**Homework 2 :**

Question 1:

1. Code in R

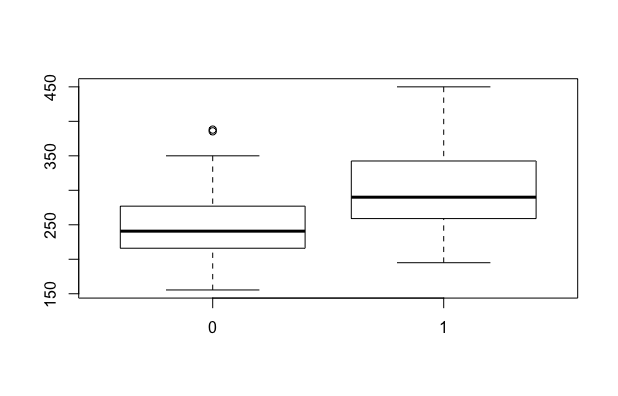
load("OregonHomes.Rdata")

ds1<-homes

ds1$ncat[ds1$Gar<=1]<-"0"

ds1$ncat[ds1$Gar>1 ]<-"1"

boxplot(ds1$Price~ds1$ncat)



The boxplot show that the mean difference is noticeable.

1. Code in R

t.test(ds1$Price~ds1$ncat,var.equal=TRUE)

The test result indicates that :

mean of x mean of y

254.3000 300.3327

There is a significant difference (almost 50K $) in the mean of 2 groups.

1. Code in R:

var.test(ds1$Price~ds1$ncat)

The result shows that the sample estimates : ratio of variances is: 1,290124.

Question 2:

1. Code in R :

aovtable <- aov(homes$Gar ~ homes$Price, data = homes)

summary(aovtable)

The results is :

Df Sum Sq Mean Sq F value Pr(>F)

homes$Gar 1 34796 34796 10.81 0.00155 \*\*

Residuals 74 238211 3219

According to the p-value = 0.00155 < 0.05 , we can conclude that there are significant differences between the 2 groups.

1. Code in R:

temp1<-lm(Price~ncat, data=ds1)

summary(temp1)

The result in the linear regression model confirms the hypothesis that stated above.

1. All the p-values in t-test, linear regression model and ANOVA test are the same because: all of them in this case measure the variances of the predictors compare to the “real” value result.

Question 3:

1. Code in R:

a2<-aov(ds1$Price~as.factor(ds1$Gar))

summary(a2)

The results:

Df Sum Sq Mean Sq F value Pr(>F)

as.factor(ds1$Gar) 3 36682 12227 3.725 0.015 \*

Residuals 72 236325 3282

The result is differ from the previous experiment when we use Garage size as a factor. The p values change to 0.015 this is still show that there is a relation between the two variable. But with this experiment the relation decreases dramatically.

1. Code in R:

TukeyHSD(a2,ordered=TRUE,conf.level = 0.95

Only the comparison between houses has 2 garages with the houses has no garage becomes **significant** for the price.

1. The difference is significant between 2 garages and 0 garages is understandable because for the couple each has his/her own cat will prefer the house with 2 spots garage. Meanwhile, 3 spots garage houses have no relation to the price compare to the houses without garage is because the houses with 3 spots garage will have totally different way for pricing. The garage spots become insignificant for the case with 3 spots garage houses.

Question 4:

1. Code in R:

temp2 <- lm(Price ~ .-ncat-index, data = ds1)

summary(temp2)

The variables that significant to the price are: Floor, Lot, StatusSold, SchoolEdison, SchoolHarris.

1. The model fitting is not that good only show 44,28% accuracy toward the real data.
2. The form of variable Gar according to the result from linear model of the whole dataset perform the best is as a factor. Because the p-value from the model is closer to the complete regression model.

Question 5:

The variable age is marked NA because the value doesn’t play a role in the model, the year values of this variable give no information! Age variable provide the information as the variable year so the model automatically eliminate the relevant of Age.

Question 6:

Price = 80.3\*Floor + 10.34\*Lot - 37.35\*StatusSold + 91.76\*SchoolEdison + 61.9\*SchoolHarris - 202.15

The price is equal to the floor + lot + schooledison + schoolharris - statussold.

Question 7:

Code from R:

temp3<-lm(Price~.-ncat-index-Year, data=ds1)

temp4 <- step(temp3, direction="both")

summary(temp4)

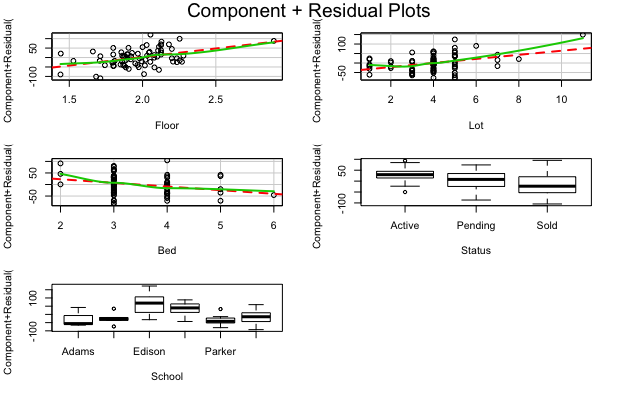
The result is that the model will focus only on how to adjust the significant variables from the original linear model. Moreover, the Bed variable now becomes significant after the model adjusting stepwise to find the best linear model. The adjusted R-squared value increase to 45.51%. All the “previous” significant variables are still remained its valuable.

Question 8:

1. Code in R:

crPlots(temp4)

Plot:



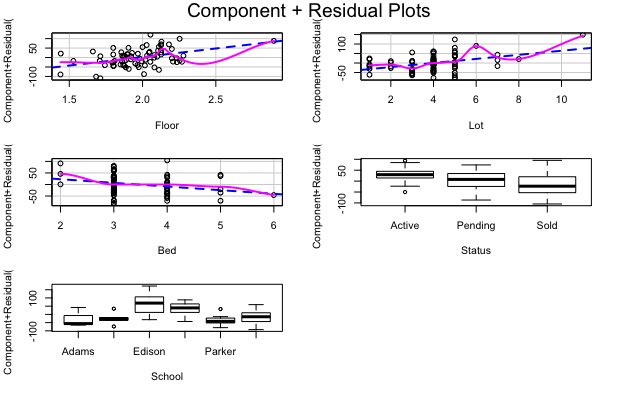
The floor and bed don’t need quadratic effects base on the green line. The Lot needs quadratic effect

1. Code in R:

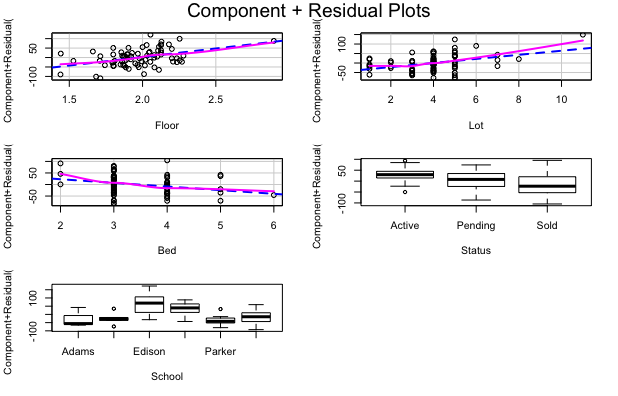
crPlots(temp4, smooth=list(span=0.25))

crPlots(temp4, smooth=list(span=0.75))

Plot for smooth parameter to 0.25



Plot for smoothing parameter to 0.75:



Smoothing parameter with span = 0.75 is the parameter setting indicates the quadratic effects more clearly.