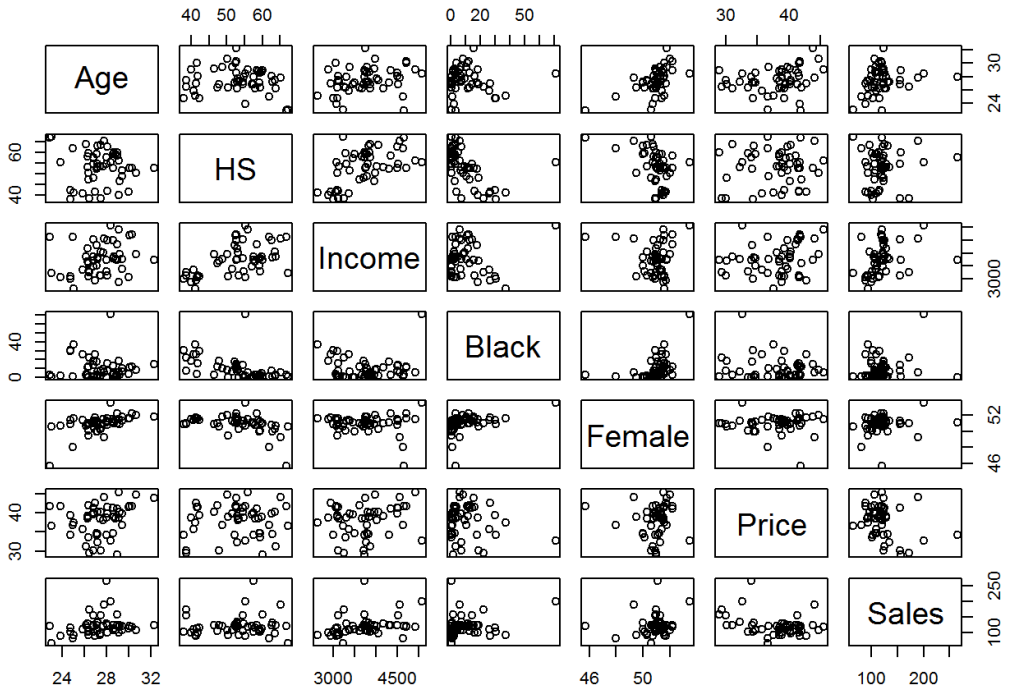
```
##     State                Age              HS            Income
##  Length:51          Min.   :22.90   Min.   :37.80   Min.   :2626
##  Class :character   1st Qu.:26.40   1st Qu.:48.30   1st Qu.:3271
##  Mode  :character   Median :27.40   Median :53.30   Median :3751
##                     Mean   :27.47   Mean   :53.15   Mean   :3764
##                     3rd Qu.:28.75   3rd Qu.:59.10   3rd Qu.:4116
##                     Max.   :32.30   Max.   :67.30   Max.   :5079
##      Black            Female          Price           Sales
##  Min.   : 0.200   Min.   :45.70   Min.   :29.00   Min.   : 65.5
##  1st Qu.: 1.600   1st Qu.:50.75   1st Qu.:34.70   1st Qu.:105.3
##  Median : 6.000   Median :51.10   Median :38.90   Median :119.0
##  Mean   : 9.992   Mean   :50.95   Mean   :38.07   Mean   :121.5
##  3rd Qu.:13.550   3rd Qu.:51.50   3rd Qu.:41.35   3rd Qu.:124.5
##  Max.   :71.100   Max.   :53.50   Max.   :45.50   Max.   :265.7
```

4.2 Identify type of variables

Qualitative variable: state
Quantitative variable: all others

4.3 Matrix of scatter plots

```
plot(cigsales[,-1])
```



Age and Income appear to be loosely positively correlated with Sales
Price appears to be negatively correlated with Sales

4.4 Multiple Linear regression model

```
lm.cigsales <- lm(Sales ~ . -State, data = cigsales)
summary(lm.cigsales)
```

```
##
## Call:
## lm(formula = Sales ~ . - State, data = cigsales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.398 -12.388  -5.367   6.270 133.213
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 103.34485  245.60719   0.421  0.67597
## Age           4.52045    3.21977   1.404  0.16735
## HS           -0.06159    0.81468  -0.076  0.94008
## Income        0.01895    0.01022   1.855  0.07036 .
## Black         0.35754    0.48722   0.734  0.46695
## Female       -1.05286    5.56101  -0.189  0.85071
## Price        -3.25492    1.03141  -3.156  0.00289 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.17 on 44 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.2282
## F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

The low value of R-squared indicates that the model does not sufficiently explain the variance of predicted variable

The low p-value of F-statistic tells us that there is a 0.6% chance that all co-efficients are zero. i.e. we can safely assume that at-least one regression coefficient is non-zero

Out of all predictors, only Price is significant at 0.01 level, and is negatively correlated with Sales (as expected)

4.5 95% confidence interval range for all coefficients

```
confint(lm.cigsales)
```

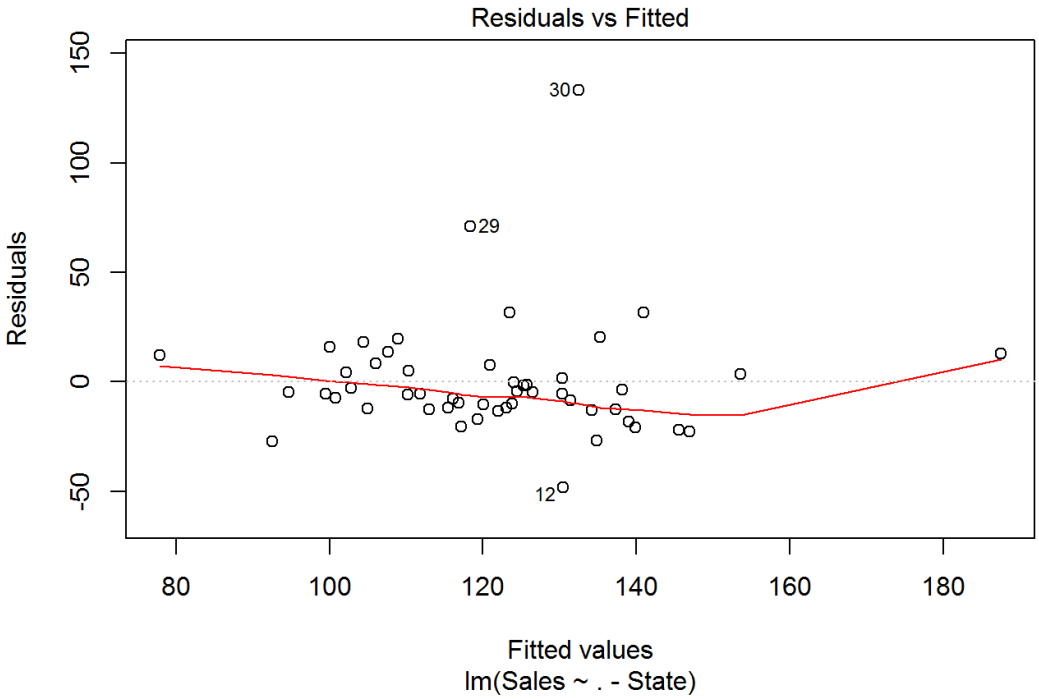
```
##                     2.5 %       97.5 %
## (Intercept) -3.916439e+02 598.33360254
## Age         -1.968565e+00  11.00946945
## HS          -1.703475e+00   1.58030249
## Income      -1.642517e-03   0.03953542
## Black       -6.243909e-01   1.33946122
## Female      -1.226033e+01  10.15461632
## Price       -5.333583e+00  -1.17625412
```

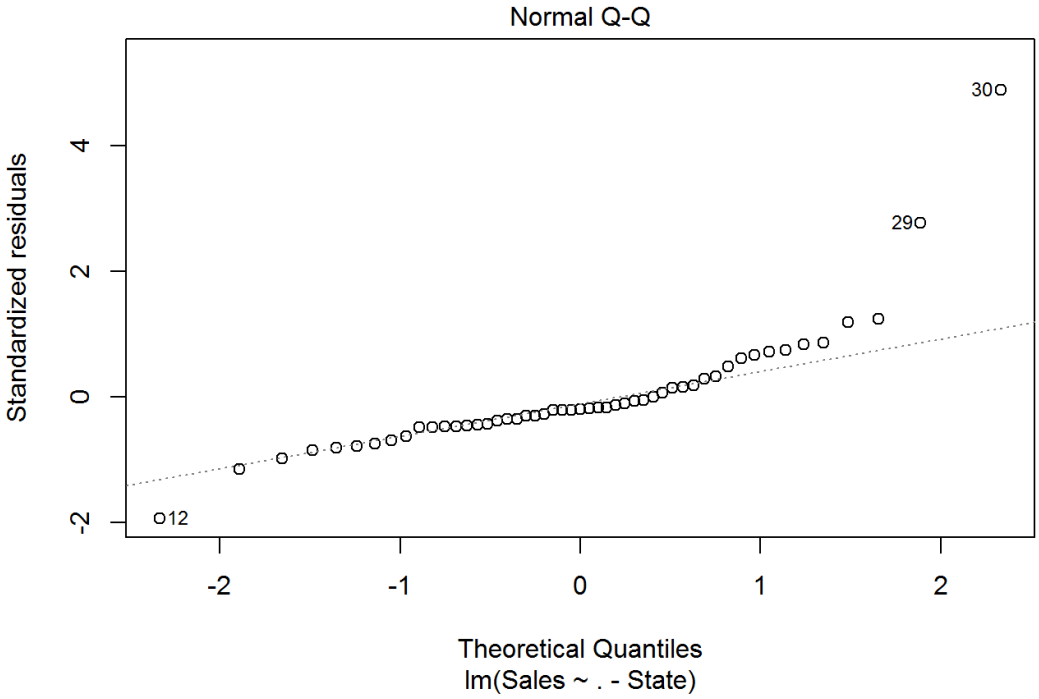
As seen from the above 95% confidence interval range, we fail to reject the null hypothesis for Female parameter. So it cannot be conclusively predicted whether the Sales will increase or reduce by including more Female candidates.

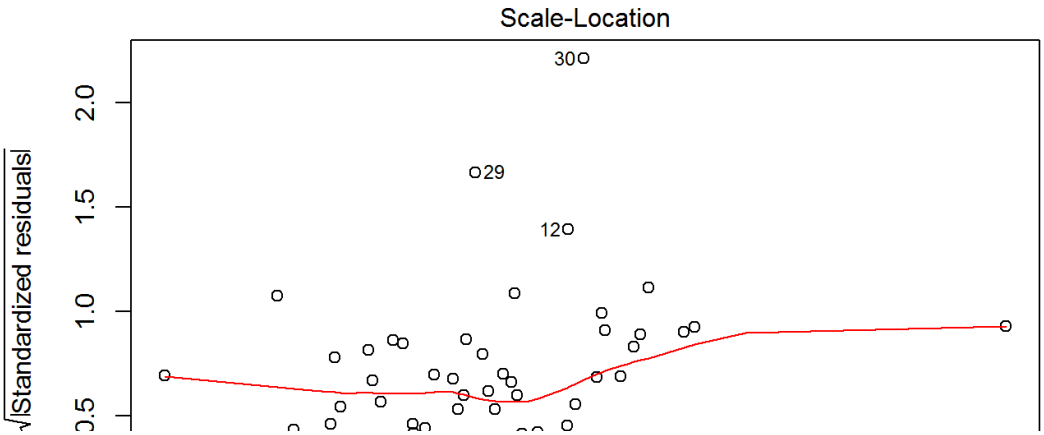
From the same argument, we can expect Sales to increase with a drop in Price, as they are both significant correlated with a negative coefficient. For each dollar drop in price, the sales increases by 3.25 units.

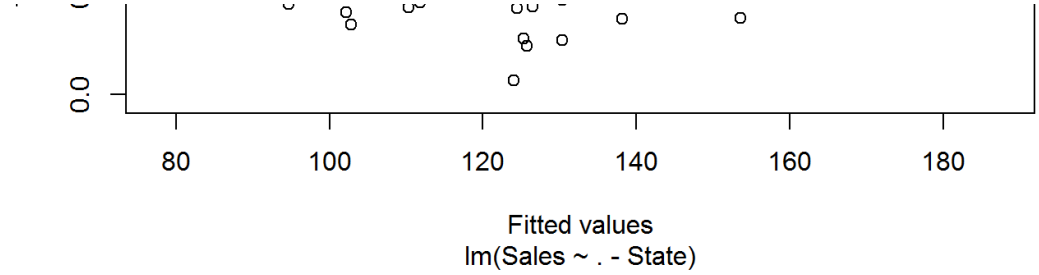
4.6 Diagnostic plots

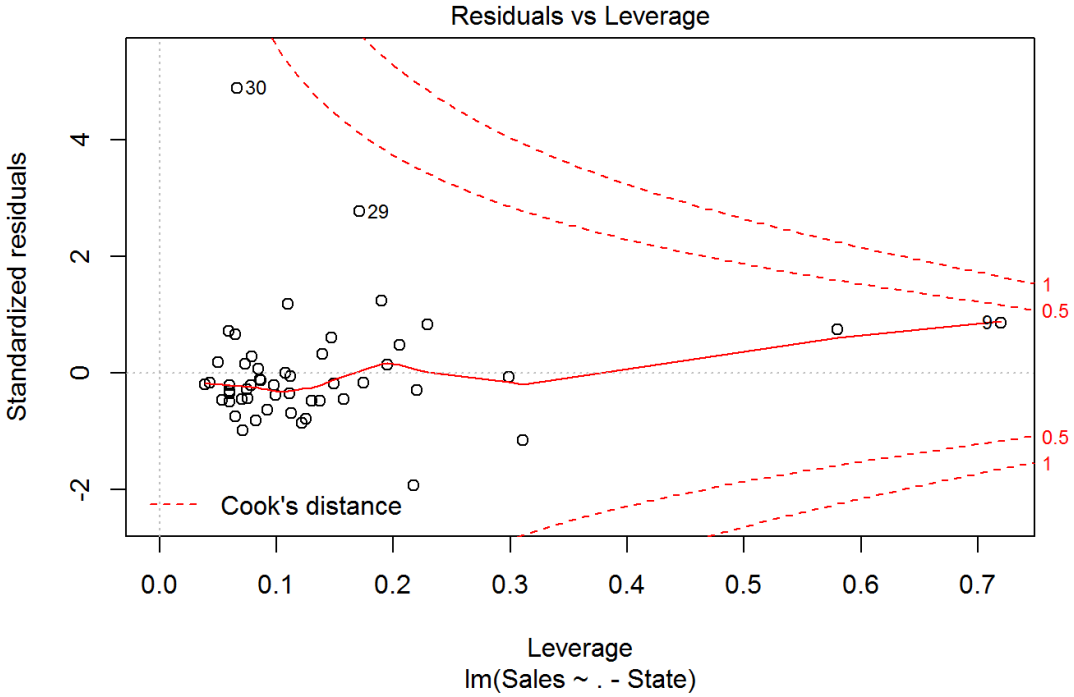
```
plot(lm.cigsales)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ . - State)
30
29
12
150
100
50
0
-50
80
100
120
140
160
180


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ . - State)
30
29
12
4
2
0
-2
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
30
29
12
2.0
1.5
1.0





Points 12,29,30 appear to be outliers

Points 9,29,30 have a high Cook's distance (they are just within the 0.5 threshold)

Strictly speaking, at 0.5 threshold Cook's distance we do not eliminate any observations. But to be sure, we may want to look at removing these observations to see if the summary of regression improves significantly.

4.7 Checking for combinations of interaction

```
# Testing for interaction between Price & Income
lm.4.7a <- lm(Sales ~ Price + Income + Price*Income, data = cigsales)
summary(lm.4.7a)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Income + Price * Income, data = cigsales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.838 -12.412  -4.109   4.742 132.929
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.576e+02  2.587e+02   0.996    0.324
## Price        -5.702e+00  6.649e+00  -0.857    0.396
## Income       -5.471e-03  6.782e-02  -0.081    0.936
## Price:Income  7.065e-04  1.730e-03   0.408    0.685
##
## Residual standard error: 28.59 on 47 degrees of freedom
## Multiple R-squared:  0.2529, Adjusted R-squared:  0.2052
## F-statistic: 5.304 on 3 and 47 DF,  p-value: 0.003122
```

As seen from p-values, the interaction is not significant

```
# Testing for interaction between Price & Income
lm.4.7b <- lm(Sales ~ Price + Age + Price*Age, data = cigsales)
summary(lm.4.7b)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Age + Price * Age, data = cigsales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.840 -18.467  -3.750   4.683 124.542
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1158.4081   714.1452  -1.622   0.1115
## Price          28.2280    17.8869   1.578   0.1212
## Age            50.5392    25.9041   1.951   0.0570 .
## Price:Age      -1.1291     0.6467  -1.746   0.0873 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.9 on 47 degrees of freedom
## Multiple R-squared:  0.2365, Adjusted R-squared:  0.1877
## F-statistic: 4.852 on 3 and 47 DF,  p-value: 0.005065
```

As seen from p-values, the interaction is not significant

---

5 F statistic and R squared

$Null\ Hypothesis\ H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots \beta_p = 0$

$Alternate\ Hypothesis\ H_1 : \beta_i \neq 0\ for\ atleast\ one\ i \in (1, 2, 3 \ldots p)$

$Unrestricted\ model\ : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots . + \beta_p X_p$

$Restricted\ model\ : Y = \beta_0$

$F = \frac{(SSR_r - SSR_u)/p}{SSR_u/n-p-1}$

$However,\ for\ the\ restricted\ model\ SSR_r = TSS$

$Dividing\ numerator\ and\ denominator\ by\ TSS,\ we\ get$

$F = \frac{(1 - \frac{SSR}{TSS})/p}{\frac{SSR}{TSS}/n-p-1}$

$But\ we\ know\ that\ R^2 = 1 - \frac{SSR}{TSS}$

$Therefore\ by\ replacing\ in\ above\ equation,\ we\ get$

$F = \frac{(R^2)/p}{(1-R^2)/n-p-1}$

$Hence\ proved.$
