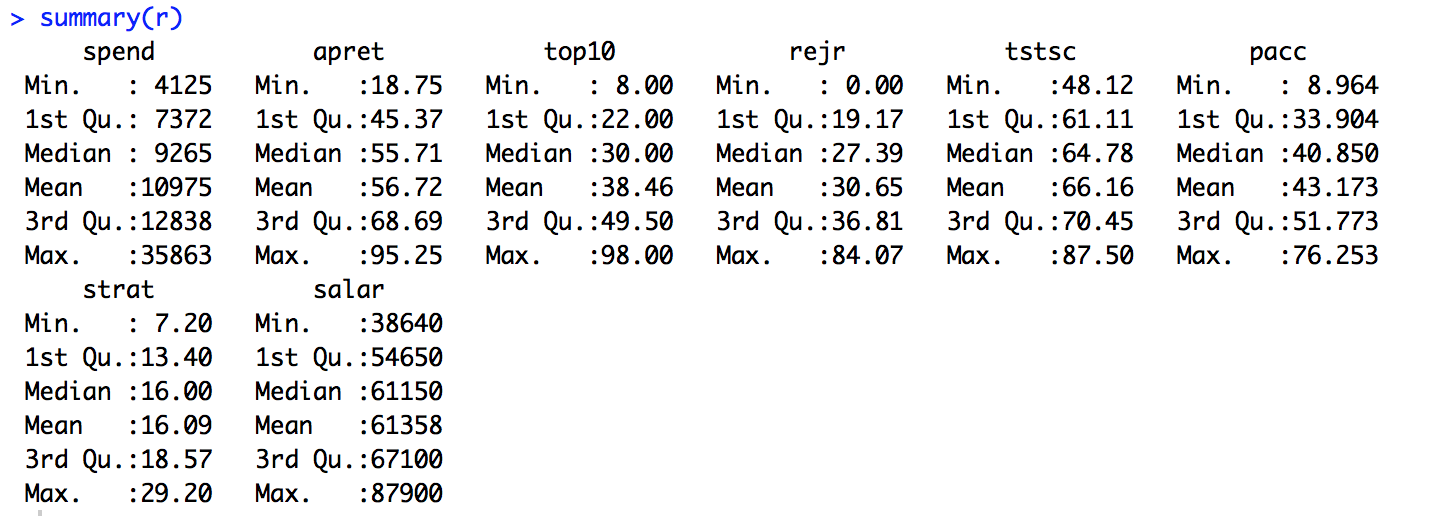
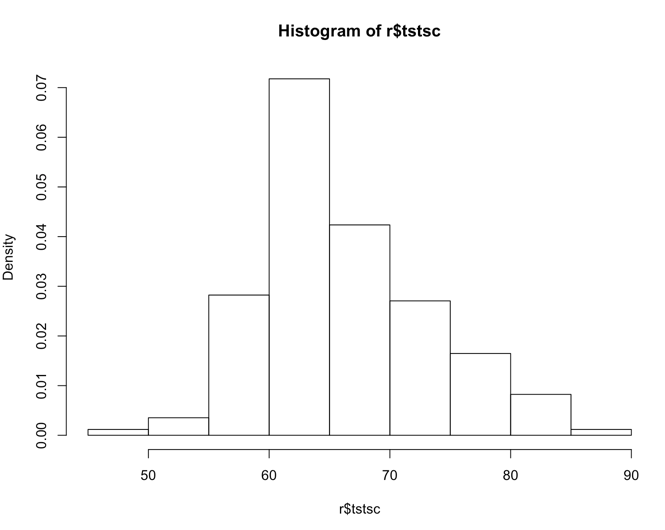
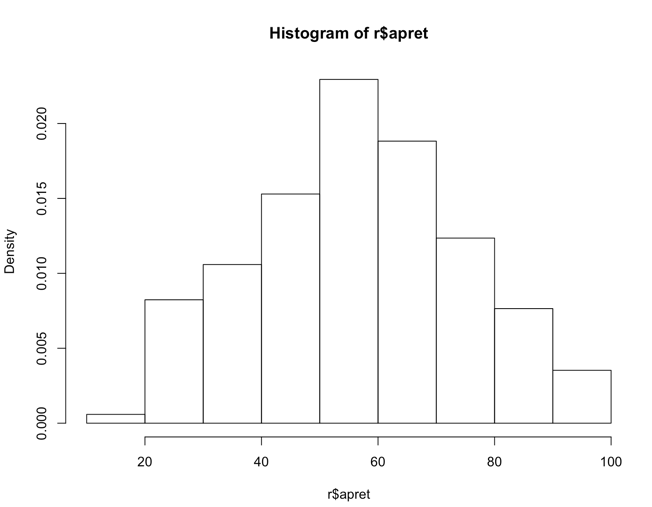
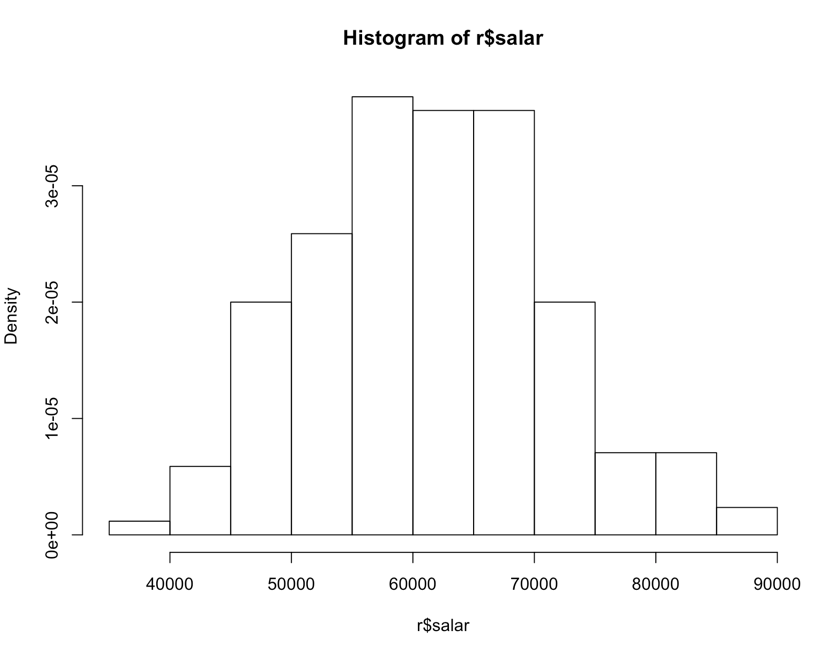
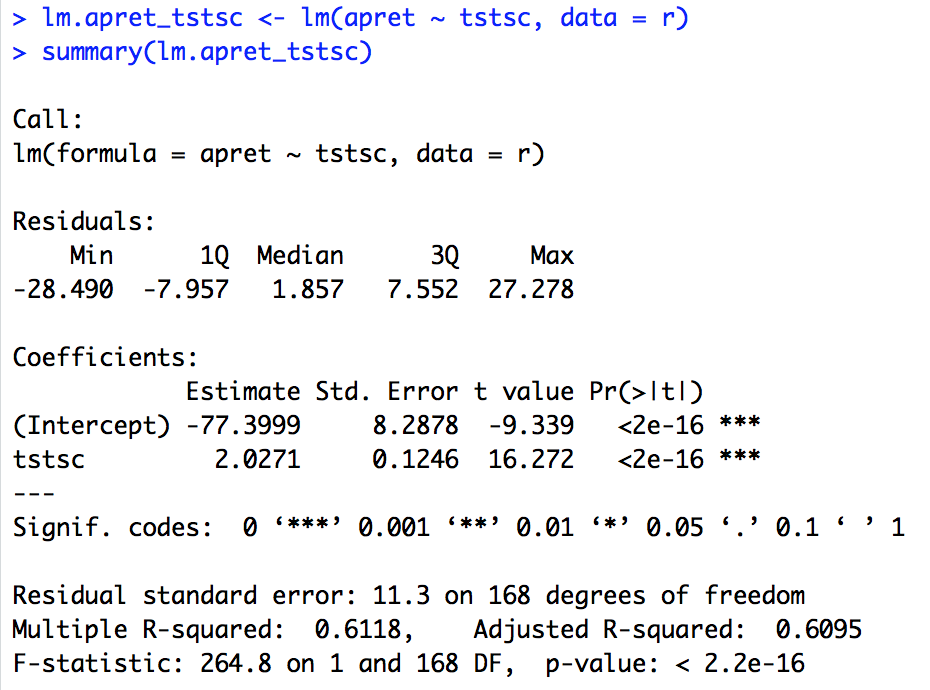
1. Descriptive statistics and plot histograms for three columns: apret, tstsc, and salar.

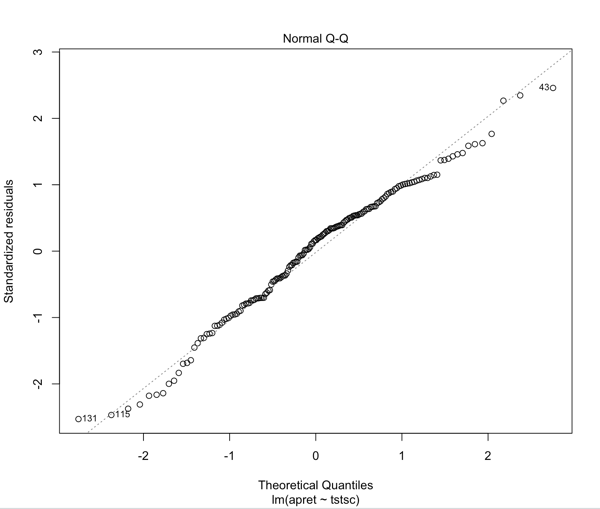
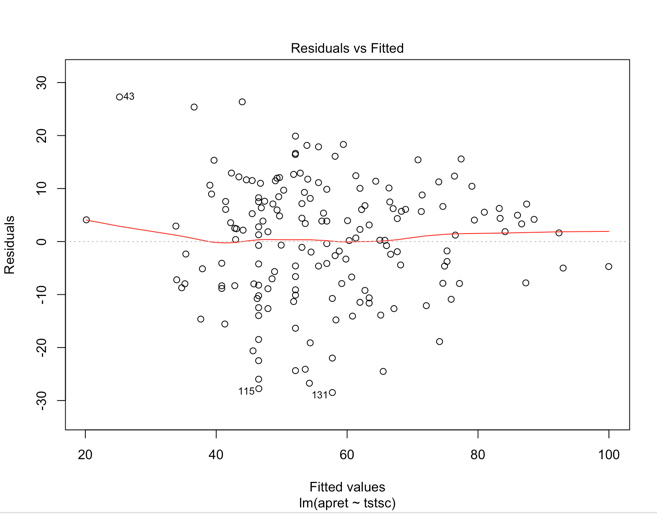


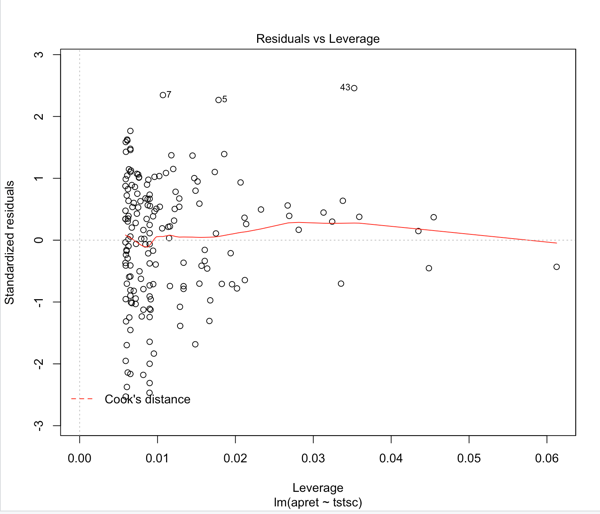
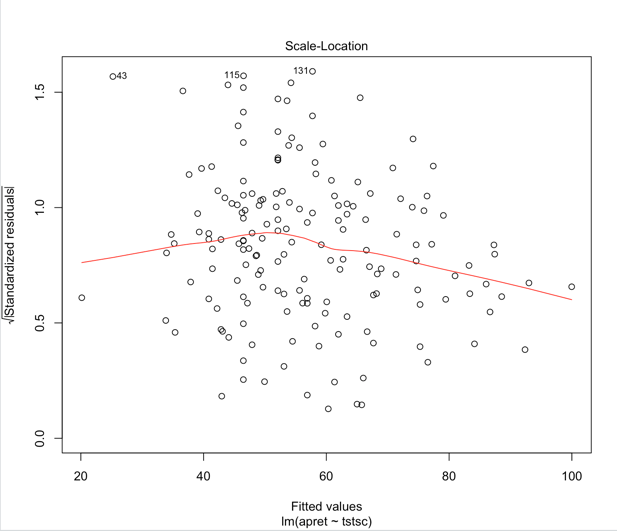




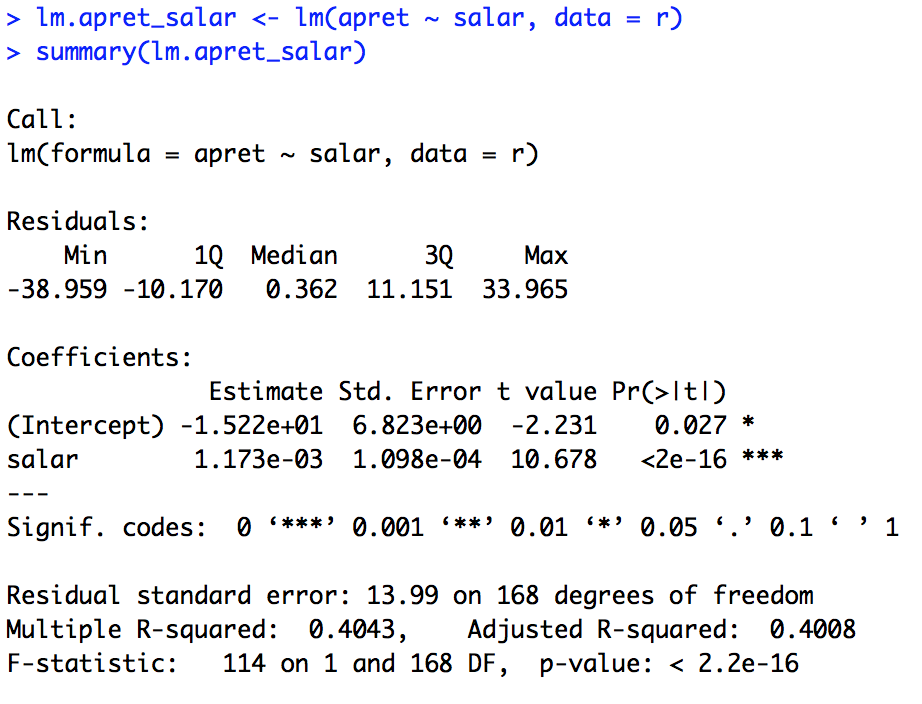
1. Linear regression of apret on tstsc.

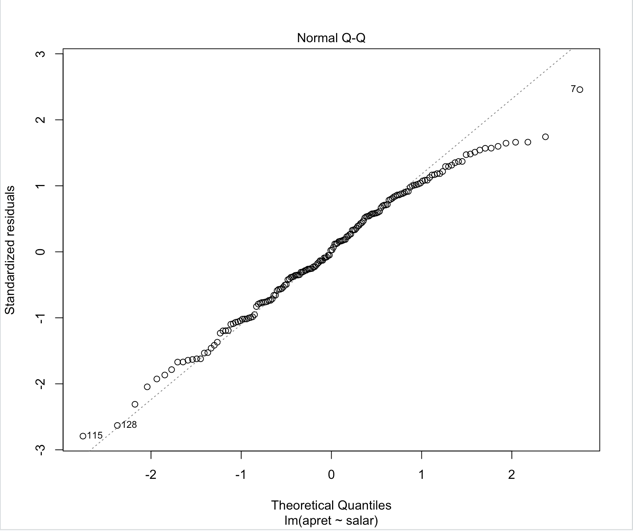
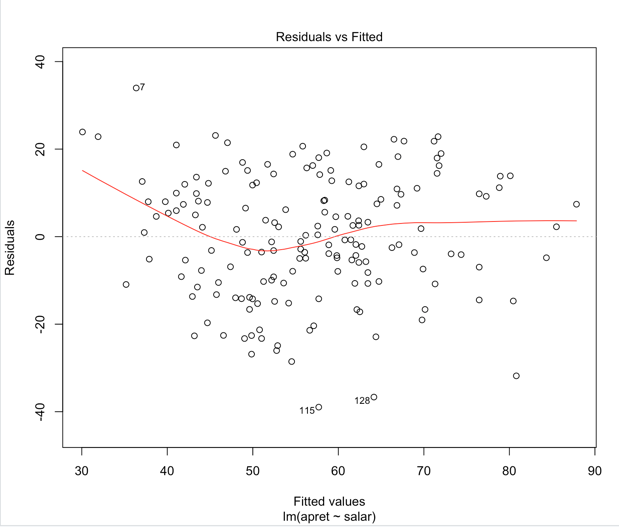


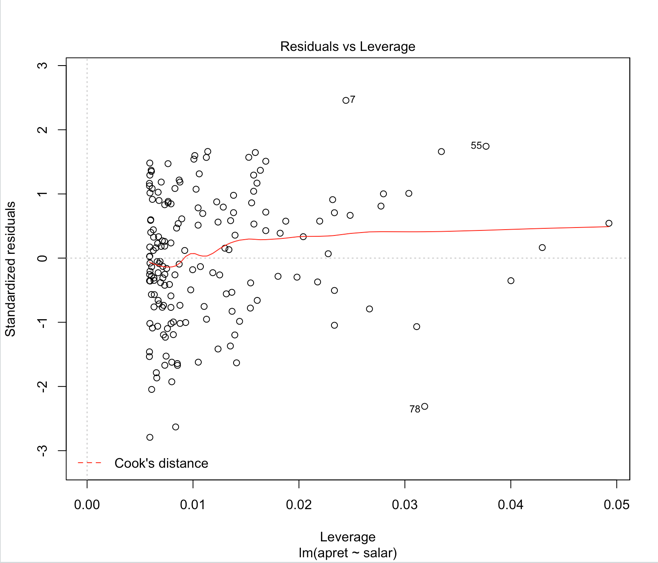
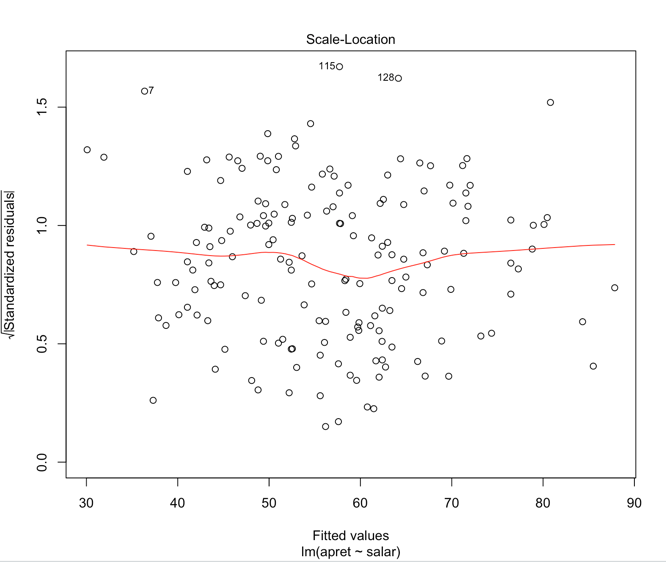




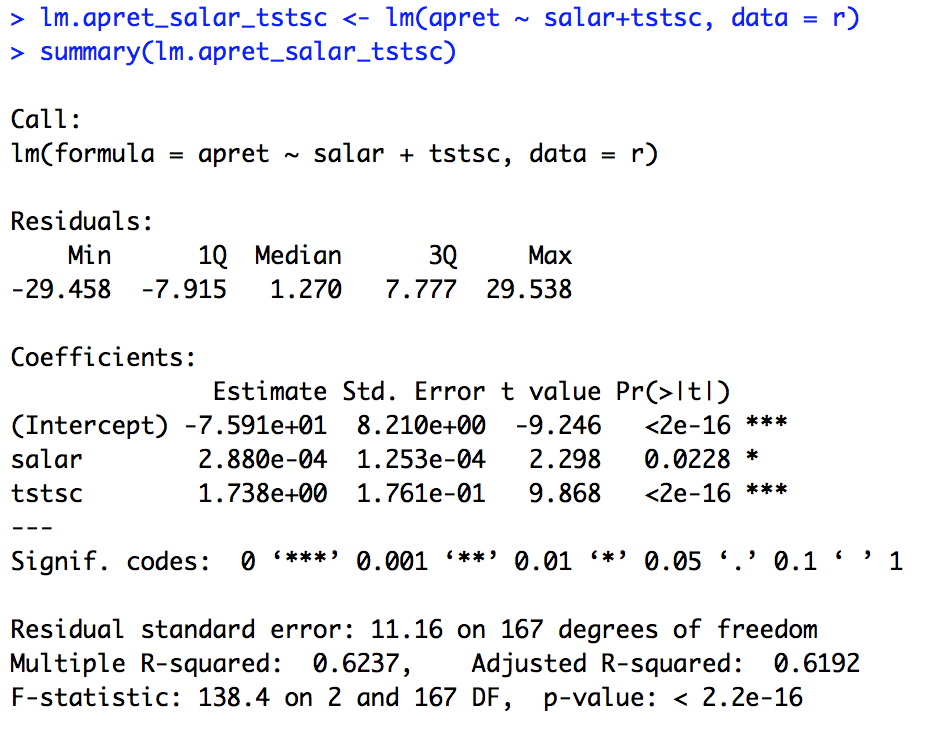
1. Linear regression of apret on salar.

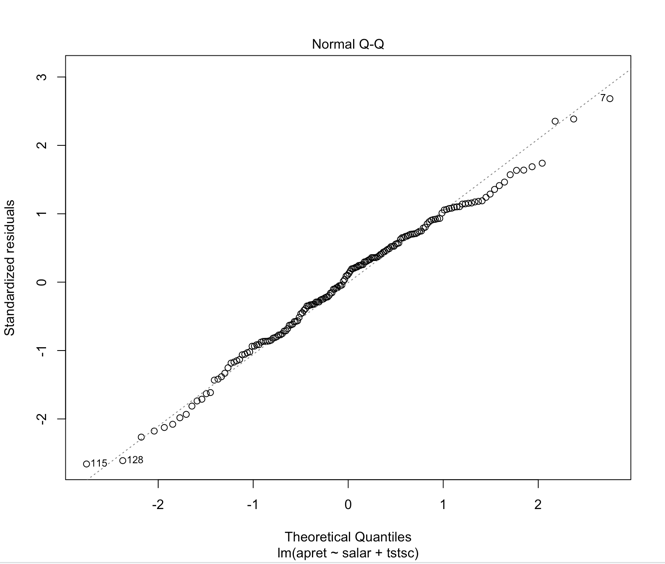
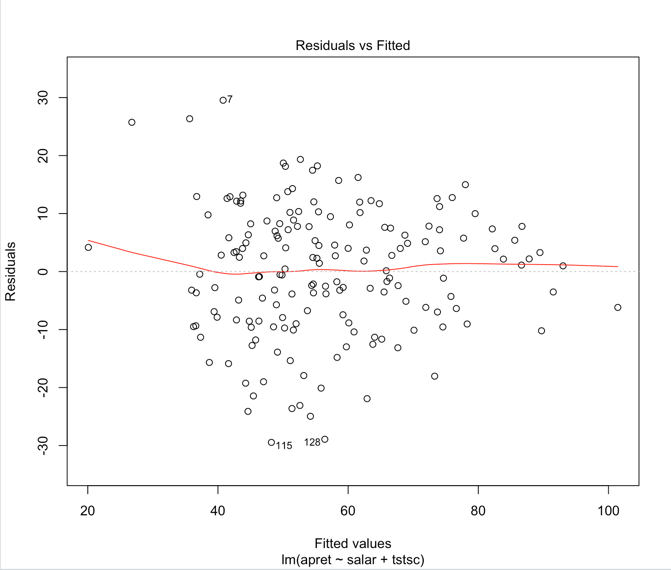


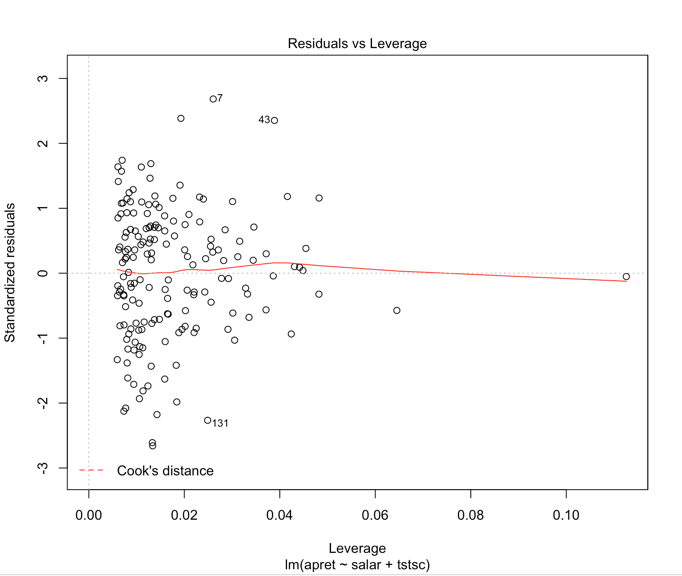
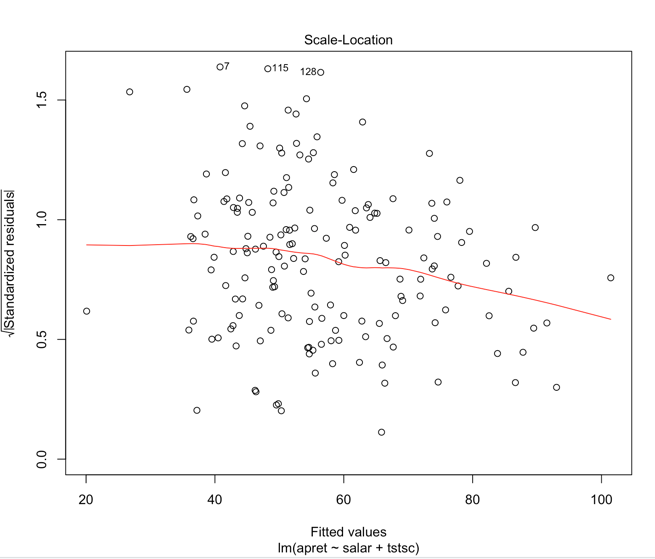




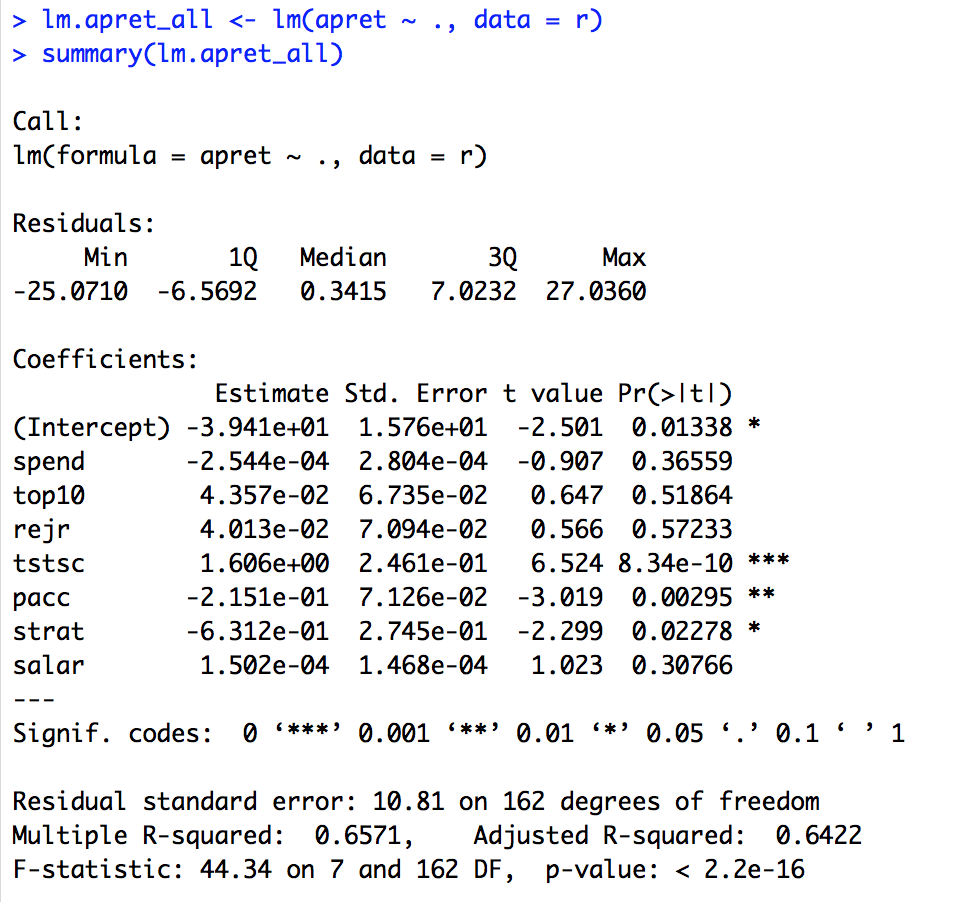
1. Linear regression of apret on tstsc and salar.

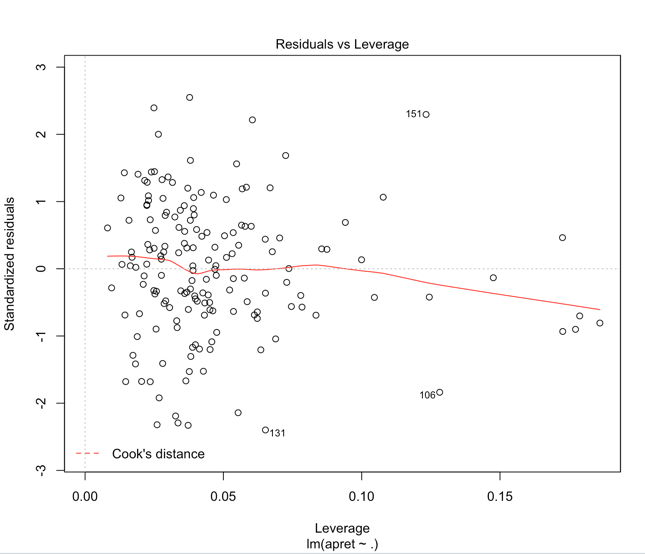
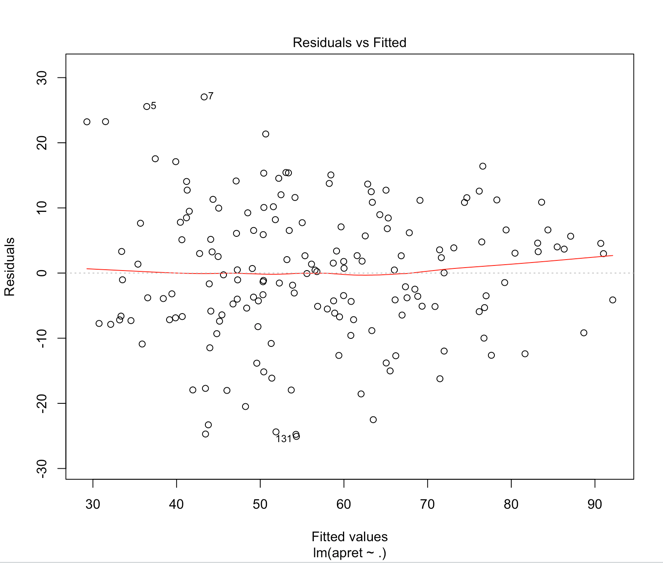
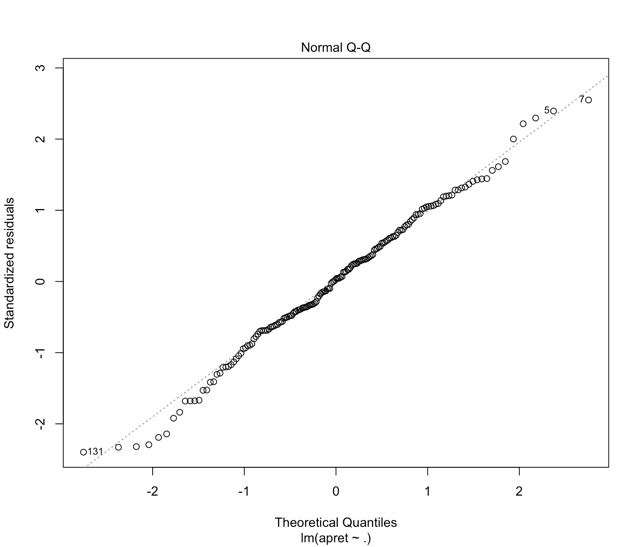
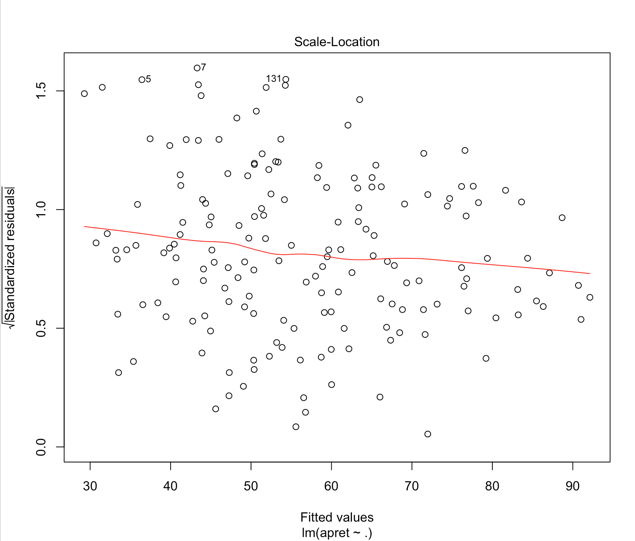




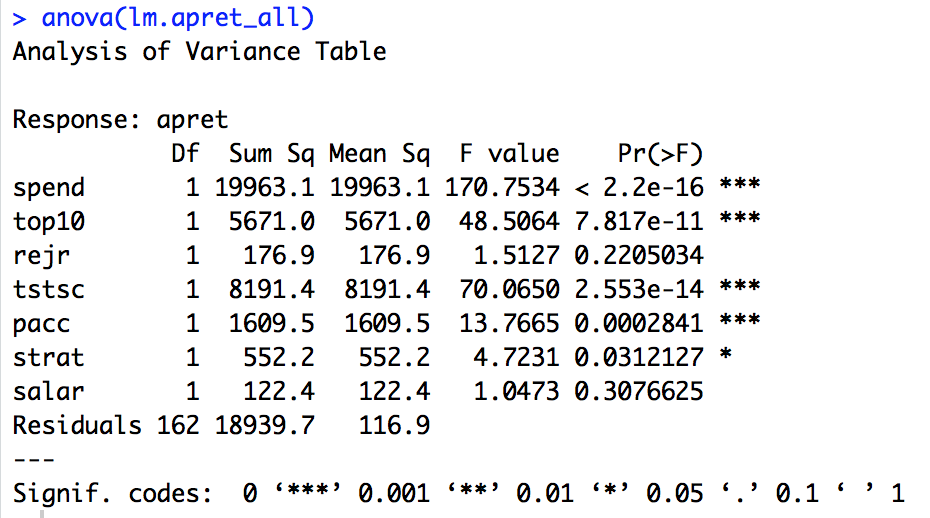


1. Linear regression of apret on all elements.

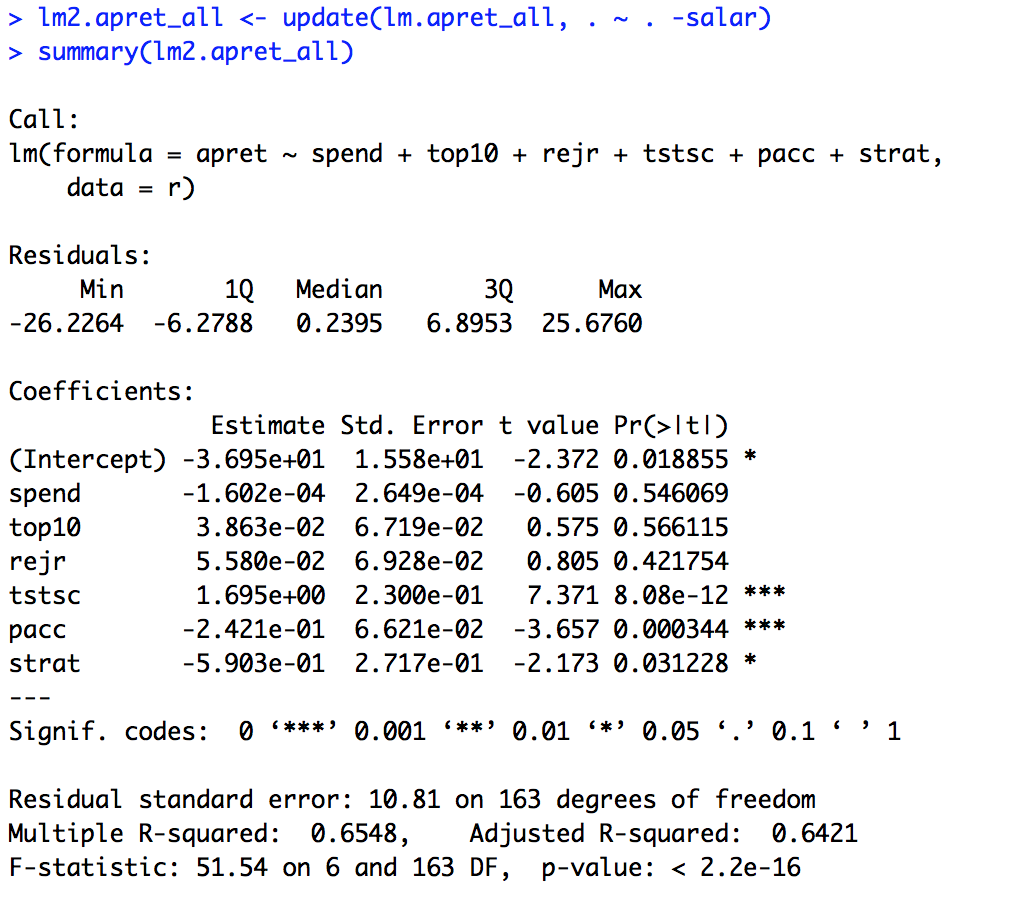




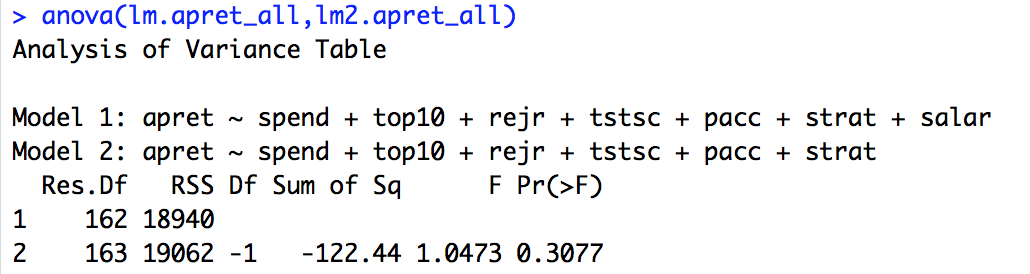
Optimize model:



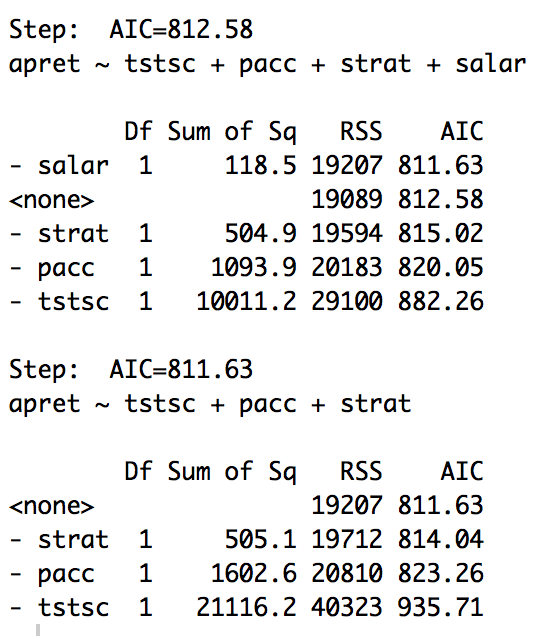
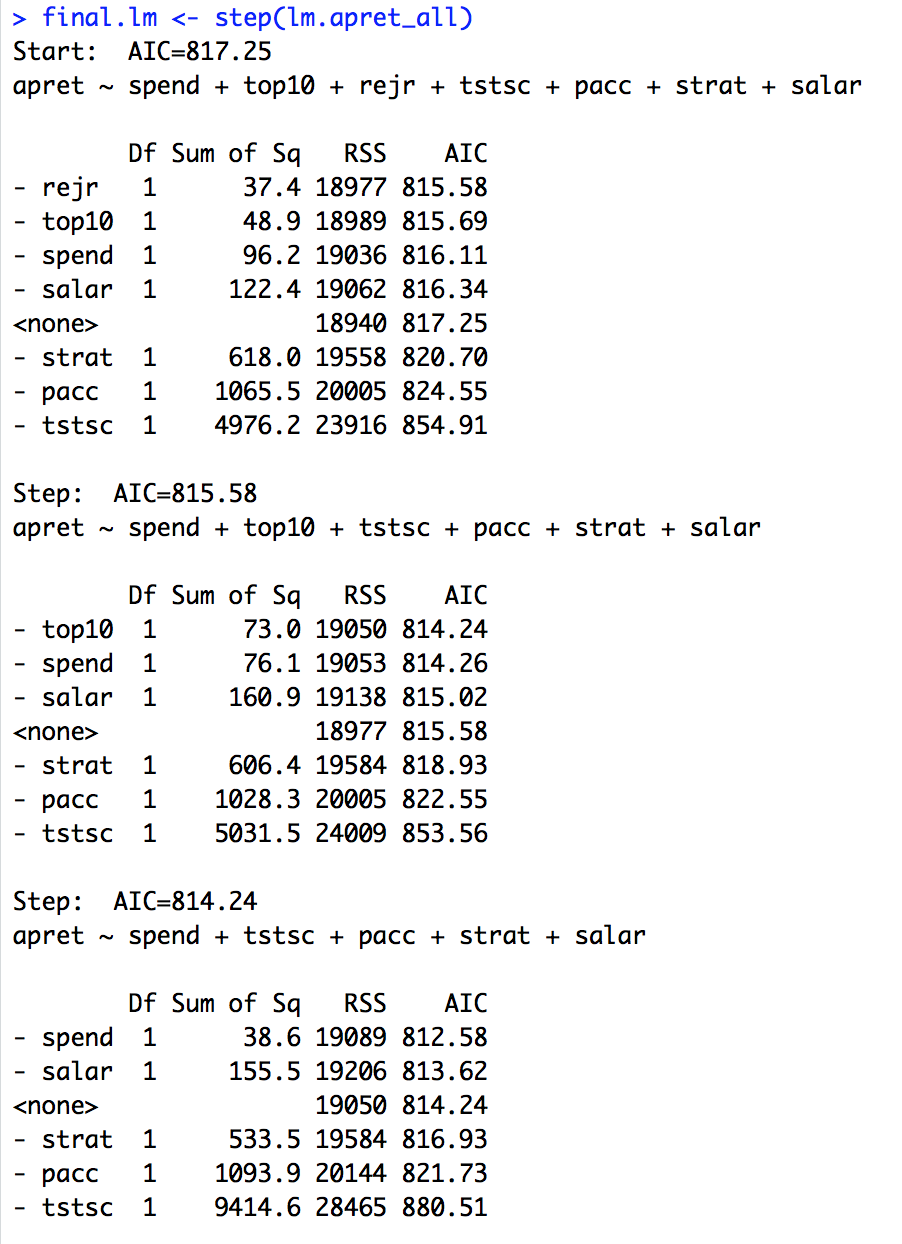
It is seen that salar has the least contribution to reducing the model fitting error, thus removing it from the model.

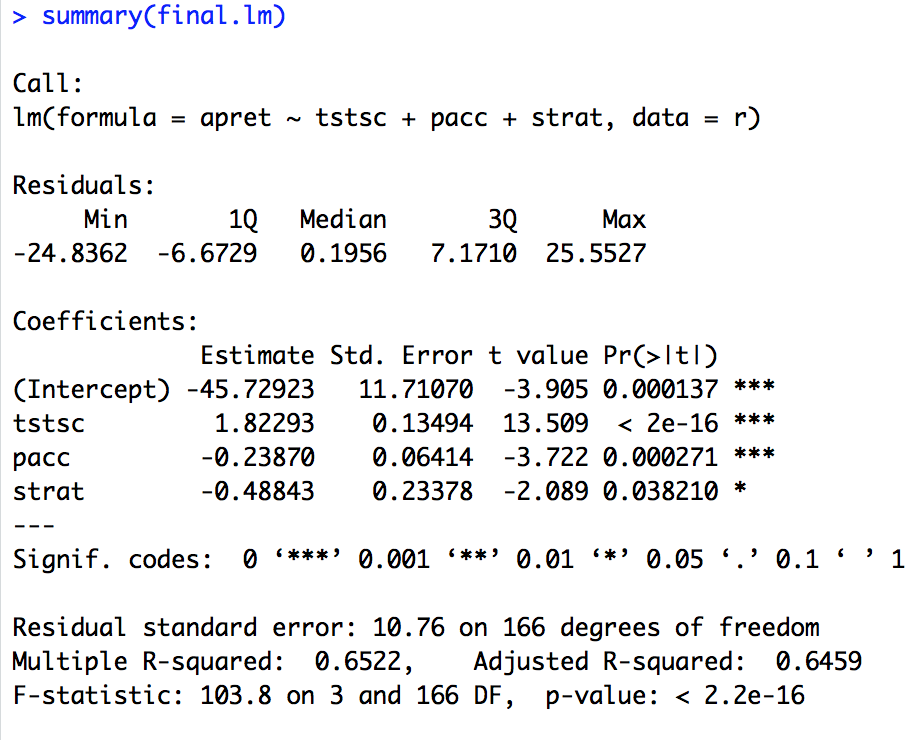


The model's fitting index has not improved. The following is a formal comparison of the two models with anova().



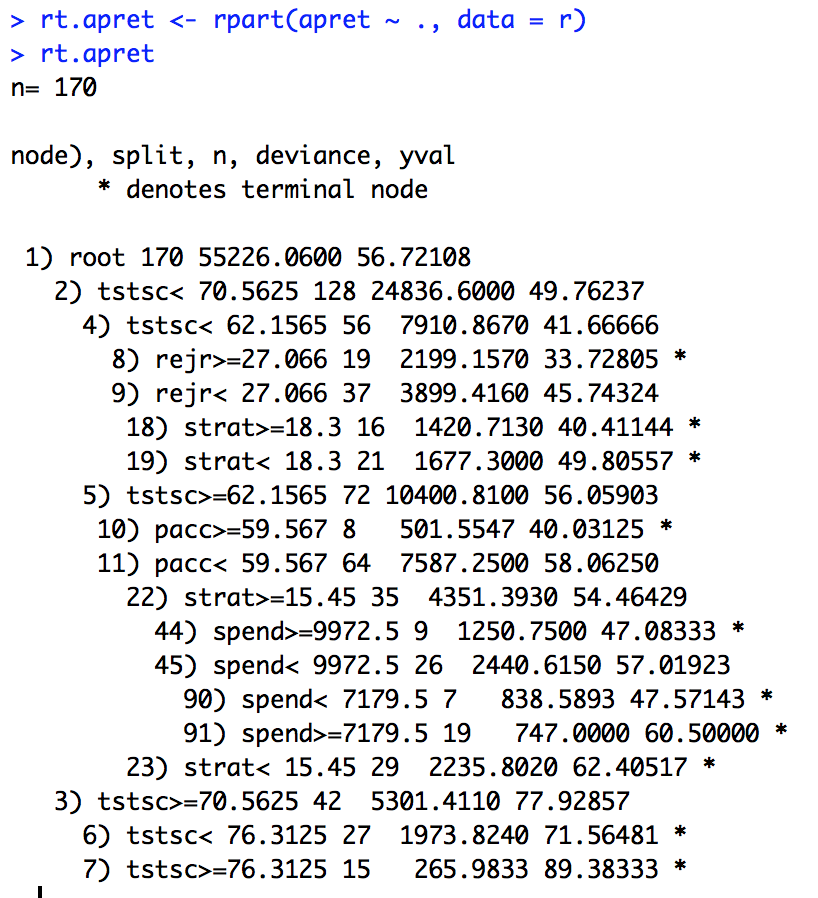
Although the sum of squared errors is reduced by 122, the probability of the two models differing is 70%. So continue to eliminate candidate coefficients. A new linear model is obtained by using the backward elimination method for the first-time model.

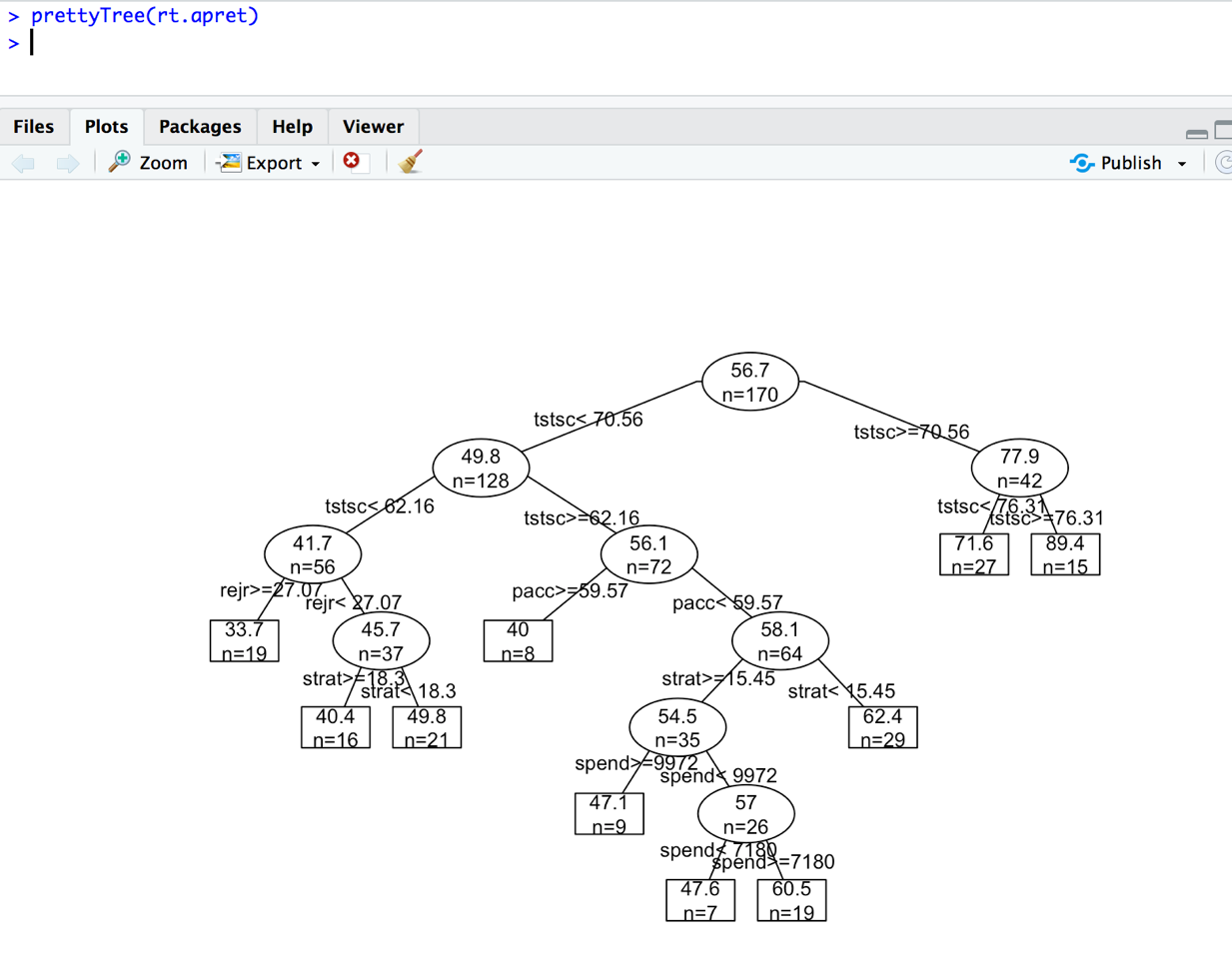




So it seems that the model is not good enough using linear regression, so than we will try another way to deal with it. Using regression tree.

Regression Tree





> 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)

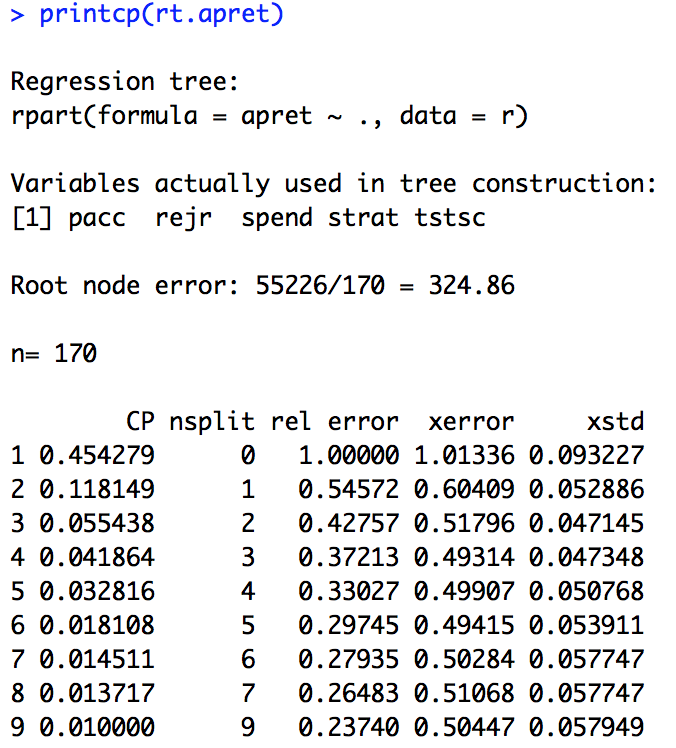
strat < 21.5 to the right, agree=0.769, adj=0.143, (0 split)

Node number 90: 7 observations

mean=47.57143, MSE=119.7985

Node number 91: 19 observations

mean=60.5, MSE=39.31579



Pruning

