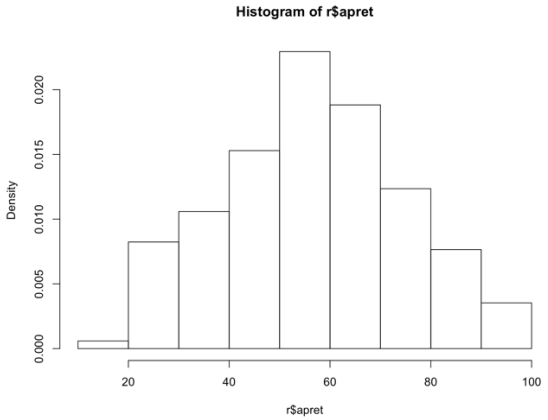
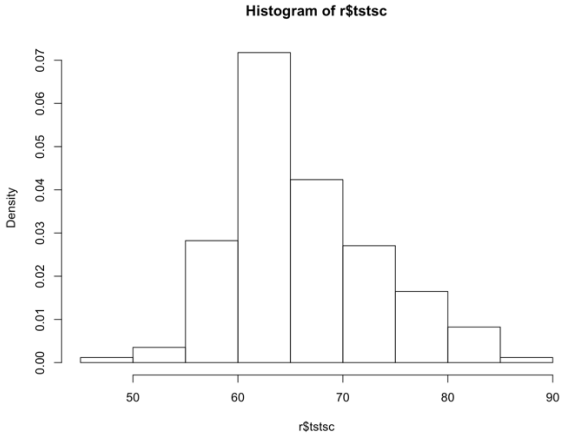
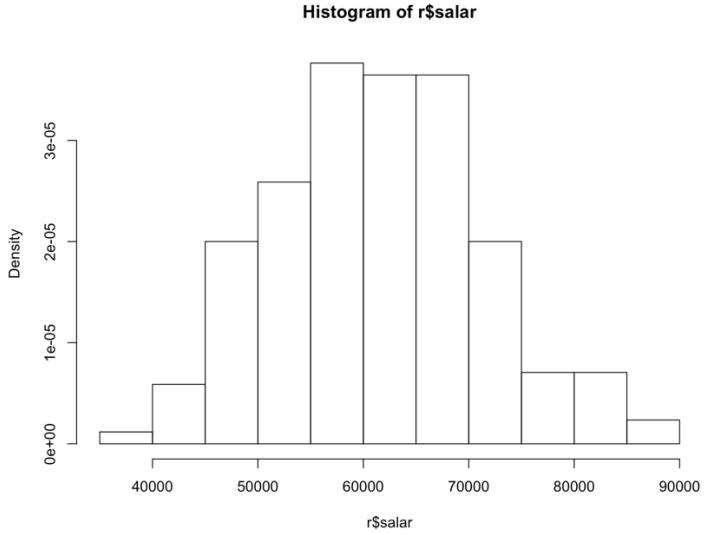
1. Descriptive statistics and plot histograms for three columns: apret, tstsc, and salar.

```
> summary(r)
     spend          apret          top10           rejr          tstsc           pacc      
 Min.   : 4125   Min.   :18.75   Min.   : 8.00   Min.   : 0.00   Min.   :48.12   Min.   : 8.964  
 1st Qu.: 7372   1st Qu.:45.37   1st Qu.:22.00   1st Qu.:19.17   1st Qu.:61.11   1st Qu.:33.904  
 Median : 9265   Median :55.71   Median :30.00   Median :27.39   Median :64.78   Median :40.850  
 Mean   :10975   Mean   :56.72   Mean   :38.46   Mean   :30.65   Mean   :66.16   Mean   :43.173  
 3rd Qu.:12838   3rd Qu.:68.69   3rd Qu.:49.50   3rd Qu.:36.81   3rd Qu.:70.45   3rd Qu.:51.773  
 Max.   :35863   Max.   :95.25   Max.   :98.00   Max.   :84.07   Max.   :87.50   Max.   :76.253  
     strat           salar       
 Min.   : 7.20   Min.   :38640  
 1st Qu.:13.40   1st Qu.:54650  
 Median :16.00   Median :61150  
 Mean   :16.09   Mean   :61358  
 3rd Qu.:18.57   3rd Qu.:67100  
 Max.   :29.20   Max.   :87900  
```

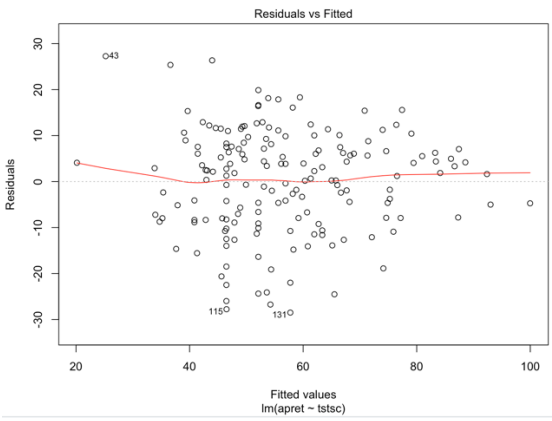


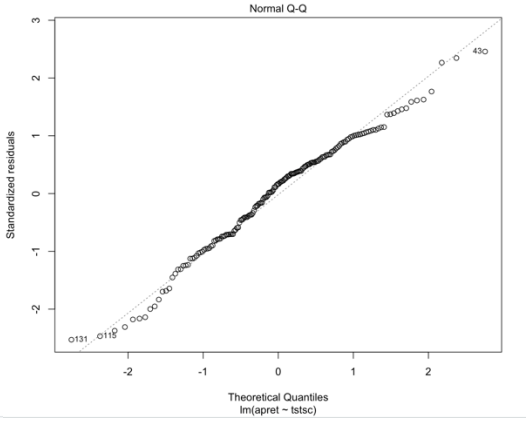


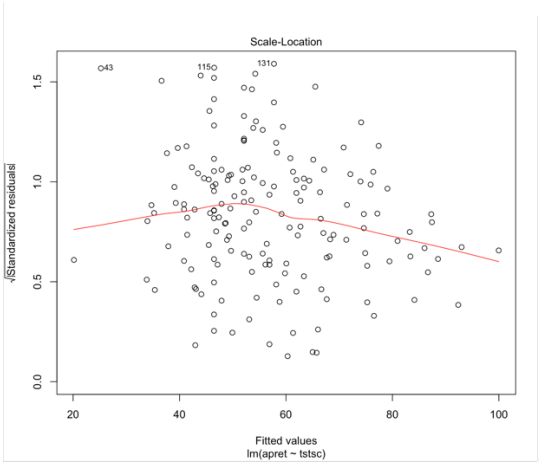


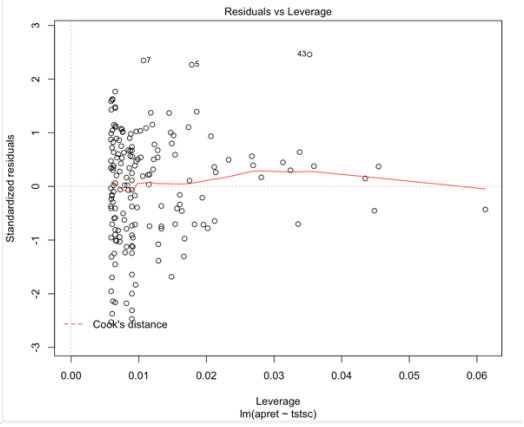
2. Linear regression of apret on tstsc.

```
> lm.apret_tstsc <- lm(apret ~ tstsc, data = r)
> summary(lm.apret_tstsc)

Call:
lm(formula = apret ~ tstsc, data = r)

Residuals:
    Min      1Q  Median      3Q     Max
-28.490  -7.957   1.857   7.552  27.278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -77.3999     8.2878  -9.339   <2e-16 ***
tstsc         2.0271     0.1246  16.272   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.3 on 168 degrees of freedom
Multiple R-squared:  0.6118,    Adjusted R-squared:  0.6095
F-statistic: 264.8 on 1 and 168 DF,  p-value: < 2.2e-16
```



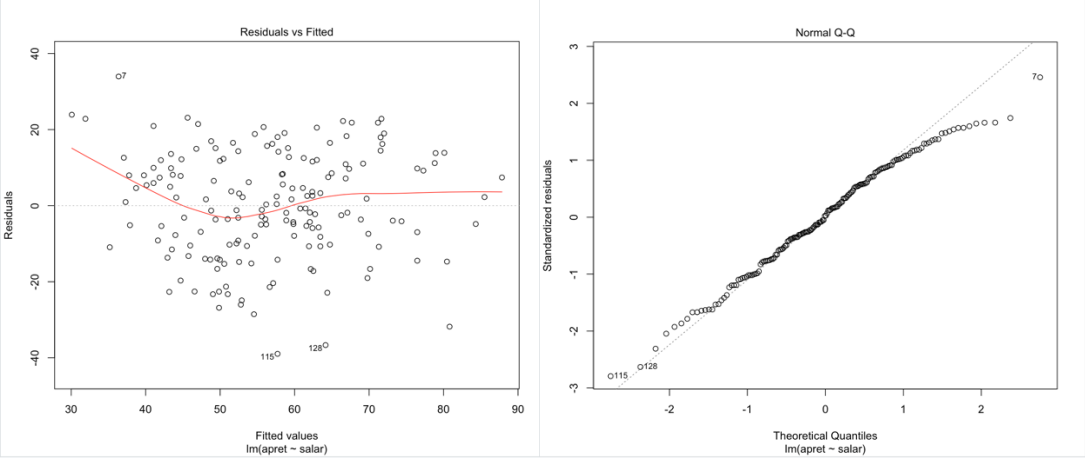


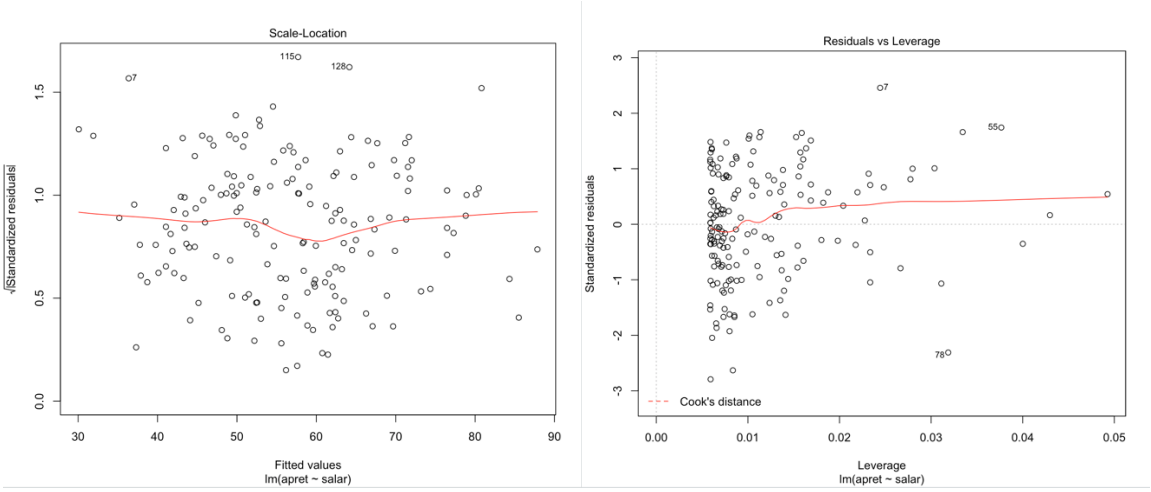




3. Linear regression of apret on salar.

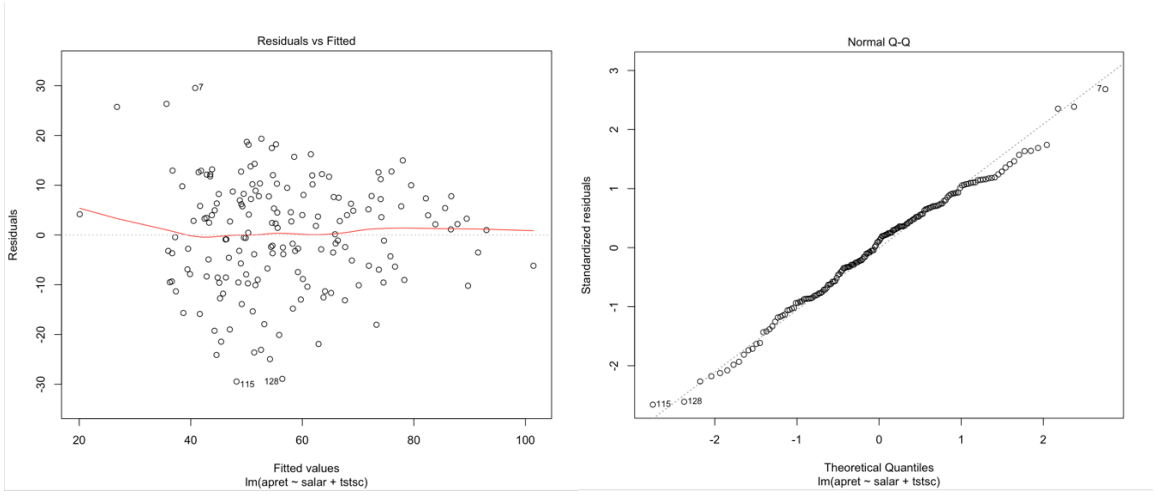
```
> lm.apret_salar <- lm(apret ~ salar, data = r)
> summary(lm.apret_salar)

Call:
lm(formula = apret ~ salar, data = r)

Residuals:
    Min      1Q  Median      3Q     Max
-38.959 -10.170   0.362  11.151  33.965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.522e+01  6.823e+00  -2.231    0.027 *
salar        1.173e-03  1.098e-04  10.678   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.99 on 168 degrees of freedom
Multiple R-squared:  0.4043,    Adjusted R-squared:  0.4008
F-statistic:   114 on 1 and 168 DF,  p-value: < 2.2e-16
```

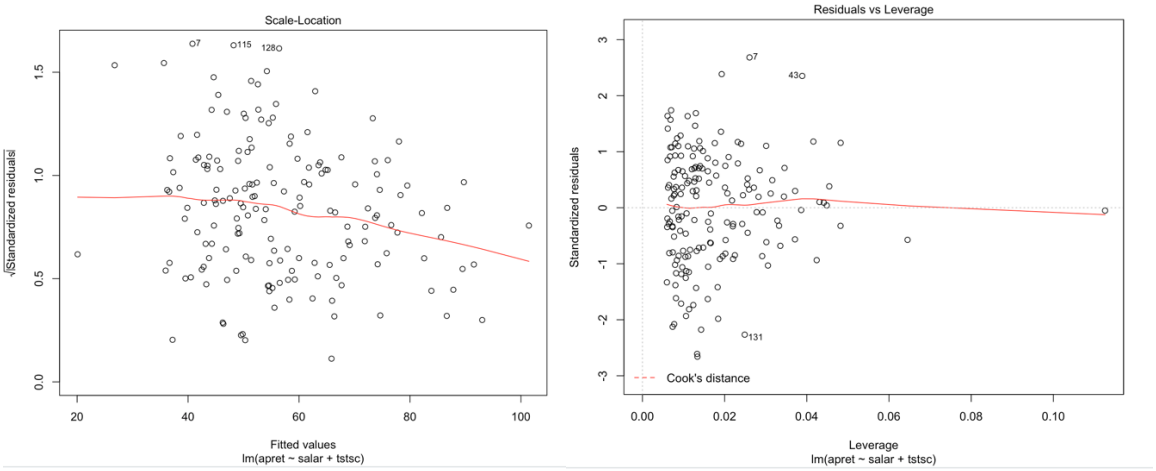




4. Linear regression of apret on tstsc and salar.

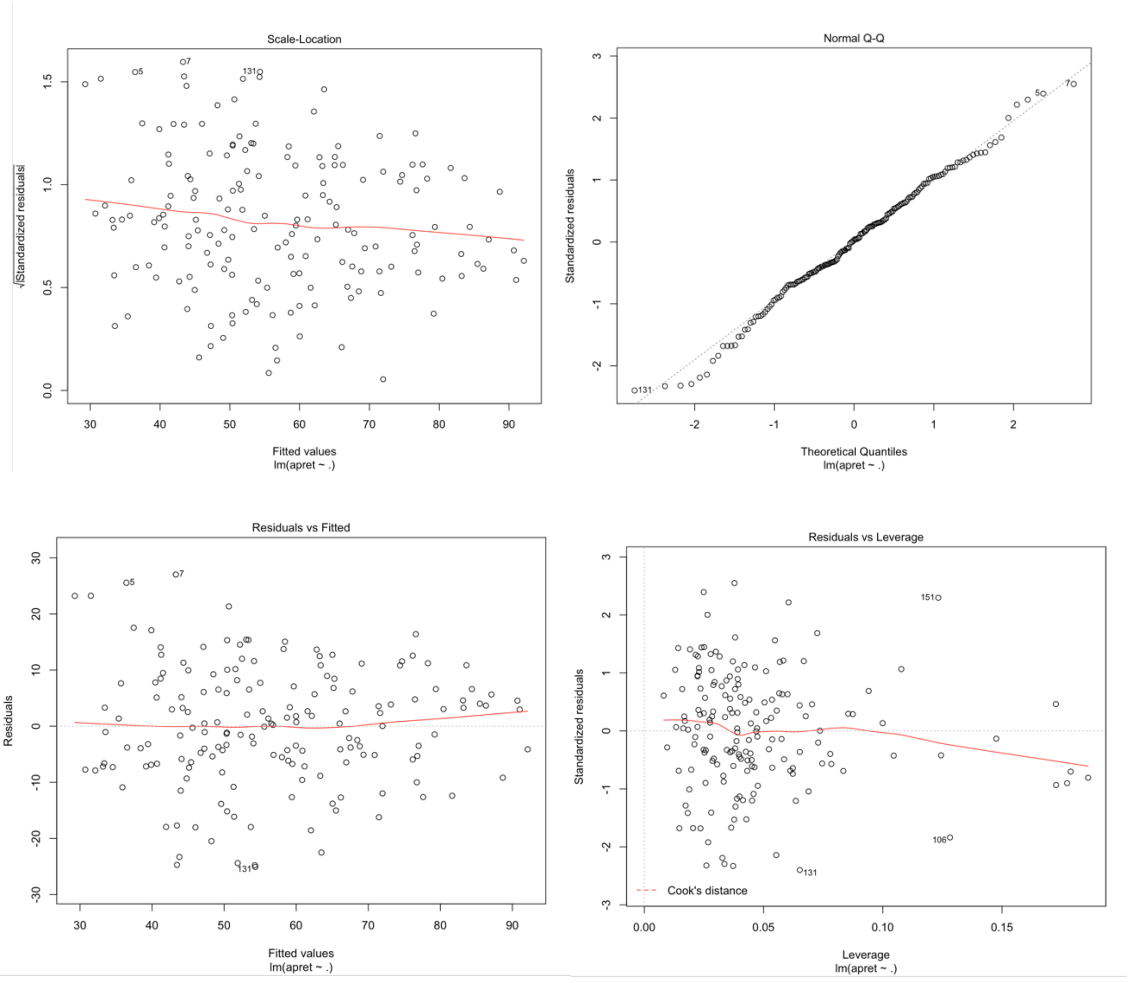
```
> lm.apret_salar_tstsc <- lm(apret ~ salar+tstsc, data = r)
> summary(lm.apret_salar_tstsc)

Call:
lm(formula = apret ~ salar + tstsc, data = r)

Residuals:
    Min      1Q  Median      3Q     Max
-29.458  -7.915   1.270   7.777  29.538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.591e+01  8.210e+00  -9.246   <2e-16 ***
salar        2.880e-04  1.253e-04   2.298   0.0228 *
tstsc        1.738e+00  1.761e-01   9.868   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.16 on 167 degrees of freedom
Multiple R-squared:  0.6237,    Adjusted R-squared:  0.6192
F-statistic: 138.4 on 2 and 167 DF,  p-value: < 2.2e-16
```





5. Linear regression of apret on all elements.

```
> lm.apret_all <- lm(apret ~ ., data = r)
> summary(lm.apret_all)

Call:
lm(formula = apret ~ ., data = r)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0710  -6.5692   0.3415   7.0232  27.0360

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.941e+01  1.576e+01  -2.501  0.01338 *
spend       -2.544e-04  2.804e-04  -0.907  0.36559
top10        4.357e-02  6.735e-02   0.647  0.51864
rejr         4.013e-02  7.094e-02   0.566  0.57233
tstsc        1.606e+00  2.461e-01   6.524 8.34e-10 ***
pacc        -2.151e-01  7.126e-02  -3.019  0.00295 **
strat       -6.312e-01  2.745e-01  -2.299  0.02278 *
salar        1.502e-04  1.468e-04   1.023  0.30766
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.81 on 162 degrees of freedom
Multiple R-squared:  0.6571,    Adjusted R-squared:  0.6422
F-statistic: 44.34 on 7 and 162 DF,  p-value: < 2.2e-16
```



Optimize model:

```
> anova(lm.apret_all)
Analysis of Variance Table

Response: apret
           Df  Sum Sq Mean Sq  F value    Pr(>F)
spend       1 19963.1 19963.1 170.7534 < 2.2e-16 ***
top10       1  5671.0  5671.0  48.5064 7.817e-11 ***
rejr        1   176.9   176.9   1.5127 0.2205034
tstsc       1  8191.4  8191.4  70.0650 2.553e-14 ***
pacc        1  1609.5  1609.5  13.7665 0.0002841 ***
strat       1   552.2   552.2   4.7231 0.0312127 *
salar       1   122.4   122.4   1.0473 0.3076625
Residuals 162 18939.7   116.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It is seen that salar has the least contribution to reducing the model fitting error, thus removing it from the model.

```
> lm2.apret_all <- update(lm.apret_all, . ~ . -salar)
> summary(lm2.apret_all)

Call:
lm(formula = apret ~ spend + top10 + rejr + tstsc + pacc + strat,
    data = r)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2264  -6.2788   0.2395   6.8953  25.6760

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.695e+01  1.558e+01  -2.372 0.018855 *
spend       -1.602e-04  2.649e-04  -0.605 0.546069
top10        3.863e-02  6.719e-02   0.575 0.566115
rejr         5.580e-02  6.928e-02   0.805 0.421754
tstsc        1.695e+00  2.300e-01   7.371 8.08e-12 ***
pacc        -2.421e-01  6.621e-02  -3.657 0.000344 ***
strat       -5.903e-01  2.717e-01  -2.173 0.031228 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.81 on 163 degrees of freedom
Multiple R-squared:  0.6548,    Adjusted R-squared:  0.6421
F-statistic: 51.54 on 6 and 163 DF,  p-value: < 2.2e-16
```

The model's fitting index has not improved. The following is a formal comparison of the two models with anova().

```
> anova(lm.apret_all,lm2.apret_all)
Analysis of Variance Table

Model 1: apret ~ spend + top10 + rejr + tstsc + pacc + strat + salar
Model 2: apret ~ spend + top10 + rejr + tstsc + pacc + strat
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    162 18940
2    163 19062 -1   -122.44 1.0473 0.3077
```

Although the sum of squared errors is reduced by 122, the probability of the two models differing is 70%. So continue to eliminate candidate coefficients. A new linear model is obtained by using the backward elimination method for the first-time model.

```
> final.lm <- step(lm.apret_all)
Start:  AIC=817.25
apret ~ spend + top10 + rejr + tstsc + pacc + strat + salar

       Df Sum of Sq   RSS    AIC
- rejr   1      37.4 18977 815.58
- top10  1      48.9 18989 815.69
- spend  1      96.2 19036 816.11
- salar  1     122.4 19062 816.34
<none>               18940 817.25
- strat  1     618.0 19558 820.70
- pacc   1    1065.5 20005 824.55
- tstsc  1    4976.2 23916 854.91

Step:  AIC=815.58
apret ~ spend + top10 + tstsc + pacc + strat + salar

       Df Sum of Sq   RSS    AIC
- top10  1      73.0 19050 814.24
- spend  1      76.1 19053 814.26
- salar  1     160.9 19138 815.02
<none>               18977 815.58
- strat  1     606.4 19584 818.93
- pacc   1    1028.3 20005 822.55
- tstsc  1    5031.5 24009 853.56

Step:  AIC=814.24
apret ~ spend + tstsc + pacc + strat + salar

       Df Sum of Sq   RSS    AIC
- spend  1      38.6 19089 812.58
- salar  1     155.5 19206 813.62
<none>               19050 814.24
- strat  1     533.5 19584 816.93
- pacc   1    1093.9 20144 821.73
- tstsc  1    9414.6 28465 880.51
```

```
Step:  AIC=812.58
apret ~ tstsc + pacc + strat + salar

       Df Sum of Sq   RSS    AIC
- salar  1     118.5 19207 811.63
<none>               19089 812.58
- strat  1     504.9 19594 815.02
- pacc   1    1093.9 20183 820.05
- tstsc  1   10011.2 29100 882.26

Step:  AIC=811.63
apret ~ tstsc + pacc + strat

       Df Sum of Sq   RSS    AIC
<none>               19207 811.63
- strat  1     505.1 19712 814.04
- pacc   1    1602.6 20810 823.26
- tstsc  1   21116.2 40323 935.71
```

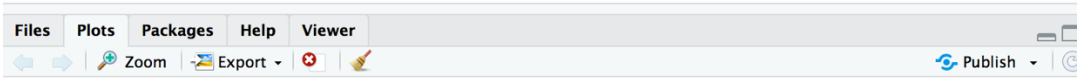
```
> summary(final.lm)

Call:
lm(formula = apret ~ tstsc + pacc + strat, data = r)

Residuals:
     Min       1Q   Median       3Q      Max
-24.8362  -6.6729   0.1956   7.1710  25.5527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.72923   11.71070  -3.905 0.000137 ***
tstsc         1.82293    0.13494  13.509  < 2e-16 ***
pacc         -0.23870    0.06414  -3.722 0.000271 ***
strat        -0.48843    0.23378  -2.089 0.038210 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.76 on 166 degrees of freedom
Multiple R-squared:  0.6522,    Adjusted R-squared:  0.6459
F-statistic: 103.8 on 3 and 166 DF,  p-value: < 2.2e-16
```

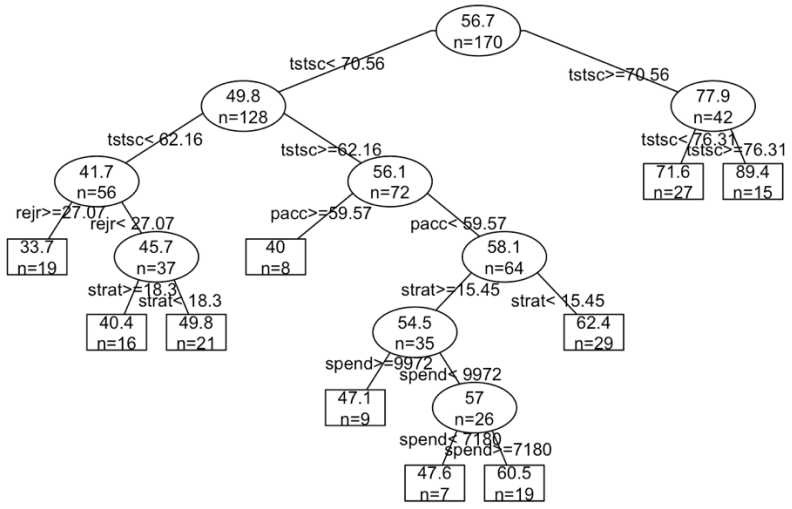
So it seems that the model is not good enough using linear regression, so than we will try another way to deal with it. Using regression tree.

Regression Tree

```
> rt.apret <- rpart(apret ~ ., data = r)
> rt.apret
n= 170

node), split, n, deviance, yval
      * denotes terminal node

 1) root 170 55226.0600 56.72108
   2) tstsc< 70.5625 128 24836.6000 49.76237
     4) tstsc< 62.1565 56  7910.8670 41.66666
       8) rejr>=27.066 19  2199.1570 33.72805 *
       9) rejr< 27.066 37  3899.4160 45.74324
        18) strat>=18.3 16  1420.7130 40.41144 *
        19) strat< 18.3 21  1677.3000 49.80557 *
     5) tstsc>=62.1565 72 10400.8100 56.05903
      10) pacc>=59.567 8   501.5547 40.03125 *
      11) pacc< 59.567 64  7587.2500 58.06250
        22) strat>=15.45 35  4351.3930 54.46429
          44) spend>=9972.5 9  1250.7500 47.08333 *
          45) spend< 9972.5 26  2440.6150 57.01923
            90) spend< 7179.5 7   838.5893 47.57143 *
            91) spend>=7179.5 19   747.0000 60.50000 *
        23) strat< 15.45 29  2235.8020 62.40517 *
   3) tstsc>=70.5625 42  5301.4110 77.92857
     6) tstsc< 76.3125 27  1973.8240 71.56481 *
     7) tstsc>=76.3125 15   265.9833 89.38333 *
 '
```

```
> prettyTree(rt.apret)
>
```





```
> summary(rt.apret)
Call:
rpart(formula = apret ~ ., data = r)
  n= 170
          CP nsplit rel error       xerror            xstd
1 0.45427927      0 1.0000000 1.0133632 0.09322742
2 0.11814924      1 0.5457207 0.6040886 0.05288551
3 0.05543765      2 0.4275715 0.5179635 0.04714462
4 0.04186442      3 0.3721338 0.4931417 0.04734799
5 0.03281592      4 0.3302694 0.4990725 0.05076818
6 0.01810840      5 0.2974535 0.4941476 0.05391125
7 0.01451132      6 0.2793451 0.5028399 0.05774678
8 0.01371684      7 0.2648338 0.5106797 0.05774691
9 0.01000000      9 0.2374001 0.5044740 0.05794875

Variable importance
tstsc top10 spend salar  rejr strat  pacc
   29    16    15    15    12     8     5

Node number 1: 170 observations,       complexity param=0.4542793
  mean=56.72108, MSE=324.8592
  left son=2 (128 obs) right son=3 (42 obs)
  Primary splits:
      tstsc < 70.5625 to the left,  improve=0.4542793, (0 missing)
      salar < 62250   to the left,  improve=0.3756453, (0 missing)
      top10 < 32.5    to the left,  improve=0.3520516, (0 missing)
      rejr  < 52.909  to the left,  improve=0.3072566, (0 missing)
      spend < 11411.5 to the left,  improve=0.2729790, (0 missing)
  Surrogate splits:
      top10 < 44.5    to the left,  agree=0.894, adj=0.571, (0 split)
      spend < 16455.5 to the left,  agree=0.876, adj=0.500, (0 split)
      salar < 66088   to the left,  agree=0.876, adj=0.500, (0 split)
      rejr  < 51.876  to the left,  agree=0.847, adj=0.381, (0 split)
      strat < 13.25   to the right, agree=0.824, adj=0.286, (0 split)

Node number 2: 128 observations,       complexity param=0.1181492
```

mean=49.76237, MSE=194.0359
left son=4 (56 obs) right son=5 (72 obs)
Primary splits:
tstsc < 62.1565 to the left, improve=0.2627138, (0 missing)
pacc < 47.846 to the right, improve=0.2133163, (0 missing)
salar < 59550 to the left, improve=0.1871468, (0 missing)
spend < 7234 to the left, improve=0.1801977, (0 missing)
top10 < 27.5 to the left, improve=0.1397898, (0 missing)
Surrogate splits:
top10 < 22.5 to the left, agree=0.797, adj=0.536, (0 split)
salar < 55550 to the left, agree=0.742, adj=0.411, (0 split)
spend < 7257 to the left, agree=0.695, adj=0.304, (0 split)
pacc < 41.7095 to the right, agree=0.648, adj=0.196, (0 split)
strat < 17.85 to the right, agree=0.625, adj=0.143, (0 split)

.....

Node number 45: 26 observations, complexity param=0.01371684
mean=57.01923, MSE=93.86982
left son=90 (7 obs) right son=91 (19 obs)
Primary splits:
spend < 7179.5 to the left, improve=0.35033220, (0 missing)
salar < 60950 to the left, improve=0.17991530, (0 missing)
strat < 16.9 to the left, improve=0.09747433, (0 missing)
pacc < 39.0555 to the right, improve=0.07759316, (0 missing)
rejr < 25.609 to the right, improve=0.04889815, (0 missing)
Surrogate splits:
rejr < 13.677 to the left, agree=0.846, adj=0.429, (0 split)
salar < 52150 to the left, agree=0.846, adj=0.429, (0 split)
pacc < 51.381 to the right, agree=0.769, adj=0.143, (0 split)
strat < 21.5 to the right, agree=0.769, adj=0.143, (0 split)
Node number 90: 7 observations
mean=47.57143, MSE=119.7985

Node number 91: 19 observations
mean=60.5, MSE=39.31579

```
> printcp(rt.apret)

Regression tree:
rpart(formula = apret ~ ., data = r)

Variables actually used in tree construction:
[1] pacc  rejr  spend strat tstsc

Root node error: 55226/170 = 324.86

n= 170

        CP nsplit rel error  xerror     xstd
1 0.454279      0   1.00000 1.01336 0.093227
2 0.118149      1   0.54572 0.60409 0.052886
3 0.055438      2   0.42757 0.51796 0.047145
4 0.041864      3   0.37213 0.49314 0.047348
5 0.032816      4   0.33027 0.49907 0.050768
6 0.018108      5   0.29745 0.49415 0.053911
7 0.014511      6   0.27935 0.50284 0.057747
8 0.013717      7   0.26483 0.51068 0.057747
9 0.010000      9   0.23740 0.50447 0.057949
```

Pruning

```
> (rt.apret_prun <- rpartXse(apret ~ ., data = r))
n= 170

node), split, n, deviance, yval
      * denotes terminal node

1) root 170 55226.0600 56.72108
  2) tstsc< 70.5625 128 24836.6000 49.76237
    4) tstsc< 62.1565 56  7910.8670 41.66666 *
    5) tstsc>=62.1565 72 10400.8100 56.05903 *
  3) tstsc>=70.5625 42  5301.4110 77.92857
    6) tstsc< 76.3125 27  1973.8240 71.56481 *
    7) tstsc>=76.3125 15   265.9833 89.38333 *
```