A real estate appraiser is interested in predicting residential home prices in a mid-western city as a function of various features. For that purpose a regression model is to be constructed from a sample

of 522 houses. Download the homes.xls data set from blackboard.

Consider the predictors

x1: lot size (square feet), x2: area (square feet), x3: number of bedrooms,

x4: number of bathrooms, x5: year of construction, x6: garage size (number of cars).

1. What are the predictors with the highest correlation?

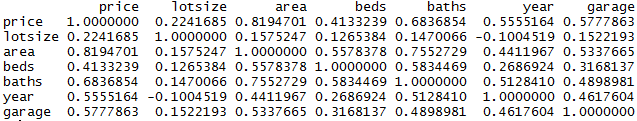
#question 1

library(readxl)

homes <- read\_excel("C:/Users/Joey Zhao/Desktop/529/2/homes.xls")

View(homes)

data1=subset(homes,select = c(price, lotsize, area, beds, baths, year,garage))

cor(data1)

Looking at the first row of this form above, we can find that area have the highest correlation with price.

1. What is the area (not lot size) of the most expensive house?

#question 2

which.max(data1$price)

data1[73,]



Therefore, the area of the most expensive house is 3857.

1. If there are outliers find the largest one (in absolute value).

#question 3

m1=lm(price~.,data=data1)

m1

library(PASWR2)

checking.plots(m1)

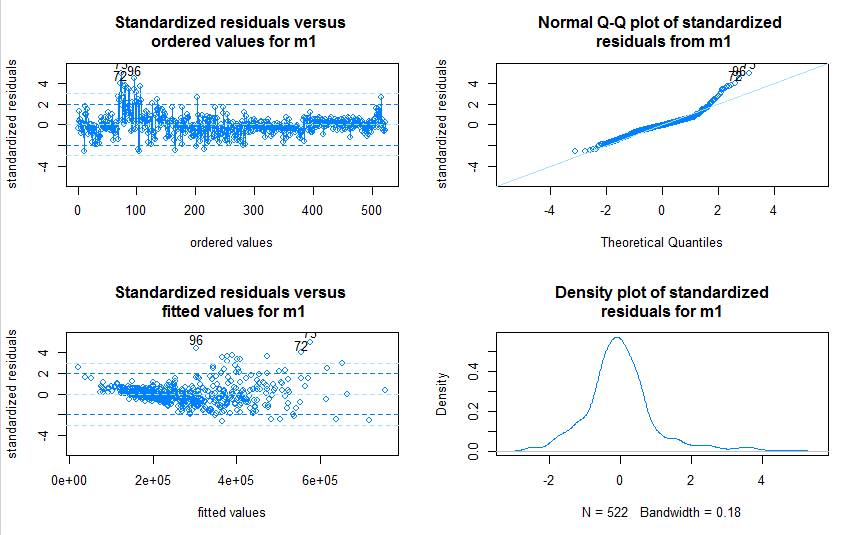
help(predict)

pre1=predict(m1,data1)

res1=(data1$price-pre1)^2

which.max(res1)

data1[73,]





The largest one is No.73

4. Find a 99% confidence interval for 2

#question 4

#the lm model is price=-3.568e+06 + 1.555e+00 lotsize + 1.257e+02 area + -1.304e+04 beds + 7.988e+03 baths + 1.780e+03 year + 2.253e+04 garage

#beta 2 is 1.257e+02 which is the coefficient of area

summary(m1)

help("confint")

confint(m1,3,0.99)



Therefore, the confidence interval for is (107.7357 , 143.7415)

5. Find a 95% confidence interval for the mean price of a house with garage for two cars, area of 2650

square feet, built in 1990, 24500 square feet size, three bedrooms, three bathrooms

#question 5

#Find a 95% confidence interval for the mean price of a house with garage for two cars, area of 2650

#square feet, built in 1990, 24500 square feet size, three bedrooms, three bathrooms

new1=data.frame(lotsize=24500,area=2650,beds=3,baths=3,year=1990,garage=2)

predict(m1,new1,interval = "conf" , level = 0.95)



Therefore, the 95% confidence interval for the mean price of this house is (362128.4 , 387712.6)

6. Find the predicted price when all predictors are equal to their median values. Fit the model with the best subset of predictors (in terms of adj-R2).

#question 6

help(median)

new2=data.frame(lotsize=median(data1$lotsize),area=median(data1$area),beds=median(data1$beds),baths=median(data1$baths),year=median(data1$year),garage=median(data1$garage))

predict(m1,new2)

254573.4

install.packages("leaps")

library(leaps)

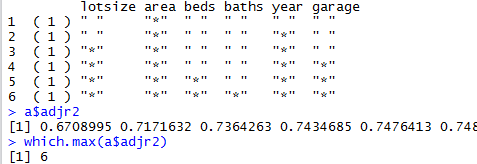
models=regsubsets(price~.,data1)

models

a=summary(models)

a$adjr2

which.max(a$adjr2)



Therefore, the predicted price is 254573.4 and the best subset is the subset with all six predictors. The linear model is

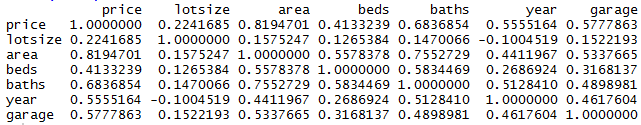
price=-3.568e+06 + 1.555e+00 lotsize + 1.257e+02 area + -1.304e+04 beds + 7.988e+03 baths + 1.780e+03 year + 2.253e+04 garage

7. Find the best and worst predictors

Fit a model with only x4, the number of bedrooms as the predictor

#question 7

cor(data1)



Therefore, the worst predictor is lotsize, which has the lowest correlation with price.

m2=lm(price~beds,data2)

m2



The model is price = 82809 + 56200 \* beds

8. Interpret the slope value b1.

Fit a full model for houses having between two to four bedrooms

#question 8

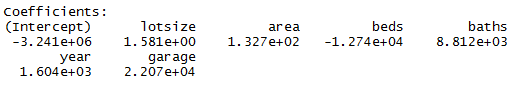
#b1, which is the coefficient of beds, means when bed increases by 1,then theoretically the price of a residential house will increase 56200 units.

data3=subset(data1,beds>=2 & beds<=4)

#full model means with six predictors

m3=lm(price~.,data3)

m3



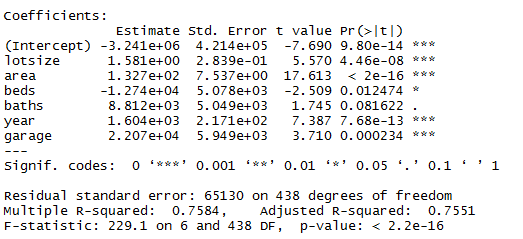
The full model is shown as above.

9. Interpret adequacy values (MSE, R2).

Take model m3 as an example, which is shown above, the meaning of MSE and R^2 are as follows:

#question 9

summary(m3)



In this form, we can find MSE=65130^2= 4241916900. MSE means mean square error. In this example MSE = SSE/438. SSE is the sum of squares of the difference between the predicted value and the sample value accordingly. Smaller MSE means better prediction accuracy.

R^2 = 0.7584 which is SSR divided by SST, means this model(m3) could explain 75.84% of the residential home price.

10. Find a 95% prediction interval for the price of a house with a garage for two cars, area of 3150

square feet, built in 1996, 26250 square feet size, two bedrooms, three bathrooms. #question 10

#in this case, we choose to use model m3, because it has higer R

new3=data.frame(lotsize=26250,area=3150,beds=2,baths=3,year=1996,garage=2)

predict(m3,new3,interval = "prediction", level = 0.95)



Therefore, the interval is (334969.9 , 595399.1)