SDS230\_Final\_Project

### Introduction

This project will examine data from Ken French’s website which provides information on the average firm size, monthly return, and beta for 49 different industries, as well as the overall market return and risk free rate. The questions of the project are two-fold: First, we wish to assess whether the returns of different industries behave similarly (do they have similar mean returns, are they correlated, do they have similar book equity to market equity ratios, etc). Second, we wish to assess whether the Capital Asset Pricing Model accurately predicts the cross sectional returns of the various industries.

CAPM is a model that predicts a portfolio’s returns given the portfolio’s market risk. We will test the CAPM using the method pioneered in Fama-Macbeth (1973) by first calculating each industry’s market beta, and then running cross sectional regressions to see whether beta explains the average returns on each industy.

### Data

The variables we plan to use and create in this dataset are:

**Year**

**Industry Percent Return** - The percent return on investment for a given industry in a given year.

**Industry** - the 49 individual industries.

**Composite Industry** - larger industry category that aggregates the 49 individual industries (e.g. TMT, Healthcare…).

**Average Firm Size** - The average market capitalization value of a firm in a given year represented in millions of dollars.

**Book Equity / Market Equity** - a measure of a company’s asset value on the books relative to market valuation.

**Market Return** - Defined as “The return on the overall theoretical market portfolio which includes all assets and having the portfolio weighted for value.”

**Risk Free Rate** - Defined as “the rate of return a hypothetical investment with no risk of financial loss”.

**Beta** - Defined as “a measure of a stock’s volatility in relation to the market” The formula for Beta is:

where = the return on an individual portfolio and = the market return.

**Composite Returns** - The percent return on investment for a given composite industry (larger industry catagory) in a given year.

**Composite BE/ME Ratio** - a measure of a company’s asset value on the books relative to market valuation by composite industry.

**Composite Beta** - the composite industry average of “volatility in relation to the market.”

**Composite Size** - the average market capitalization of firms within a composite industry over a year.

The data sets we collected were formatted in such a way that required extensive data cleaning before beginning our analysis. Below are the steps we took:

**First**, we needed to resolve cases of incomplete data across all of our data sets.

**Second**, the “returns” and “size” data was provided given on a monthly level from 1926 to 2009, and we needed to aggregate on a yearly level by applying the average function on all the months in each given year.

**Third**, the data provided in the four different data sets, “BE” (Book to Equity), “size” (Market Value), “returns” (returns in percent for the given period), “rf” (market return for that period), was organized horizontally, with years as rows, each industry (e.g. Agric, Food, Soda…) as columns, and then returns by year and industry in the middle. We needed to aggregate the various industry columns names into one “industry” column, which required transposing each data set.

**Fourth**, in the “BE”, “size”, and “returns” data sets, there were 50 unique industries. When we took a closer look, it was clear that the industries were quite specific, so we needed to aggregate these industries into larger industry categories which we called “Composite Industries” (e.g. Healthcare, Industrials, and TMT). This industry aggregation process required averaging book to equity ratios, returns, sizes of each industry within Composite Industries.

**Fifth**, we created a new “beta” variable based on the formula indicated above for each composite industry for each year.

**Sixth**, once we formatted each of the four data sets in similar manners, we merged them into one data frame to be able to run tests and analyses on the data.

### Data Cleaning

First, across all datasets, when an observation has a firm size, percent return, or beta ratio of zero, the dataset reads -99.99. We will begin by transforming all of these values to zero.

BE[BE== -99.99] <- NA  
size[size == -99.99] <- NA   
returns[returns == -99.99] <- NA   
rf[rf == -99.99] <- NA

Next, we will reformat the year variable to include only the year as oppossed to year-month while finding the yearly averages for the “BEME” and “size” dataframes, while fining total yearly return for the “returns” dataframe. We then adjust the dataframes so the years line up correctly.

size$Year <- substr(size$Year, 1, 4) #Strip off all ofthe month values   
size <- aggregate(.~Year, data = size, mean, na.action = na.pass)  
returns$Year <- substr(returns$Year, 1, 4)  
returns <- aggregate(.~Year, data = returns, total\_return, na.action = na.pass)  
rf$Year <- substr(rf$Year,1,4)  
rf <- aggregate(.~Year, data= rf, total\_return, na.action = na.pass)  
rf$mktret <- rf$Mkt.RF+rf$RF #This creates the column for market return - done by adding back the risk free rate to the risk free market return

We want the data at each time click to be the data that would have been available to someone at that time. Because size and BE/ME are backwards looking calculations but predictive variables of returns, they contain information from the future that would not have been avaialble to someone trying to make a prediction. Therefore, we lag both of these variables by one year.

Before using our datasets, we need to convert them from wide form to long form. Using the melt function, we alter our dataframe such that it transforms from a 89x50 dataframe to a 4361x3 dataframe. We then adjust the names of the columns, eliminate repeated columns, and repeat for our other dataframes (returns and size). This step and the following couple will have muted code to save space. See the RMD file for the code that does this.

Once we have transformed the dataframes into the appropriate form, we can then merge the datasets together into the final, usable dataframe. Because year and industry are repeated so many times in each of the datasets, to merge, we must create a YearInd variable that represents an industry in a unique year. After merging, we delete the repeated columns and rename the remaining ones.

To expand our realm of analysis, we have decided to create a number of variables: composite industry (categorical) and Beta (continuous). Below is the series of code used to create our new variables. We begin with our composite industry variable which recodes our 49 independent industries into eight overarching industries (Healthcare, Industrials, TMT, Financials, Consumer, Utilities, Commodities, and Other).

Next, we create Beta for each industry, which represents the covariance of market returns and a given industry returns divided by the variance in market returns.

df <- merge(df, rf, by = "Year")  
df$Year <- as.factor(df$Year)  
df$Beta <- NA  
for(i in unique(df$Ind)){  
 temp <- df[df$Ind == i,]  
 a <- (cov(temp$returns,temp$mktret))/(var(temp$mktret))  
 df$Beta[df$Ind==i] <- a  
}

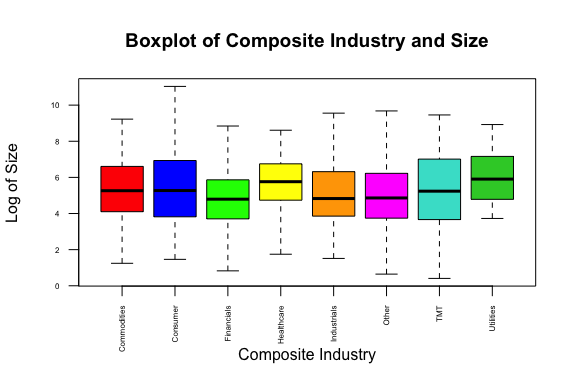
Here, we create all of the composite variables for our dataset such as Composite Industry Returns, Composite Size, and Composite BE/ME ratio. We then merge these into a new dataframe.

indreturnscomp <- aggregate(df$returns ~ df$CompInd + df$Year, df, mean)  
colnames(indreturnscomp) <- c("CompInd", "Year", "Composite Returns")  
aggcompret <- merge(df, indreturnscomp, all.x=T, by = c("Year", "CompInd"))  
  
#composite industry size   
indsizecomp <- aggregate(df$size ~ df$CompInd + df$Year, df, mean)  
colnames(indsizecomp) <- c("CompInd", "Year", "Composite Size")  
#indsizecomp  
aggcompsize <- merge(aggcompret, indsizecomp, all.x=T, by = c("Year", "CompInd"))  
  
#composite BE/ME ratio  
indbemecomp <- aggregate(df$BEMERatio ~ df$CompInd + df$Year, df, mean)  
colnames(indbemecomp) <- c("CompInd", "Year", "Composite BE/ME Ratio")  
aggcompbeme <- merge(aggcompsize, indbemecomp, all.x=T, by = c("Year", "CompInd"))  
  
#composite beta and final dataframe  
indbetacomp <- aggregate(df$Beta ~ df$CompInd + df$Year, df, mean)  
colnames(indbetacomp) <- c("CompInd", "Year", "Composite Beta")  
aggcompfin <- merge(aggcompbeme, indbetacomp, all.x=T, by = c("Year", "CompInd"))  
  
#final dataframe  
aggcompfin$Year <- as.numeric(aggcompfin$Year)

### Graphics

First, we examine a boxplot between firm size and composite industry to examine if there are noticeable differences in the log size between the composite industries. We would need to use a t-test of some sort to examine whether these differences were statistically significant. We added 1 inside the log because some of the firm sizes are technically zero, which causes an error given that we are taking the natural log.

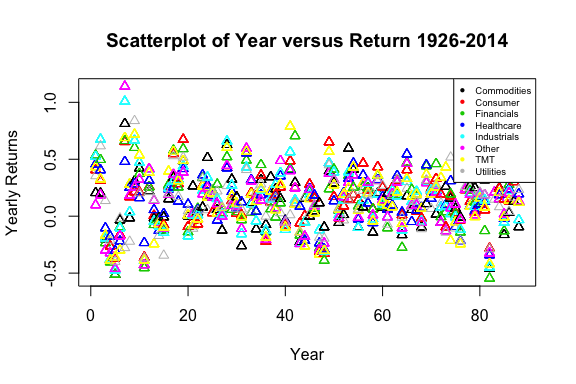
boxplot(log(df$size + 1)~df$CompInd, main = "Boxplot of Composite Industry and Size", col = c("red", "blue", "green", "yellow","orange", "magenta", "turquoise", "lime green"), ylab = "Log of Size", xlab = 'Composite Industry', las = 2, cex.axis = .5)



Overall, the industries appear to have relatively similar means and variances and there are no observable skews within any of the industries.

We now examine the returns for each composite industry by year from 1926-2014.

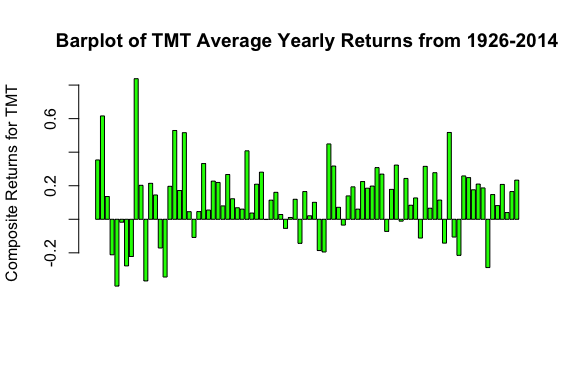
aggcompfin$CompInd <- as.factor(aggcompfin$CompInd)  
plot(aggcompfin$`Composite Returns` ~ aggcompfin$Year, main = "Scatterplot of Year versus Return 1926-2014", pch = 24, col = aggcompfin$CompInd, xlab = "Year", ylab = "Yearly Returns")  
legend("topright", legend= levels(aggcompfin$CompInd), pch=16, cex = 0.6, col = c(1:9))



Overall, there seems to be a lot of variance in the returns throughout the entire period. Given the number of composite industries, it is difficult to see any single trend in the graph. As such, we will have to run regression analyses by industry to verify if there are trends for each.

Now, we will compare specific composite industries to observe how our continuous variables have changed throughout time. We have elected two composite industries at random (TMT and Commodities) in order to limit the number of graphs we are making. We will also limit the number of graphs we print for space purposes.

#barplot(single$Composite\_BEME\_Ratio\_TMT, main = "Barplot of TMT BE/ME Ratio from 1926-2014", col = "blue", ylab = "Composite BE/ME for TMT")  
barplot(single$Composite\_Returns\_TMT, main = "Barplot of TMT Average Yearly Returns from 1926-2014", col = "green", ylab = "Composite Returns for TMT")

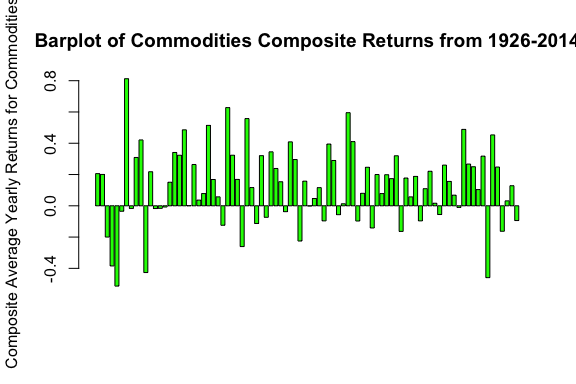


#barplot(single$Composite\_Size\_TMT, main = "Barplot of TMT Composite Size from 1926-2014", col = "green", ylab = "Composite Average Size for TMT")

We conclude from the three graphs that 1) TMT company valuations spiked in the 1940s, 2) Average yearly returns in the TMT industry were net positive from 1926-2014, and 3) the market value of TMT companies decreased during the Great Depression and subsequently grew from 1940s to 1970s.

Now, we examine the same graphs for the Commodities industry.

#barplot(single$Composite\_BEME\_Ratio\_Com, main = "Barplot of Commodities BE/ME Ratio from 1926-2014", col = "green", ylab = "Composite BE/ME for Commodities")  
barplot(single$Composite\_Returns\_Com, main = "Barplot of Commodities Composite Returns from 1926-2014", col = "green", ylab = "Composite Average Yearly Returns for Commodities")



#barplot(single$Composite\_Size\_Com, main = "Barplot of Commodities Size from 1926-2014", col = "green", ylab = "Composite Size for Commodities")

We conclude from the three graphs that 1) Commodities company valuations spiked in the 1930s before the worst of the Great Depression, 2) Average yearly returns in the Commodities industry were net positive from 1926-2014, and 3) the market value of Commodities companies grew consistently from 1940s to 2010s.

### Basic Tests

Because our data does not include any two-level categorical groups, we find it more useful to examine the correlations between our different continous composite variables which we made graphs for above. However, since our new dataframe only has two industries, we can use a basic t-test to compare the four composite variables. Again, we removed the box plots to save space.

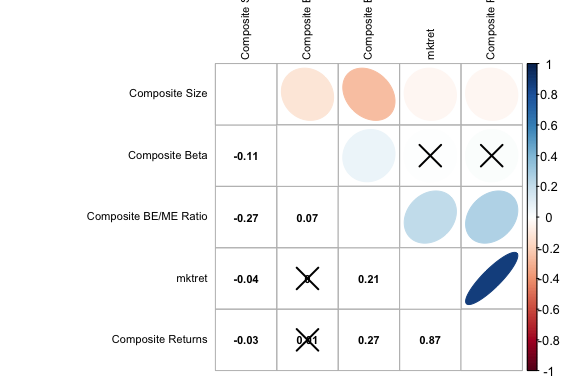
t.test(single$Composite\_Returns\_TMT, single$Composite\_Returns\_Com)

##   
## Welch Two Sample t-test  
##   
## data: single$Composite\_Returns\_TMT and single$Composite\_Returns\_Com  
## t = -0.35297, df = 172.05, p-value = 0.7245  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.08092492 0.05637315  
## sample estimates:  
## mean of x mean of y   
## 0.1140227 0.1262986

#t.test(single$Composite\_Size\_TMT, single$Composite\_Size\_Com)  
#t.test(single$Composite\_BEME\_Ratio\_TMT, single$Composite\_BEME\_Ratio\_Com)

Based on the t-test, Composite Returns between TMT and Commodities are not statistically significantly different at any reasonable level of alpha. By contrast, size for the two industries is statistically significant at alpha = 0.05 as the p-value is 0.04651. Based on the boxplot, it appears that TMT is larger than commodities for size. Finally, the BE/ME Ratio between the two industries is statistically significantly different at the 0.05 alpha level as the p-value is 0.02033. Based on the boxplot, the BE/ME Ratio for Commodities is larger than TMT. This makes intuitive sense because tech stocks are typically valued at higher multiples of their revenue because more future growth is priced in.

#Now we check for significant correlations at the 95% level and plot them.   
sigcorr <- cor.mtest(aggcompfin[,c(9, 11:14)], conf.level = .95)  
corrplot.mixed(cor(aggcompfin[,c(9, 11:14)]), lower.col="black", upper = "ellipse", tl.col = "black", number.cex=.7,   
 order = "hclust", tl.pos = "lt", tl.cex=.7, p.mat = sigcorr$p, sig.level = .05) #Adjust the top axis text

 Overall, the correlations between variables are relatively low. The highest correlation is between market return and composite returns for all of the industries. Composite Beta and Composite BE/ME ratio both have a low negative correlation with composite size, but the correlation is statistically significant at the 95% level. Other than these several correlations, the others are too low to hold much interpretation. We will proceed by analyzing these three correlations through bootstraps and permutation tests.

### Permutation Test

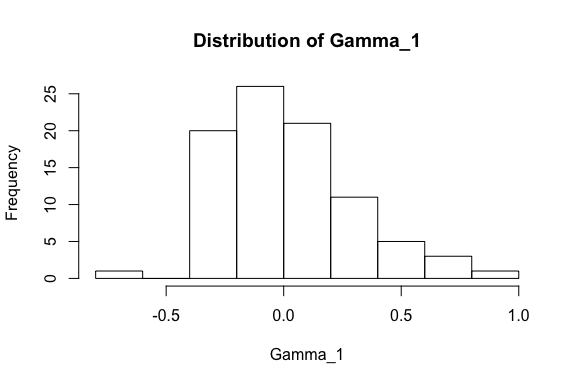
We now conduct a permutation test between composite industry return and composite BE/ME in order to show a non-zero statistically significant correlation between these two variables. The code is not included to save space.

Based on the permuted sample correlations, the actual correlation between composite industry return and composite BE/ME appears to be significant at any reasonable level of alpha. [add small analysis]

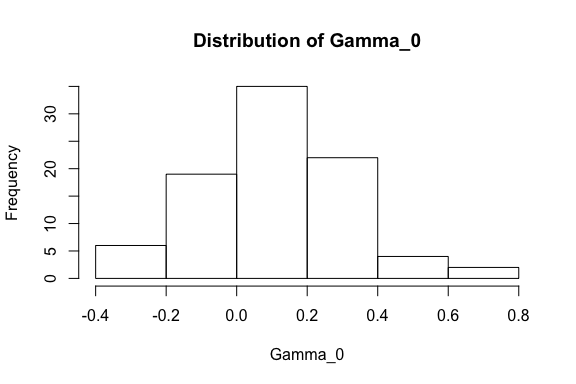
### Regression Testing CAPM and Bootstrap

As mentinoed in the introduction, CAPM is a model for returns that says the excess returns of a portfolio (returns greater than the risk free rate) depend solely on the amount of market risk associated with that portfolio. Market risk can be estimated by calculating a portfolio’s beta with the market, as we have already done by taking the covariance of the market and each portfolio scaled by the standard deviation of each. Now, to test the CAPM we will use the model , to represent a single cross section of returns for all the industries in our data set. We will run this regression for each cross section and then estimate and from the means we get from each iteration of the cross sectional regression. We can then test the CAPM by testing whether the means for and are equal to 0 using a t test. If the returns are fully explained by market risk, or , then should be approximately equal to 0 (technically it should be equal to the risk free rate, but since that is so close to 0 we can approximate it as such for ease of calculation). This is because if a portfolio is compensated for reasons other than market risk, will capture those other sources of return by being larger than 0. Therefore, we will reject the CAPM if is statistically significantly different from 0. We will also test if is significantly different from 0, because if market risk predicts returns then its coefficient () should be different from 0.

intercepts <- c()  
coefficients <- c()  
for (i in 1:length(unique(df$Year))){  
 year <- unique(df$Year)[i]  
 #print(df$returns[df$Year == year])  
 temp\_df <- na.omit(data.frame(df$returns[df$Year == year], df$Beta[df$Year == year]))  
 names(temp\_df) <- c('returns', 'beta')  
 intercepts <- append(intercepts, as.double(lm(temp\_df$returns ~ temp\_df$beta)$coefficients[1]))  
 coefficients <- append(coefficients, as.double(lm(temp\_df$returns ~ temp\_df$beta)$coefficients[2]))  
}  
hist(coefficients, main = 'Distribution of Gamma\_1', xlab = 'Gamma\_1')



hist(intercepts, main = 'Distribution of Gamma\_0', xlab = 'Gamma\_0')



means <- c(mean(intercepts), mean(coefficients))  
sds <- c(sd(intercepts), sd(coefficients))  
tstats <- c(mean(intercepts) / (sd(intercepts)/sqrt(length(intercepts))), mean(coefficients) / (sd(coefficients)/sqrt(length(coefficients))))  
results\_table <- data.frame(c('Gamma\_0', 'Gamma\_1'), means, sds, tstats)  
names(results\_table) <- c('Variable', 'Mean', 'SE', 't-stat')  
  
print('The following table contains the resulting means, standard errors, t statistics and P-values for Gamma\_0 and Gamma\_1')

## [1] "The following table contains the resulting means, standard errors, t statistics and P-values for Gamma\_0 and Gamma\_1"

results\_table

## Variable Mean SE t-stat  
## 1 Gamma\_0 0.10856513 0.2035720 5.0028070  
## 2 Gamma\_1 0.02577306 0.2777121 0.8705876

First, we printed a histogram of to make sure the distribution appeared symmetrical and approximately normally distributed (we need this assumption for the t test). It does appear to From the table, we can see that our test clearly rejects the CAPM. At a significance level of alpha = 0.05, the critical value for a 2 sided test would be approximately 2, meaning a t stat of 5 for means we reject the null that it is equal to 0, and therefore we reject that the CAPM fully explains the cross sectional returns. Similarly, we see that the t stat for is much lower than the critical value of 2 (or even the critical value for a 1 sided test, which would be 1.645). Therefore, we fail to reject the null that is different from 0. Again, this is evidence against the CAPM.

Just to be sure of our results, we will now create a bootstrapped confidence interval for the true mean of .

samp\_mean <- function(x, i) {  
 mean(x[i])  
}  
results <- boot(intercepts, statistic=samp\_mean,   
 R=1000)  
# get 95% confidence interval   
boot.ci(results, type="bca")

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 1000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = results, type = "bca")  
##   
## Intervals :   
## Level BCa   
## 95% ( 0.0684, 0.1546 )   
## Calculations and Intervals on Original Scale

We can see here too that 0 is not contained in the bootstrapped confidence interval for , confirming what we observed in the t test earlier.

Now, we will attempt to find the best model to predict returns, given the data we have. This means we will use the industries, composite industries, BE/ME, and size measures in addition to beta. Some important things to note: measures of size and book equity to market equity will be lagged by 1 time step because that data would not have been available at the time of prediction. We should also note that when calculating beta, we used the entire sample of returns, which we did to get the most accurate measure of each industry’s beta. However, that inevitably includes forward looking data, meaning that in practice the predictive power of beta will likely be lower out of sample. We will be doing a time series regression this time (as opposed to cross sectional), meaning that we won’t do a seperate regression for each year. Instead, our aim will be to predict returns in any time period given the variables. Note: For space purposes, only the final model in the stepwise regression will be printed.

data <- na.omit(df)  
#mod1 <- lm(data$returns ~ data$size + data$BEMERatio + data$Beta + data$CompInd + data$Ind + data$size\*data$Beta + data$size\*data$BEMERatio + data$BEMERatio\*data$Beta + data$CompInd\*data$size, data=data)  
#ols\_step\_backward\_p(mod1, prem = 0.05)  
mod2 <- lm(data$returns ~ data$CompInd + data$BEMERatio\*data$Beta, data=data)  
summary(mod2)

##   
## Call:  
## lm(formula = data$returns ~ data$CompInd + data$BEMERatio \* data$Beta,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.56132 -0.17768 -0.00031 0.16543 2.26552   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.089109 0.038918 2.290 0.0221 \*  
## data$CompIndConsumer 0.044871 0.019717 2.276 0.0229 \*  
## data$CompIndFinancials 0.019047 0.022603 0.843 0.3995   
## data$CompIndHealthcare 0.058519 0.023349 2.506 0.0122 \*  
## data$CompIndIndustrials 0.043752 0.020110 2.176 0.0297 \*  
## data$CompIndOther 0.008321 0.020321 0.409 0.6822   
## data$CompIndTMT 0.049626 0.021773 2.279 0.0227 \*  
## data$CompIndUtilities 0.007989 0.035895 0.223 0.8239   
## data$BEMERatio 0.004399 0.027163 0.162 0.8714   
## data$Beta -0.033615 0.031578 -1.064 0.2872   
## data$BEMERatio:data$Beta 0.052648 0.023348 2.255 0.0242 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2932 on 3509 degrees of freedom  
## Multiple R-squared: 0.03373, Adjusted R-squared: 0.03097   
## F-statistic: 12.25 on 10 and 3509 DF, p-value: < 2.2e-16

Using Backwards Stepwise regression, we eliminate the industry categorical variable, size, and all interaction terms except for the interaction between BEMERatio and Beta (this is why BEME and Beta are included despite them being insignificant on their own. The model automatically includes them because we include the interaction). There are also some composite industries that aren’t significant, but because there are some significant industries within the composite industry categorical variable we have to include them all in the regression.

Looking at the residual plots, we see that the residuals are approximately symmetric however there are some outliers. The residuals also aren’t perfectly normally distributed as we can see from the quantile plot, suggesting that our fit isn’t great and we should interpret model results keeping this in mind.

### ANOVA, Logit, Multinomial, or Webscraping

For this portion of the project, we will use ANOVA to examine the differences of composite industry over composite size. We are curious about the pairs of differences for Composite Industry with regards to the Composite Size variable. Thus, our ANOVA will focus on the composite size variable, although we may choose another one for some

Earlier, we examined the boxplot of composite industry by size and it appeared that the boxplots were relatively even. We start by examing the sample standard deviations across composite industries to see if we pass our assumptions for ANOVA (ie. the max to min ratio of standard deviations is less than 2).

print("SD by Genre")

## [1] "SD by Genre"

(sds <- tapply(aggcompfin$`Composite Size`, aggcompfin$CompInd, sd))

## Commodities Consumer Financials Healthcare Industrials Other   
## 1097.8887 3362.6497 1055.5974 755.9716 922.8877 1250.9710   
## TMT Utilities   
## 1253.0774 1718.6425

print("Ratio of Max/Min Sample SD")

## [1] "Ratio of Max/Min Sample SD"

round(max(sds)/min(sds),1)

## [1] 4.4

The ratio is 4.4. Thus, our assumption for ANOVA is not satisfied. We will continue with the ANOVA analysis, but we will also use non-parametric tests to see if the variances are the same across composite industries. Specifically, we believe it best to use the Kruskal-Wallace test as it makes no assumptions of normality within each group or that the variances be equal (have a ratio less than 2). We examined our normal quantile plots before, and saw that the data was relatively normally distributed, but not for all genres. Thus, Welch’s ANOVA is not a solid choice in this scenario.

We begin with ANOVA and examine residual plots to see if our assumptions are met:

aov1 <- aov(aggcompfin$`Composite Size` ~ aggcompfin$CompInd)  
#We use Tukey comparisons to see differences in the mean size between composite industries  
#TukeyHSD(aov1)  
#par(mar=c(5,11,4,1))  
#plot(TukeyHSD(aov1), las=1)  
  
#We examine our residuals for the ANOVA model  
#myResPlots2(aov1, label = "Size Composite Industry")

*Analysis:* Examining the normal quantile plot we see that our errors are not normally distributed and the fits vs studentized residuals plot shows evidence of heteroskedasticity. Therefore, we will use a transformation to hopefully have a more accurate test.

Because size is measured on a dollar scale, we find it best to use a log transformation. We confirm this with the Box-Cox procedure, which has been muted to save space.

#Now, let's see what transformation is suggested  
#trans1 <- boxCox(aov1)  
#trans1$x[which.max(trans1$y)]  
#The suggested transformation is a log. Let's see what happens to our ANOVA model if we use the transformation. The transformation makes sense given that size is on a dollar scale.  
transsize <- log(aggcompfin$`Composite Size`)  
  
print("SD by Genre")

## [1] "SD by Genre"

(sds <- tapply(transsize, aggcompfin$CompInd, sd))

## Commodities Consumer Financials Healthcare Industrials Other   
## 1.403026 2.056953 1.553958 1.010922 1.481583 1.544122   
## TMT Utilities   
## 1.205141 1.396993

print("Ratio of Max/Min Sample SD")

## [1] "Ratio of Max/Min Sample SD"

round(max(sds)/min(sds),1)

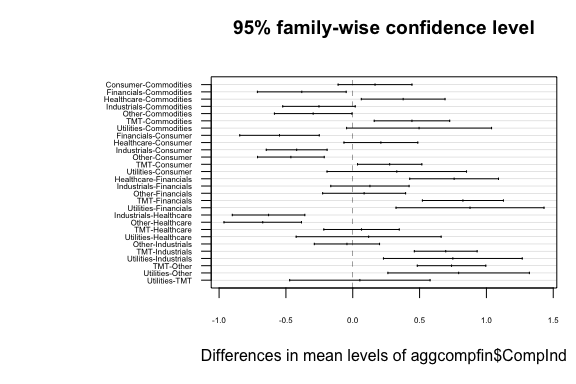
## [1] 2

#Now the ratio is less than 2 and our assumption for ANOVA is satisfied. We will poroceed with the transformed model  
aov2 <- aov(transsize ~ aggcompfin$CompInd)  
summary(aov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## aggcompfin$CompInd 7 403 57.54 24.53 <2e-16 \*\*\*  
## Residuals 4304 10094 2.35   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Analysis:* Our second model, using our transformed model in which we took the log of size, shows statistically significant differences between composite industries given that the p-value is lower than any reasonable alpha.

#pairwise.t.test(transsize, aggcompfin$CompInd)   
#TukeyHSD(aov2)  
par(mar=c(5,11,4,1))  
plot(TukeyHSD(aov2), las=1, cex.axis = 0.5)



*Analysis:* The errors of our new ANOVA model still do not appear normally distributed, but the fits vs residuals plot shows less evidence of heteroskedasticity. This is an improvement from the prior model. The graph of Tukey comparison confidence intervals shows that there are statistically significant differences for mean firm size between composite industries for the majority of pairs.

Before we continue with our non-parametric tests, we will use Bartlett’s and Levene’s test to see if variances are the equal. We are not using Bartlett’s test because it assumes a normal distribution for transsize, which we know is not the case.

leveneTest(transsize, aggcompfin$CompInd)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 7 60.6 < 2.2e-16 \*\*\*  
## 4304   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Analysis:* Once again, the F-value is very high and the p-value is much lower than any reasonable alpha meaning we reject the null hypothesis that the variances are homogenous (Levene’s test makes no assumptions about normality which means its more appropriate for analyzing pctreturn).

Based on the results of Bartlett’s test, we find it best to use Kruskal’s test as our non-parametric test as it makes no assumptions of normality or equality of variances. Now, we proceed with our non-parametric tests.

#Now, with Kruskal's test  
kruskal.test(transsize ~ aggcompfin$CompInd)

##   
## Kruskal-Wallis rank sum test  
##   
## data: transsize by aggcompfin$CompInd  
## Kruskal-Wallis chi-squared = 213.83, df = 7, p-value < 2.2e-16

#Comparing to One-Way ANOVA  
summary.aov(aov(transsize ~ aggcompfin$CompInd))

## Df Sum Sq Mean Sq F value Pr(>F)   
## aggcompfin$CompInd 7 403 57.54 24.53 <2e-16 \*\*\*  
## Residuals 4304 10094 2.35   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Analysis:* The chi-squared value for the Kruskal-Wallis test is high and the p-value is small enough to reject the null hypothesis at any reasonable level of alpha meaning that at least one group median is different from others. As stated earlier, The Kurskal-Wallis test is a good choice here given that the results are statistically significant and our original data is not normally distributed within each genre and the variances are not the same between groups (based on the results from Levene and Bartlette tests).

### Final Comments

At the beginning of this project we set out with two main goals: to explore the variability in returns and characteristics of different industries over time, and to test the CAPM. Regarding the first objective, we observed that returns were actually very similar for most industries in the long run. From our tests, we didn’t see much statistically dignificant difference in the mean returns since 1926 of each industry. However, we did notice variablility in some of the other metrics, such as size or BE/ME. For example, we noticed that TMT companies have a statistically significantly lower BE/ME ratio than commodities, meaning that the market gives them a higher valuation relative to their tangible assets. We also noticed that over time, certain companies, especially technology companies, tend to be larger in market value than firms in other sectors.

We also rejected the CAPM by exploring whether market risk was able to explain the yearly cross section of industry returns. We found that there were unexplained sources of return in the model, meaning that either markets do not efficiently compensate investors for the risk they take or there are other sources of risk we did not include in the model (technically one could also argue that our estimates for beta were wrong, but it is hard to imagine that a small amount of estimation error could cause a t stat of 5). Either way, we showed with a high degree of statistical certainty that CAPM doesn’t work.