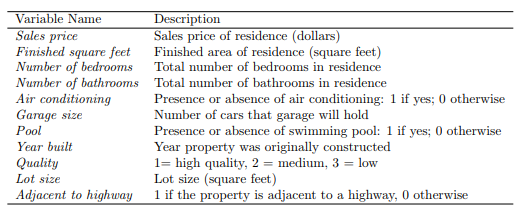
Assignment 3

Due: Wednesday, October 9, 2019

This assignment relates to the ‘Real-estate’ data set. People buying or selling houses would like to know how much they can expect to get, or pay, for a property. This is also a concern for those who are making mortgage loans, or for those taxing real estate (and who are more likely to commission statistical studies than individual home-owners). The price of a house depends on its physical characteristics, including size, features, quality of construction, age, etc. It also depends on location, and current market characteristics. You are approached by a research group which has a data on a sample of residential sales in a midwestern city; the variables are described in Table below.



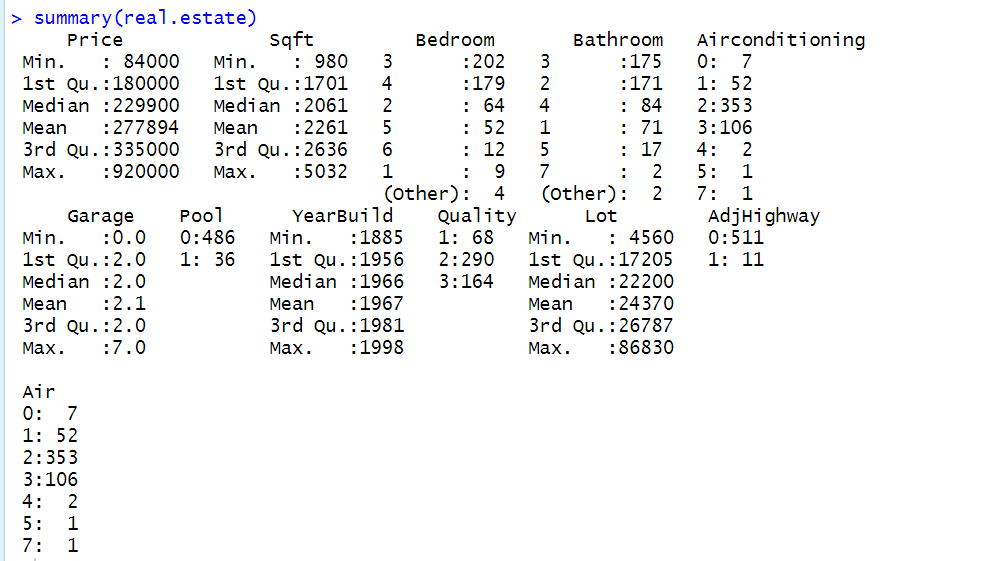
1. Read the data into R. Call the loaded data “real.estate”.

real.estate <- read.csv("real-estate.csv",row.names = "ID")



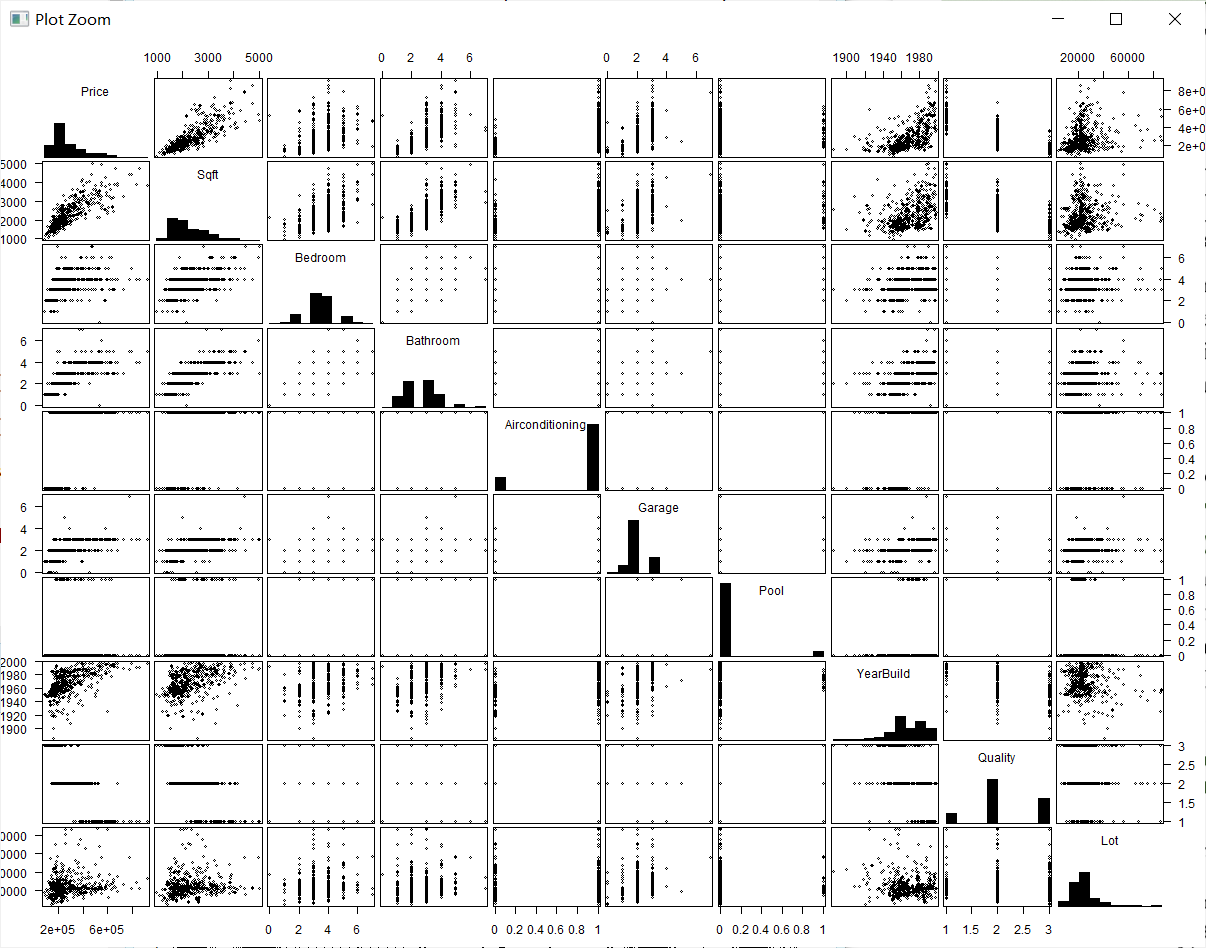
1. Answer the following sub-questions
2. Use the “summary()” function to identify the types of variables. Which variables are categorical? Which variables are quantitative? Are there any concerns in the summary table? Explain.

All of the variables are quantitative. None of the variables are categorical. Yes. For the quantitative variable, the summary function will provides the minimum, maximum, quartiles, and mean. While for the categorical variable, the summary function only displays the frequencies in each category.



1. Use the “pairs()” or “gpairs()” function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first then columns of a matrix A using A [,1:10]. Is there any interesting patterns? Which variables seem associated with the sales price? Explain.

Yes, Price is almost proportional to Sqft, Bedroom, Bathroom, Garage, Airconditon, YearBuild, Quality, Lot. But, what’s interesting is it’s negative proportional to the Pool. All the variables seem associated with the sales price. Because the scatterplot depicts the positive or negative proportional relationship.



1. Use the “as.factor” function to regenerate categorical variables.

real.estate$Bathroom <- as.factor(real.estate$Bathroom)

real.estate$Bedroom <- as.factor(real.estate$Bedroom)

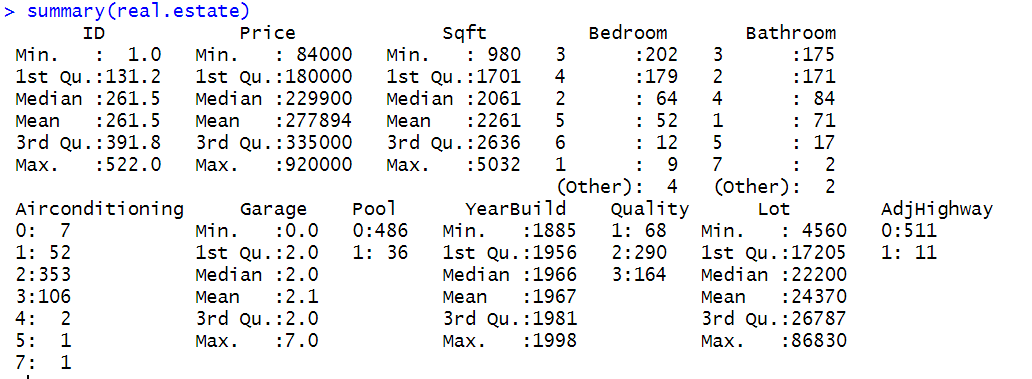
real.estate$Airconditioning <- as.factor(real.estate$Garage)

real.estate$Pool <- as.factor(real.estate$Pool)

real.estate$Quality <- as.factor(real.estate$Quality)

real.estate$AdjHighway <- as.factor(real.estate$AdjHighway)

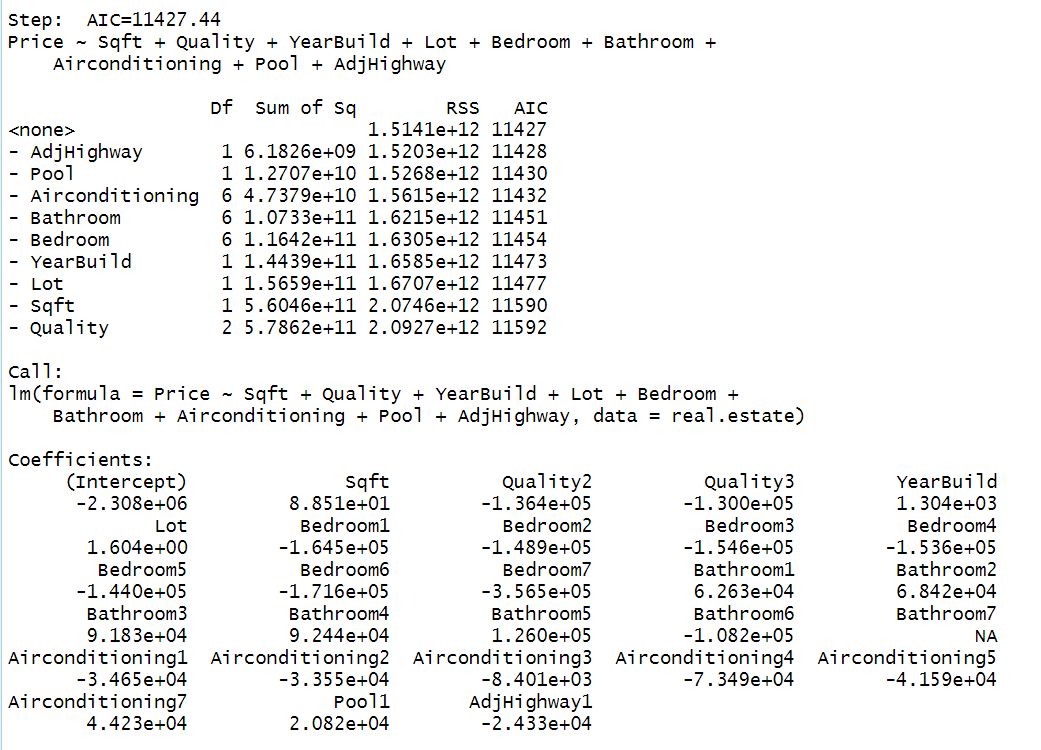
summary(real.estate)



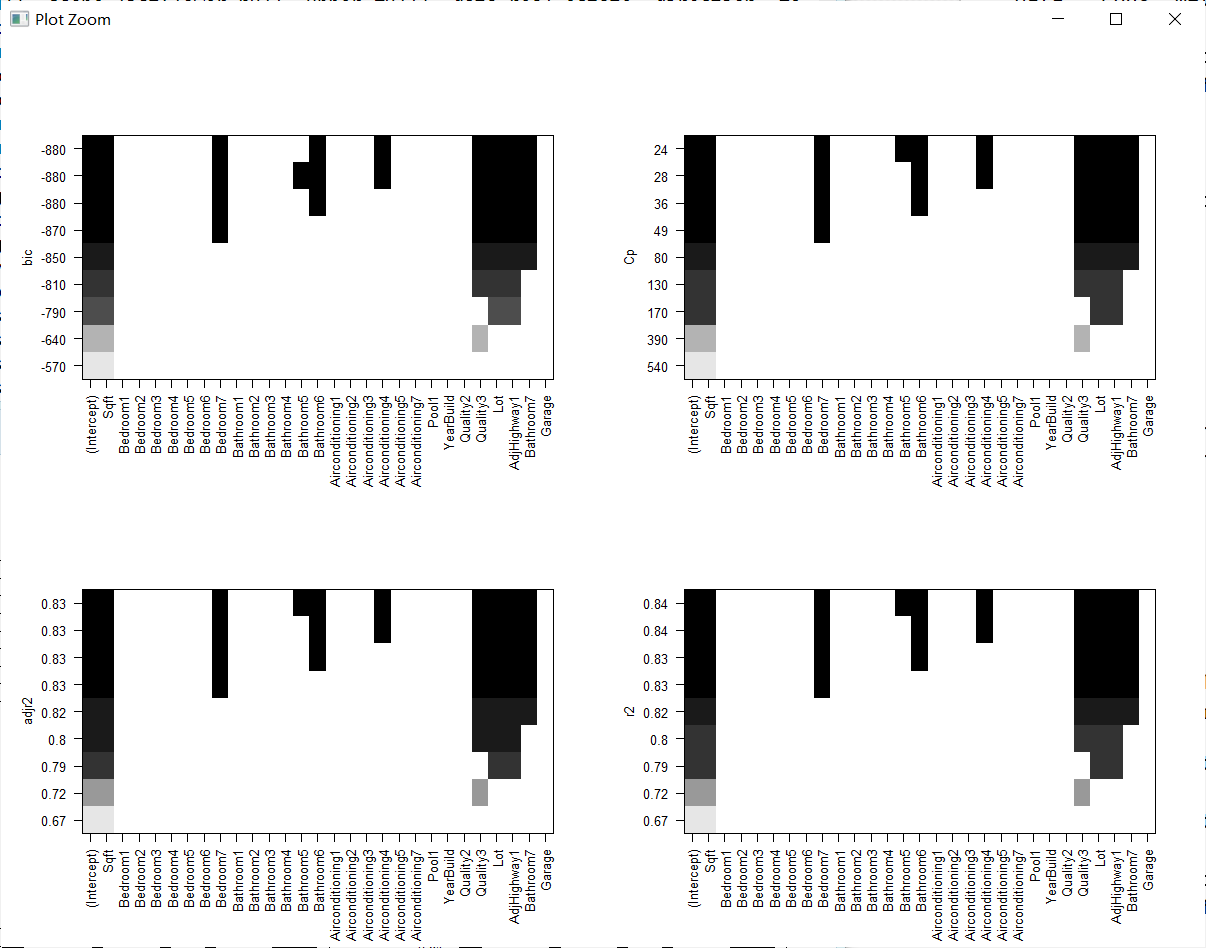
1. Fit the models and address the following when building that model:
2. Fit the null model and the full model



1. Find ***the best sets*** of predictors using the stepwise procedures.



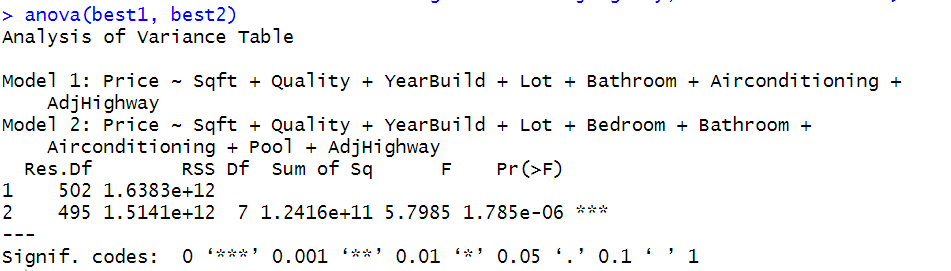
1. Find ***the best sets*** of predictors using the best subset approach.



Bic: 7peaks; Cp:7peaks; Adjr2: 9 peaks; r2: 9 peaks

Price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom + AdjHighway + Air + Airconditioning + Bathroom + Pool

1. Considering the models in part (ii) and part(iii), choose the best model.



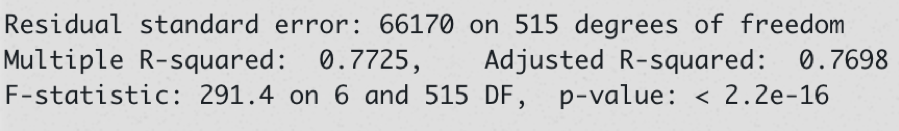
best=lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Bedroom +

Bathroom + Airconditioning + Pool + AdjHighway, data = real.estate)

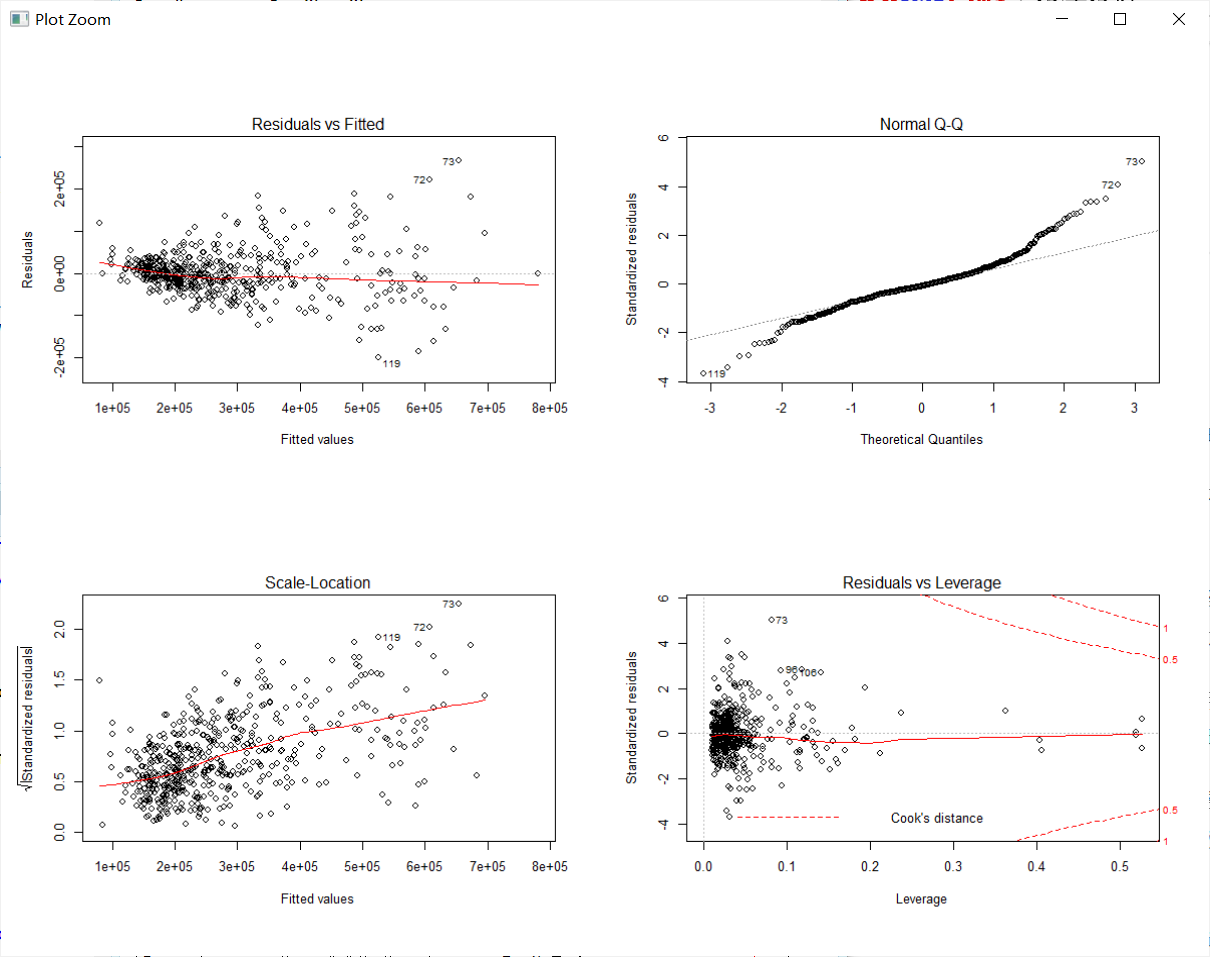
1. Interpret the coefficients of the best model ***in this context***.

The Price increases with the Sqft, the Beadroom. And the closer with the highway, the lower the Price becomes. And the price decays with the Airconditioning. The worse it is, the lower the price is. Price is Proportional to the Sqft, YearBuild, Lot, Bathroom, Pool. And it is negative proportional to the Airconditioning, Quality, Bedroom.

1. Evaluate the best model.



1. Check the assumptions using “plot()” function.



1. The residuls are distributed with normal distribution form the top two figures.
2. But the homoscedasticity may be violated from the third plot.
3. Continue exploring the data and provide a brief summary of what you discover.

shrinkage <- function(fit,k=10){

require(bootstrap)

# define functions

theta.fit <- function(x,y){lsfit(x,y)}

theta.predict <- function(fit,x){cbind(1,x)%\*%fit$coef}

# matrix of predictors

x <- fit$model[,2:ncol(fit$model)]

# vector of predicted values

y <- fit$model[,1]

results <- crossval(x,y,theta.fit,theta.predict,ngroup=k)

r2 <- cor(y, fit$fitted.values)\*\*2 # raw R2

r2cv <- cor(y,results$cv.fit)\*\*2 # cross-validated R2

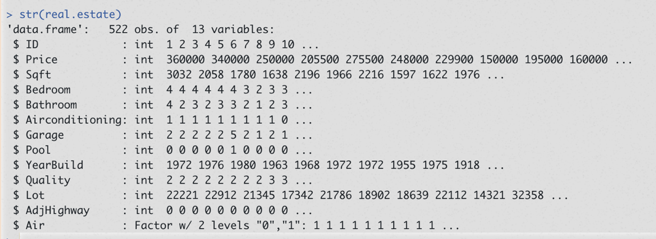
cat("Original R-square =", r2, "\n")

cat(k, "Fold Cross-Validated R-square =", r2cv, "\n")

cat("Change =", r2-r2cv, "\n")

}





In the previous problems, we examine our model with my training data. This will be too optimistic. So here we use cross-validation to examine our model. We can see the original R-square is larger than the newer ten-fold cross-validated R-square. So we can redo our previous model selection with cross-validation. And we will reselect the best model.