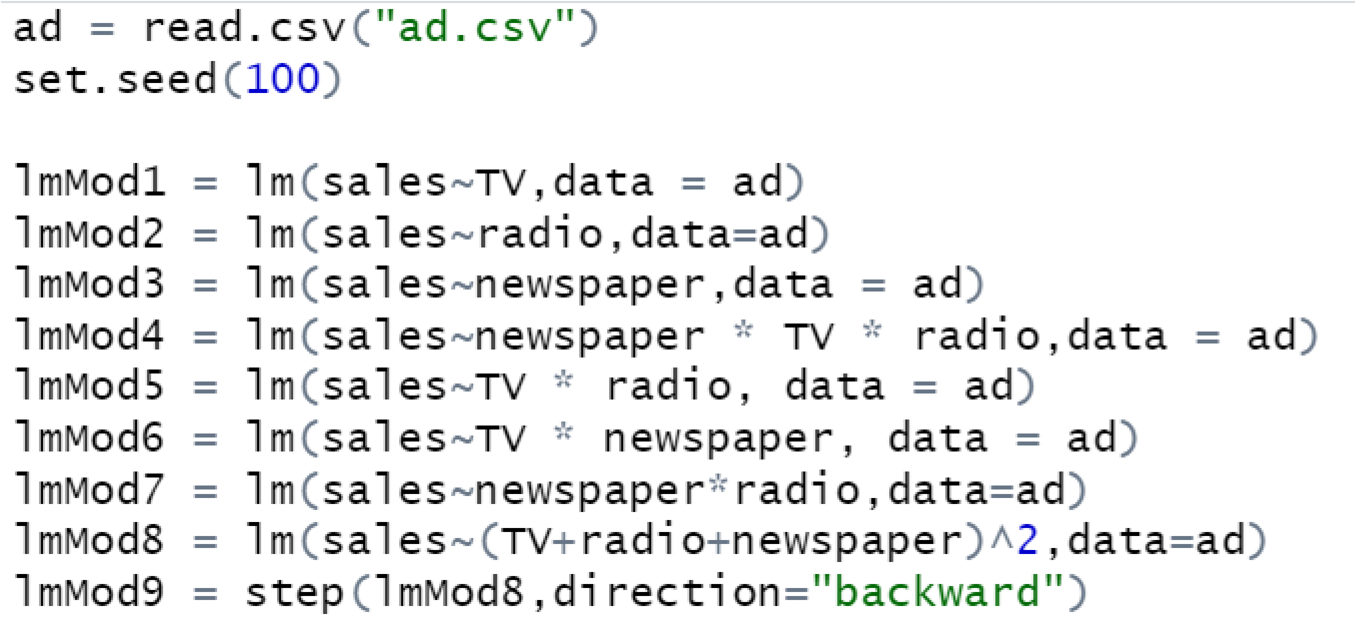
R통계분석 12주차 과제 보고서

60191697 최 솔

1. 예측 모델 생성 및 모델 검증



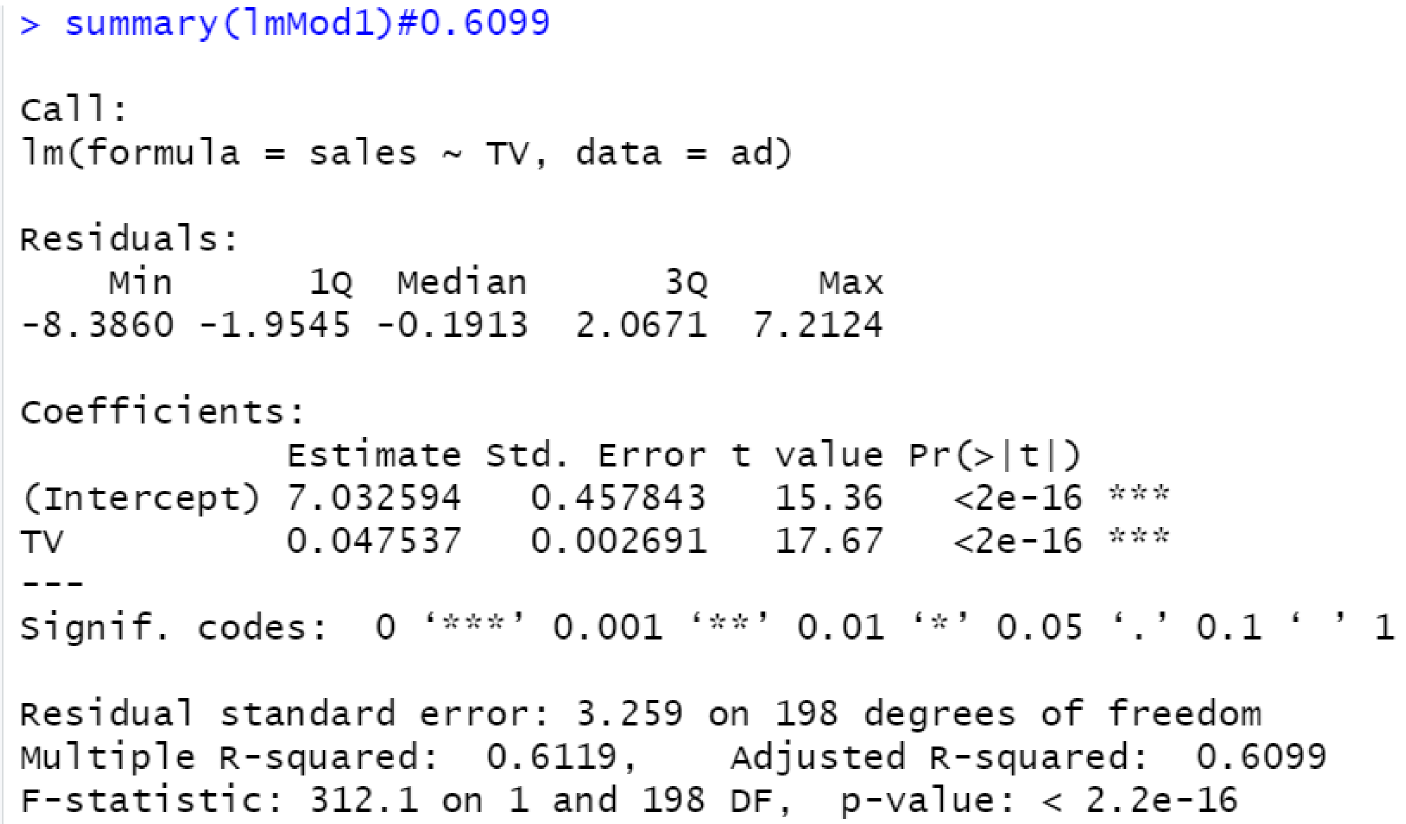
```
ad = read.csv("ad.csv")
set.seed(100)

lmMod1 = lm(sales~TV,data = ad)
lmMod2 = lm(sales~radio,data=ad)
lmMod3 = lm(sales~newspaper,data = ad)
lmMod4 = lm(sales~newspaper * TV * radio,data = ad)
lmMod5 = lm(sales~TV * radio, data = ad)
lmMod6 = lm(sales~TV * newspaper, data = ad)
lmMod7 = lm(sales~newspaper*radio,data=ad)
lmMod8 = lm(sales~(TV+radio+newspaper)^2,data=ad)
lmMod9 = step(lmMod8,direction="backward")
```

<그림 1>

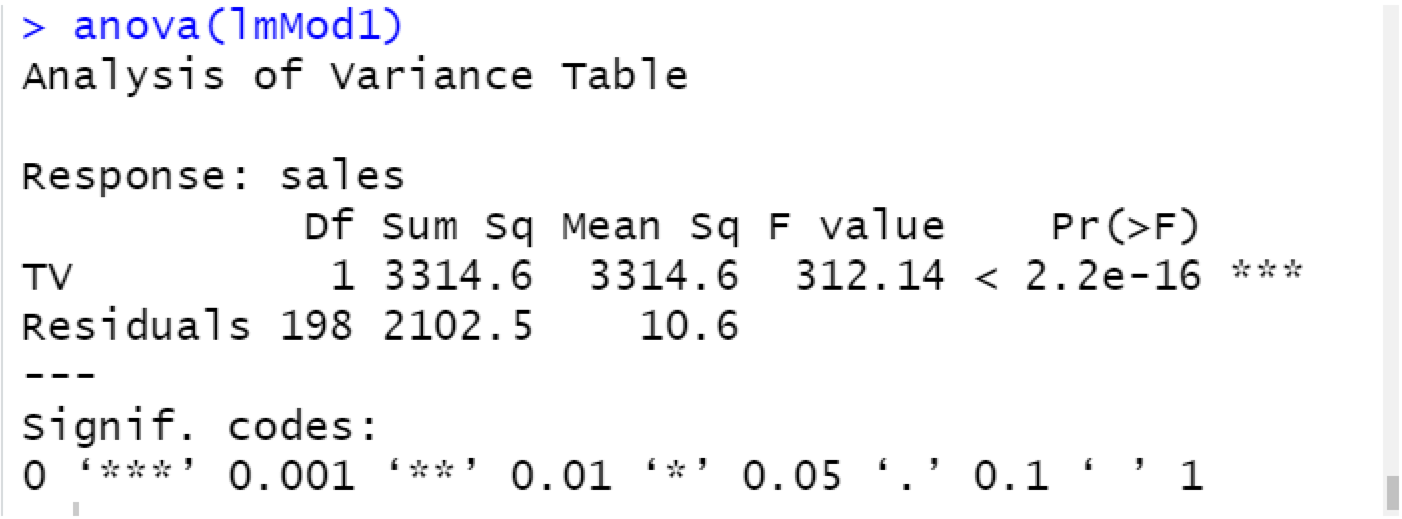
위 <그림 1>의 코드처럼 lm과 다중 회귀분석의 formula를 이용해 총 9개의 모델을 생성하였다. 그리고, 다음에서 summary와 anova를 이용해 9개 모델을 검증하고, 이후 그래프로 가시화 하였다.

해당 내용은 다음과 같다.

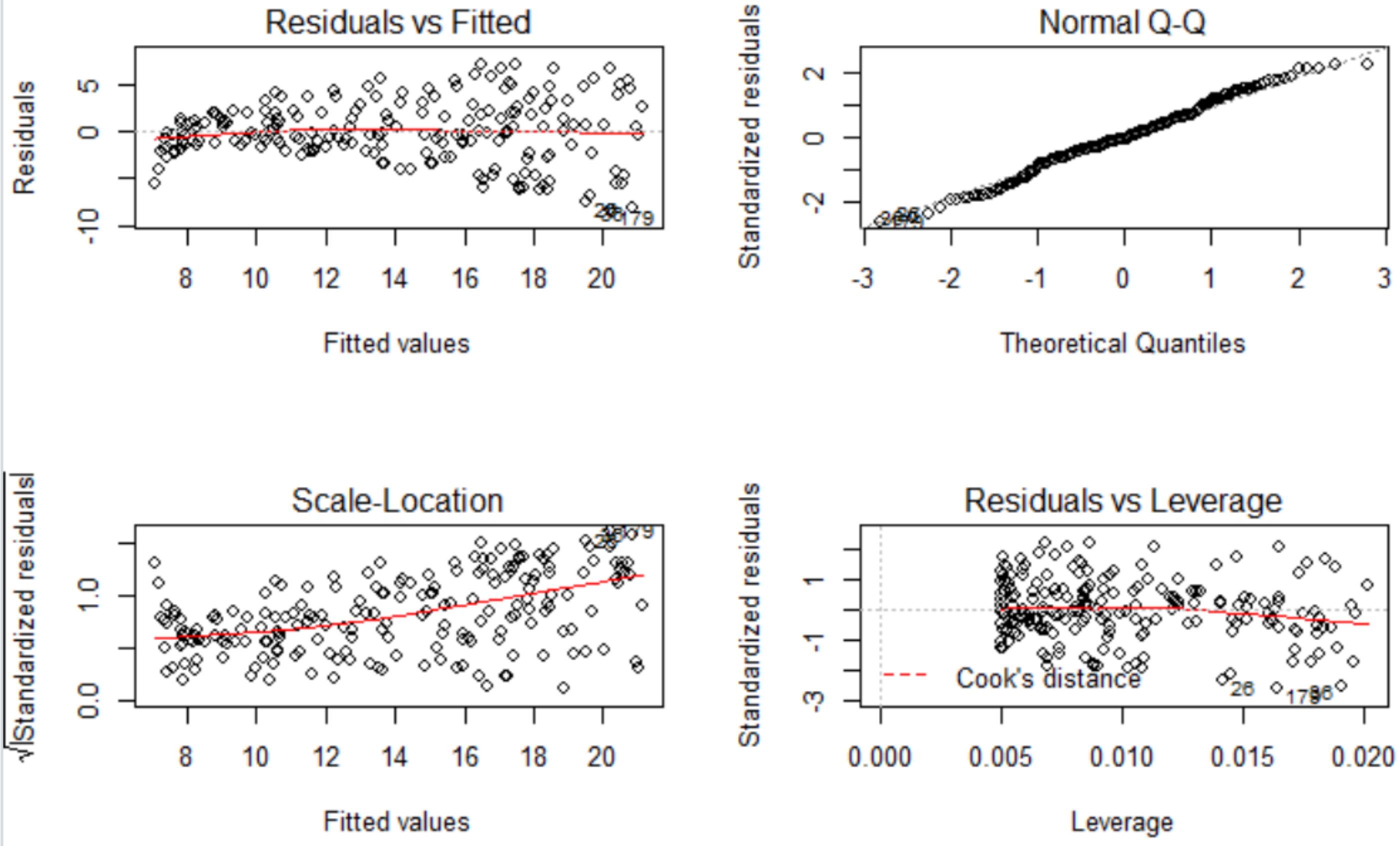
1) lmMod1



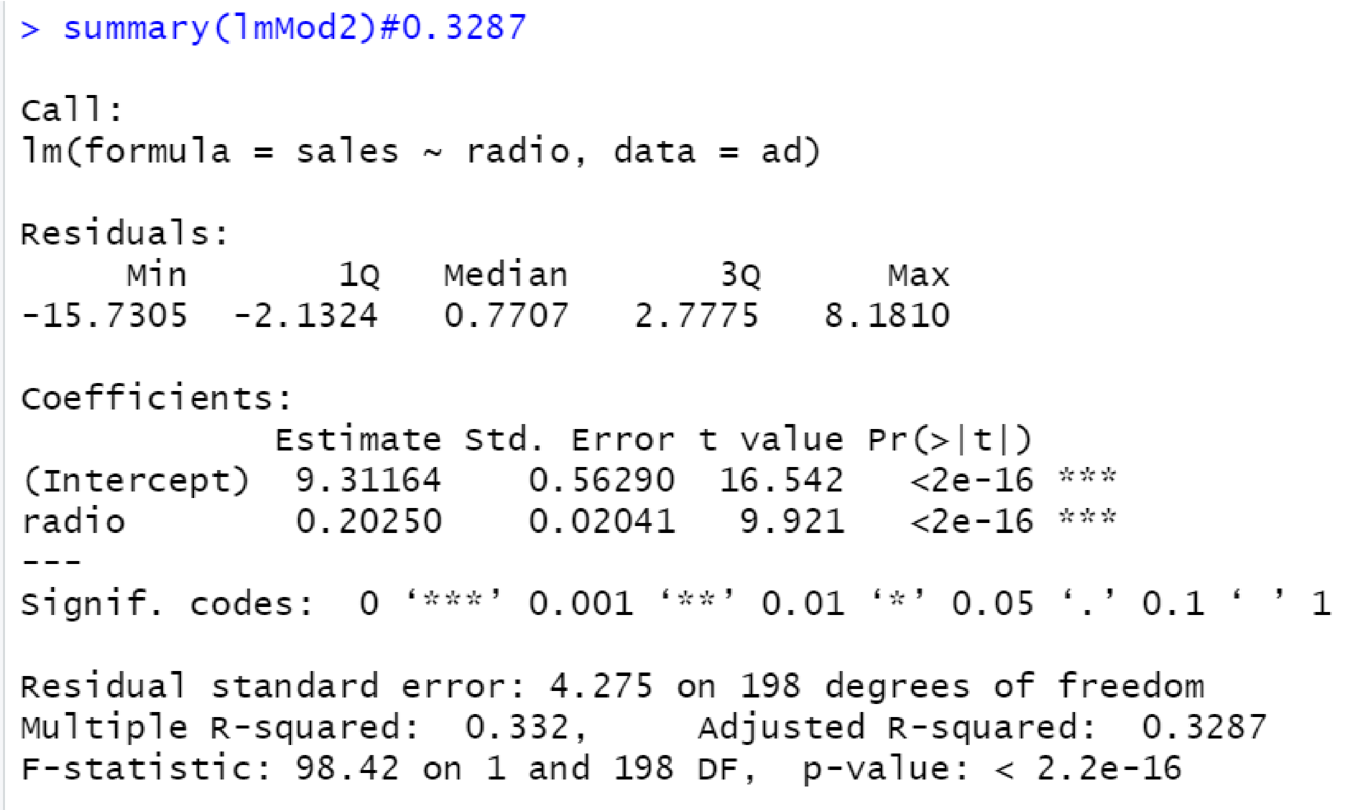
```
> summary(lmMod1)#0.6099

Call:
lm(formula = sales ~ TV, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3860 -1.9545 -0.1913  2.0671  7.2124 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,	Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```



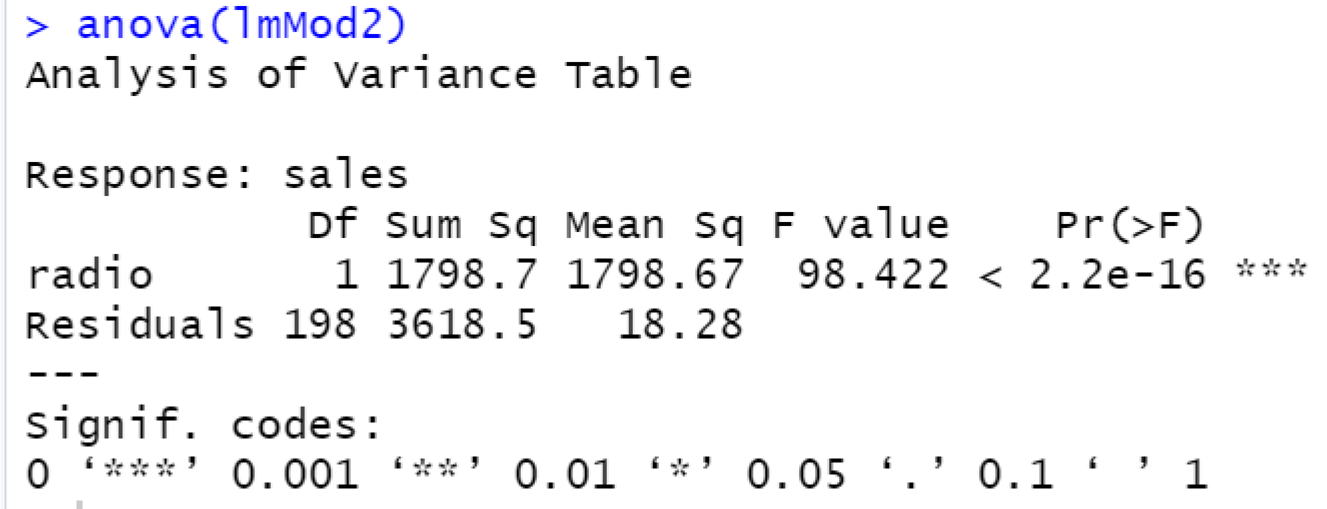
```
> anova(lmMod1)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)    
TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
Residuals 198 2102.5    10.6                      
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



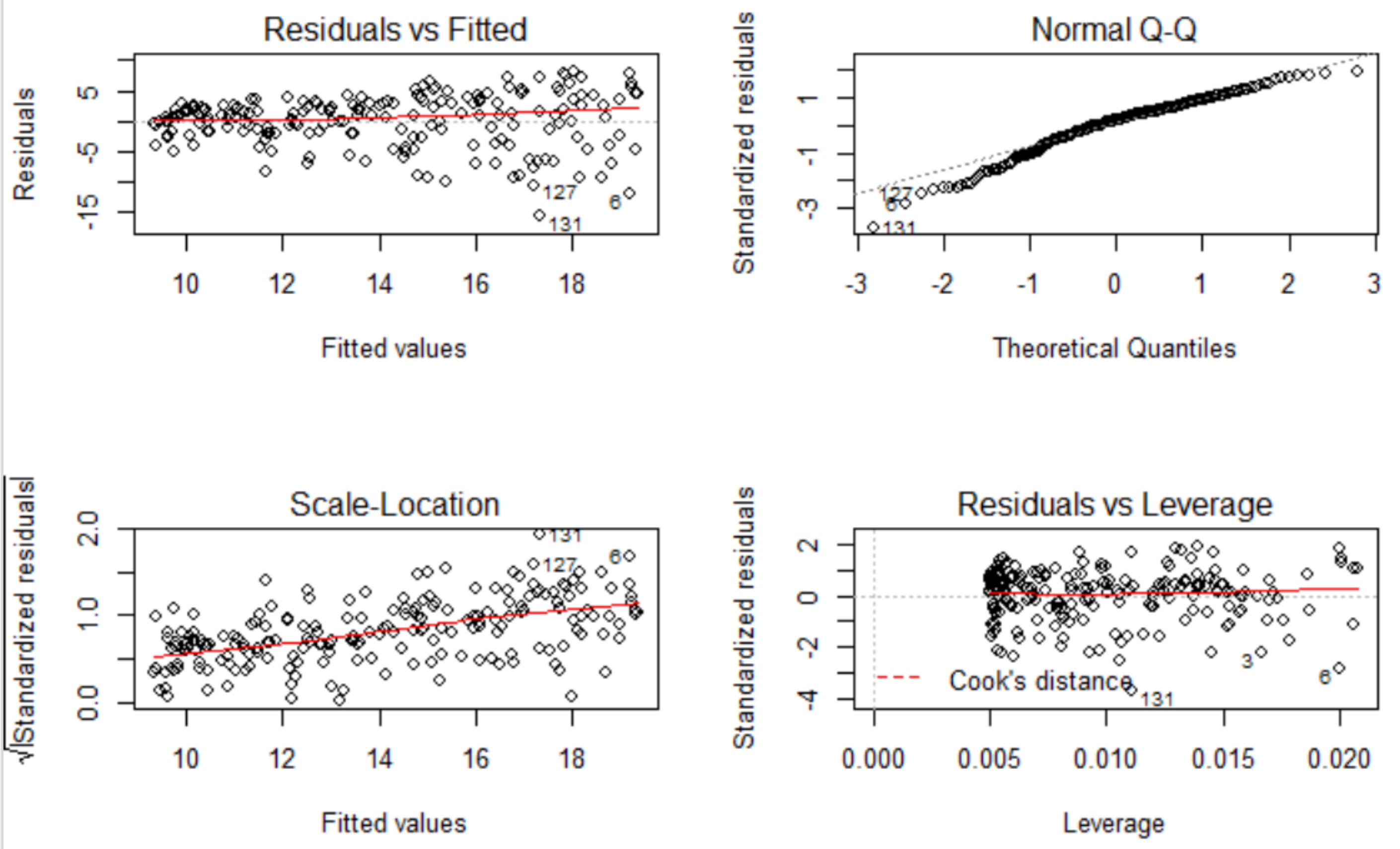
2) lmMod2



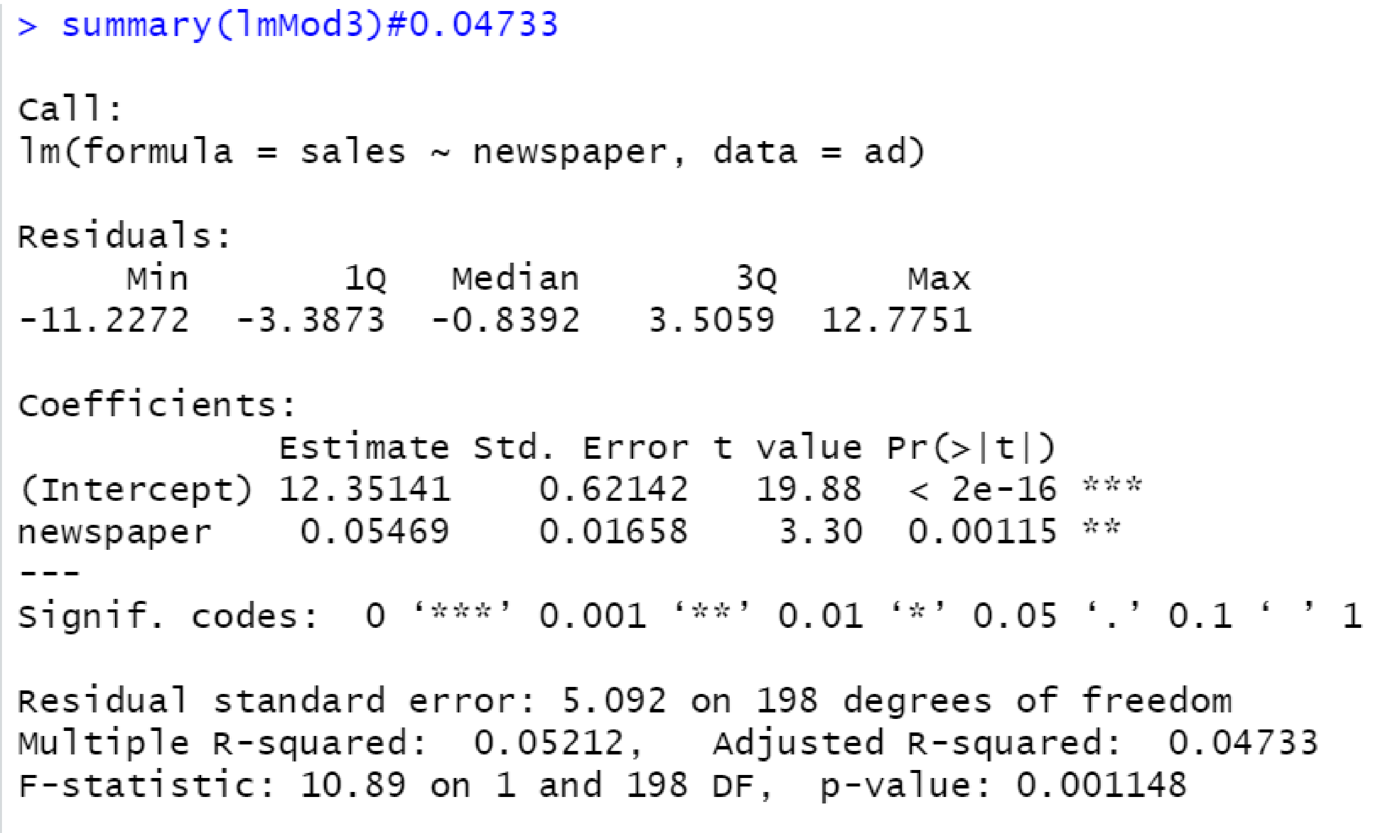
```
> summary(lmMod2)#0.3287

Call:
lm(formula = sales ~ radio, data = ad)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.7305  -2.1324   0.7707   2.7775   8.1810 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,	Adjusted R-squared:  0.3287 
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```



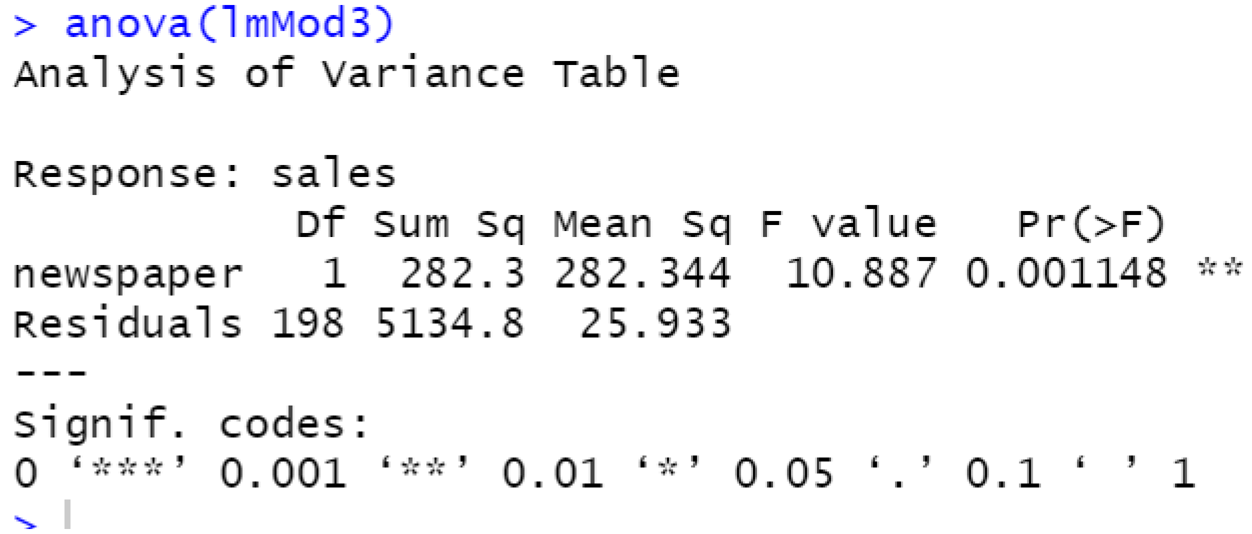
```
> anova(lmMod2)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)    
radio       1 1798.7 1798.67  98.422 < 2.2e-16 ***
Residuals 198 3618.5   18.28                      
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



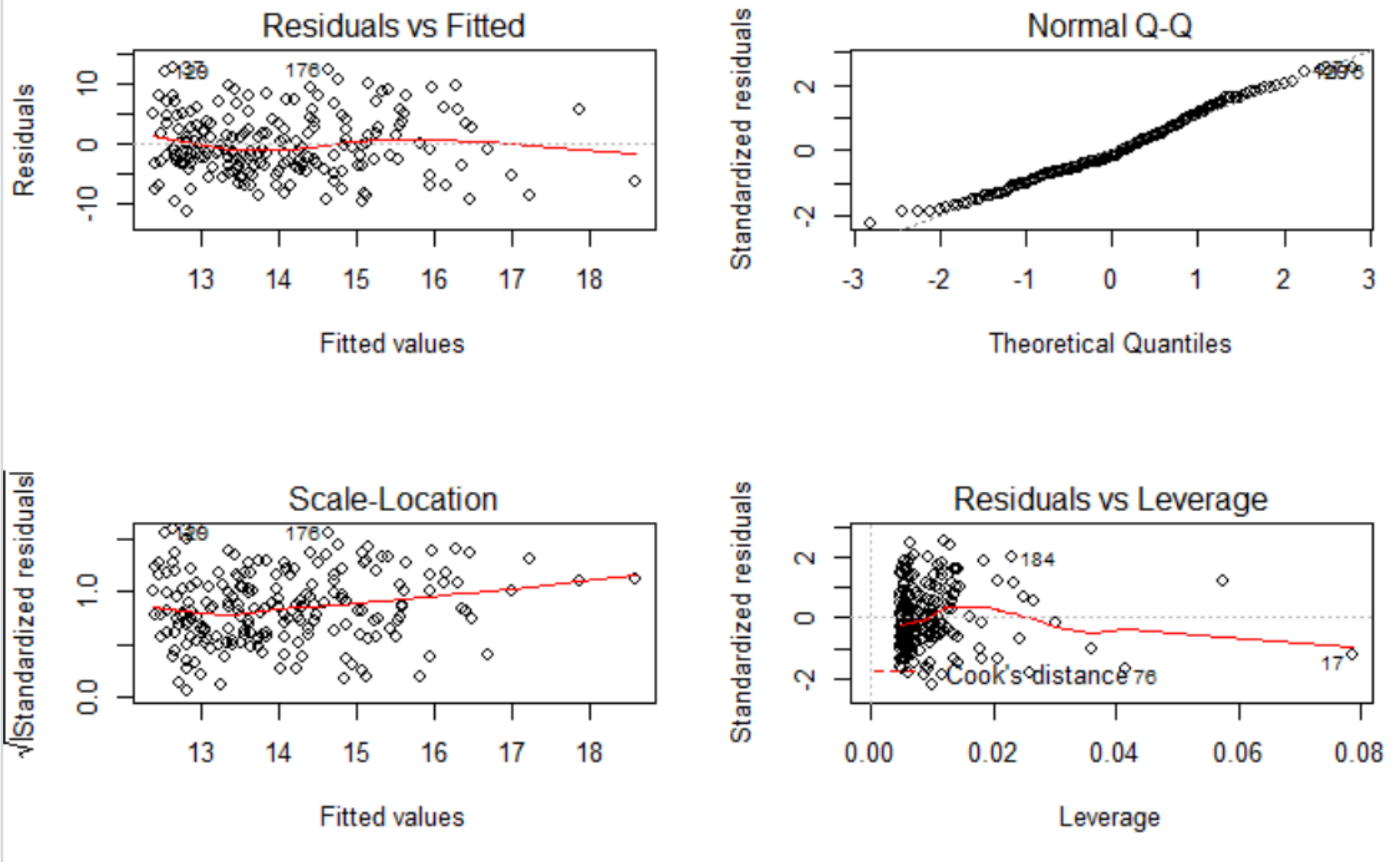
3) lmMod3



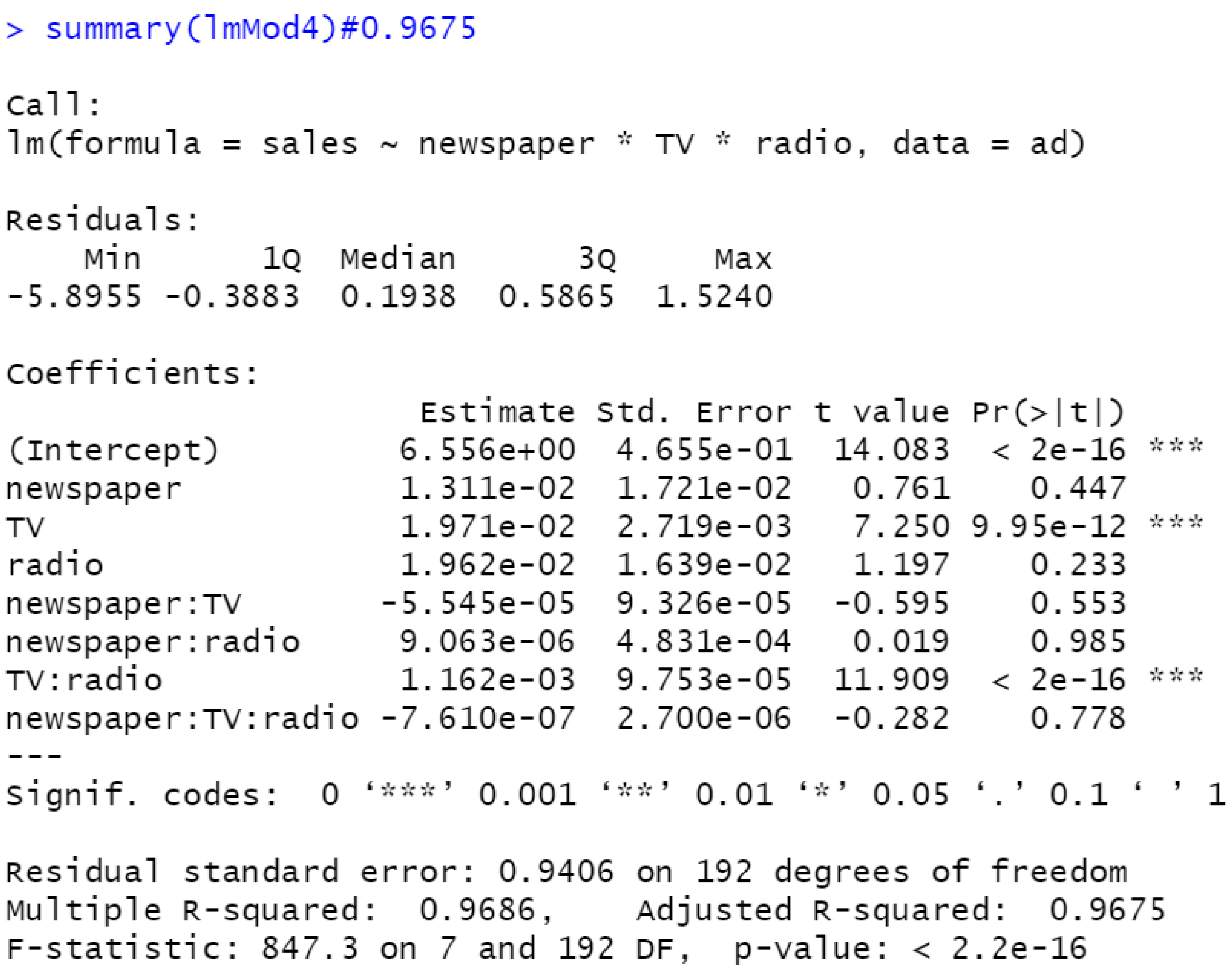
```
> summary(lmMod3)#0.04733

Call:
lm(formula = sales ~ newspaper, data = ad)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.2272  -3.3873  -0.8392   3.5059  12.7751 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,	Adjusted R-squared:  0.04733 
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```



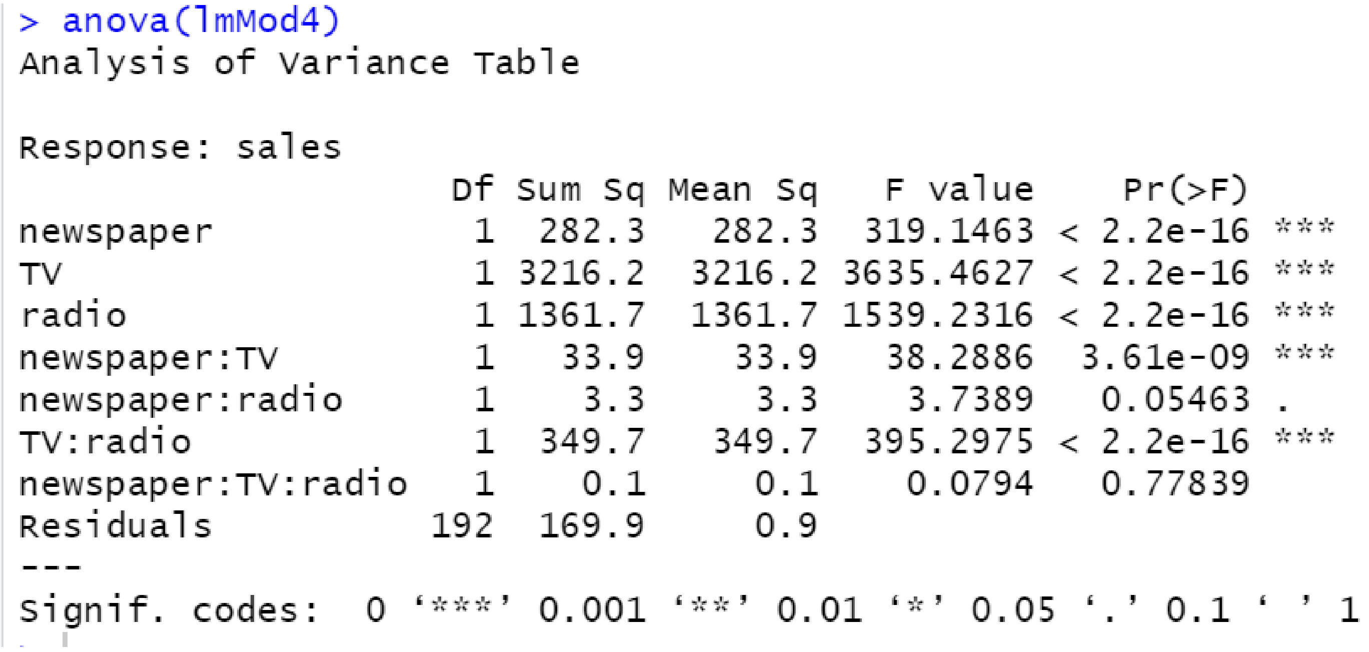
```
> anova(lmMod3)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value   Pr(>F)   
newspaper   1  282.3 282.344  10.887 0.001148 **
Residuals 198 5134.8  25.933                    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```



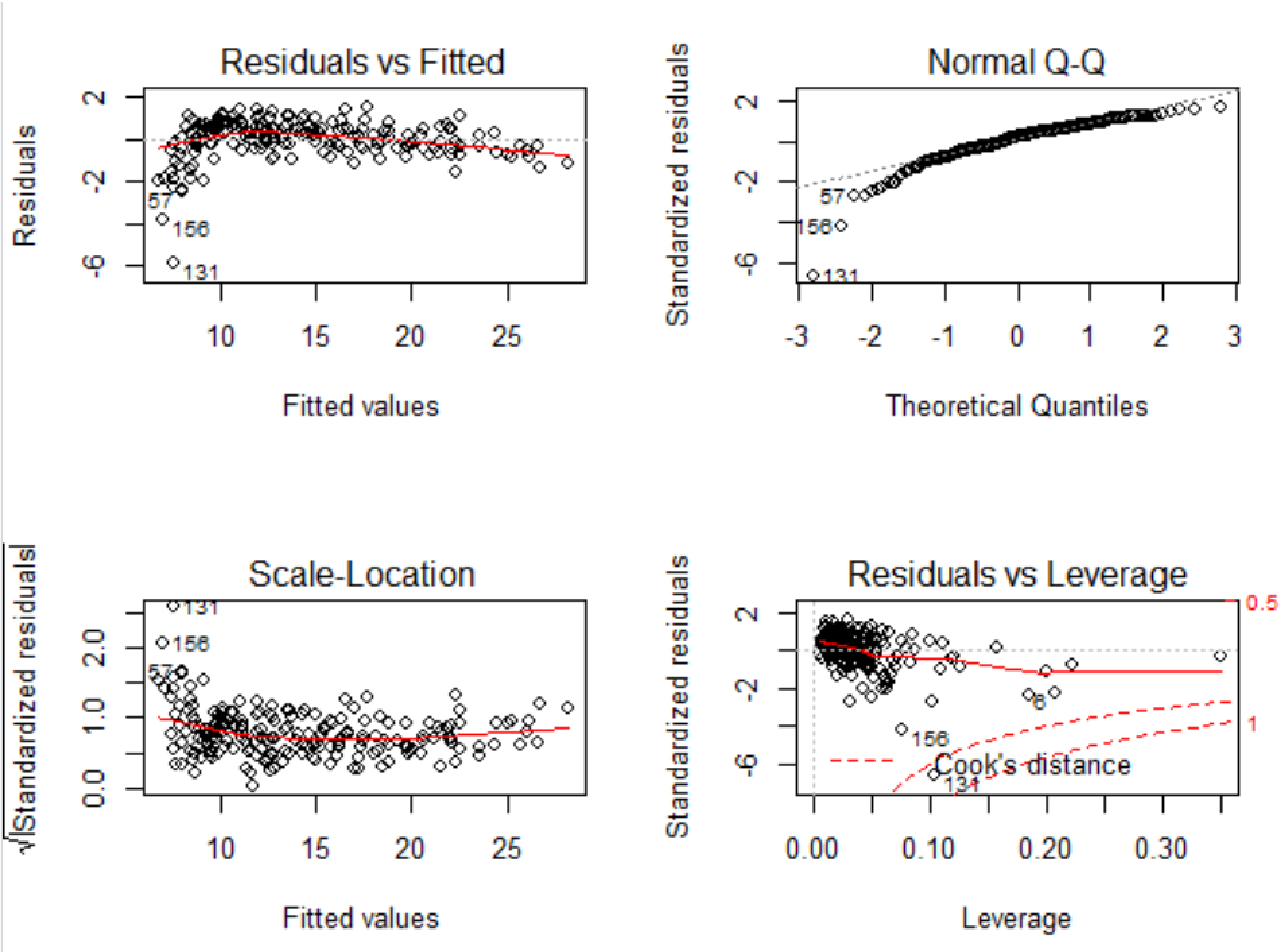
4) lmMod4



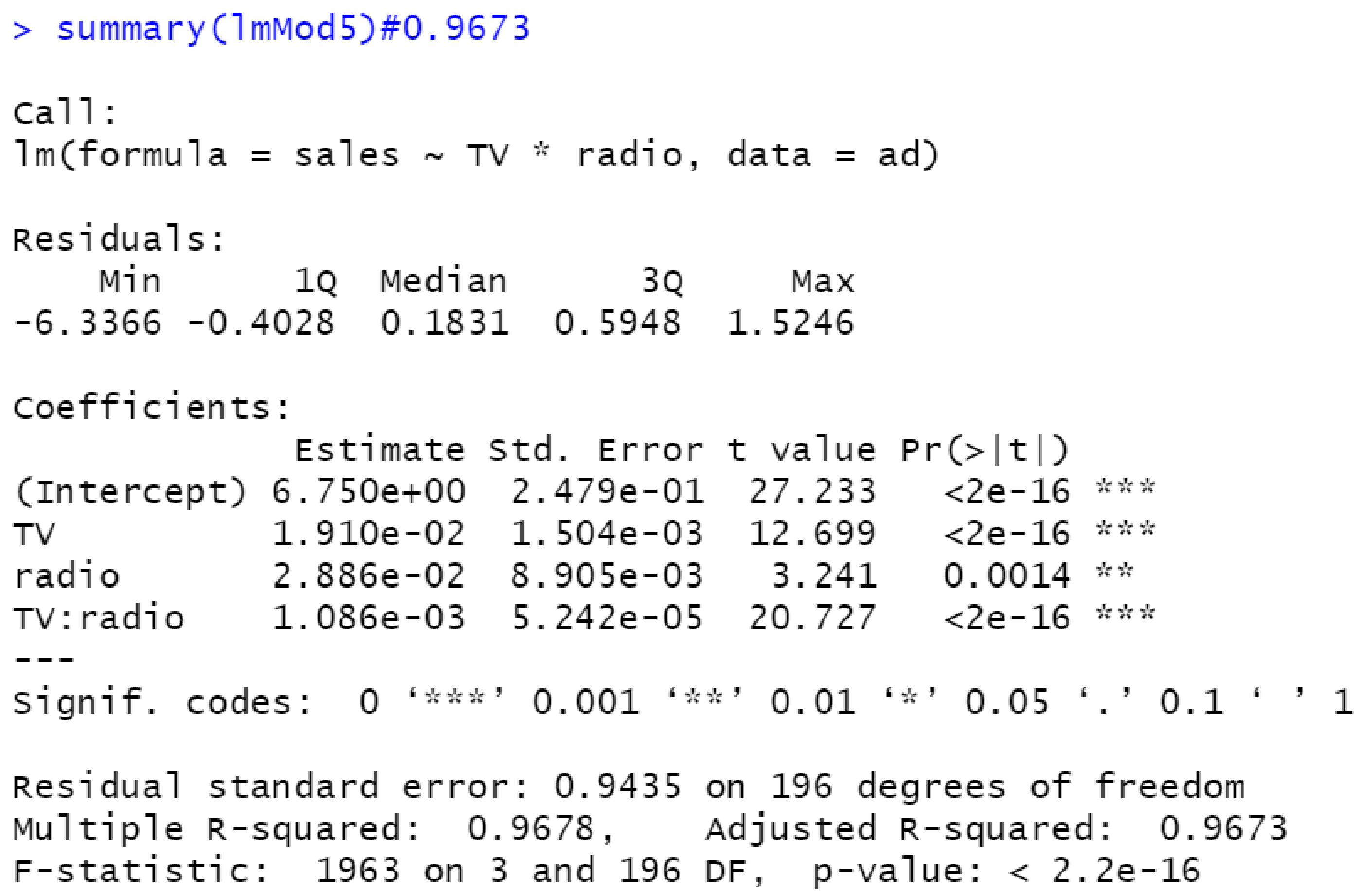
```
> summary(lmMod4)#0.9675

Call:
lm(formula = sales ~ newspaper * TV * radio, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8955 -0.3883  0.1938  0.5865  1.5240 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6.556e+00  4.655e-01  14.083  < 2e-16 ***
newspaper           1.311e-02  1.721e-02   0.761    0.447    
TV                  1.971e-02  2.719e-03   7.250 9.95e-12 ***
radio               1.962e-02  1.639e-02   1.197    0.233    
newspaper:TV       -5.545e-05  9.326e-05  -0.595    0.553    
newspaper:radio     9.063e-06  4.831e-04   0.019    0.985    
TV:radio            1.162e-03  9.753e-05  11.909  < 2e-16 ***
newspaper:TV:radio -7.610e-07  2.700e-06  -0.282    0.778    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9406 on 192 degrees of freedom
Multiple R-squared:  0.9686,	Adjusted R-squared:  0.9675 
F-statistic: 847.3 on 7 and 192 DF,  p-value: < 2.2e-16
```



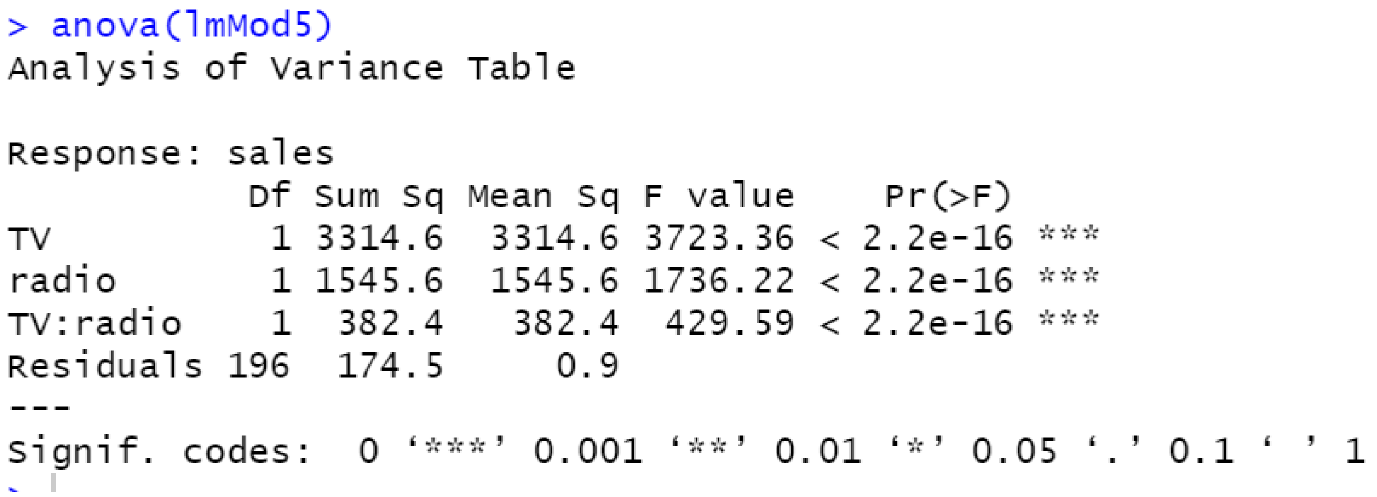
```
> anova(lmMod4)
Analysis of Variance Table

Response: sales
                    Df Sum Sq Mean Sq   F value    Pr(>F)    
newspaper            1  282.3   282.3  319.1463 < 2.2e-16 ***
TV                   1 3216.2  3216.2 3635.4627 < 2.2e-16 ***
radio                1 1361.7  1361.7 1539.2316 < 2.2e-16 ***
newspaper:TV         1   33.9    33.9   38.2886  3.61e-09 ***
newspaper:radio      1    3.3     3.3    3.7389   0.05463 .  
TV:radio             1  349.7   349.7  395.2975 < 2.2e-16 ***
newspaper:TV:radio   1    0.1     0.1    0.0794   0.77839    
Residuals          192  169.9     0.9                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Residuals vs Fitted
Residuals
2
-2
-6
57
156
131
10
15
20
25
Fitted values
Normal Q-Q
Standardized residuals
2
-2
-6
57
156
131
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
131
156
57
10
15
20
25
Fitted values
Residuals vs Leverage
Standardized residuals
2
-2
-6
0.5
1
6
156
131
Cook's distance
0.00
0.10
0.20
0.30
Leverage

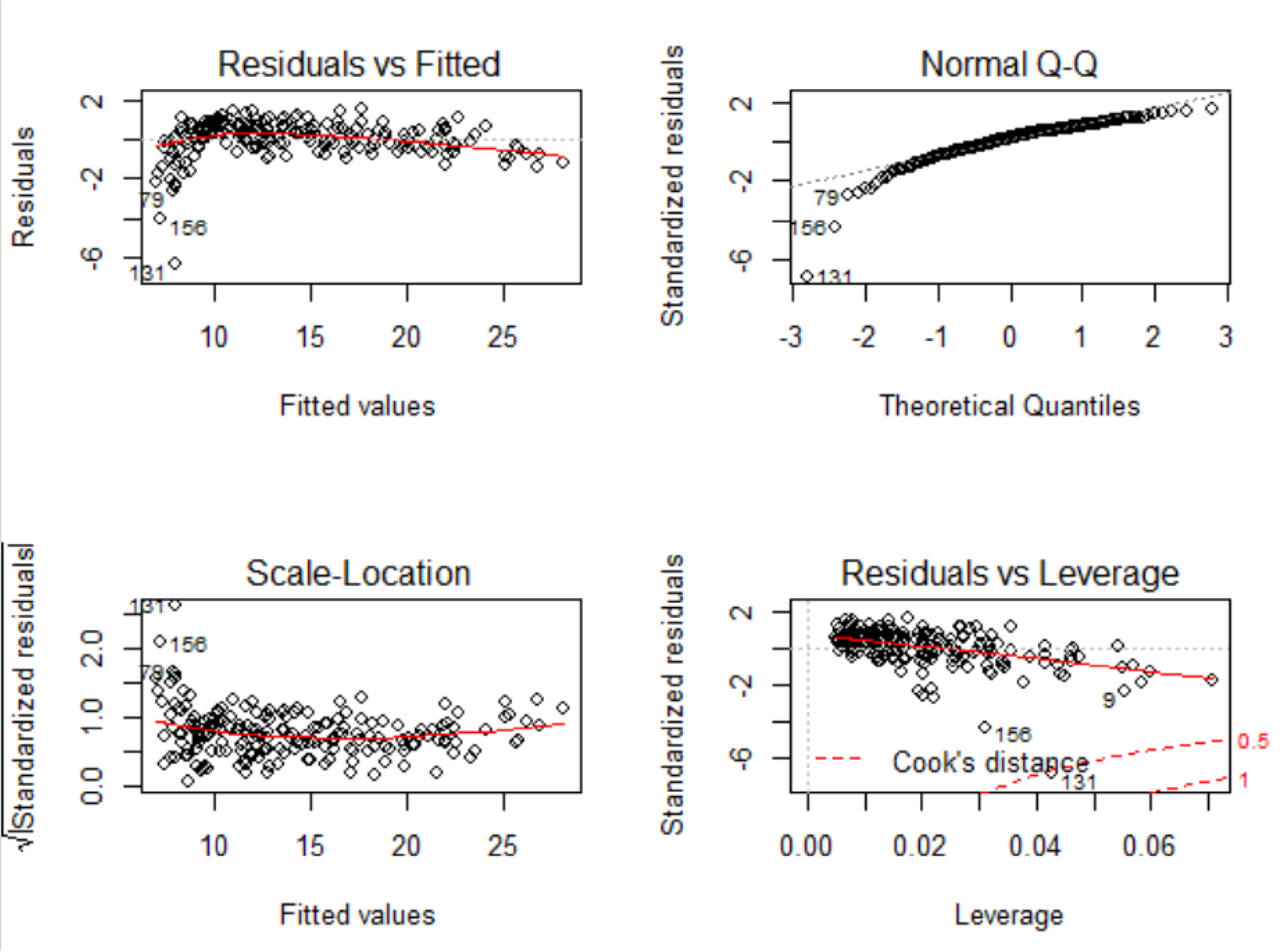
5) lmMod5



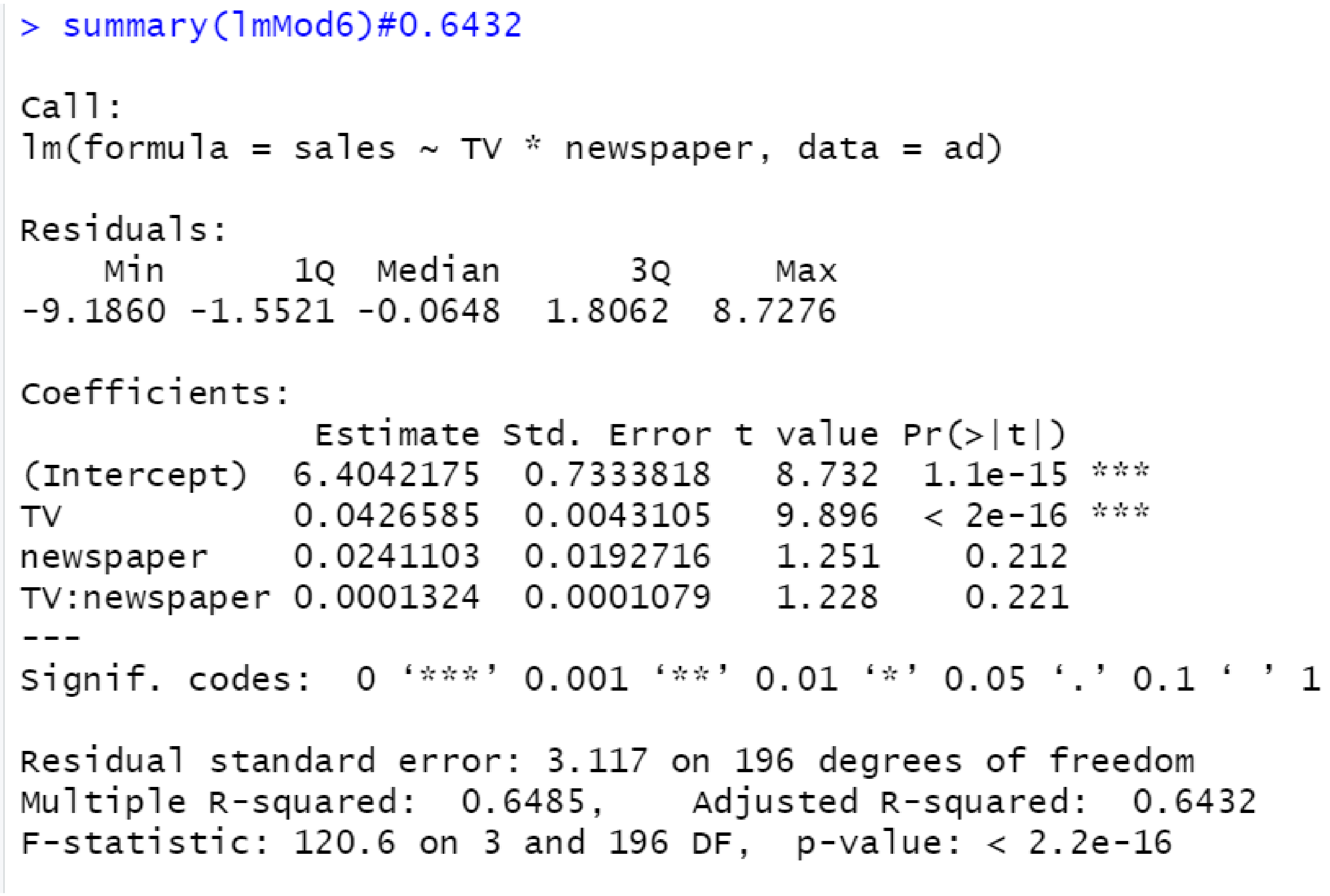
```
> summary(lmMod5)#0.9673

Call:
lm(formula = sales ~ TV * radio, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```



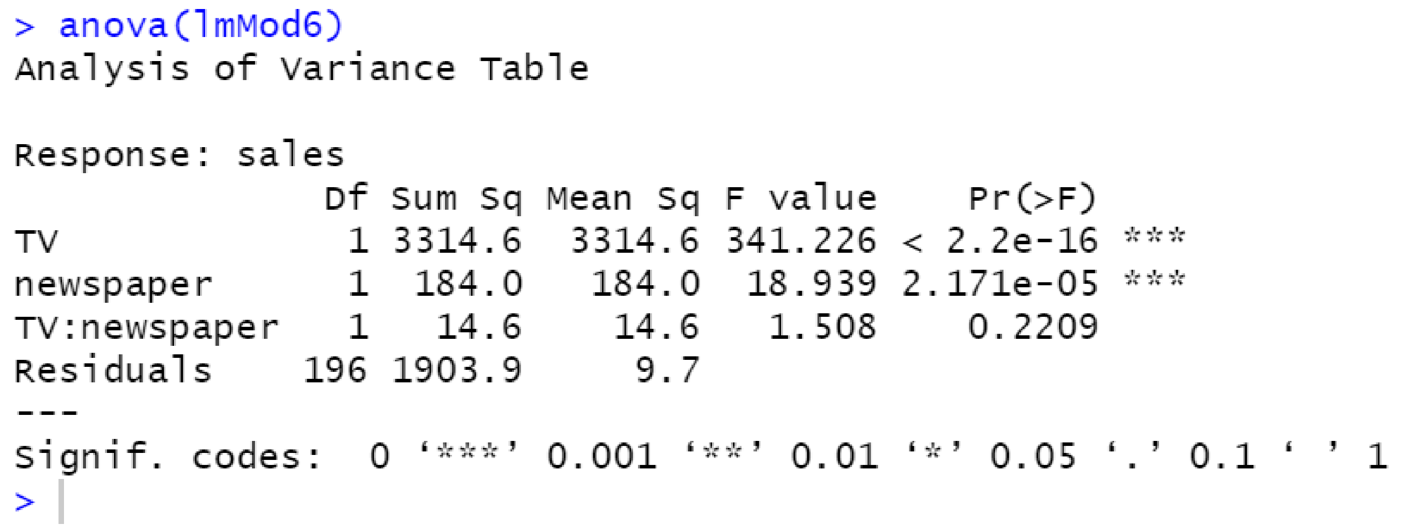
```
> anova(lmMod5)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)    
TV          1 3314.6  3314.6 3723.36 < 2.2e-16 ***
radio       1 1545.6  1545.6 1736.22 < 2.2e-16 ***
TV:radio    1  382.4   382.4  429.59 < 2.2e-16 ***
Residuals 196  174.5     0.9                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Residuals vs Fitted
Residuals
Fitted values
2
-2
-6
10
15
20
25
79
156
131
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
3
79
156
131
Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
131
156
79
Residuals vs Leverage
Leverage
0.00
0.02
0.04
0.06
Cook's distance
0.5
1
9
156
131

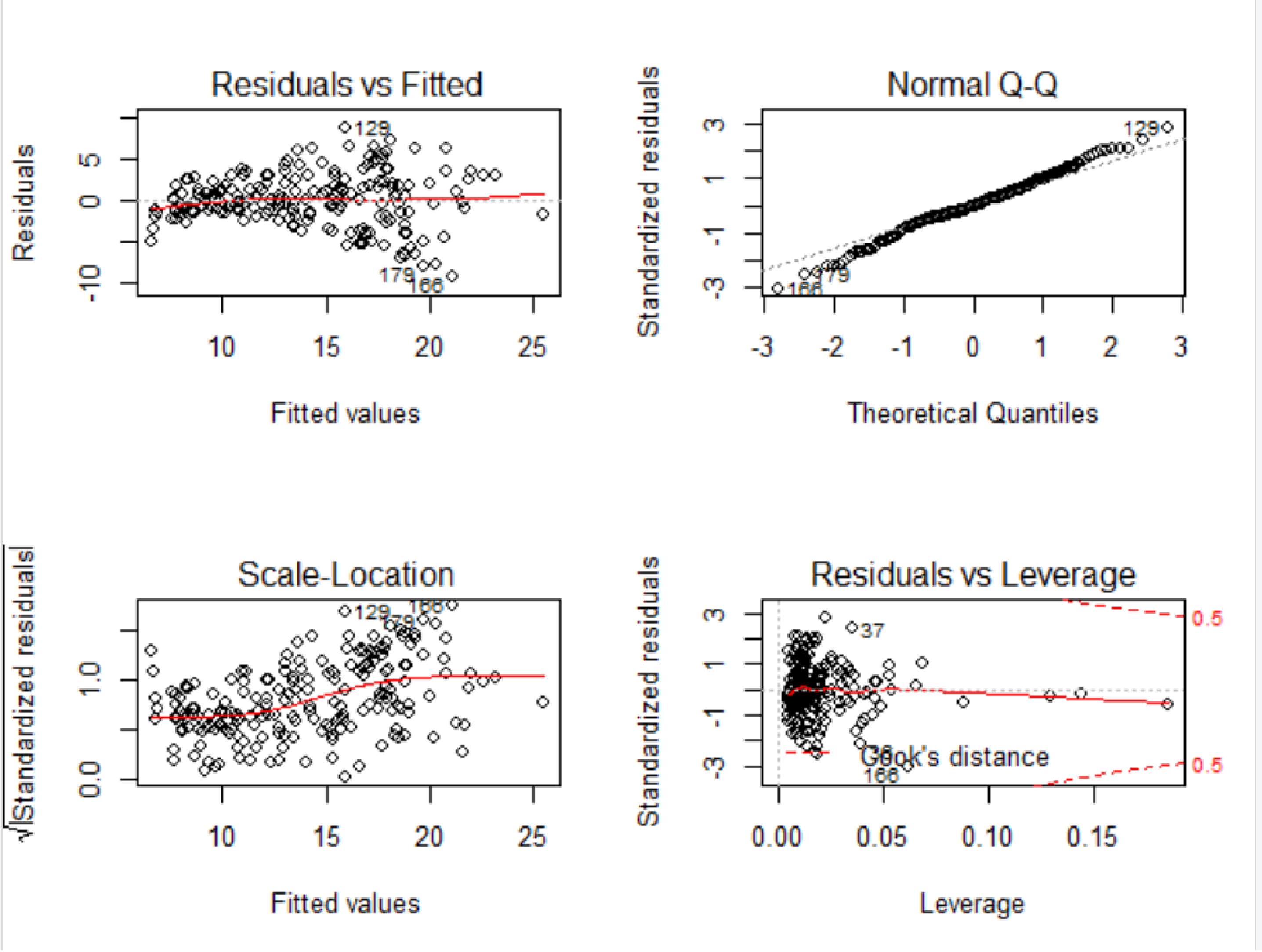
6) lmMod6



```
> summary(lmMod6)#0.6432

Call:
lm(formula = sales ~ TV * newspaper, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1860 -1.5521 -0.0648  1.8062  8.7276 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.4042175  0.7333818   8.732  1.1e-15 ***
TV           0.0426585  0.0043105   9.896  < 2e-16 ***
newspaper    0.0241103  0.0192716   1.251    0.212    
TV:newspaper 0.0001324  0.0001079   1.228    0.221    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.117 on 196 degrees of freedom
Multiple R-squared:  0.6485,	Adjusted R-squared:  0.6432 
F-statistic: 120.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

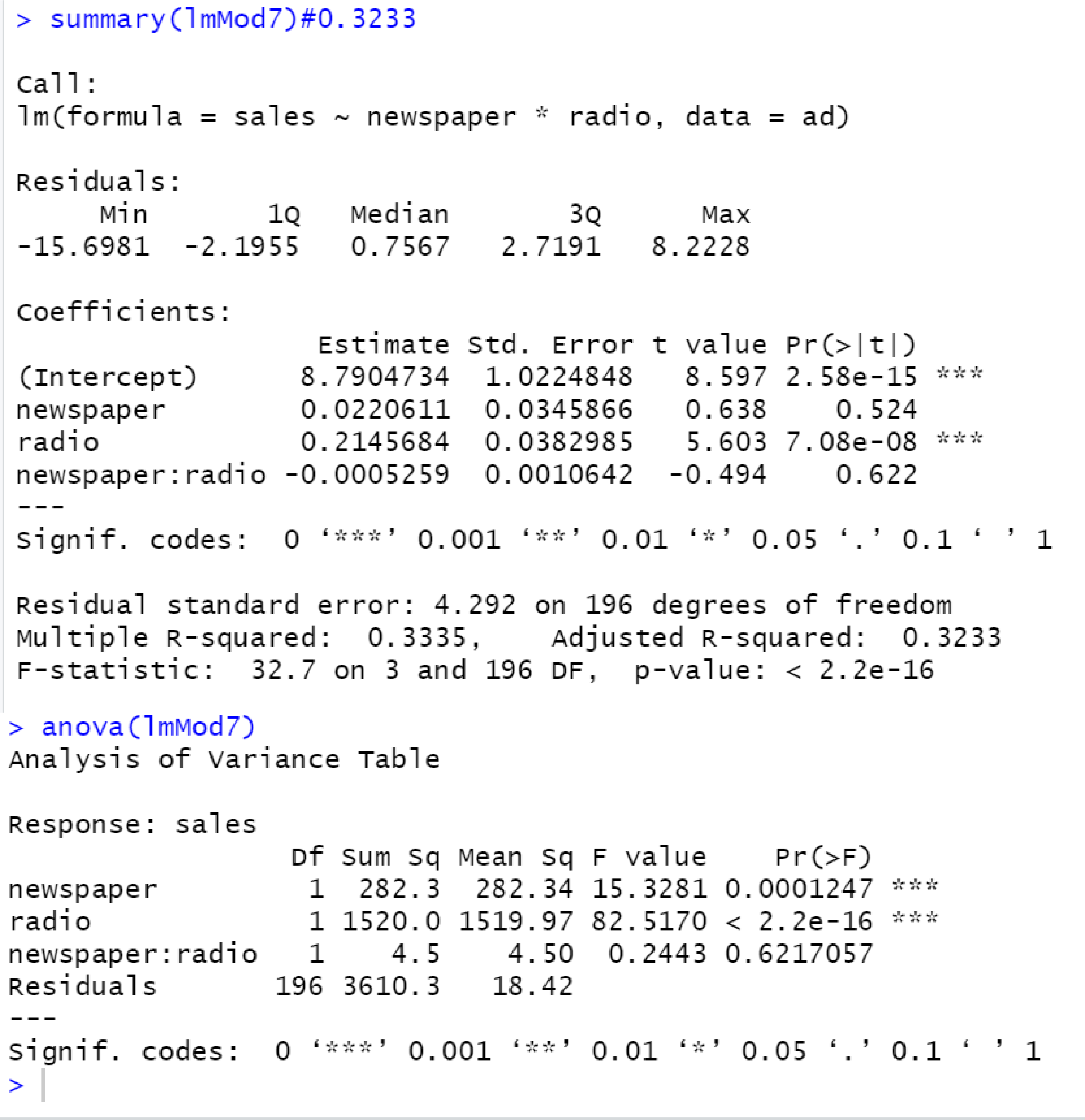


```
> anova(lmMod6)
Analysis of Variance Table

Response: sales
              Df Sum Sq Mean Sq F value    Pr(>F)    
TV             1 3314.6  3314.6 341.226 < 2.2e-16 ***
newspaper      1  184.0   184.0  18.939 2.171e-05 ***
TV:newspaper   1   14.6    14.6   1.508    0.2209    
Residuals    196 1903.9     9.7                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

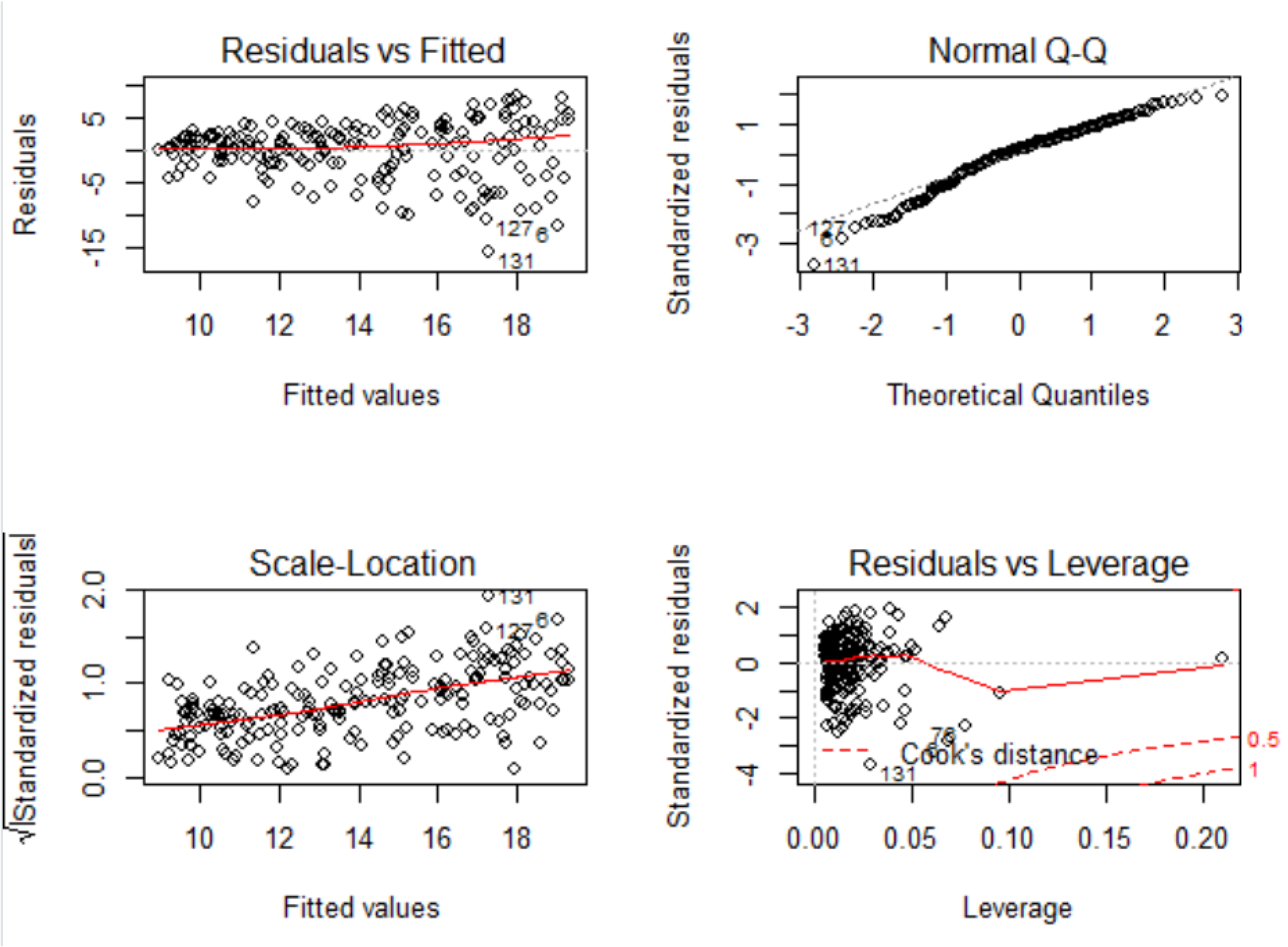


Residuals vs Fitted
Residuals
5
0
-10
10
15
20
25
Fitted values
129
179
166
Normal Q-Q
Standardized residuals
3
1
-1
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
129
179
166
Scale-Location
√|Standardized residuals|
1.0
0.0
10
15
20
25
Fitted values
129
179
166
Residuals vs Leverage
Standardized residuals
3
1
-1
-3
0.00
0.05
0.10
0.15
Leverage
37
Cook's distance
166
0.5
0.5

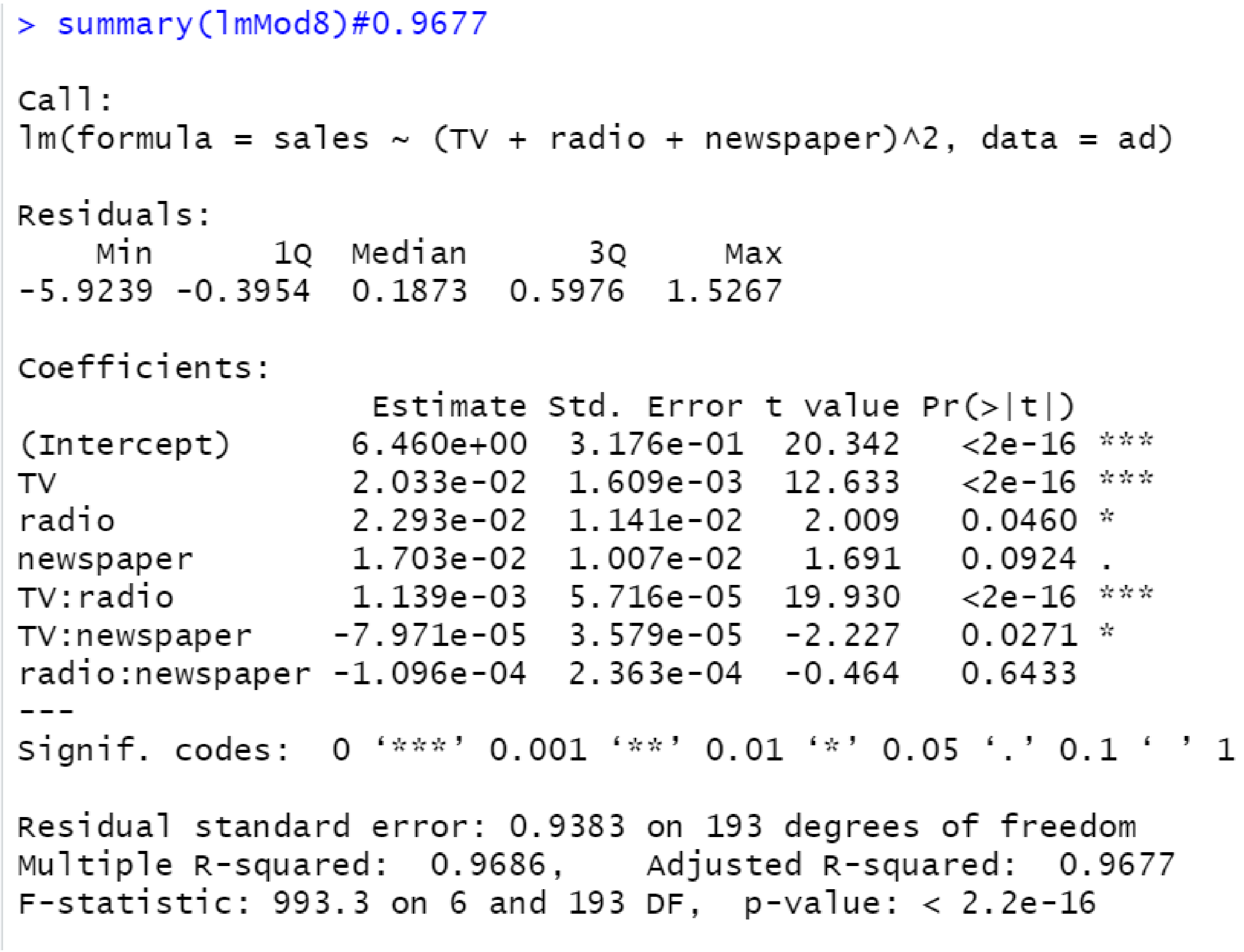
7) lmMod7



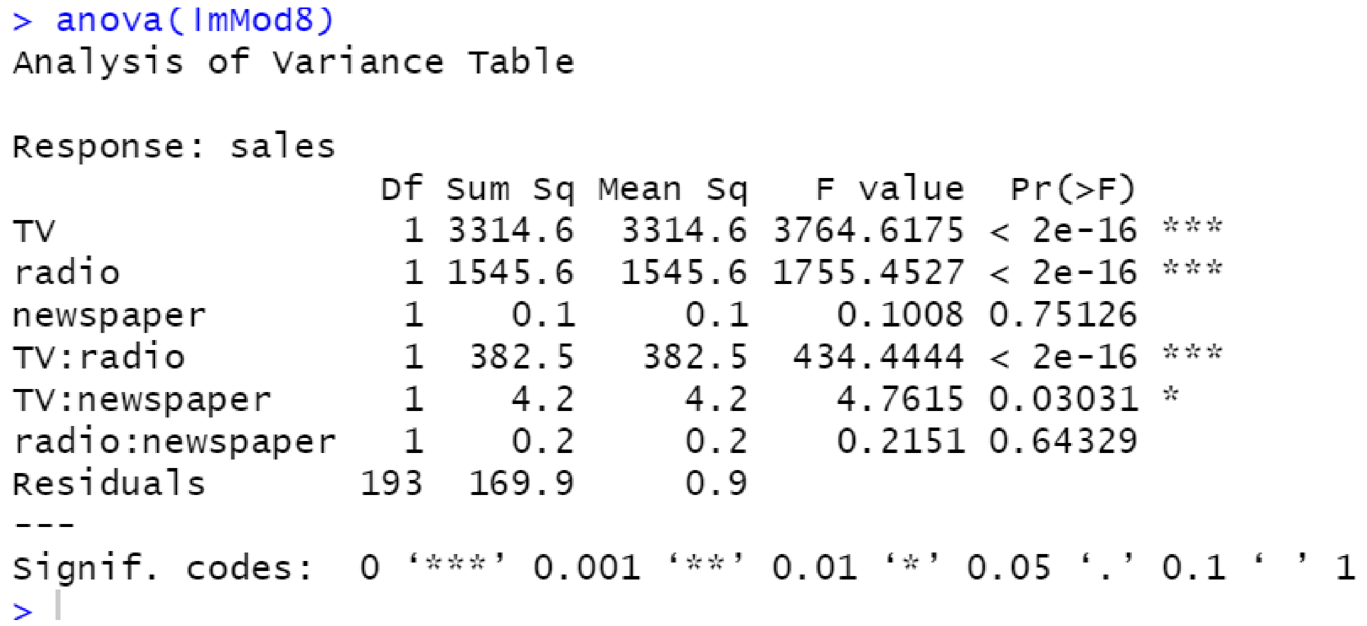
```
> summary(lmMod7)#0.3233

Call:
lm(formula = sales ~ newspaper * radio, data = ad)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.6981  -2.1955   0.7567   2.7191   8.2228 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      8.7904734  1.0224848   8.597 2.58e-15 ***
newspaper        0.0220611  0.0345866   0.638    0.524    
radio            0.2145684  0.0382985   5.603 7.08e-08 ***
newspaper:radio -0.0005259  0.0010642  -0.494    0.622    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.292 on 196 degrees of freedom
Multiple R-squared:  0.3335,	Adjusted R-squared:  0.3233 
F-statistic:  32.7 on 3 and 196 DF,  p-value: < 2.2e-16

> anova(lmMod7)
Analysis of Variance Table

Response: sales
                 Df Sum Sq Mean Sq F value    Pr(>F)    
newspaper         1  282.3  282.34 15.3281 0.0001247 ***
radio             1 1520.0 1519.97 82.5170 < 2.2e-16 ***
newspaper:radio   1    4.5    4.50  0.2443 0.6217057    
Residuals       196 3610.3   18.42                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```


Residuals vs Fitted
Residuals
5
-5
-15
10
12
14
16
18
Fitted values
127
6
131
Normal Q-Q
Standardized residuals
1
-1
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
127
6
131
Scale-Location
√|Standardized residuals|
2.0
1.0
0.0
10
12
14
16
18
Fitted values
131
127
6
Residuals vs Leverage
Standardized residuals
2
0
-2
-4
0.00
0.05
0.10
0.15
0.20
Leverage
76
131
Cook's distance
0.5
1

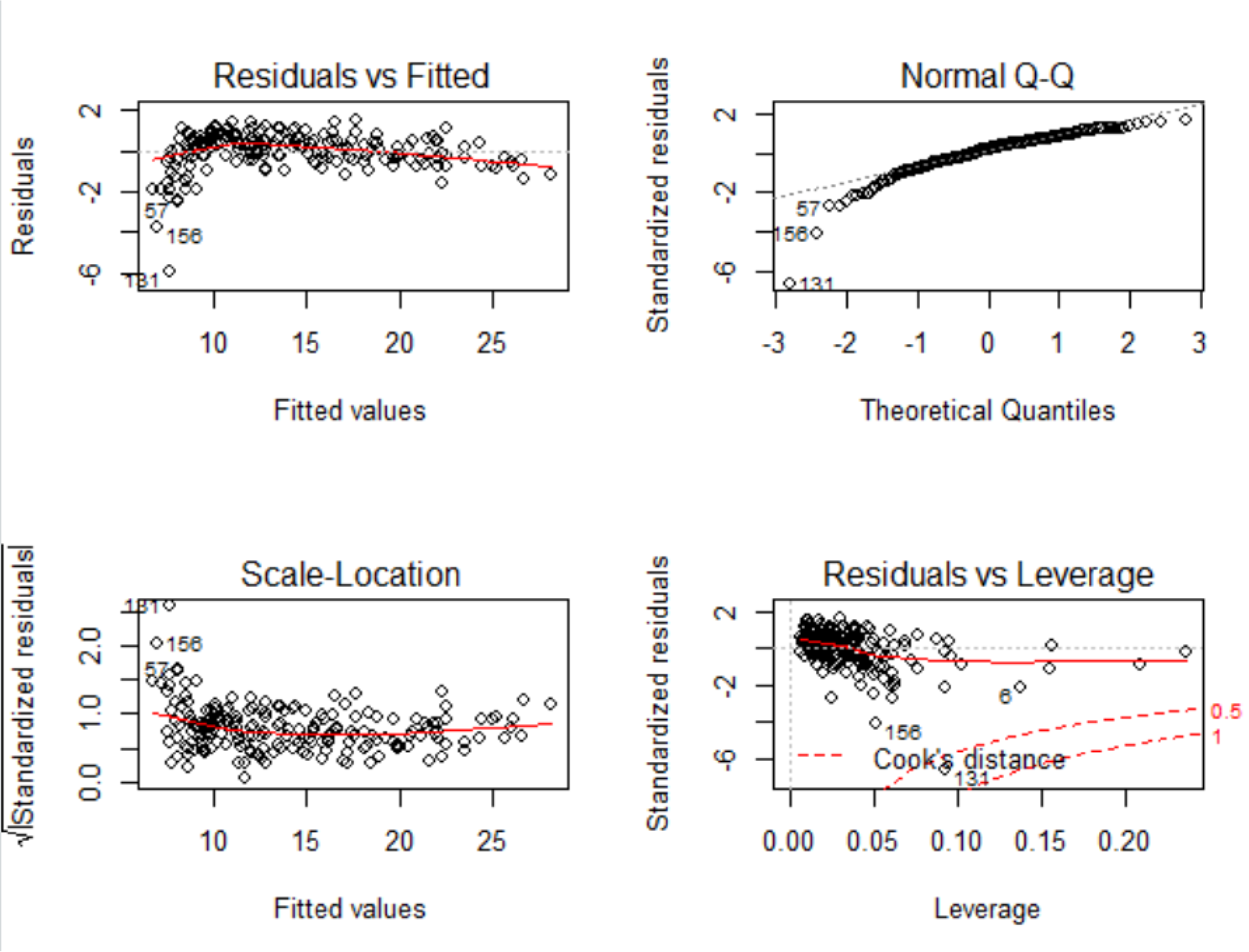
8) lmMod8



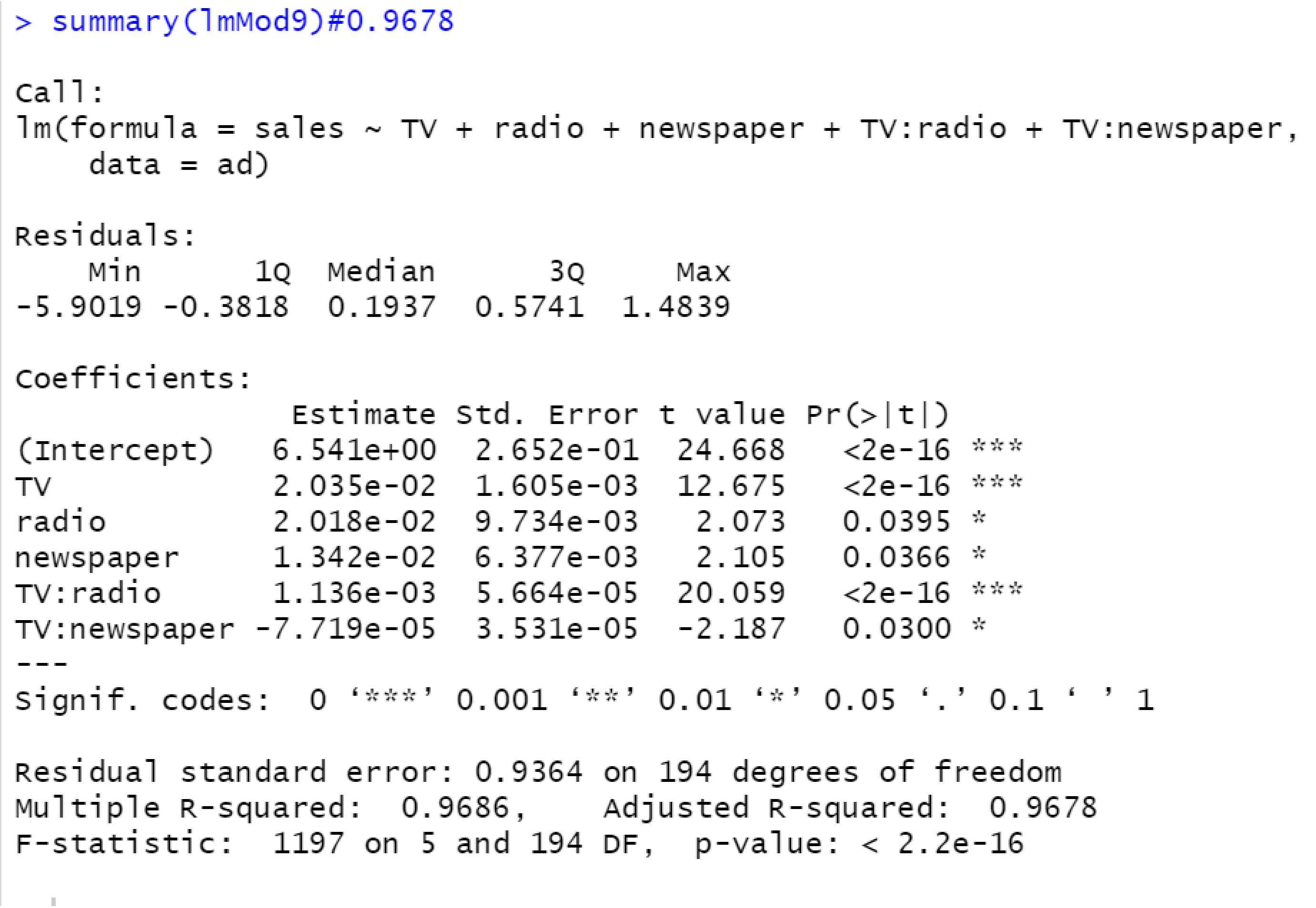
```
> summary(lmMod8)#0.9677

Call:
lm(formula = sales ~ (TV + radio + newspaper)^2, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9239 -0.3954  0.1873  0.5976  1.5267 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.460e+00  3.176e-01  20.342   <2e-16 ***
TV               2.033e-02  1.609e-03  12.633   <2e-16 ***
radio            2.293e-02  1.141e-02   2.009   0.0460 *  
newspaper        1.703e-02  1.007e-02   1.691   0.0924 .  
TV:radio         1.139e-03  5.716e-05  19.930   <2e-16 ***
TV:newspaper    -7.971e-05  3.579e-05  -2.227   0.0271 *  
radio:newspaper -1.096e-04  2.363e-04  -0.464   0.6433    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9383 on 193 degrees of freedom
Multiple R-squared:  0.9686,	Adjusted R-squared:  0.9677 
F-statistic: 993.3 on 6 and 193 DF,  p-value: < 2.2e-16
```



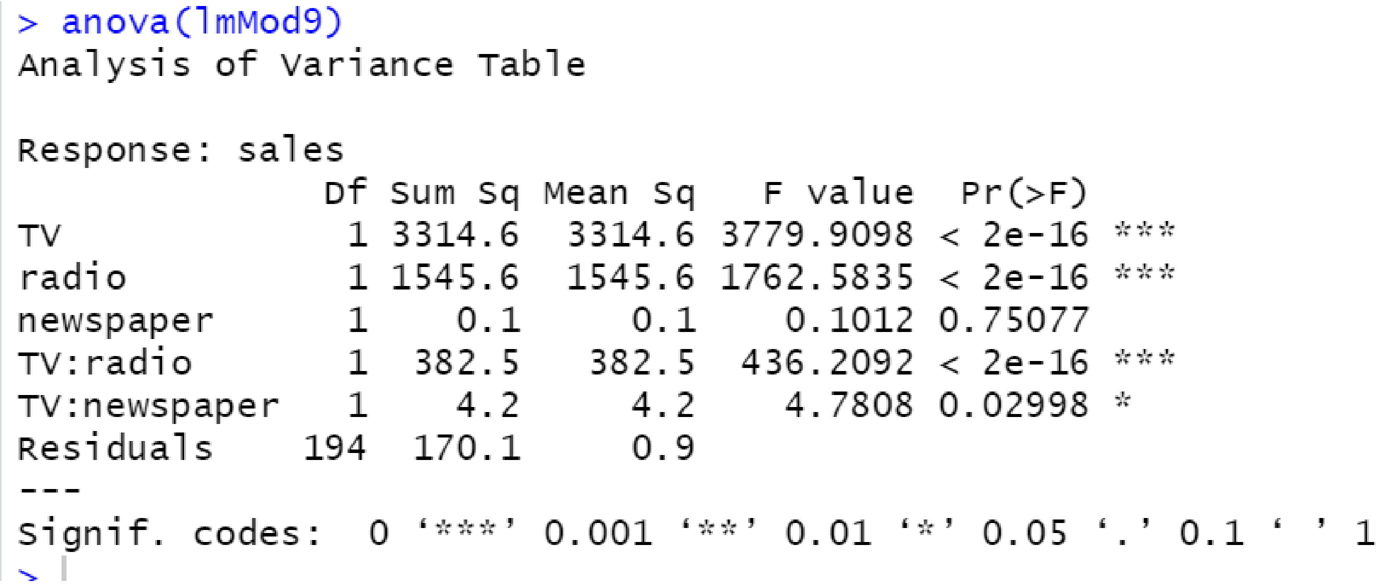
```
> anova(lmMod8)
Analysis of Variance Table

Response: sales
                 Df Sum Sq Mean Sq   F value  Pr(>F)    
TV                1 3314.6  3314.6 3764.6175 < 2e-16 ***
radio             1 1545.6  1545.6 1755.4527 < 2e-16 ***
newspaper         1    0.1     0.1    0.1008 0.75126    
TV:radio          1  382.5   382.5  434.4444 < 2e-16 ***
TV:newspaper      1    4.2     4.2    4.7615 0.03031 *  
radio:newspaper   1    0.2     0.2    0.2151 0.64329    
Residuals       193  169.9     0.9                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```


Residuals vs Fitted
Residuals
Fitted values
2
-2
-6
10
15
20
25
57
156
131
Normal Q-Q
Standardized residuals
Theoretical Quantiles
2
-2
-6
-3
-2
-1
0
1
2
3
57
156
131
Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
10
15
20
25
131
156
57
Residuals vs Leverage
Standardized residuals
Leverage
2
-2
-6
0.00
0.05
0.10
0.15
0.20
6
156
131
Cook's distance
0.5
1

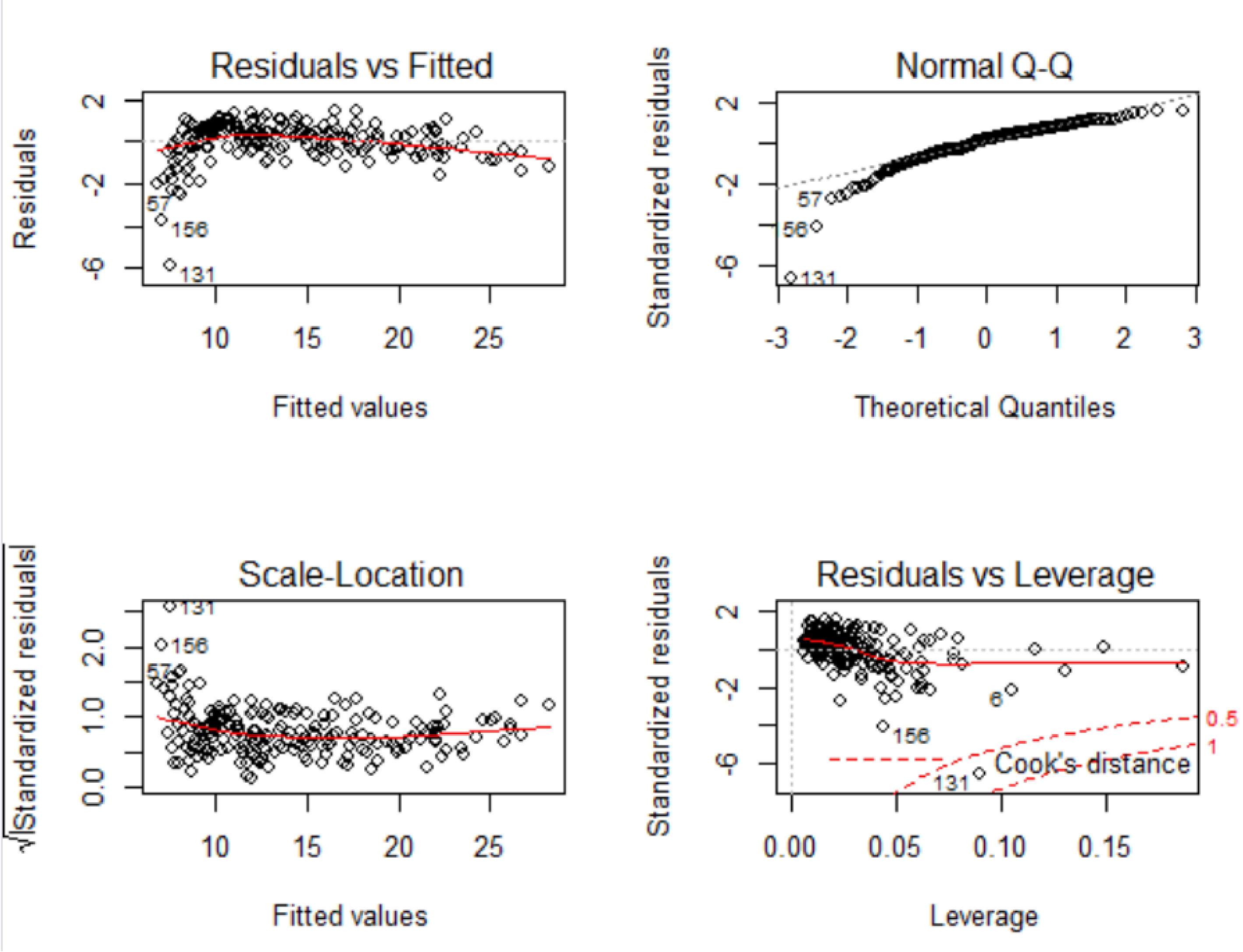
9) lmMod9



```
> summary(lmMod9)#0.9678

Call:
lm(formula = sales ~ TV + radio + newspaper + TV:radio + TV:newspaper, 
    data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9019 -0.3818  0.1937  0.5741  1.4839 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.541e+00  2.652e-01  24.668   <2e-16 ***
TV            2.035e-02  1.605e-03  12.675   <2e-16 ***
radio         2.018e-02  9.734e-03   2.073   0.0395 *  
newspaper     1.342e-02  6.377e-03   2.105   0.0366 *  
TV:radio      1.136e-03  5.664e-05  20.059   <2e-16 ***
TV:newspaper -7.719e-05  3.531e-05  -2.187   0.0300 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9364 on 194 degrees of freedom
Multiple R-squared:  0.9686,	Adjusted R-squared:  0.9678 
F-statistic:  1197 on 5 and 194 DF,  p-value: < 2.2e-16
```

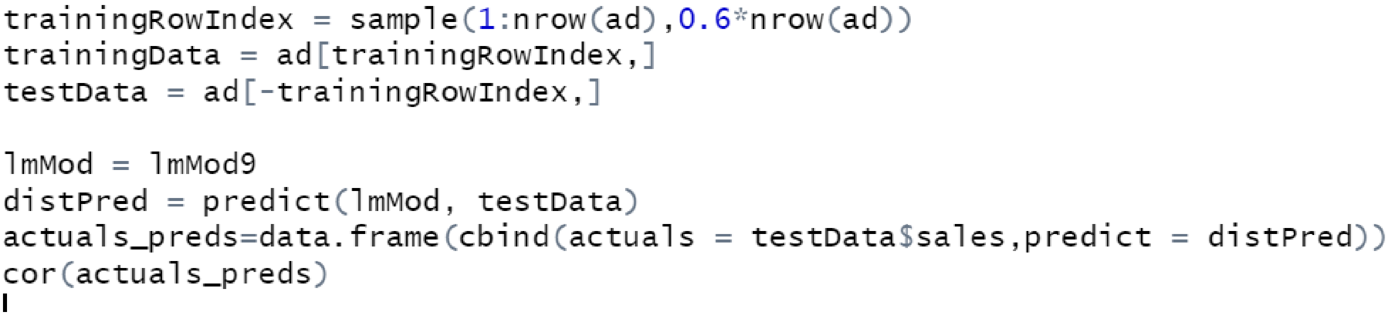


```
> anova(lmMod9)
Analysis of Variance Table

Response: sales
              Df Sum Sq Mean Sq   F value  Pr(>F)    
TV             1 3314.6  3314.6 3779.9098 < 2e-16 ***
radio          1 1545.6  1545.6 1762.5835 < 2e-16 ***
newspaper      1    0.1     0.1    0.1012 0.75077    
TV:radio       1  382.5   382.5  436.2092 < 2e-16 ***
TV:newspaper   1    4.2     4.2    4.7808 0.02998 *  
Residuals    194  170.1     0.9                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```


Residuals vs Fitted
Residuals
2
-2
-6
57
156
131
10
15
20
25
Fitted values
Normal Q-Q
Standardized residuals
2
-2
-6
57
56
131
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
131
156
57
10
15
20
25
Fitted values
Residuals vs Leverage
Standardized residuals
2
-2
-6
6
156
131
Cook's distance
0.5
1
0.00
0.05
0.10
0.15
Leverage

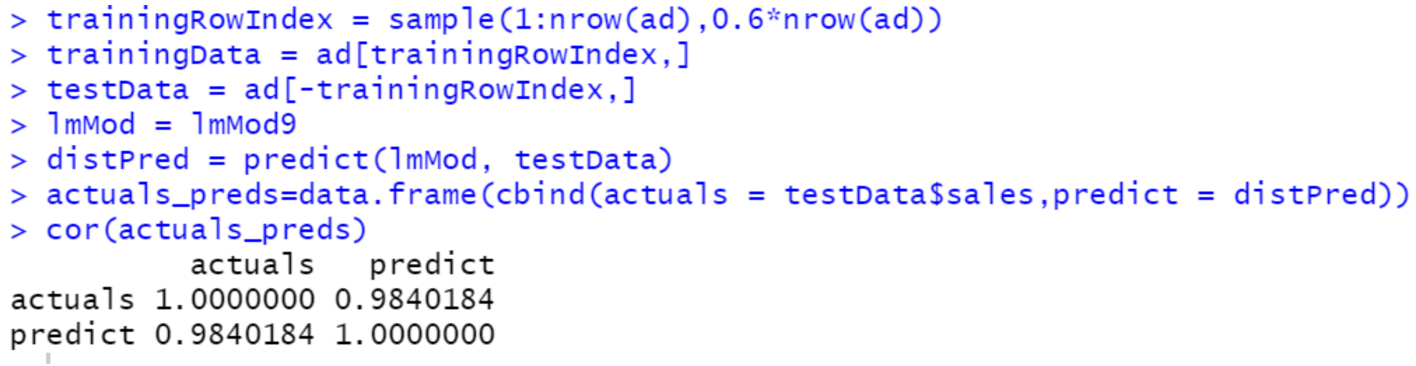
2. 예측 및 검정 수행

위에서 구한 내용을 이용해 가장 정확도가 큰 예측 모델을 구한다. Adjusted R-squared값으로 추적 회귀선이 관측값을 가장 잘 설명하는 모델을 찾아야 하므로, 각 모델 별 Adjusted R-squared 값을 이용한다. 수행 결과, lmMod9 모델이 가장 Adjusted R-Squared 값이 높음을 알 수 있다.



```
trainingRowIndex = sample(1:nrow(ad),0.6*nrow(ad))
trainingData = ad[trainingRowIndex,]
testData = ad[-trainingRowIndex,]

lmMod = lmMod9
distPred = predict(lmMod, testData)
actuals_preds=data.frame(cbind(actuals = testData$sales,predict = distPred))
cor(actuals_preds)
```

<그림 2>



```
> trainingRowIndex = sample(1:nrow(ad),0.6*nrow(ad))
> trainingData = ad[trainingRowIndex,]
> testData = ad[-trainingRowIndex,]
> lmMod = lmMod9
> distPred = predict(lmMod, testData)
> actuals_preds=data.frame(cbind(actuals = testData$sales,predict = distPred))
> cor(actuals_preds)
          actuals   predict
actuals 1.0000000 0.9840184
predict 0.9840184 1.0000000
```

<그림 3>

이후, 그림 2의 코드 처럼 6:4로 training data, testing data를 나누어 testing data를 이용해 예측을 수행, 예측 된 값에 대한 검정을 수행한다. 결과는 그림 3과 같다.