> house<-read.csv(file.choose(),header=TRUE)

> attach(house)

>

> #FOR QUESTION 1

>

> #part A

>

> hist(PRICE)

>



> #the variable PRICE is not uniformly disributed.

> #It is skewed to the right because mean is greater than median.

> #part B

>

> summary(PRICE)

Min. 1st Qu. Median Mean 3rd Qu. Max.

54000 78000 96000 106300 120000 215000

>

> #from the summary table, we can see that the mean is 106300 which is more than the median, which is 96000.

> #it does corresponds with answer in part A because the histogram is right skewed.

>

> #part C

>

> boxplot(PRICE~FEAT)

>



> #there are a few outliers for FEAT 3 and FEAT 4. And Feat 2

> #some of the plots are not normally distributed

> #Plot for FEAT 3 seems like skewed to the right

>

>

The price of the houses also tends to increase overall when we get more features as the median is continuously increasing.

> #part D

>

> qqnorm(SQFT)

> qqline(SQFT)

>



>

> #this qqplot is approximately normally distributed.

> #however, there seems to be some points that deviated away from the line.

>

This makes this not normally distributed overall.

> #part E

> cor(cbind(PRICE,SQFT,AGE,FEAT,TAX),use="pairwise.complete.obs")

PRICE SQFT AGE FEAT TAX

PRICE 1.0000000 0.84479510 -0.16867888 0.4047432 0.8756647

SQFT 0.8447951 1.00000000 -0.03965489 0.3910691 0.8585828

AGE -0.1686789 -0.03965489 1.00000000 -0.2872913 -0.2918422

FEAT 0.4047432 0.39106911 -0.28729134 1.0000000 0.4558414

TAX 0.8756647 0.85858280 -0.29184225 0.4558414 1.0000000

>

> #from the correlation matrix, we can see that variable TAX is the most highly correlated with PRICE with correlation of 0.8756647.

>

>

>

> #part F

>

> model1=lm(PRICE~SQFT+AGE+FEAT+TAX)

> summary(model1)

Call:

lm(formula = PRICE ~ SQFT + AGE + FEAT + TAX)

Residuals:

Min 1Q Median 3Q Max

-47011 -6478 208 6232 62200

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4846.28 11108.71 -0.436 0.66419

SQFT 39.36 10.12 3.890 0.00025 \*\*\*

AGE -41.34 207.89 -0.199 0.84305

FEAT 1592.81 2483.71 0.641 0.52373

TAX 54.28 16.67 3.257 0.00184 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17260 on 61 degrees of freedom

(51 observations deleted due to missingness)

Multiple R-squared: 0.8291, Adjusted R-squared: 0.8179

F-statistic: 73.98 on 4 and 61 DF, p-value: < 2.2e-16

> #from the summary..

> #value of R^2 = 0.8291, which is almost close to 1 (this is good)

> #value of R^2 adjusted = 0.8179 , which is also almost close to 1. (this is also good)

> #the p-value : < 2.2e-16, is very small and close to zero. (which is good)

>

>

> #part G

>

> #from part F, for every increment of 1 point of FEAT, there will be 1592.81 increase in PRICE, provided all other variables are held the same.

>

> #part H

>

> dim(house)

[1] 117 5

>

>

> qt(0.975,61)

[1] 1.999624

>

> #since the t test statistic value is |3.257| which is more than critical t value which is |1.999624|, we reject H0 and accept the alternative. Therefore, TAX is a statistically significant variable.

>

>

I also asked you to ‘State your null and alternative hypotheses.’

**QUESTION 2**

> car1<-read.table(file.choose(),header=TRUE,sep=",")

> #the file name that I use here is "cars"

> #I will take a subset of the data that has 4 cylinders or more

> car2<-subset(car1,subset=Cylinders>3)

> attach(car2)

> #QUESTION 2

> #part A

>

> scatterplot=data.frame(MPG,Ratio,Horse,Displace,Cylinders)

> pairs(scatterplot,upper.panel=NULL)

>

> #From the scatterplot, we can say that some of the data do not have a strong linear relationship and it seems to have many outliers in each plot.

> #Variables Cylinders seems to have a strong linear relationship with all other variables.

We are trying to predict MPG as the Y variable, so it should be listed last so it goes in the bottom right hand corner as MPG on the y axis of all of the plots.

With MPG, there seems to be no relationship with Ratio, a strong linear relationship with Horse, a somewhat curved relationship with Displace, and a somewhat linear relationship with Cylinders. There are no obvious outliers on any of these plots.

> #part B

>

> lnMPG = log(MPG)

> hist(MPG)

>

>

> hist(lnMPG)





> #the histogram for MPG is seemed to be more right skewed because of the shape.

> #the histogram for log transformation of MPG seemed to be more symmetry than MPG and approximately more normally distributed.

>

>

I asked you to ‘Make 2 histograms, one on top of the other.’

> #part C

>

> model1=lm(MPG~Ratio+Horse+Displace+Cylinders)

> summary(model1)

Call:

lm(formula = MPG ~ Ratio + Horse + Displace + Cylinders)

Residuals:

Min 1Q Median 3Q Max

-4.8775 -2.1632 -0.4747 2.5626 5.2324

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 65.10001 6.52863 9.971 4.89e-11 \*\*\*

Ratio -5.61913 1.87393 -2.999 0.00541 \*\*

Horse -0.12418 0.04513 -2.751 0.00997 \*\*

Displace -0.03729 0.02354 -1.584 0.12365

Cylinders -0.72421 0.97588 -0.742 0.46379

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.044 on 30 degrees of freedom

Multiple R-squared: 0.8053, Adjusted R-squared: 0.7793

F-statistic: 31.02 on 4 and 30 DF, p-value: 2.863e-10

> model2=lm(lnMPG~Ratio+Horse+Displace+Cylinders)

> summary(model2)

Call:

lm(formula = lnMPG ~ Ratio + Horse + Displace + Cylinders)

Residuals:

Min 1Q Median 3Q Max

-0.19462 -0.08282 -0.01340 0.09181 0.21620

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.8207049 0.2489407 19.365 < 2e-16 \*\*\*

Ratio -0.2258693 0.0714541 -3.161 0.00358 \*\*

Horse -0.0048325 0.0017210 -2.808 0.00868 \*\*

Displace -0.0015871 0.0008975 -1.768 0.08719 .

Cylinders -0.0336921 0.0372108 -0.905 0.37245

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1161 on 30 degrees of freedom

Multiple R-squared: 0.8326, Adjusted R-squared: 0.8102

F-statistic: 37.29 on 4 and 30 DF, p-value: 3.074e-11

> #the R^2 value for model2 is 0.8326, which is higher than model1.

> #the R^2 adjusted value for model2 is 0.8102, which is higher than model1, that is 0.7793.

> #the F p-value for model2 is also smaller (3.074e-11) as compared to model1 (2.863e-10)

> #this means that model2 is better than model1.

> #As expected, model2 is preferable than model1 because itt can be seen from answer in part B.

>

>

Why is this expected? You need to explain this better.

> #part D

>

>

> givendata<-data.frame(Ratio=3,Horse=100,Displace=200,Cylinders=6)

> predict(model2,givendata,interval="prediction",level=.90)

fit lwr upr

1 3.140284 2.93795 3.342617

>

> exp(3.140284)

[1] 23.11043

>

> exp(2.93795)

[1] 18.87711

>

> exp(3.342617)

[1] 28.29307

>

> #the predicted MPG is 23.11043

> #90% prediction interval for MPG is (18.87711,28.29307)

>

> #part E

>

> #It can be interpreted that we are 90% certain that the Miles per Gallon will be between 18.88 and 28.29 for given data set.

>

We are 90% confident that an individual car with the Ratio of 3, Horse of 100, Displace of 200, and Cylinders of 6 will have predicted MPG between 18.877 and 28.293.

> #part F

>

> #We can say that 81.02% of the variation in predicting the y value is high, which means it is good. All other X variables is a good predictor of Y.

81.02% of the variation in log MPG is being explained by Ratio, Horse, Displace, and Cylinders, once we have taken a penalty for the number of x variables we are using.

> #part G

>

> qt(.90,30)

[1] 1.310415

If you were going to find the critical t value here, you would want 0.95 and not 0.90 as the area to the left. Since you have 0.05 as the area left over in each tail, which leaves alpha = 0.10 total.

>

> #since the t(Displace) is 1.768 which is more than t critical of 1.310415, and the p-value of 3.074e-11 which is so small (smaller than what?), we reject H0 and accept H1.

> #Variable Displace is a statistically significant variable.

>

I also asked you to ‘State your null and alternative hypotheses.’

> #part H

>

> model4=lm(lnMPG~Ratio+Horse+Cylinders)

> res1=residuals(model4)

> model5=lm(Displace~Ratio+Horse+Cylinders)

> res2=residuals(model5)

>

> cor(res1,res2)

[1] -0.3072215

>

> cor(lnMPG,Displace)

[1] -0.788916

>

> #it seems like they have a weak negative correlation. This means that the relationship between log MPG and Displace is low with the presence of other variables.

>