ProblemSet2

Amy Richardson

2/23/2021

## Problem Set 1 - Quantitative Analysis

Working in:  
C:-Users-Amy Richardson-Dropbox-VT - PhD-ENGE 5714 - Quant-R Practice-ENGE5714-Problem Set 2

GitHub [link](https://github.com/AmyRichardsonVT/ENGE5714/blob/main/Problem%20Set%202/ProblemSet2.Rmd) to Code

library(car)  
library(psych)  
library(ggplot2)  
library(tidyverse)  
library(broom)

# ——Part 1——

Create a correlation table between %2Y All, %4Y All, %4EngTot4yr, and %CSSTEMH4yr. In text, report which correlations are statistically significant (if any) and comment on the strength of any relationships you observe. Run parametric and non-parametric versions.

Read in Data and Create a DataFrame of salient columns.

all\_degrees <- read.csv("PS\_Degrees.csv", header = TRUE)  
  
names(all\_degrees)[1] <- "div\_name"  
percent\_degrees <- all\_degrees %>%  
 select("Div\_num", "X.2r", "X.4yr", "X.EngTot4yr", "X.CSSTEMH4yr" ) %>%  
 transform(X.2r = X.2r \* 100) %>%  
 transform(X.4yr = X.4yr \* 100) %>%  
 transform(X.EngTot4yr = X.EngTot4yr \* 100) %>%  
 transform(X.CSSTEMH4yr = X.CSSTEMH4yr \* 100)

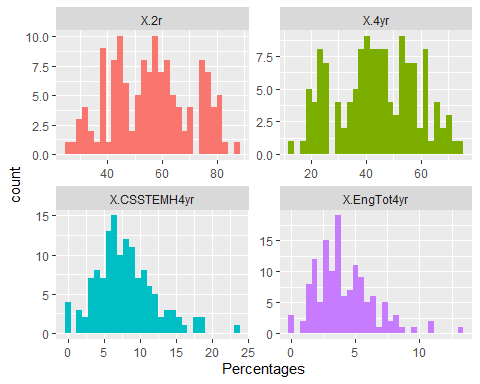
* Define variables:
  + X.2r = Percent of total degrees awarded from 2-yr public VA institutions, 2012-13 through 2016-17
  + X.4yr = Percent of total degrees awarded from 4-yr institutions, 2012-13 through 2016-17
  + X.EngTot4yr = Percent of total degrees from 4-yr institutions during five-year time period that were Engineering degrees
  + X.CSSTEMH4yr = Percent of all STEMH degrees awarded from 4-yr institutions that were CS/IS during five-year time period

Remove rows with NA since the entire row contains NA values

percent\_degrees <- percent\_degrees[-c(135:144), ]

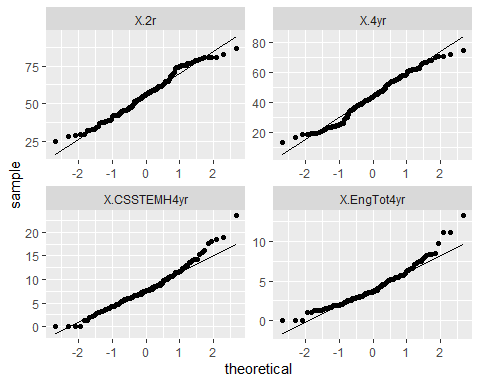
Check to see if the data is normally distributed. First Plot this data.

degree\_df\_long <- percent\_degrees %>%   
 pivot\_longer(cols = X.2r:X.CSSTEMH4yr, names\_to = "Variable", values\_to = "Percentages")  
  
degree\_df\_long %>%   
 ggplot(aes(x = Percentages, fill = Variable)) +  
 geom\_histogram() +  
 facet\_wrap(Variable ~., scales = "free") +  
 theme(legend.position = "none")



Run Q-QPlots

degree\_df\_long %>%   
 ggplot(aes(sample=Percentages)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 facet\_wrap(Variable ~ ., scales = "free")



Run correlation

my\_correlations <- percent\_degrees %>% select(X.2r, X.4yr, X.EngTot4yr, X.CSSTEMH4yr) %>% cor()  
print(my\_correlations)

## X.2r X.4yr X.EngTot4yr X.CSSTEMH4yr  
## X.2r 1.0000000 -1.0000000 -0.4780848 -0.2299198  
## X.4yr -1.0000000 1.0000000 0.4780848 0.2299198  
## X.EngTot4yr -0.4780848 0.4780848 1.0000000 0.2522311  
## X.CSSTEMH4yr -0.2299198 0.2299198 0.2522311 1.0000000

Get p-values

cor.test(percent\_degrees$X.2r, percent\_degrees$X.EngTot4yr)

##   
## Pearson's product-moment correlation  
##   
## data: percent\_degrees$X.2r and percent\_degrees$X.EngTot4yr  
## t = -6.2538, df = 132, p-value = 5.161e-09  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.5990996 -0.3357154  
## sample estimates:  
## cor   
## -0.4780848

cor.test(percent\_degrees$X.2r, percent\_degrees$X.CSSTEMH4yr)

##   
## Pearson's product-moment correlation  
##   
## data: percent\_degrees$X.2r and percent\_degrees$X.CSSTEMH4yr  
## t = -2.7143, df = 132, p-value = 0.007529  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.38451524 -0.06277922  
## sample estimates:  
## cor   
## -0.2299198

cor.test(percent\_degrees$X.4yr, percent\_degrees$X.EngTot4yr)

##   
## Pearson's product-moment correlation  
##   
## data: percent\_degrees$X.4yr and percent\_degrees$X.EngTot4yr  
## t = 6.2538, df = 132, p-value = 5.161e-09  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3357154 0.5990996  
## sample estimates:  
## cor   
## 0.4780848

cor.test(percent\_degrees$X.4yr, percent\_degrees$X.CSSTEMH4yr)

##   
## Pearson's product-moment correlation  
##   
## data: percent\_degrees$X.4yr and percent\_degrees$X.CSSTEMH4yr  
## t = 2.7143, df = 132, p-value = 0.007529  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.06277922 0.38451524  
## sample estimates:  
## cor   
## 0.2299198

cor.test(percent\_degrees$X.CSSTEM, percent\_degrees$X.EngTot4yr)

##   
## Pearson's product-moment correlation  
##   
## data: percent\_degrees$X.CSSTEM and percent\_degrees$X.EngTot4yr  
## t = 2.9947, df = 132, p-value = 0.003281  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.08633575 0.40451613  
## sample estimates:  
## cor   
## 0.2522311

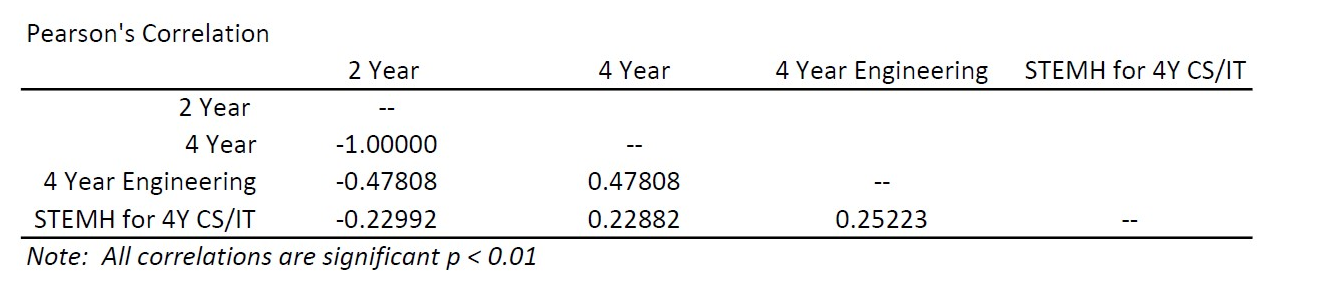
Run Spearman’s Corelation

my\_spearman\_correlations <- percent\_degrees %>% select(X.2r, X.4yr, X.EngTot4yr, X.CSSTEMH4yr) %>% cor(method="spearman")  
print(my\_spearman\_correlations)

## X.2r X.4yr X.EngTot4yr X.CSSTEMH4yr  
## X.2r 1.000000 -1.000000 -0.477266 -0.198430  
## X.4yr -1.000000 1.000000 0.477266 0.198430  
## X.EngTot4yr -0.477266 0.477266 1.000000 0.239102  
## X.CSSTEMH4yr -0.198430 0.198430 0.239102 1.000000

### Report on Results

All of the variables are continuous and from inspection are normally distributed so we will use Pearson’s r when addressing correlation. Below is a compiled table of the results:



Correlation Results

Interpretation:

* The percentage of degrees earned at a 2 year institutions are perfectly, inversely correlated with the degrees earned at a 4 year institutions. This makes sense because there were only two types of institutions in the data set, 2-year and 4-year. The sum of the two percentages from each county is 100%.
* Percentage of total Engineering degrees from 4-year institutions was significantly correlated with percentage to degrees from 2-year institutions (r = -0.4773) and degrees from 4-year institutions (r = 0.4773). The values indicate that there is a medium effect size. These values are the same but opposite sign. This follows the previous rationale given the relationship between percentage of 2 year and 4 year degrees.
* Percentage of all STEMH degree from a 4-year that were CS/IT were significantly related to percentage of 2-year degrees (r = -0.1984), percentage of 4-year degrees (r = 0.1294), and percentages of Engineering degrees from 4-year institutions. The values indicate that there is a small effect size.

# ——Part 2——

Run a simple regression using principal salary in a county to predict teacher salary in that county. Interpret the results of the model and discuss how good the fit is. Discuss in practical terms how to interpret the regression coefficient. Evaluate if there are any outliers and adjust the model if needed.

Read in Data and Create a DataFrame of salient columns.

Note that in the original data set the following sections were removed:

* Governor’s Schools
* Special Education Regional Programs
* Career and Technical (Vocational) Education Regional Programs
* Regional Alternative Education Programs

prinData <- read.csv("Prin\_Salary.csv", header = TRUE)  
teachData <- read.csv("Teach\_Salary.csv", header = TRUE)

Clean both DFs so that we have only the columns we need. We will only look at 2005 and 2016

prinData\_sub <- prinData %>%  
 select(div\_name, FY2005P, FY2016P)   
   
teachData\_sub <- teachData %>%  
 select(div\_name, FY2005T, FY2016T)

Need to change columns containing salary to integers

prinData <- prinData%>%  
 mutate\_at(vars(starts\_with("FY20")), as.numeric)

Join Principle & Teacher data with full join.

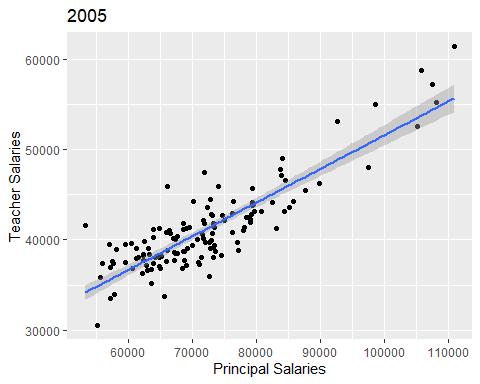
salary\_data <- inner\_join(prinData\_sub,teachData\_sub)

Some columns have salaries < 0, replace them with NA

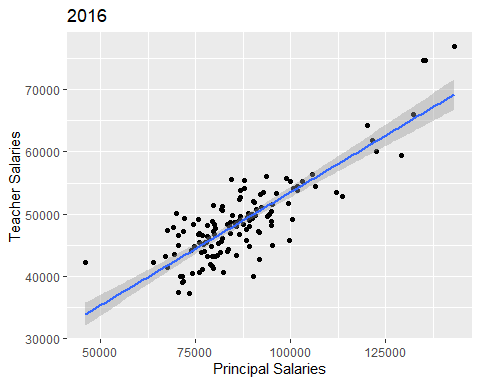
salary\_data [salary\_data <= 0] <- NA

Create Plots for 2005 and 2016

salary\_data %>%   
 ggplot(aes(x = FY2005P, y = FY2005T)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "2005", x = "Principal Salaries", y = "Teacher Salaries")



salary\_data %>%   
 ggplot(aes(x = FY2016P, y = FY2016T)) +  
 geom\_point() +  
 geom\_smooth(method = "lm")+  
 labs(title = "2016", x = "Principal Salaries", y = "Teacher Salaries")



Create Model for 2005

fit\_salary\_2005 <- lm(FY2005T ~ FY2005P, data = salary\_data)  
summary(fit\_salary\_2005)

##   
## Call:  
## lm(formula = FY2005T ~ FY2005P, data = salary\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5518.3 -1595.5 -243.9 1463.4 7381.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.426e+04 1.390e+03 10.26 <2e-16 \*\*\*  
## FY2005P 3.734e-01 1.908e-02 19.57 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2436 on 128 degrees of freedom  
## Multiple R-squared: 0.7495, Adjusted R-squared: 0.7476   
## F-statistic: 383 on 1 and 128 DF, p-value: < 2.2e-16

Create Model for 2016

fit\_salary\_2016 <- lm(FY2016T ~ FY2016P, data = salary\_data)  
summary(fit\_salary\_2016)

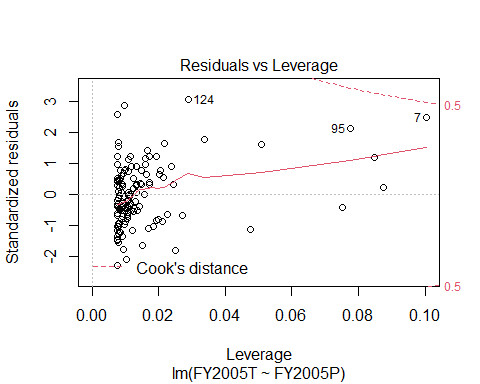
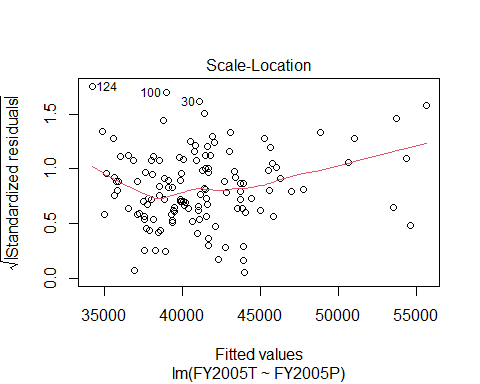
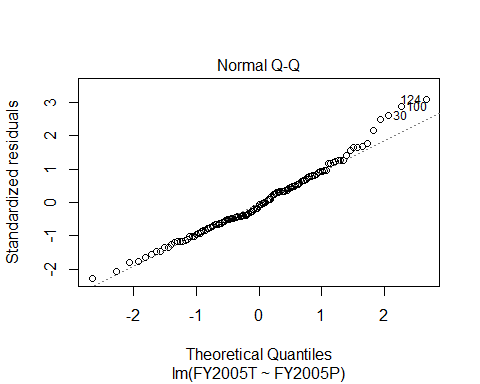
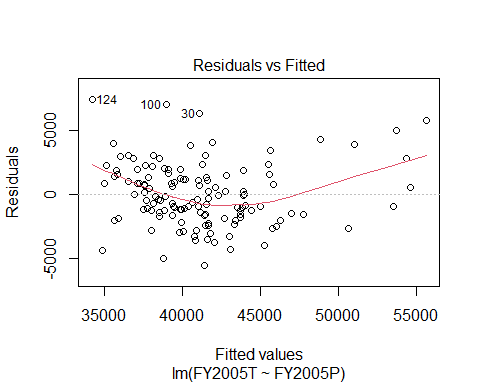
##   
## Call:  
## lm(formula = FY2016T ~ FY2016P, data = salary\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9932.8 -2885.3 220.9 2109.7 8378.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.717e+04 1.859e+03 9.237 7e-16 \*\*\*  
## FY2016P 3.636e-01 2.108e-02 17.247 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3636 on 128 degrees of freedom  
## Multiple R-squared: 0.6992, Adjusted R-squared: 0.6968   
## F-statistic: 297.5 on 1 and 128 DF, p-value: < 2.2e-16

Look for Outliers and/or Influential Cases 2005

salary\_data$residual05 <- resid(fit\_salary\_2005)  
salary\_data$standardized.residuals05 <- rstandard(fit\_salary\_2005)  
salary\_data$studentized.residuals05 <- rstudent(fit\_salary\_2005)  
salary\_data$cooks.distance05 <-cooks.distance(fit\_salary\_2005)  
salary\_data$dfbeta05 <- dfbeta(fit\_salary\_2005)  
salary\_data$dffit05 <- dffits(fit\_salary\_2005)  
salary\_data$leverage05 <- hatvalues(fit\_salary\_2005)  
salary\_data$covariance.ratios05 <- covratio(fit\_salary\_2005)  
  
fit\_salary\_df <- salary\_data %>%   
 mutate(large.residual05 = case\_when(standardized.residuals05 > 2 | standardized.residuals05 < -2 ~ TRUE,  
 abs(standardized.residuals05) <= 2 ~ FALSE))  
  
fit\_salary\_df %>% filter(large.residual05 == TRUE) %>% head()

## div\_name FY2005P FY2016P FY2005T FY2016T residual05  
## 1 Arlington County Public Schools 110868 143092 61407 76942 5750.309  
## 2 Fauquier County Public Schools 71900 99861 47429 55298 6321.144  
## 3 Russell County Public Schools 65615 71393 33725 39102 -5036.329  
## 4 Alexandria City Public Schools 105745 135566 58759 74664 5014.998  
## 5 Covington City Public Schools 66102 69973 45913 50119 6969.848  
## 6 Portsmouth City Public Schools 72710 82151 35892 51310 -5518.272  
## standardized.residuals05 studentized.residuals05 cooks.distance05  
## 1 2.489123 2.541657 0.34554358  
## 2 2.605343 2.666819 0.02631164  
## 3 -2.078413 -2.106122 0.02225084  
## 4 2.143814 2.174825 0.19313864  
## 5 2.875812 2.961838 0.04105001  
## 6 -2.274462 -2.312778 0.02013132  
## dfbeta05.(Intercept) dfbeta05.FY2005P dffit05 leverage05  
## 1 -1.048097e+03 1.523884e-02 0.8488619 0.100349239  
## 2 5.193859e+01 -4.079527e-05 0.2348106 0.007692976  
## 3 -1.827471e+02 1.994419e-03 -0.2137667 0.010196749  
## 4 -7.684376e+02 1.125288e-02 0.6305024 0.077531223  
## 5 2.376701e+02 -2.548784e-03 0.2951022 0.009829521  
## 6 -2.544520e+01 -2.407272e-04 -0.2040358 0.007722851  
## covariance.ratios05 large.residual05  
## 1 1.0224536 TRUE  
## 2 0.9179924 TRUE  
## 3 0.9581732 TRUE  
## 4 1.0235281 TRUE  
## 5 0.8976074 TRUE  
## 6 0.9426403 TRUE

plot(fit\_salary\_2005)

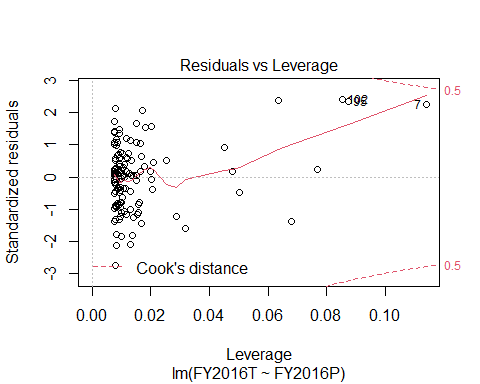
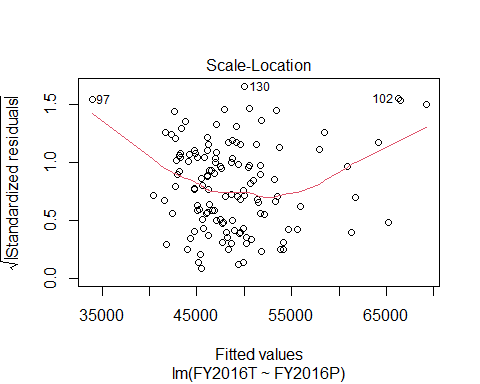
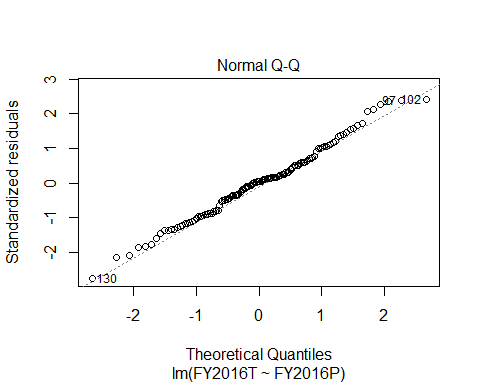
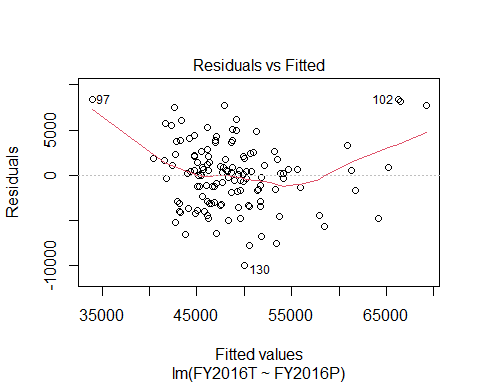


Look for Outliers and/or Influential Cases 2016

salary\_data$residual16 <- resid(fit\_salary\_2016)  
salary\_data$standardized.residuals16 <- rstandard(fit\_salary\_2016)  
salary\_data$studentized.residuals16 <- rstudent(fit\_salary\_2016)  
salary\_data$cooks.distance16 <-cooks.distance(fit\_salary\_2016)  
salary\_data$dfbeta16 <- dfbeta(fit\_salary\_2016)  
salary\_data$dffit16 <- dffits(fit\_salary\_2016)  
salary\_data$leverage16 <- hatvalues(fit\_salary\_2016)  
salary\_data$covariance.ratios16 <- covratio(fit\_salary\_2016)  
  
fit\_salary\_df <- salary\_data %>%   
 mutate(large.residual16 = case\_when(standardized.residuals16 > 2 | standardized.residuals16 < -2 ~ TRUE,  
 abs(standardized.residuals16) <= 2 ~ FALSE))  
  
fit\_salary\_df %>% filter(large.residual16 == TRUE) %>% head()

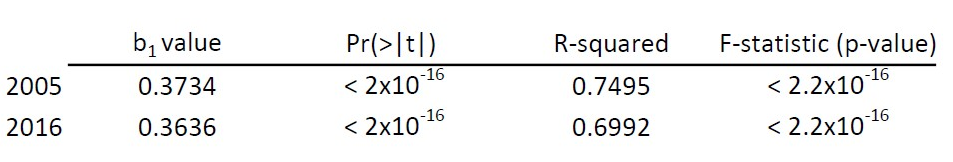
## div\_name FY2005P FY2016P FY2005T FY2016T  
## 1 Arlington County Public Schools 110868 143092 61407 76942  
## 2 Campbell County Public Schools 77138 99601 38773 45822  
## 3 Isle Of Wight County Public Schools 79694 84448 43147 55609  
## 4 Madison County Public Schools 68567 91706 38675 42770  
## 5 Alexandria City Public Schools 105745 135566 58759 74664  
## 6 Buena Vista City Public Schools 62469 46084 39833 42310  
## residual05 standardized.residuals05 studentized.residuals05 cooks.distance05  
## 1 5750.309 2.4891229 2.5416571 0.3455435769  
## 2 -4290.481 -1.7698196 -1.7848665 0.0147161600  
## 3 -870.772 -0.3595578 -0.3583315 0.0007400962  
## 4 -1188.469 -0.4900218 -0.4885623 0.0010191482  
## 5 5014.998 2.1438140 2.1748246 0.1931386436  
## 6 2246.240 0.9284302 0.9279261 0.0057962510  
## dfbeta05.(Intercept) dfbeta05.FY2005P dffit05 leverage05  
## 1 -1.048097e+03 1.523884e-02 0.84886186 0.100349239  
## 2 6.489288e+01 -1.363899e-03 -0.17301708 0.009309036  
## 3 2.314106e+01 -4.154746e-04 -0.03834206 0.011319746  
## 4 -2.741829e+01 2.527432e-04 -0.04501303 0.008417161  
## 5 -7.684376e+02 1.125288e-02 0.63050237 0.077531223  
## 6 1.133954e+02 -1.331644e-03 0.10761003 0.013270183  
## covariance.ratios05 residual16 standardized.residuals16  
## 1 1.0224536 7735.669 2.260085  
## 2 0.9757866 -7569.878 -2.095669  
## 3 1.0253660 7727.145 2.133550  
## 4 1.0205929 -7751.052 -2.140781  
## 5 1.0235281 8194.317 2.359016  
## 6 1.0156526 8378.288 2.381224  
## studentized.residuals16 cooks.distance16 dfbeta16.(Intercept)  
## 1 2.297549 0.32835654 -1.366269e+03  
## 2 -2.124227 0.02921917 2.259826e+02  
## 3 2.164028 0.01810395 1.152819e+02  
## 4 -2.171633 0.01958568 5.002611e+01  
## 5 2.402591 0.26640679 -1.207797e+03  
## 6 2.426252 0.19272943 1.135153e+03  
## dfbeta16.FY2016P dffit16 leverage16 covariance.ratios16 large.residual16  
## 1 1.649857e-02 0.8238115 0.113919783 1.0567369 TRUE  
## 2 -3.280173e-03 -0.2450345 0.013131428 0.9599036 TRUE  
## 3 -6.372978e-04 0.1930019 0.007891448 0.9523602 TRUE  
## 4 -1.267921e-03 -0.2007699 0.008474772 0.9524434 TRUE  
## 5 1.469658e-02 0.7434241 0.087378447 1.0183881 TRUE  
## 6 -1.227327e-02 0.6325936 0.063652396 0.9908771 TRUE

plot(fit\_salary\_2016)



### Report on Results

For both the 2005 and 2016 the linear regression model significantly predicts a teacher’s salary given the principal’s salary. The table below shows the pertinent values.

  
Interpretation:

* The positive b1 value indicates a direct relationship between principal and teacher salaries - as one increases/decreases that causes an increase/decrease in the other. This model indicates that with a $100 increase in the principal salary the teachers salary would increase $37.34 in 2005 and $36.36 in 2016.
* In both years presented the R-squared value is relatively large suggesting that the the model overall predicts the teacher’s salary significantly well.
* The Pf(>|t|) values indicated above reject the null hypothesis that there is no relationship between principal and teacher salary.
* The significance value of the F statistic was less than 0.001 which indicates that there is less than a 0.1% change that an F-ratio this large would happen if the null hypothesis were true.
* There no indication of outliers or influential cases in both years since there are no residuals greater less than -2 or greater than 2. This is confirmed by the dataframe and the plots.

# ——Part 3——

Randomly divide the group of counties from (2) in two halves. Run a simple regression on each half. Compare the results from each of these two samples with what you know the population answers to be and discuss any differences you see in these two estimates of the true values and if the confidence intervals from these samples encompass the results of the original population regression from (2). Plot scatterplots of these different datasets with models and use ggplot to add the lines on too.

Let’s take our joined dataframe (salary\_data) and split the 130 observations into two ramdom groups.

set.seed(37645)   
dummy\_sep <- rbinom(nrow(salary\_data), 1, 0.5)   
  
salary\_sample1 <- salary\_data[dummy\_sep == 0, ]  
salary\_sample2 <- salary\_data[dummy\_sep == 1, ]

Create a Linear Model on each set for 2005

fit\_salary1\_2005 <- lm(FY2005T ~ FY2005P, data = salary\_sample1)  
summary(fit\_salary1\_2005)

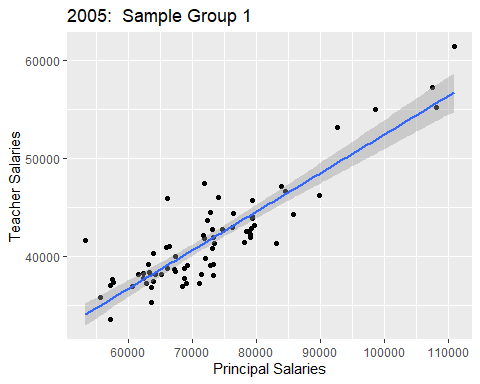
##   
## Call:  
## lm(formula = FY2005T ~ FY2005P, data = salary\_sample1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4446.2 -1588.0 -221.3 1197.8 7521.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.302e+04 1.844e+03 7.06 1.1e-09 \*\*\*  
## FY2005P 3.941e-01 2.503e-02 15.74 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2430 on 68 degrees of freedom  
## Multiple R-squared: 0.7847, Adjusted R-squared: 0.7816   
## F-statistic: 247.9 on 1 and 68 DF, p-value: < 2.2e-16

fit\_salary2\_2005 <- lm(FY2005T ~ FY2005P, data = salary\_sample2)  
summary(fit\_salary2\_2005)

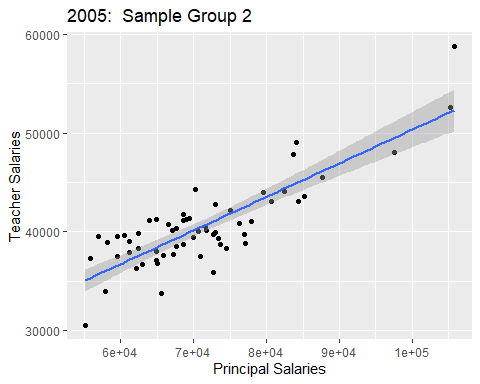
##   
## Call:  
## lm(formula = FY2005T ~ FY2005P, data = salary\_sample2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5160.5 -1494.3 -342.9 1689.0 6467.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.631e+04 2.100e+03 7.767 1.51e-10 \*\*\*  
## FY2005P 3.402e-01 2.920e-02 11.650 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2407 on 58 degrees of freedom  
## Multiple R-squared: 0.7006, Adjusted R-squared: 0.6954   
## F-statistic: 135.7 on 1 and 58 DF, p-value: < 2.2e-16

Create Plots for 2005

salary\_sample1 %>%   
 ggplot(aes(x = FY2005P, y = FY2005T)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "2005: Sample Group 1", x = "Principal Salaries", y = "Teacher Salaries")



salary\_sample2 %>%   
 ggplot(aes(x = FY2005P, y = FY2005T)) +  
 geom\_point() +  
 geom\_smooth(method = "lm")+  
 labs(title = "2005: Sample Group 2", x = "Principal Salaries", y = "Teacher Salaries")



Confidence Intervals of Samples (compare to Population b0 = 14260 and b1= 0.374)

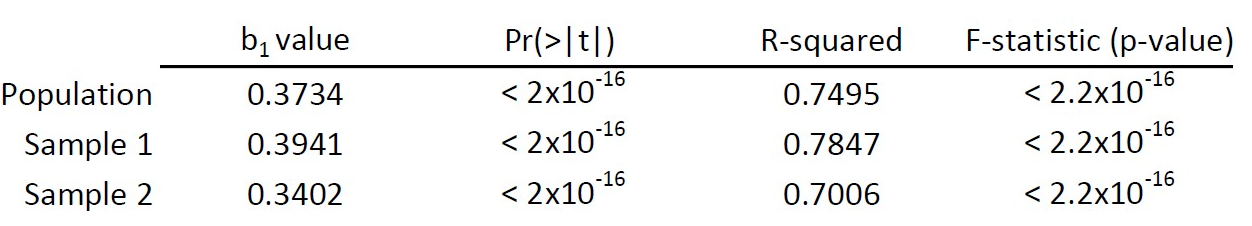
confint(fit\_salary1\_2005, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 9338.923271 1.669778e+04  
## FY2005P 0.344128 4.440153e-01

confint(fit\_salary2\_2005, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 1.210957e+04 2.051874e+04  
## FY2005P 2.817734e-01 3.986917e-01

### Report on Results

  
Interpretation:

* The table above compares the samples with the population.
* While there are differences in the regression coefficients from the two samples and the population, the 95% confidence intervals for each sample include the regression coefficients of the population.

# ——Part 4——

Run a multiple regression. Choose an outcome variable of interest from the one of the degree percentage variables. Choose one predictor variable from the tables provided to pretend it is the theory variable. Choose two additional predictor variables from the tables to be the “additional” possible predictor variables. Interpret the results of the models and discuss how good a fit are the models, including whether or not the second level model is better than the first. Discuss in practical terms how to interpret the regression coefficient. Evaluate if there are any outliers and adjust the model if needed. Compute relevant diagnostics, produce and interpret plots, etc. to demonstrate your knowledge of various means of checking assumptions.

Read in the new data, select the relevant columns, ensure the data are numbers, and join data.

* I will use the following to predict the composite index for each county:
  + X.STEMHall = Percent of total degrees awarded during five-year time period that were STEM-H degree
  + HS2015 = On-time high school graduation rate, [2015]
  + Percent of children ages 0-17 living below 50% of the federal poverty level [2011-15]
  + Comp\_index\_2014\_2016 = Virginia Composite Index [2014-2016] (CI) is a formula determined by the Virginia General Assembly to calculate localities’ ability to pay for K-12 education the quality of the state Standards of Learning (SOL).

VA\_chars <- read.csv("VA\_characteristics.csv", header = TRUE)  
  
VA\_chars\_sub <- VA\_chars %>%  
 select(Locality, HS2015, Comp\_index\_2014\_2016, PerPovertyBelow50)   
  
degree\_sub <- all\_degrees %>%  
 select(div\_name, X.STEMHall)   
  
names(VA\_chars\_sub)[1] <- "div\_name"  
  
VA\_chars\_sub <- VA\_chars\_sub %>%  
 mutate\_at(vars(starts\_with ("HS")), as.numeric)   
  
VA\_chars\_sub <- VA\_chars\_sub %>%  
 mutate\_at(vars(starts\_with ("Comp")), as.numeric)  
  
VA\_chars\_sub <- VA\_chars\_sub %>%  
 mutate\_at(vars(starts\_with ("PerPo")), as.numeric)  
  
  
char\_degree <- inner\_join(VA\_chars\_sub, degree\_sub, by = "div\_name")  
  
char\_degree <- char\_degree[-c(29), ]

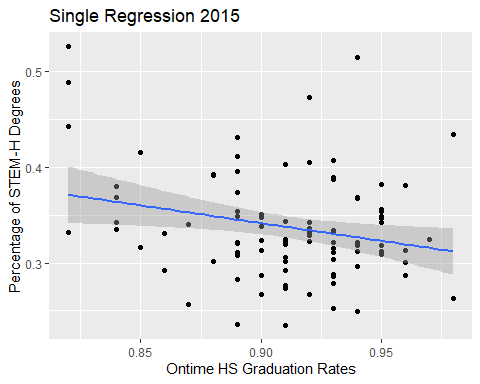
Using the Hierarchical method I am going to start using ontime high school graduation rates to predict the percent of STEM-H degrees in higher education institutions.

fit\_single\_degree <- lm(X.STEMHall ~ HS2015, data = char\_degree)  
summary(fit\_single\_degree)

##   
## Call:  
## lm(formula = X.STEMHall ~ HS2015, data = char\_degree)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.109218 -0.032029 -0.008957 0.027886 0.187994   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6749 0.1400 4.821 5.62e-06 \*\*\*  
## HS2015 -0.3702 0.1537 -2.409 0.018 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05471 on 92 degrees of freedom  
## Multiple R-squared: 0.05933, Adjusted R-squared: 0.0491   
## F-statistic: 5.802 on 1 and 92 DF, p-value: 0.018

Plot linear regression

char\_degree %>%   
 ggplot(aes(x = HS2015, y = X.STEMHall)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "Single Regression 2015", x = "Ontime HS Graduation Rates", y = "Percentage of STEM-H Degrees")



Create a multiple linear regression model to predict the percent of STEM-H degrees and higher education using Ontime HS graduation rates, Virginia Composite Index, and percent of children ages 0-17 living below 50% of the federal poverty level.

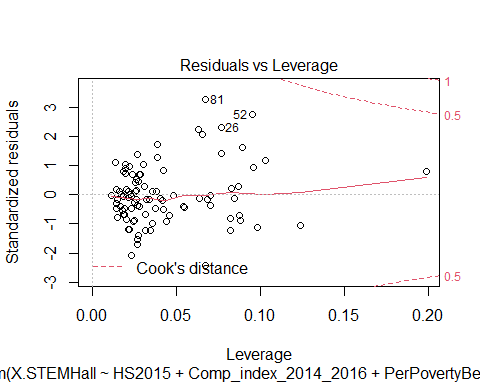
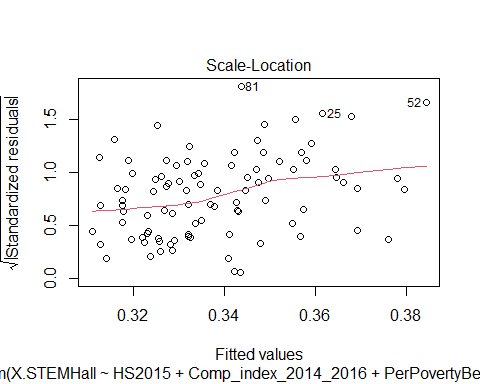
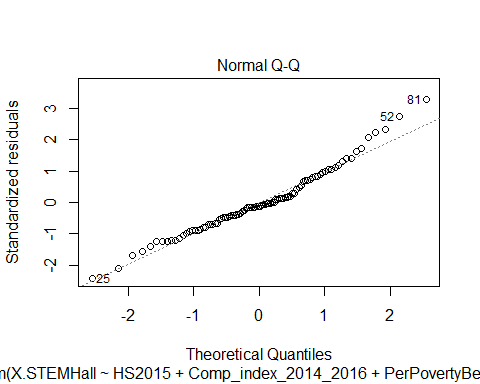
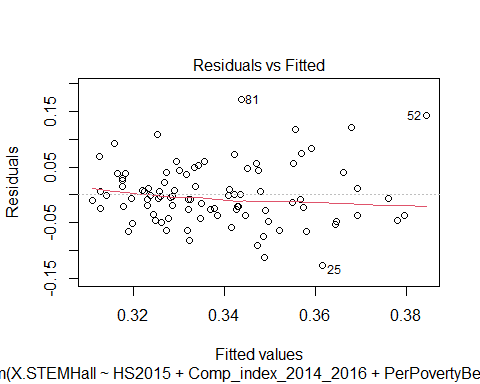
fit\_multiple\_degree <- lm(formula = X.STEMHall ~ HS2015 + Comp\_index\_2014\_2016 + PerPovertyBelow50, data = char\_degree)  
summary(fit\_multiple\_degree)

##   
## Call:  
## lm(formula = X.STEMHall ~ HS2015 + Comp\_index\_2014\_2016 + PerPovertyBelow50,   
## data = char\_degree)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.126676 -0.036390 -0.007052 0.034501 0.171266   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.545824 0.155712 3.505 0.000713 \*\*\*  
## HS2015 -0.254442 0.168026 -1.514 0.133454   
## Comp\_index\_2014\_2016 0.010376 0.038357 0.271 0.787378   
## PerPovertyBelow50 0.002319 0.001241 1.869 0.064920 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05426 on 90 degrees of freedom  
## Multiple R-squared: 0.09475, Adjusted R-squared: 0.06457   
## F-statistic: 3.14 on 3 and 90 DF, p-value: 0.02919

char\_degree$residual <- resid(fit\_multiple\_degree)  
char\_degree$standardized.residuals <- rstandard(fit\_multiple\_degree)  
char\_degree$studentized.residuals <- rstudent(fit\_multiple\_degree)  
char\_degree$cooks.distance <-cooks.distance(fit\_multiple\_degree)  
char\_degree$dfbeta <- dfbeta(fit\_multiple\_degree)  
char\_degree$dffit <- dffits(fit\_multiple\_degree)  
char\_degree$leverage <- hatvalues(fit\_multiple\_degree)  
char\_degree$covariance.ratios <- covratio(fit\_multiple\_degree)  
  
fit\_multiple\_df <- char\_degree %>%   
 mutate(large.residual = case\_when(standardized.residuals > 2 | standardized.residuals < -2 ~ TRUE,  
 abs(standardized.residuals) <= 2 ~ FALSE))  
  
fit\_multiple\_df %>% filter(large.residual == TRUE) %>% head()

## div\_name HS2015 Comp\_index\_2014\_2016 PerPovertyBelow50 X.STEMHall  
## 1 Charles City County 0.98 0.44 10.4 0.4341  
## 2 Cumberland County 0.91 0.28 19.1 0.2348  
## 3 Dickenson County 0.82 0.27 12.0 0.4889  
## 4 Henry County 0.89 0.24 11.6 0.2362  
## 5 Lee County 0.82 0.19 19.5 0.5264  
## 6 Scott County 0.94 0.19 15.1 0.5149  
## residual standardized.residuals studentized.residuals cooks.distance  
## 1 0.1089482 2.077233 2.117033 0.07593258  
## 2 -0.1266758 -2.417356 -2.485948 0.10558595  
## 3 0.1210915 2.322225 2.381746 0.11178266  
## 4 -0.1125588 -2.098911 -2.140256 0.02627950  
## 5 0.1420308 2.752020 2.859653 0.19971396  
## 6 0.1712664 3.268408 3.462157 0.19324507  
## dfbeta.(Intercept) dfbeta.HS2015 dfbeta.Comp\_index\_2014\_2016  
## 1 -7.663652e-02 8.217472e-02 5.964690e-04  
## 2 4.044282e-02 -4.069889e-02 2.570274e-03  
## 3 9.109094e-02 -9.630724e-02 -2.793889e-03  
## 4 -1.011116e-02 7.288238e-03 7.002511e-03  
## 5 7.492170e-02 -8.227160e-02 -5.160625e-03  
## 6 -8.352044e-02 9.523371e-02 -1.632369e-02  
## dfbeta.PerPovertyBelow50 dffit leverage covariance.ratios  
## 1 3.376765e-04 0.5616767 0.06576202 0.9196101  
## 2 -6.953889e-04 -0.6683201 0.06740294 0.8572239  
## 3 -1.103764e-04 0.6858172 0.07656531 0.8844322  
## 4 -6.394304e-05 -0.3306058 0.02330495 0.8759230  
## 5 4.360399e-04 0.9287441 0.09541462 0.8133088  
## 6 6.288087e-04 0.9313109 0.06747695 0.6764827  
## large.residual  
## 1 TRUE  
## 2 TRUE  
## 3 TRUE  
## 4 TRUE  
## 5 TRUE  
## 6 TRUE

plot(fit\_multiple\_degree)



Check for outliers

fit\_multiple\_df$large.residual <- fit\_multiple\_df$standardized.residuals > 2 | fit\_multiple\_df$standardized.residuals < -2   
fit\_multiple\_df[fit\_multiple\_df$large.residual,c("div\_name", "HS2015", "Comp\_index\_2014\_2016", "PerPovertyBelow50", "X.STEMHall", "standardized.residuals")]

## div\_name HS2015 Comp\_index\_2014\_2016 PerPovertyBelow50 X.STEMHall  
## 19 Charles City County 0.98 0.44 10.4 0.4341  
## 25 Cumberland County 0.91 0.28 19.1 0.2348  
## 26 Dickenson County 0.82 0.27 12.0 0.4889  
## 45 Henry County 0.89 0.24 11.6 0.2362  
## 52 Lee County 0.82 0.19 19.5 0.5264  
## 81 Scott County 0.94 0.19 15.1 0.5149  
## 93 Wise County 0.92 0.25 17.8 0.4731  
## standardized.residuals  
## 19 2.077233  
## 25 -2.417356  
## 26 2.322225  
## 45 -2.098911  
## 52 2.752020  
## 81 3.268408  
## 93 2.237307

There are 94 observations and 7 (7.4%) of them have large residuals. We will investigate this further.

Leverage, Cooks Distance, and Covariance Ratio for the 7 cases

fit\_multiple\_df[fit\_multiple\_df$large.residual,c("cooks.distance", "leverage", "covariance.ratios", "standardized.residuals")]

## cooks.distance leverage covariance.ratios standardized.residuals  
## 19 0.07593258 0.06576202 0.9196101 2.077233  
## 25 0.10558595 0.06740294 0.8572239 -2.417356  
## 26 0.11178266 0.07656531 0.8844322 2.322225  
## 45 0.02627950 0.02330495 0.8759230 -2.098911  
## 52 0.19971396 0.09541462 0.8133088 2.752020  
## 81 0.19324507 0.06747695 0.6764827 3.268408  
## 93 0.08465984 0.06336596 0.8880398 2.237307

* Analysis
  + Since none of the cases has a Cook’s distance greater than 1, none of these cases are having an undue influence on the model.
  + The average leverage for this model is 0.043 [(k+1)/n] so we are looking for values three times as large (0.129). Only one case is over twice as large (row 52).
  + The covariance ratio limits for this model are 0.872 and 1.128 [1 +/- (3(k+1))/n]. The only case that falls outside of this range is row 81 however given the Cook’s distance for this case we will leave it in the model.

Assessing the Assumption of Independence

dwt(fit\_multiple\_degree)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.07681541 2.150978 0.482  
## Alternative hypothesis: rho != 0

Assumption of Multicollinearity

vif(fit\_multiple\_degree)

## HS2015 Comp\_index\_2014\_2016 PerPovertyBelow50   
## 1.214825 1.130747 1.256578

mean(vif(fit\_multiple\_degree))

## [1] 1.200716

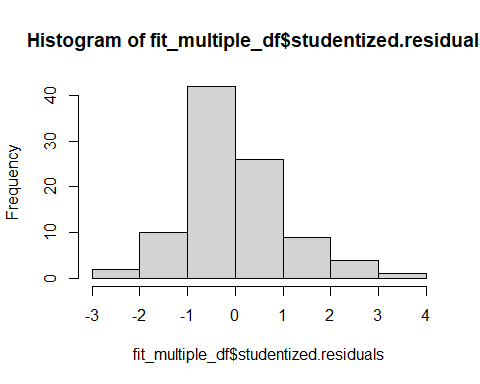
1/vif(fit\_multiple\_degree)

## HS2015 Comp\_index\_2014\_2016 PerPovertyBelow50   
## 0.8231641 0.8843712 0.7958122

* Analysis
  + Since none of the cases have a VIF greater than 10 there is no cause for concern.
  + The average VIF is not substantially greater than 1 so the regression is not biased
  + The tolerance is not below 0.2. We can safely conclude that there is no collinieary within out data.

Assumption about Residuals

hist(fit\_multiple\_df$studentized.residuals)

 By inspection of the histogram and qqplot the sample looks like a normal distribution.

### Report on Results

Although the assumptions have been met and there are no outliers and influential cases, this model is not good for predicting percent of STEM-H degrees in higher education institutions. The single-linear regression model has an r-squared value of 0.05 however the the Pf(>|t|) and the significance value of the F-statistic are both less than 0.05 suggesting that that the null hypothesis is not valid. When additional predictor variables were added to the model the R-squared value increased slightly (0.06) but the low value indicates that the model is not significant.

The practical significance of the regression coefficients are as follows:

* The negative b1 value associated with high school graduation rates indicates an indirect relationship between high school graduation rates and percent of STEM-H degrees - as one increases/decreases that causes the opposite motion (decrease/increase) in the other. This model indicates that with a 4% increase in high school graduation rates the percent of STEM-H degrees would decrease by 1%.
* The positive b1 value associated with Composite Index indicates a direct relationship between Composite Index and percent of STEM-H degrees - as one increases/decreases that causes an increase/decrease in the other. This model indicates that with a 1 point increase in Composite Index the percent of STEM-H degrees would increase by 0.01.
* The positive b1 value associated with Perfect of children living below poverty level indicates a direct relationship between the percent living under the poverty level and percent of STEM-H degrees - as one increases/decreases that causes an increase/decrease in the other. This model indicates that with a 1 point increase in percent poverty the percent of STEM-H degrees would increase by 0.002.

Logically this doesn’t make much sense particularly with the last predictor which is another indication that the model is not a good fit.

xxxxxxx