Homework\_4\_FINAL

Lincoln Bohn

11/4/2021

Homework Part 1

rep(0, 5) # Replicate 0, 5 times

## [1] 0 0 0 0 0

rep(1:3, 4) # Replicate 1-3, 4 times

## [1] 1 2 3 1 2 3 1 2 3 1 2 3

rep(4:6, c(1, 2, 3))# Replicate 4-6, repeat 1st variable 1 time, repeat 2nd variable 2 times, repeat 3rd variable 3 times

## [1] 4 5 5 6 6 6

My\_Variable <- 3 # Defines My\_Variable as 3  
 print(My\_Variable) # Print My\_Variable

## [1] 3

seq(1,10, by = 2) # Sequence 1-10, incrementally by 2

## [1] 1 3 5 7 9

vector\_1 <- c(1, 2, 3, 4, 5) # Create a vector variable vector\_1, defines each row as it aligns inside c(...)  
 print(vector\_1) # Print vector\_1

## [1] 1 2 3 4 5

Homework Part 2

data(mtcars) # Imports mtcars data  
names(mtcars) # Names of mt cars variables

## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"  
## [11] "carb"

ncol(mtcars) # Number of columns in mtcars

## [1] 11

nrow(mtcars) # Number of rows in mtcars

## [1] 32

#3 How many rows and how many variables belong to this dataframe? What does this dataset represent and where does it come from?

There are 32 rows (observations) and 11 columns (variables) | The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

#4 Descriptive statistics by function

#Seperate functions for descriptive statistics in mtcars, mpg column  
  
min(mtcars$mpg)

## [1] 10.4

median(mtcars$mpg)

## [1] 19.2

mean(mtcars$mpg)

## [1] 20.09062

max(mtcars$mpg)

## [1] 33.9

#4B Summary of common descriptive statistics

#Summary of common descriptive statistics   
summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

#5 Creates a tables showing the summary statistics of mpg by cyl.

mtcars\_df <- mtcars # mtcars\_df is defined with data from mtcars for later manipulation if needed  
 aggregate(mpg ~ cyl, data = mtcars\_df, summary) # Computes summary of mpg by cyl, using mtcars\_df, with a summary of descriptive statistics

## cyl mpg.Min. mpg.1st Qu. mpg.Median mpg.Mean mpg.3rd Qu. mpg.Max.  
## 1 4 21.40000 22.80000 26.00000 26.66364 30.40000 33.90000  
## 2 6 17.80000 18.65000 19.70000 19.74286 21.00000 21.40000  
## 3 8 10.40000 14.40000 15.20000 15.10000 16.25000 19.20000

min\_ag <- # Defines variable min\_ag as the minimum mpg explained by cyl.  
 aggregate(mpg ~ cyl,  
 data = mtcars\_df,   
 FUN = min  
 )  
min\_ag # Prints min ag

## cyl mpg  
## 1 4 21.4  
## 2 6 17.8  
## 3 8 10.4

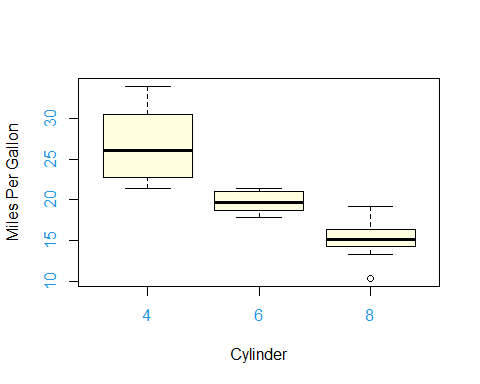
max\_ag <- # Defines variable max\_ag as the maximum mpg explained by cyl  
 aggregate(mpg ~ cyl,  
 data = mtcars\_df,   
 FUN = max  
 )  
max\_ag # Prints max ag

## cyl mpg  
## 1 4 33.9  
## 2 6 21.4  
## 3 8 19.2

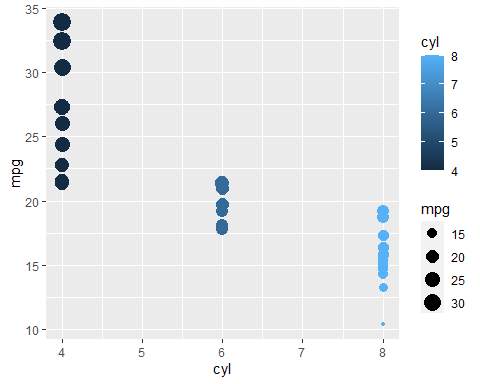
The least fuel efficient 4 cyl car is: Volvo 142E 21.4  
  
The least fuel efficient 6 cyl car is: Merc 280C 17.8  
  
The least fuel efficient 8 cyl car is: Cadillac Fleetwood & Lincoln Continental 10.4  
  
The most fuel efficient 8 cyl car is: Pontiac Firebird 19.2

#6 Create a boxplot showing MPG by cylinder

mpg\_cyl\_bplot <- # Defines variable as a boxplot of mpg by cyl, from mtcars, with some renaming and recoloring  
 boxplot(mpg ~ cyl, data = mtcars, xlab = "Cylinder", ylab = "Miles Per Gallon", col = "light yellow", col.axis = "12")

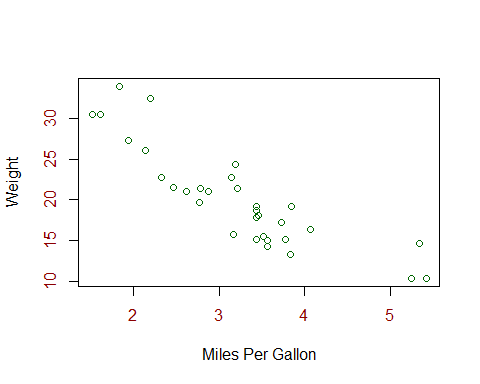


library(ggplot2)  
 # Testing ggplot package   
ggplot(mtcars) +  
 geom\_point(mapping = aes(x = cyl, y = mpg, color = cyl, size = mpg))



#7 Create a scatterplot of MPG versus weight

# Scatterplot of wt and mpg, renamed axis and colors  
plot(mtcars$wt,   
 mtcars$mpg,   
 xlab = "Miles Per Gallon",   
 ylab = "Weight",  
 col = "Dark Green",  
 col.axis = "darkred")



In my view, Miles Per Gallon is negatively correlated with Car Weight. The strength of this relationship looks above R= -.5

#8 What is the correlation coefficient between MPG and weight? Does this confirm your answer to the previous question?

# Prints correlation coefficient of the two variables  
cor(mtcars$wt,   
 mtcars$mpg)

## [1] -0.8676594

Yes. I believed it would be over .5, it is stronger than my estimate. I was being too conservative.

#9 Create a linear regression that predicts MPG as a function of cylinders, displacement, horsepower, and weight.

# Defines mpgreg as the linear model of mpg as a function of cyl, displacement, horsepower, and weight, from mtcars data  
mpgreg <-  
 lm(mpg ~ cyl + disp + hp + wt,  
 data = mtcars)  
# Prints summary of mpgreg variable  
summary(mpgreg)

##   
## Call:  
## lm(formula = mpg ~ cyl + disp + hp + wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0562 -1.4636 -0.4281 1.2854 5.8269   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.82854 2.75747 14.807 1.76e-14 \*\*\*  
## cyl -1.29332 0.65588 -1.972 0.058947 .   
## disp 0.01160 0.01173 0.989 0.331386   
## hp -0.02054 0.01215 -1.691 0.102379   
## wt -3.85390 1.01547 -3.795 0.000759 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.513 on 27 degrees of freedom  
## Multiple R-squared: 0.8486, Adjusted R-squared: 0.8262   
## F-statistic: 37.84 on 4 and 27 DF, p-value: 1.061e-10

#10 How strong is the model overall? Which variables were significant predictors of MPG?

The model is quite strong with an adjusted R-squared of 0.82, weight is a significant predictor at the strongest level possible, followed by Cylinder at the .05 level. Horsepower was not found to be a significant predictor of MPG

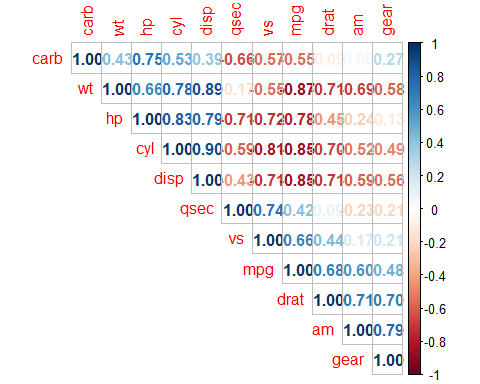
Homework Part 3

#11 Create a plot of the correlation matrix between all the variables in the mtcars dataset.

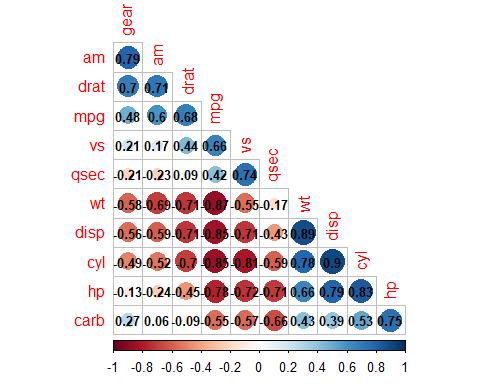
library(corrplot) #loads corrplot in this chunk

## corrplot 0.92 loaded

mtcars\_cor <- cor(mtcars\_df) #defines correlation of all variables in mtcars  
  
 corrplot(mtcars\_cor, # using correlations from mtcars data, defines the graphic printed below.  
 method = 'number', # Shows correlation coefficient strength in graphic  
 type = "upper", #  
 order = "hclust")



# Another correlation matrix graphic from researching ??corrplot  
   
 corrplot(mtcars\_cor,  
 p.mat = ,  
 method = 'circle',  
 type = 'lower',  
 insig='blank',  
 addCoef.col ='black',  
 number.cex = 0.8,  
 order = 'AOE',  
 diag=FALSE)



#11B What does this uncover about the correlations between the explanatory variables in the model you constructed for #9?

This graphic uncovers the strength of the correlation we did from #9, If we go to a signifcant correlation like mpg ~ cyl, we can see -.85, which is a strong relationship. I would suspect this. Lets see mpg ~ wt, -.87, a little stronger, but also expected. Finally, mpg ~ hp also shows a -.78 strong negative relationship, but from #9 p=.102 which is telling us that there is a large chance that gross horsepower is not a good explanatory for mpg, because it could be due to another, possibly random, factor

#12 Re-create the regression that you ran for #6 in the SPSS homework assignment within R and interpret the results.

library(foreign) #library "foreign" package  
  
  
df\_spss\_BB <- # Defines df\_spss\_BB as reading an SPSS file, and using its original labels imported into the data frame  
 read.spss("CPUC\_Broadband\_CTs.sav", use.value.label=TRUE, to.data.frame=TRUE)

#12B Construct another multivariate regression that predicts Adoption Rate, this time including at least five independent variables, two of which should be ethnic/racial variables. How strong is the model overall? Which variables are statistically significant predictors?

spss\_BB\_lm <- # Defines variable as linear model of adoption rate as explained by other variables  
 lm(AdoptionRate ~ medage + MHI + White + Hispanic + pct25wbach, data = df\_spss\_BB)  
 # Summary of desctriptive statistics for the linear model  
summary(spss\_BB\_lm)

##   
## Call:  
## lm(formula = AdoptionRate ~ medage + MHI + White + Hispanic +   
## pct25wbach, data = df\_spss\_BB)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78227 -0.07462 0.01350 0.09151 0.52559   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.445e-01 1.836e-02 29.653 < 2e-16 \*\*\*  
## medage -1.245e-03 4.067e-04 -3.062 0.00221 \*\*   
## MHI 2.891e-06 1.128e-07 25.621 < 2e-16 \*\*\*  
## White -6.649e-02 1.308e-02 -5.085 3.8e-07 \*\*\*  
## Hispanic -1.136e-01 1.364e-02 -8.324 < 2e-16 \*\*\*  
## pct25wbach 3.067e-01 2.000e-02 15.331 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1593 on 5383 degrees of freedom  
## (48 observations deleted due to missingness)  
## Multiple R-squared: 0.4334, Adjusted R-squared: 0.4329   
## F-statistic: 823.6 on 5 and 5383 DF, p-value: < 2.2e-16

Adjusted R-squared values are the same across spss and R calculations. spss = .433, .4329 in the model above, so even further decimal points, though they are not necessary. Therefore, the model is medium strength, showing our calculations are correct across two programs!   
Nearly All variables are significant at the strongest level, this was true in both programs. In both programs, medage was slightly less significant, but it is negligible at p = .002, whereas in this program it is p=0.00221.