MTH 325, Lab #3, Spring 2023 Name _____

Instructions: Follow along with the tutorial portion of the lab. Replicate the code examples in R on your own, along with the demonstration. Then use those examples as a model to answer the questions/perform the tasks that follow. Copy and paste the results of your code to answer questions where directed. Submit your response file and the code used (both for the tutorial and part two). Your code file and your lab response file should each include your name inside.

Multiple Regression

We can look at our regression methods using the mtcars data set which has many numerical variables we can use to model mpg.

We'll start with backward selection. Begin with the full model.

```
 8  data(mtcars)
 9
10  fit<-lm(mpg~.,data=mtcars)
11  summary(fit)
```

We can indicate to the linear model function to use all the variables except mpg with just the . after the tilda.

As we saw in a previous lab, the variables suffer from collinearity issues and we start out with none of the specific coefficients having P-values less than 0.05. Start eliminating variables one at a time, beginning with the variable with the highest P-value.

```
13  fit<-lm(mpg~disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars)
14  summary(fit)
```

You can remove just one variable with the call lm(mpg~.-cyl, data=mtcars). This command uses all the variables except cyl. But since we're likely to need to remove additional variables, it may just be easier, as I did here, to list out the remaining variables and delete them from the function call one at a time.

Continue removing variables until all the remaining variables are statistically significant.

```
16  fit<-lm(mpg~disp+hp+drat+wt+qsec+am+gear+carb,data=mtcars)
17  summary(fit)
18
19  fit<-lm(mpg~disp+hp+drat+wt+qsec+am+gear,data=mtcars)
20  summary(fit)
21
22  fit<-lm(mpg~disp+hp+drat+wt+qsec+am,data=mtcars)
23  summary(fit)
24
25  fit<-lm(mpg~disp+hp+wt+qsec+am,data=mtcars)
26  summary(fit)
27
28  fit<-lm(mpg~hp+wt+qsec+am,data=mtcars)
29  summary(fit)
```

```
30
31  fit<-lm(mpg~wt+qsec+am,data=mtcars)
32  summary(fit)
33
34  fit<-lm(mpg~wt+qsec+am+0,data=mtcars)
35  summary(fit)
36
```

```
Call:
lm(formula = mpg ~ wt + qsec + am + 0, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8820 -1.5401 -0.4246  1.6623  4.1711

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
wt    -3.1855     0.4828  -6.598 3.13e-07 ***
qsec   1.5998     0.1021  15.665 1.09e-15 ***
am     4.2995     1.0241   4.198 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.497 on 29 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9858
F-statistic:   741 on 3 and 29 DF,  p-value: < 2.2e-16
```

Forward selection works similarly, but you add in variables one at time. Start with a one-variable model. Then add in one variable at a time. Retain it if it is statistically significant. Remove it and try another if it is not. Continue until you have tried and/or eliminated all the variables.

We can also automate this process to get us closer to the final model more quickly.

```
36
37  library(MASS)
38
39  fit<-lm(mpg~.,data=mtcars)
40  summary(fit)
41  step.model <- stepAIC(fit, direction = "both", trace = FALSE)
42  summary(step.model)
43
```

The direction can be “forward”, “backward” or “both”. The model we end up with here is similar to our best backward model, but with the intercept still in the model. But this process eliminated most of the steps. This can be extremely helpful if you have many, many variables.

We can use the best subset method to find good combinations of variables we might not otherwise find in the stepwise approach.

```
43
44  library(leaps)
45
46  models <- regsubsets(mpg~., data = mtcars, nvmax = 10)
47  summary(models)
48
```

```
Selection Algorithm: exhaustive
          cyl disp hp  drat wt  qsec vs  am  gear carb
1  ( 1 )  " " " "  " " " " "*" " "  " " " " " "  " "
2  ( 1 )  "*" " "  " " " " "*" " "  " " " " " "  " "
3  ( 1 )  " " " "  " " " " "*" "*"  " " "*" " "  " "
4  ( 1 )  " " " "  "*" " " "*" "*"  " " "*" " "  " "
5  ( 1 )  " " "*"  "*" " " "*" "*"  " " "*" " "  " "
6  ( 1 )  " " "*"  "*" "*" "*" "*"  " " "*" " "  " "
7  ( 1 )  " " "*"  "*" "*" "*" "*"  " " "*" "*"  " "
8  ( 1 )  " " "*"  "*" "*" "*" "*"  " " "*" "*"  "*"
9  ( 1 )  " " "*"  "*" "*" "*" "*"  "*" "*" "*"  "*"
10  ( 1 ) "*" "*"  "*" "*" "*" "*"  "*" "*" "*"  "*"
```

The output tells you which variables are in or out in the best models.

This output tells you which combination of variables is best with the specified number of variables. The three-variable model matches our best three-variable model from backward selection. We can then test these models for further analysis.

Once you've selected your best model, follow through with the diagnostic tests we examined in the last lab, looking for whether we follow model assumptions and outliers.

In addition to our diagnostic graphs, you can conduct a hypothesis test to look for outliers in your dataset among the residuals. Below, I've used the Rosner test since it allows us to test for more than one outlier at a time.

```
55
56  library(EnvStats)
57  rosnerTest(mtcars$residuals, k = 3)
```

If you elect to remove outliers, rerun your model and examine any changes in the model to determine if the outcome is significantly different.

In some datasets, we may want to look at a scatterplot with more than one variable at a time. We can only do this with three variables, but sometimes this can be useful.

```
60  library("scatterplot3d")
61  scatterplot3d(trees)
```

There are several packages that can produce 3D scatterplots. Another option is plotly. This package has a lot functionality, but it can take some practice to master the ins and outs of the syntax, and if you want to use it in a publication, there are restrictions on that. An example using the mtcars dataset is shown below. 3D graphs can be useful when they are not static, but can be harder to reader when they don't rotate easily.

```r
library(plotly)

mtcars$am[which(mtcars$am == 0)] <- 'Automatic'
mtcars$am[which(mtcars$am == 1)] <- 'Manual'
mtcars$am <- as.factor(mtcars$am)

fig <- plot_ly(mtcars, x = ~wt, y = ~hp, z = ~qsec, color = ~am, colors = c('#BF382A', '#0C4B8E'))
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'Weight'),
                                   yaxis = list(title = 'Gross horsepower'),
                                   zaxis = list(title = '1/4 mile time')))

fig
```

Sources for more options and features for the two 3D scatterplots are linked in the reference list.

Tasks

1. Use the data in **325lab3data.xlsx** file. Import the data into R and conduct a thorough multiple regression analysis of the data. (Recall that House is not a variable. You may need to remove it from the dataset.) Choose one or more selection methods to determine the subset of variables to use to model the selling price from the other variables. Assess the variables using a correlation table/correlation plot or pairplot. Select the best model that avoids overfitting. Conduct a thorough diagnostic test of the residuals including outliers and influential points tests. Determine whether to remove any outliers, if any, or not. Clearly explain your process and reasoning along the way. Present your final model. Use your model to predict the selling price of a home with 2500 square feet, 7 rooms, 10 years old, and with an attached garage. Include a discussion of all supporting graphs and hypothesis tests.

Next week: we will have class time to work on the tutorial assignment.

References:

1. Discovering Statistics Using R. Andy Field, Jeremy Miles, Zoe Field. (2012)
2. http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf
3. https://book.stat420.org/applied_statistics.pdf
4. https://scholarworks.montana.edu/xmlui/bitstream/handle/1/2999/Greenwood_Book_Version3_CC_optimized.pdf?sequence=7&isAllowed=y
5. https://www.rstudio.com/resources/cheatsheets/
6. http://r-statistics.co/Outlier-Treatment-With-R.html
7. https://statsandr.com/blog/outliers-detection-in-r/
8. https://rpubs.com/mpfoley73/501093
9. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/
10. https://towardsdatascience.com/selecting-the-best-predictors-for-linear-regression-in-r-f385bf3d93e9
11. http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization
12. https://plotly.com/r/3d-scatter-plots/