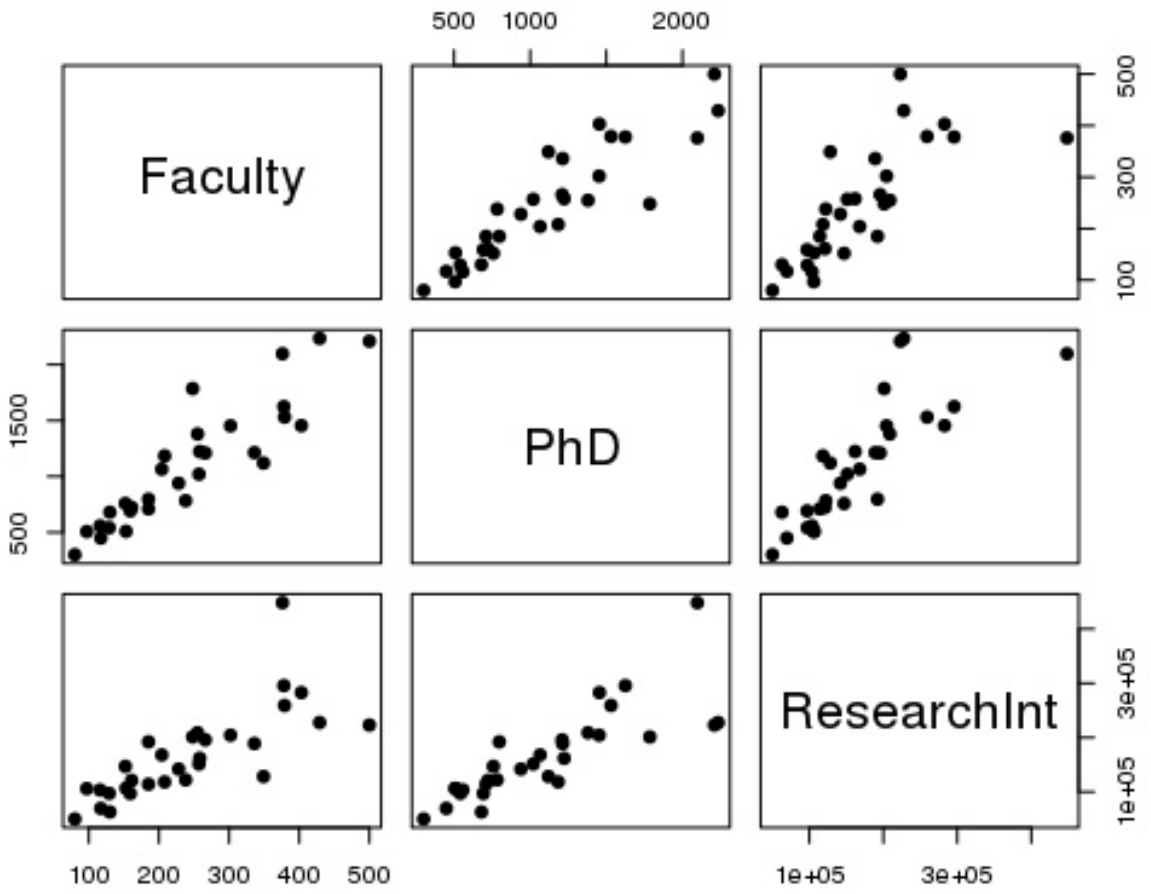
3.13

a.) All 3 have strong positively correlated relationships. Faculty and PhD seem to be more strongly positively correlated than Faculty and Research $ or PhD and Research $.



Calculate Correlation Matrix -> PhD Student Count & Faculty Count are 90% correlation. Faculty has less impact than PhD student count with Research $ (76.48% vs 81.74% respectively).

```
> corMatrix <- cor(researchClean[, c(3,5,6)])
> print(corMatrix)
              Faculty       PhD ResearchInt
Faculty     1.0000000 0.9036829   0.7648421
PhD         0.9036829 1.0000000   0.8174254
ResearchInt 0.7648421 0.8174254   1.0000000
```

b.)

PhD is significant (p value = 0.0125) but Faculty is not (p value = 0.5842).

However, this does not mean that simply increasing the number of PhD students will directly drive increase research funds.
Typically, there might be more research funding where there are more PhD students, but there are also underlying variables influencing this relationship, for example: where there are more PhD students, there might be more professors to teach the class load, and there are more professors working on projects, thus they submit more funding proposals and overall get more grant money. Thus, if all other factors are not in place, simply letting in more PhD students may not have the intended effect.
Because we know there is an issue of multicollinearity (above we note PhD and Faculty are strongly positively correlated) - where predictor variables are not independent and are related to each other, skewing impact on the response variable and it could be artificially inflating the p value.

```
> researchLM <- lm(researchClean$ResearchInt ~ researchClean$PhD + researchClean$Faculty)
> summary(researchLM)

Call:
lm(formula = researchClean$ResearchInt ~ researchClean$PhD +
    researchClean$Faculty)

Residuals:
   Min     1Q Median     3Q    Max
-90804 -16921  -2921  13605 159743

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          23525.91   22034.47   1.068   0.2951
researchClean$PhD      107.14      40.06   2.675   0.0125 *
researchClean$Faculty  107.13     193.39   0.554   0.5842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49040 on 27 degrees of freedom
Multiple R-squared:  0.6719,    Adjusted R-squared:  0.6476
F-statistic: 27.65 on 2 and 27 DF,  p-value: 2.923e-07
```

c.)
Partial Correlation -

- Research & Faculty: 0.106
- Research & PhD: 0.458

T statistic: As you can see the PhD t value > t critical value, thus it is significant just as noted above in the regression

- PhD t value: 2.675
- Faculty t value: 0.554

- T critical value: 2.04

```
> pcor(researchClean[, c(3,5,6)], method = c("pearson"))
$estimate
              Faculty       PhD ResearchInt
Faculty     1.0000000 0.7504387   0.1060117
PhD         0.7504387 1.0000000   0.4576694
ResearchInt 0.1060117 0.4576694   1.0000000

$p.value
                 Faculty          PhD ResearchInt
Faculty     0.000000e+00 2.756307e-06  0.58415554
PhD         2.756307e-06 0.000000e+00  0.01254632
ResearchInt 5.841555e-01 1.254632e-02  0.00000000

$statistic
              Faculty      PhD ResearchInt
Faculty     0.0000000 5.899769   0.5539747
PhD         5.8997693 0.000000   2.6746821
ResearchInt 0.5539747 2.674682   0.0000000

$n
[1] 30

$gp
[1] 1

$method
[1] "pearson"
```

3.14

a.) Correlation Matrix

```
> no <- c(1,2,3,4,5,6,7,8,9,10)
> x1 <- c(31,46,40,49,38,49,31,38,33,42)
> x2 <- c(1.85, 2.80, 2.20, 2.85, 1.80, 2.80, 1.85, 2.30, 1.60, 2.15)
> y <- c(4.20, 7.28, 5.60, 8.12, 5.46, 7.42, 3.36, 5.88, 4.62, 5.88)
>
> salary <- data.frame(x1,x2,y)
>
> #a.) #Correlation Matrix
> #R -> x1 & x2
> x1x2 <- cor(salary[, c(1,2)])
> print(x1x2)
          x1        x2
x1 1.0000000 0.9132577
x2 0.9132577 1.0000000
> #r -> y & x1      y & x2
> yx1yx2 <- c(cor(salary[3], salary[1]), cor(salary[3], salary[2]))
> print(yx1yx2)
[1] 0.9708553 0.9040219
>
> #partial corrleation coeff - r yx2|x1
> # 0.178
> (yx1yx2[2] - (yx1yx2[1]* x1x2[1,2])) / sqrt((1-yx1yx2[1] ^ 2) * (1-x1x2[1,2] ^ 2))
[1] 0.1780173
> (yx1yx2[1] - (yx1yx2[2]* x1x2[1,2])) / sqrt((1-yx1yx2[2] ^ 2) * (1-x1x2[1,2] ^ 2))
[1] 0.8340515
```

b.) Standardized Regression Coeff B's: Beta hat star matrix = R-inverse * r

```
> #b.) B = R-1r
> Rinv = inv(x1x2)
> c(((Rinv[1,1]*yx1yx2[1])+(Rinv[1,2]*yx1yx2[2])),((Rinv[2,1]*yx1yx2[1])+(Rinv[2,2]*yx1yx2[2])))
[1] 0.8752107 0.1047290
```

c.) Unstandardized LS estimates -> Beta1_star = beta1*(sd(x1)/sd(y))
Beta Coefficients:

- B0 = -2.606
- B1 = 0.1922
- B2 = 0.341

Beta Star Values Match Above part b.)

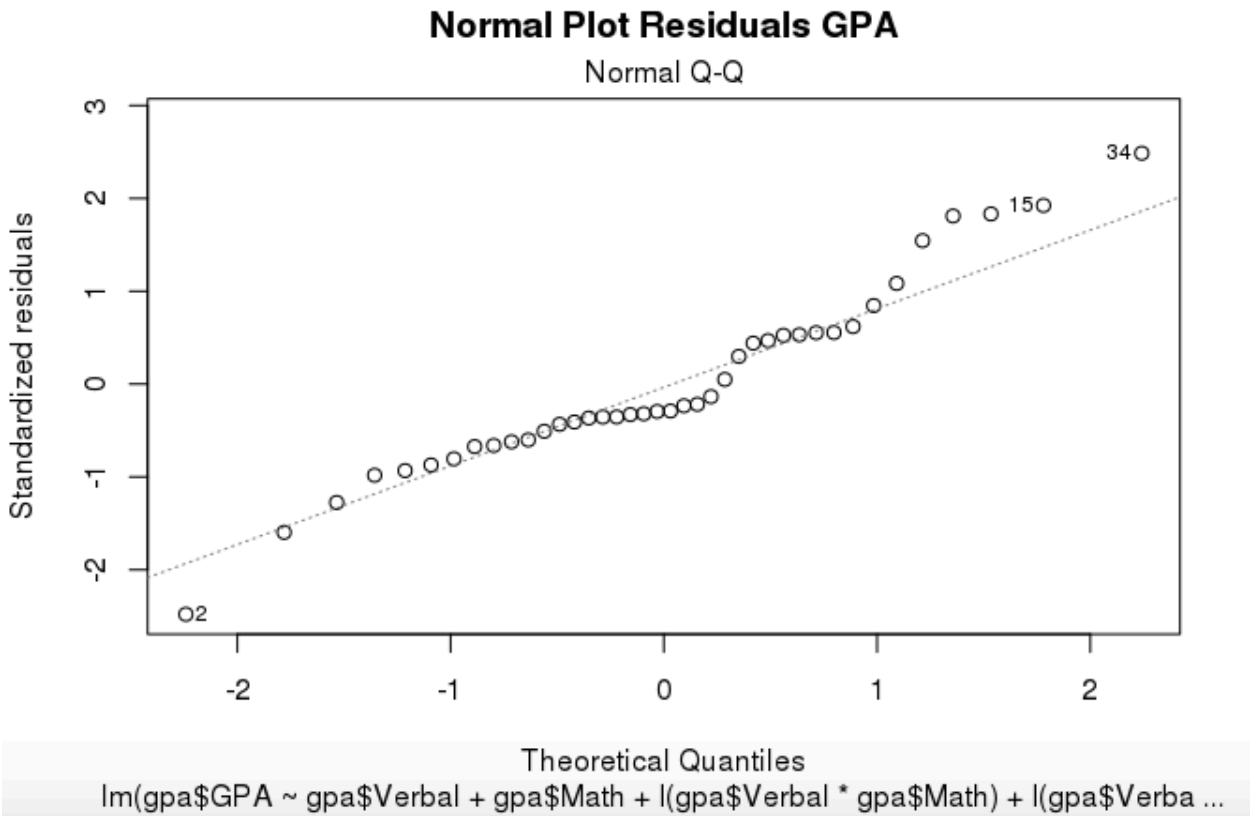
- B1* = 0.875
- B2* = 0.104

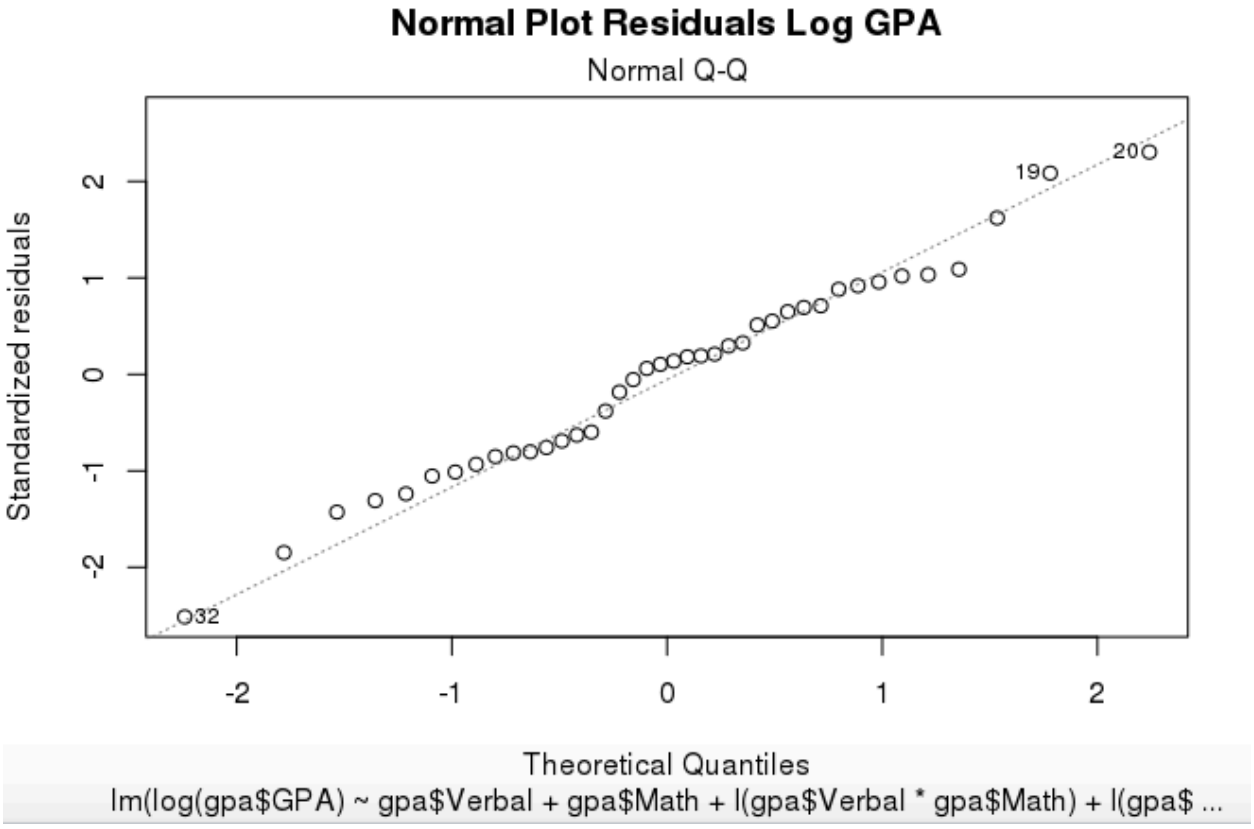
```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    -2.6062       0.1922       0.3406

> b0 <- as.numeric(b$coefficients[1])
> paste("coeff 0",b0)
[1] "coeff 0 -2.60615124622428"
> b1 <- as.numeric(b$coefficients[2])
> paste("coeff 1",b1)
[1] "coeff 1 0.192240808183646"
> b2 <- as.numeric(b$coefficients[3])
> paste("coeff 2",b2)
[1] "coeff 2 0.340626649249346"
>
> # B1 * Sx1/Sy
> b1* std(x1)/std(y)
[1] 0.8752107
>
> # B2 * Sx2/Sy
> b2* std(x2)/std(y)
[1] 0.104729
```

d.)
Now that Betas are standardized and without units, we can compare B1* and B2*. B1* is 0.875, thus we see it has a higher impact on the response variable y

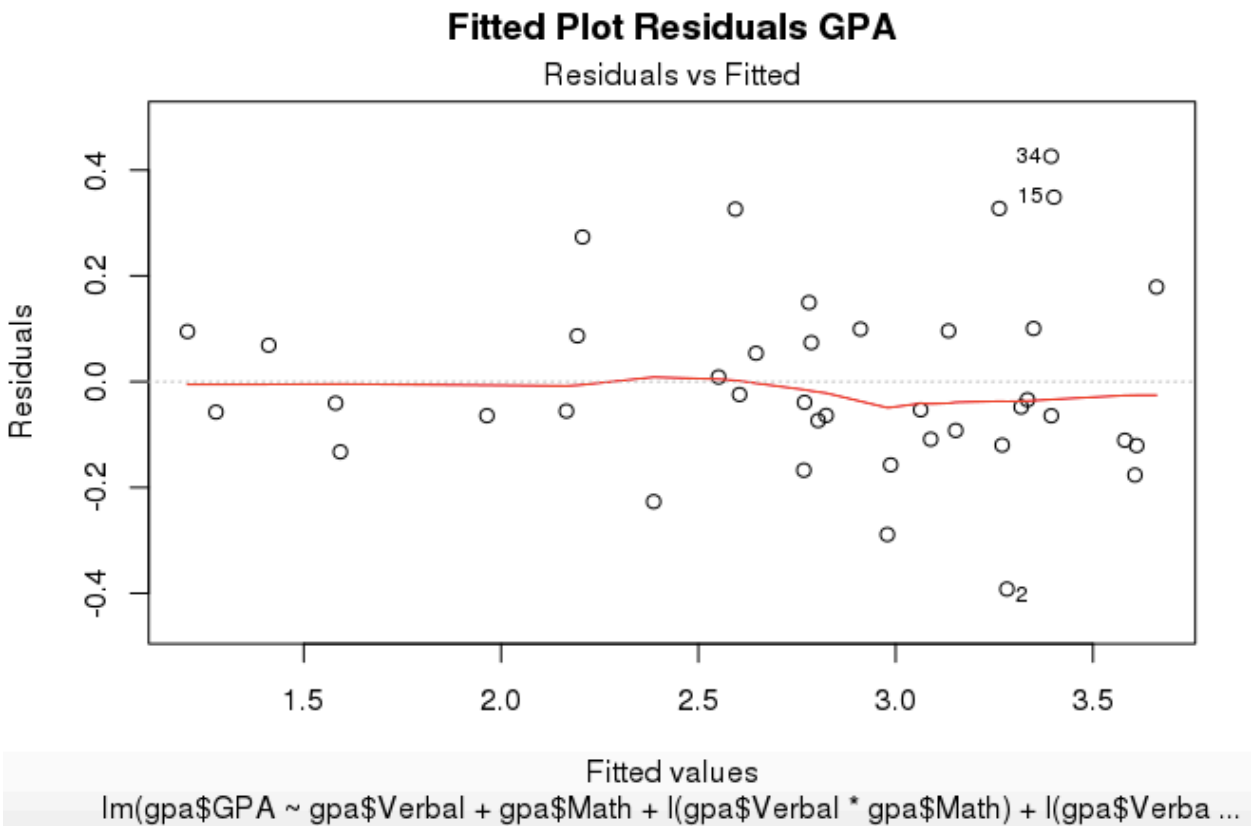
4.4

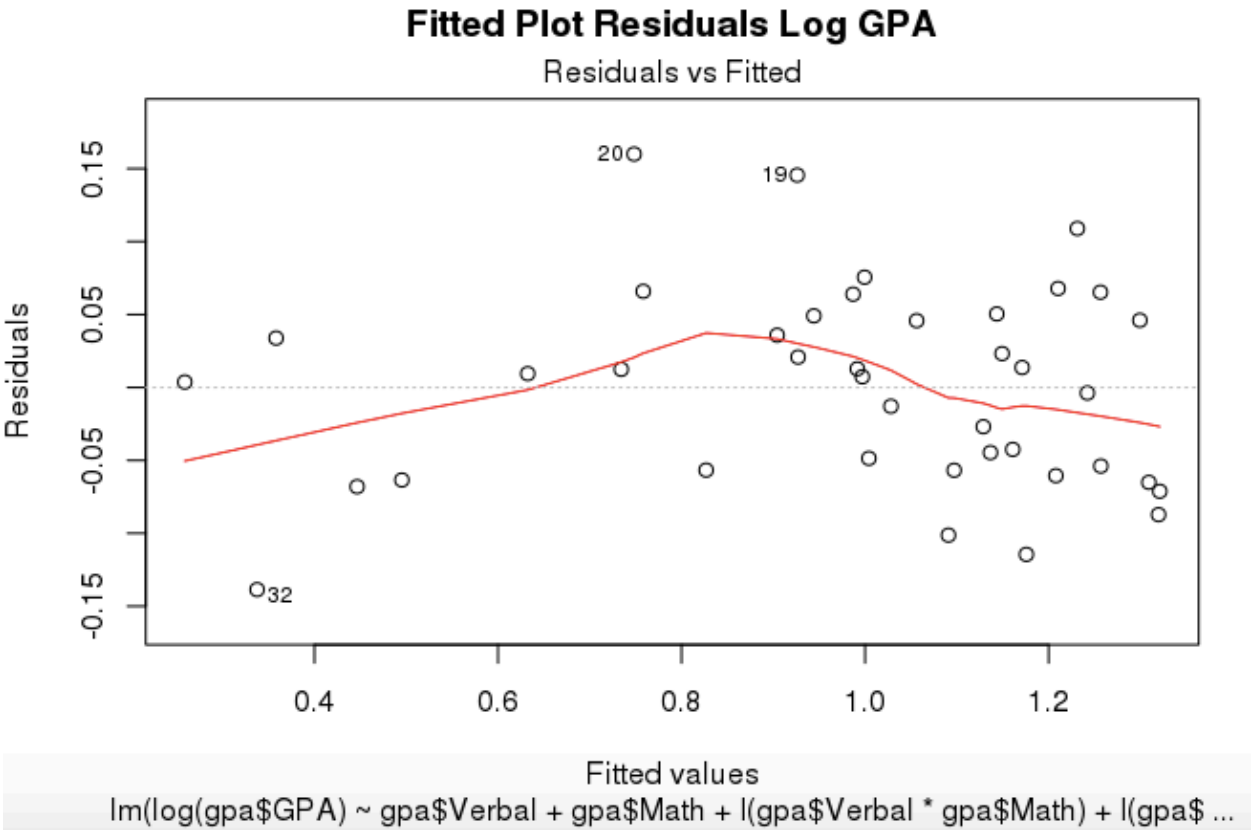
a. + b.) The Log transformation has improved the homoscedasticity


Normal Plot Residuals GPA
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(gpa$GPA ~ gpa$Verbal + gpa$Math + I(gpa$Verbal * gpa$Math) + I(gpa$Verba ...
34
15
2
-2
-1
0
1
2
3



Fitted Plot Residuals -> The log transformation does seem to help reduce the funneling from the non - log GPA model


Fitted Plot Residuals GPA
Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
34
15
2
1.5
2.0
2.5
3.0
3.5
Fitted values
lm(gpa$GPA ~ gpa$Verbal + gpa$Math + I(gpa$Verbal * gpa$Math) + I(gpa$Verba ...



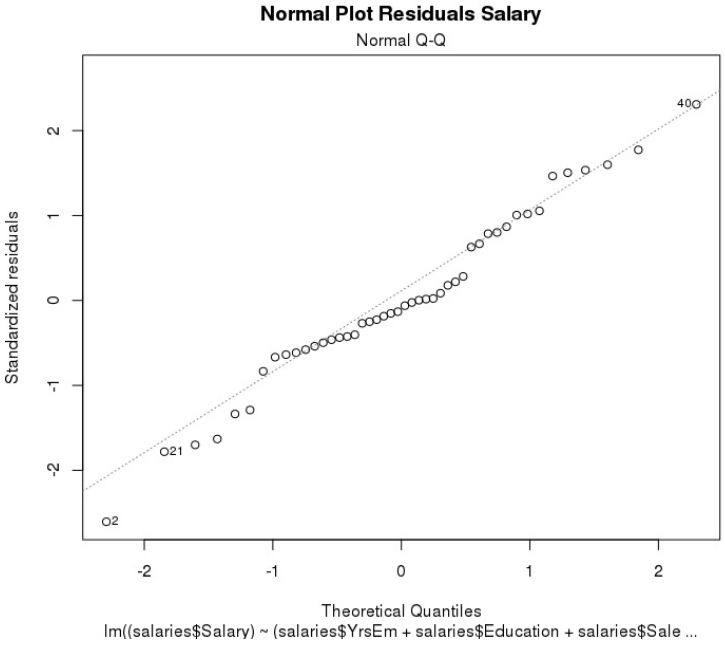
4.5

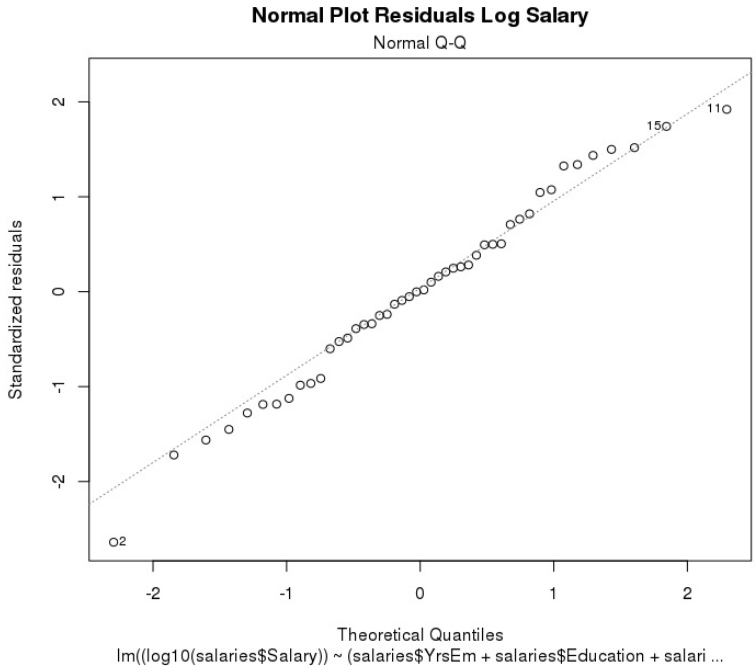
- Since Cook's distance is > than the threshold value (4/n-(p+1)), These universities are outliers and influencing variables most likely because they are well known and are more likely to receive additional grand funding for a variety of underlying other variables

```
> researchLM <- lm(researchClean$ResearchInt ~ researchClean$Faculty + researchClean$PhD)
>
> stdres <- rstandard(researchLM)
> cat("Standard residuals : ", stdres)
Standard residuals :  3.651581 -1.006193 0.215068 0.380574 0.2928507 1.228307 -2.099762 0.666384 -0.1509935 -1.860682 1.357926 1.322016 0.1785693
 0.545465 -1.154302 -0.3713471 -1.002534 -0.1375913 0.1658425 -0.1116019 0.06087808 0.2573966 -0.3071559 -0.4223388 -0.190253 -0.01123799 -1.1937
7 -0.2289631 -0.3100899 0.04018446>
> cat("Outliers : ",researchClean[abs(stdres)>2,]$University)
Outliers : MIT GaTech>
> cook <- cooks.distance(researchLM)
> print(cook)
           1            2            3            4            5            6            7            8            9           10
1.140101e+00 1.688762e-01 1.166544e-03 5.096460e-03 1.038926e-03 4.841874e-02 4.200536e-01 1.692423e-02 4.028740e-04 2.976933e-01
          11           12           13           14           15           16           17           18           19           20
1.370335e-01 2.684938e-02 5.707627e-04 6.247735e-03 3.663364e-02 2.628166e-03 2.665642e-02 2.577261e-04 7.892818e-04 2.377181e-04
          21           22           23           24           25           26           27           28           29           30
6.835526e-05 1.941144e-03 4.016666e-03 2.267992e-03 5.820058e-04 5.357640e-06 1.132149e-01 1.707586e-03 2.951434e-03 4.186510e-05
>
> cook_distance_threshold <- 4/(30-(2+1))
> cat("Influencing Variables :",researchClean[as.integer(which(cook > cook_distance_threshold)),"University"])
Influencing Variables : MIT Stanford GaTech UIUC
>
```

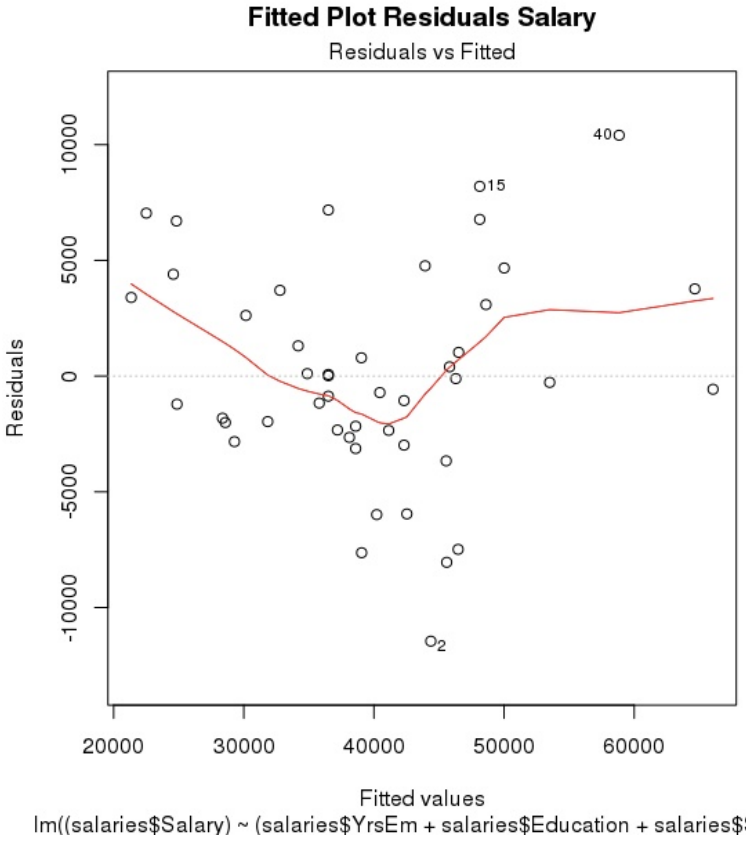
4.6

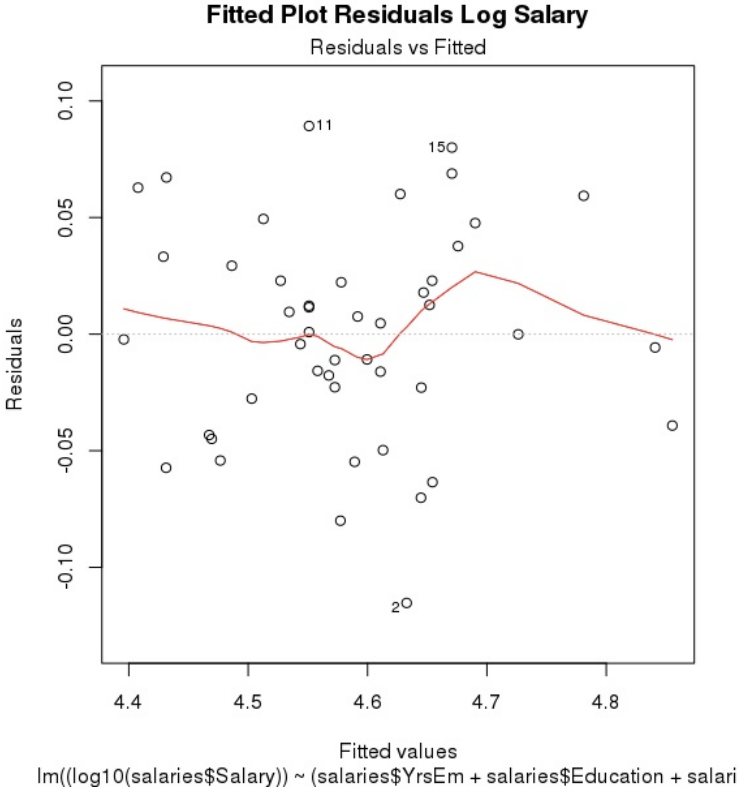
a.) Log transformation has improved normality as it is more linear close to the line





b.) Log also improves the homoscedasticity as the first plot looks like it funnels out wider across the y, whereas after the log transform it looks more randomly dispersed across a parallel band around x axis


Fitted Plot Residuals Salary
Residuals vs Fitted
Residuals
Fitted values
lm((salaries$Salary) ~ (salaries$YrsEm + salaries$Education + salaries$Sale


Fitted Plot Residuals Log Salary
Residuals vs Fitted
Residuals
Fitted values
lm((log10(salaries$Salary)) ~ (salaries$YrsEm + salaries$Education + salari