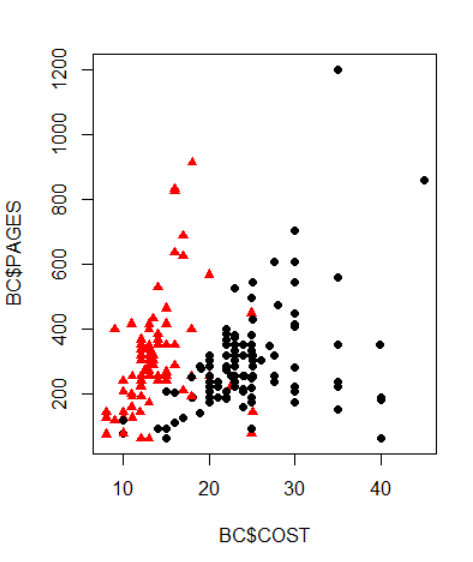
#1

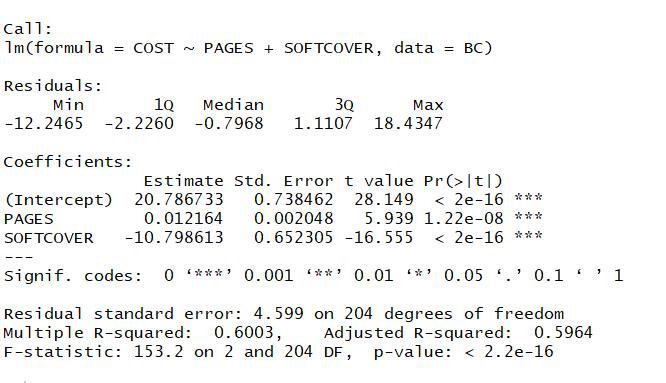
(a)

The blacks are hardcover, the reds are softcover. It is obviously that when a book has same pages, hard cover trends to right (expensive side) while soft cover trends to left (cheap) side. The similarity is they both have exceptions and pages are usually under 400 pages.



(b)

COST = Beta\_0 + Beta\_1 \* PAGES + Beta\_2 \* SOFTCOVER +e



If PAGES is held fixed, cost decreases $10.8 for a hard cover change to soft cover;

If the cover is fixed, cost increases $0.012 for per page which add into book.

(What is the difference between

“Comment on the overall significance of the parameter estimates and the model”

And

“interpret the slope parameter estimates associated with PAGES and SOFTCOVER in the context of the problem” in (c)?

)

(c)

Books with soft cover,

COST = 10 + 0.012 \* PAGES

Books with hard cover

COST = 20.8 + 0.012 \* PAGES

Averagely, every page add into a book will increase 0.012 dollars, which is based on using hard cover. If you use a soft cover instead of it, averagely you will save 10.8 dollars for each.

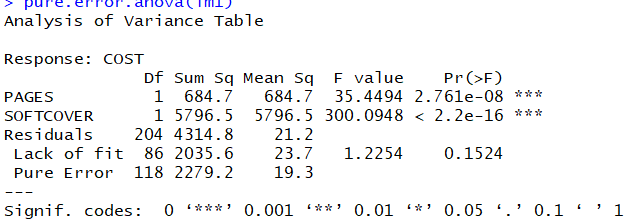
(d)

**Lack of fit**

Do F statistic to test the hypotheses

H0: the relationship is linear

Ha: the relationship is not linear



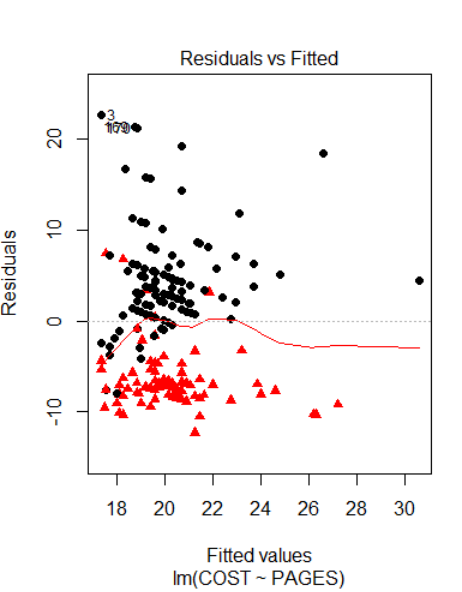
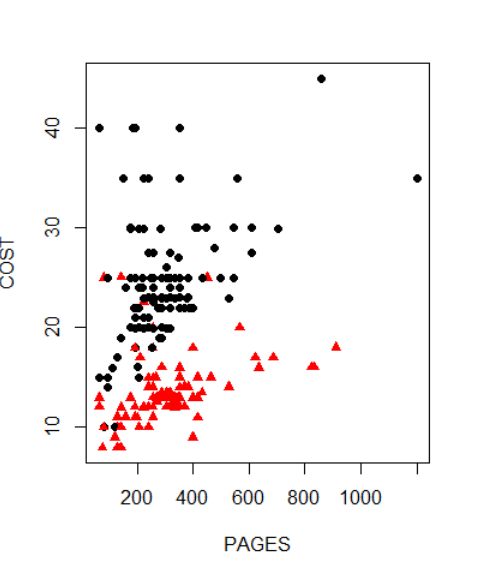


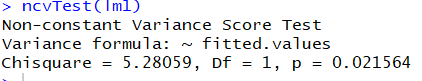
F – statistic = 1.2254 < 1.386, so we fail to reject H0. We conclude the relationship is linear when level of significance is at 5%.

**constant variance**

H0: Variance is constant

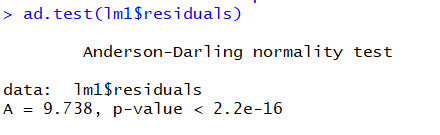
Ha: Variance is not constant





p-value less that a significance level of 0.05, therefore we can reject the null hypothesis that the variance of the residuals is constant.

**normality assumption using Anderson-Darling normality test procedure**

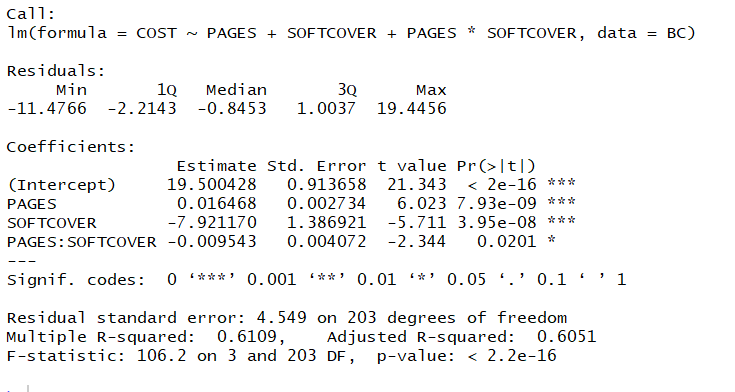


(e)

No, there are some outliers and just minus $10.8 for softcover is too general for cost.

(f)

Huh, First one is just use these two variances together.



COST = 19.5 + 0.016 PAGES – 7.9 SOFTCOVER – 0.0095 PAGES \* SOFTCOVER

The cost is influenced by number of pages and quality of cover. The average cost on pages are different when use different cover. When use hardcover, each page is more expensive, when use softcover, not only the fixed cost decrease but also each page is cheaper.

(I also do not sure the difference of meanings between

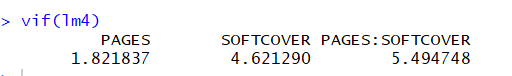
“Comment on the overall significance of the parameter estimates and the model”

And

“interpret the slope parameter estimates associated with PAGES and SOFTCOVER in the context of the problem” in (g)

)

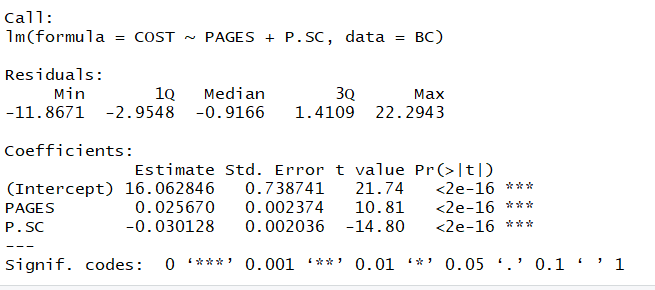
**multicollinearity problem?**



There has one vif value bigger than 5, so it has a multicollinearity problem.

The second one is create a new column named PAGE-SOFT

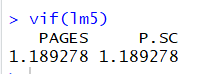




COST = 16.06 + 0.026 PAGES – 0.03 PAGES \* SOFTCOVER

Similar conclusion with the first one. Because if it is softcover, the cost for each page would decrease (here is negative), so the meaning of the cheaper page is included in the fixed cost part.

**multicollinearity problem?**



This one is better in this aspect. VIF values both smaller than 5, so it does not have a multicollinearity problem.

(g)

Hadrcover (SOFTCOVER = 0)

COST = 16.06 + 0.026 PAGES

Softcover (SOFTCOVER = 1)

COST = 16.06 – 0.004 PAGES.

For a hardcover book, the fixed cost is $16.06, every page add into will increase $0.026.

For a softcover book, the fixed cost is $16.06, every page add into will decrease $0.004, the decreasing shows the cost of cover decreasing.

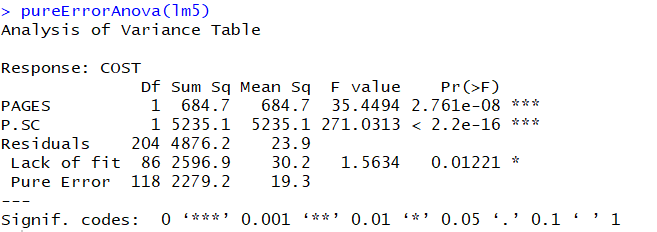
(h)

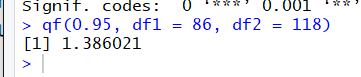
**Lack of fit**

Do F statistic to test the hypotheses

H0: the relationship is linear

Ha: the relationship is not linear



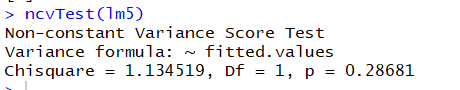


F – statistic = 1.563406 > 1.386, so we reject H0. We conclude the relationship is not linear when level of significance is at 5%.

**constant variance**

H0: Variance is constant

Ha: Variance is not constant



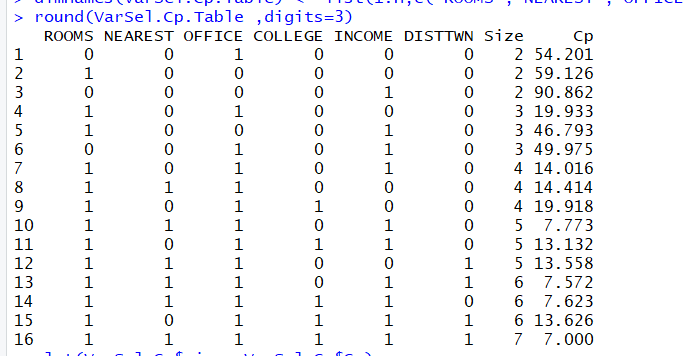
Test has a p-value more that a significance level of 0.05, therefore we fail to reject the null hypothesis that the variance of the residuals is constant.

(i)

I prefer to (f), it does show the fact: the hardcover books not only the cover are better, but also the quality of pages is better. Plus it does not have a multicollinearity problem.

#2

A)

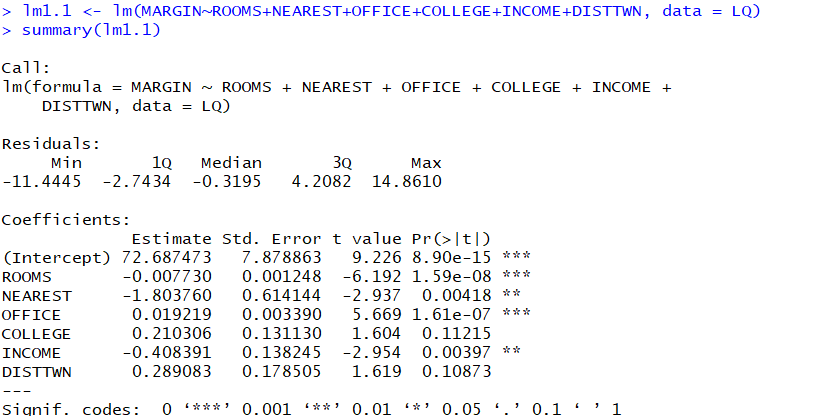


The 6 variables model has smallest Cp value. So it might be the best.

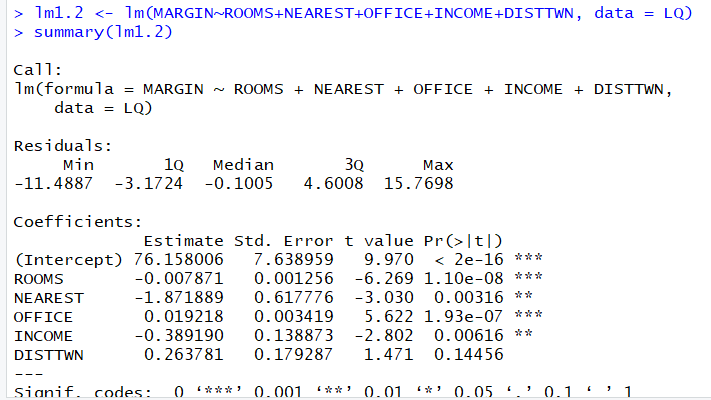
B)

Backwards Elimination

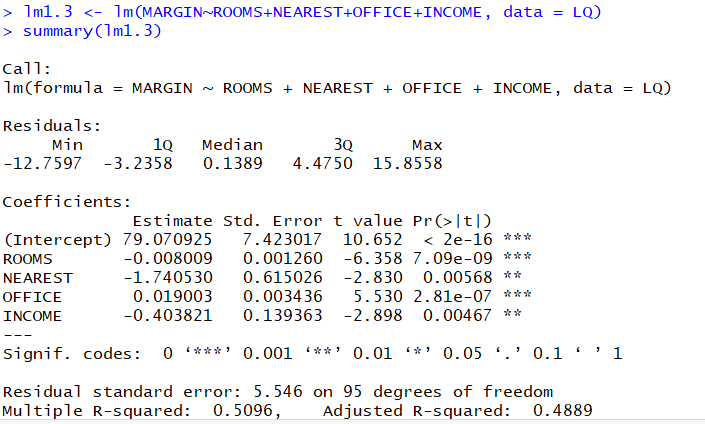
10% level



Delete COLLEGE



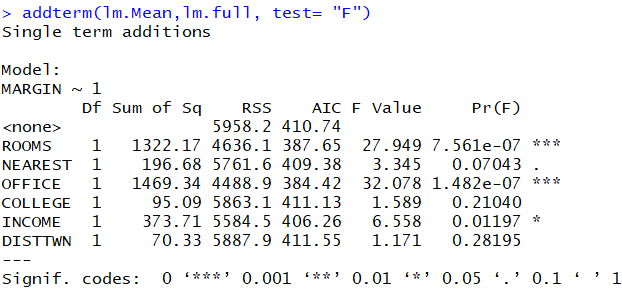
Delete DISTTWN



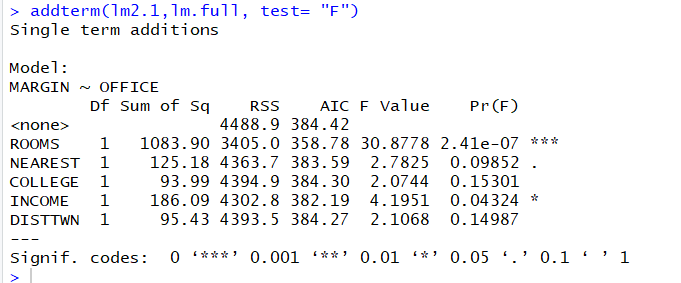
All Ps are smaller than 0.1, all remaining variables are significant, so select ROOMS, NEAREST, OFFICE and INCOME as variables.

C)

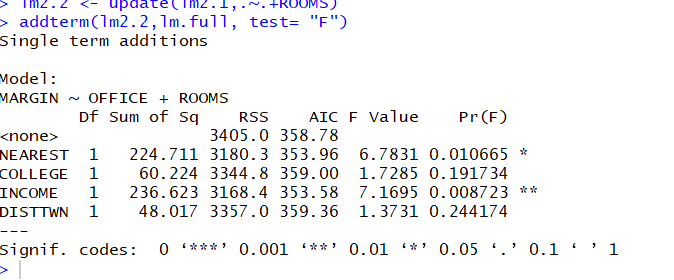
Level is 25%.



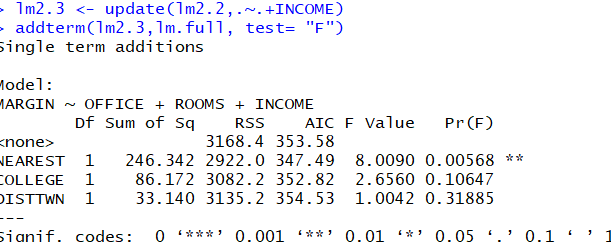
Select the smallest OFFICE to start.



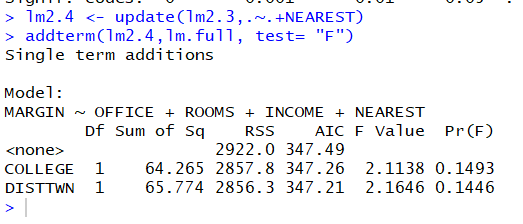
Select ROOMS to continue.



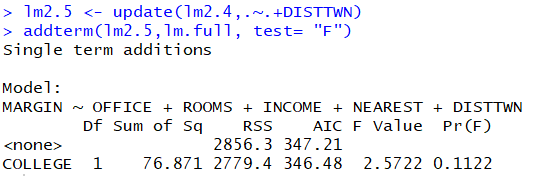
Select INCOME to continue.



Select NEAREST to continue.



Select DISTTWN to continue.



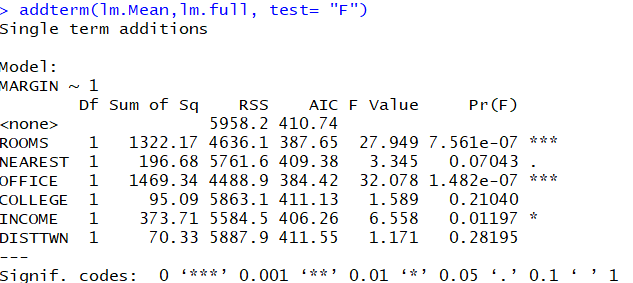
Select COLLEGE to continue.

All variables are added into. None is bigger than 25%, so all variables are significant, so select all of them.

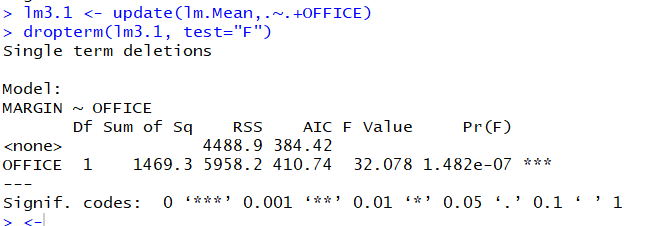
D)

Levels both are 25%

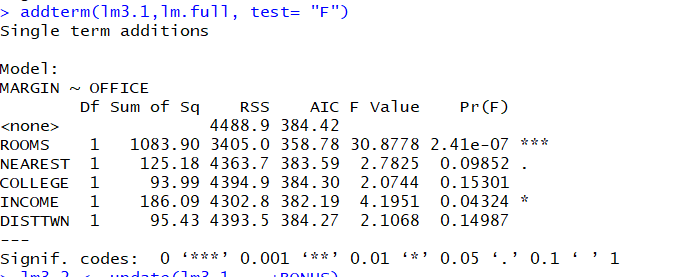
1



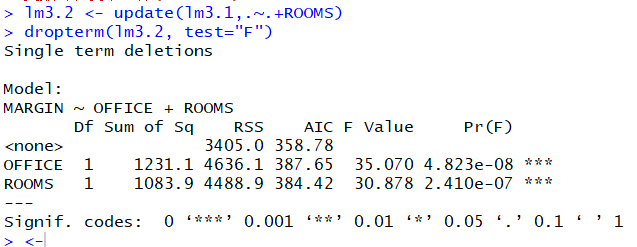
Add OFFICE.



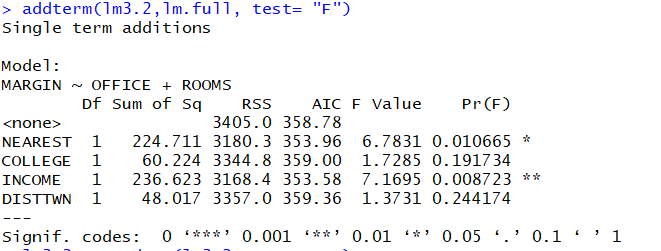
2



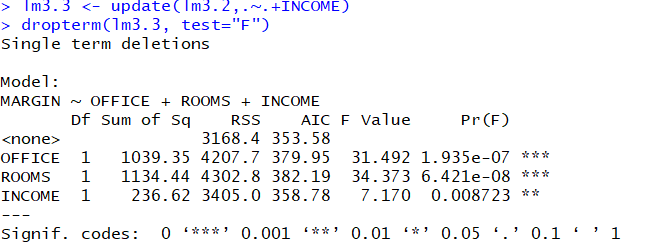
Add ROOMS



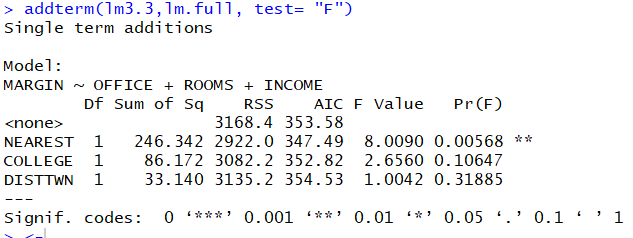
3



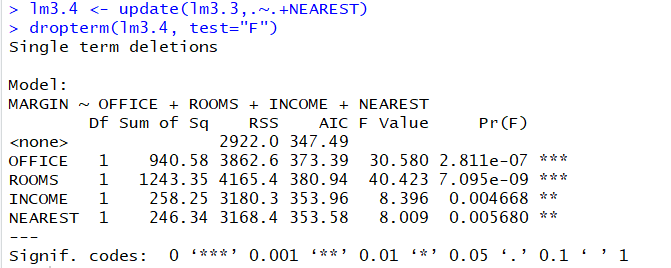
Add INCOME



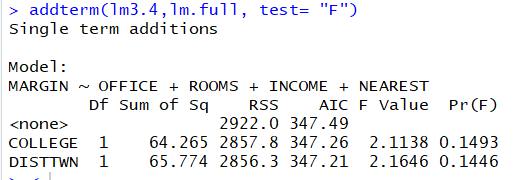
4



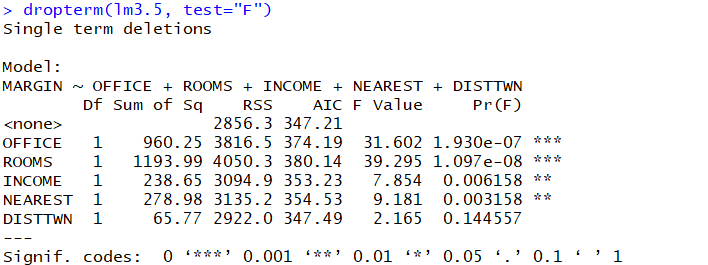
Add NEAREST



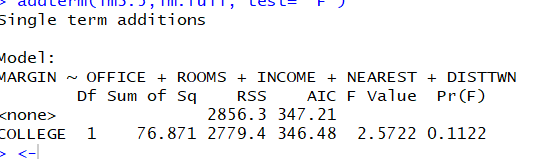
5.



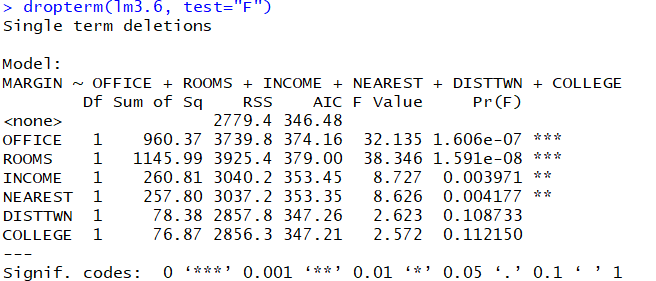
Add DISTTWN



6



Add COLLEGE



All variables are added into. None is bigger than 25% when add into, none is smaller than 0.25 when I want to delete. So all variables are significant, so select all of them.

E)

I might choose all variables, because in the 4 methods, there are 3 results suggest me select all variables.