HW 1 STATS 402

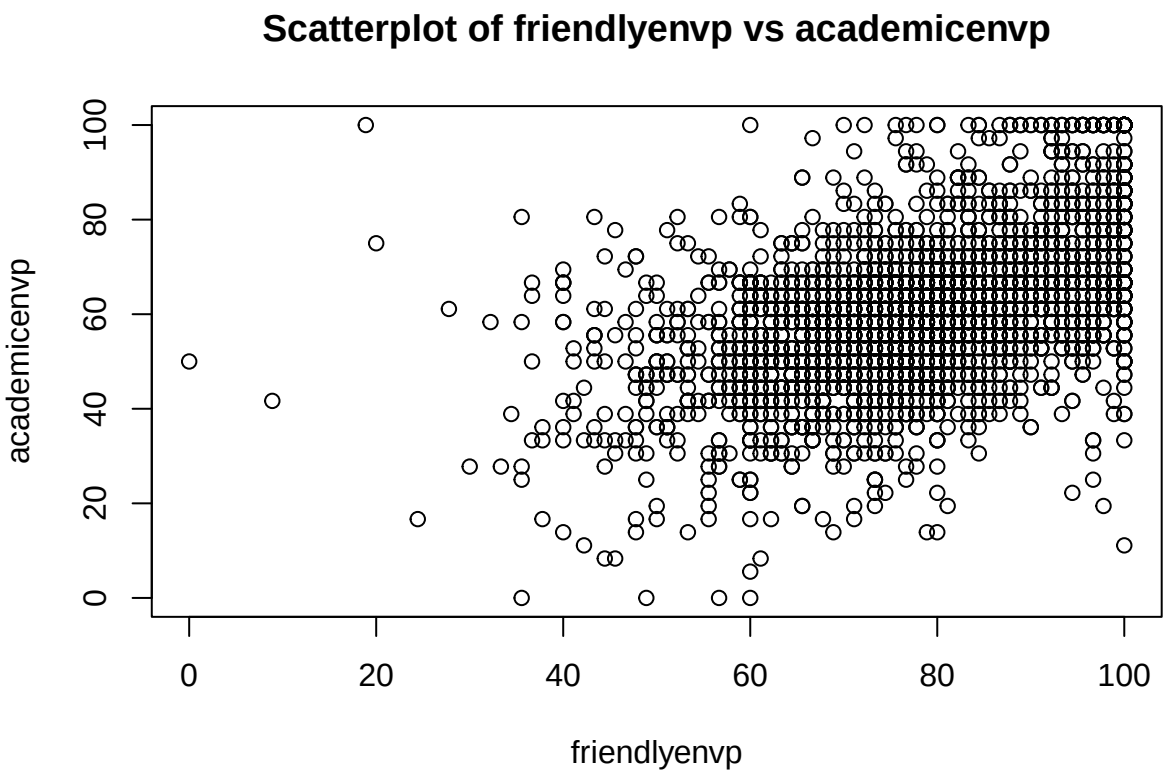
Clayton Chan

2024-10-13

1a.

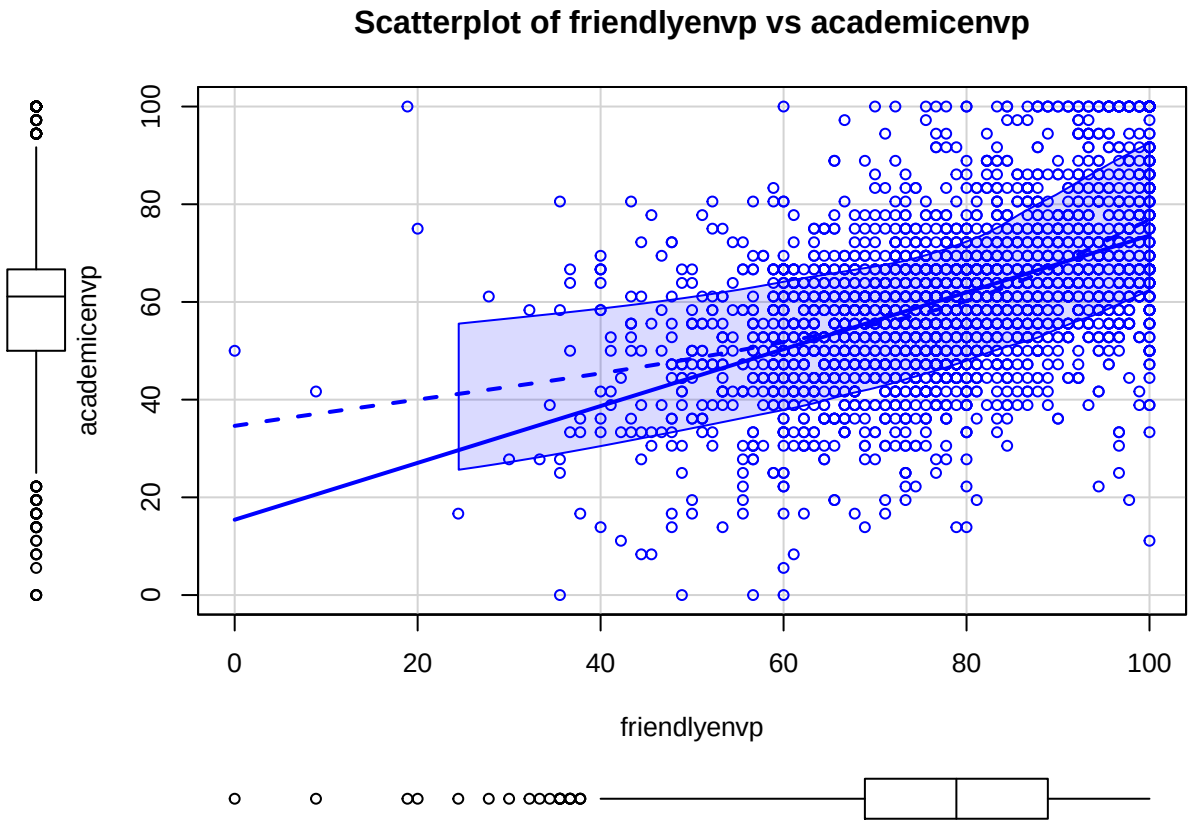
Before we start our analysis, we see that some of the rows have NA values so in order to clean our data, we will drop all the NA rows.

```
##   friendlyenvp academicenvp
## 1    100.00000    100.00000
## 2     88.88889     75.00000
## 3     75.55556           NA
## 4     51.11111           NA
## 5     95.55556     69.44444
## 6    100.00000    100.00000
```



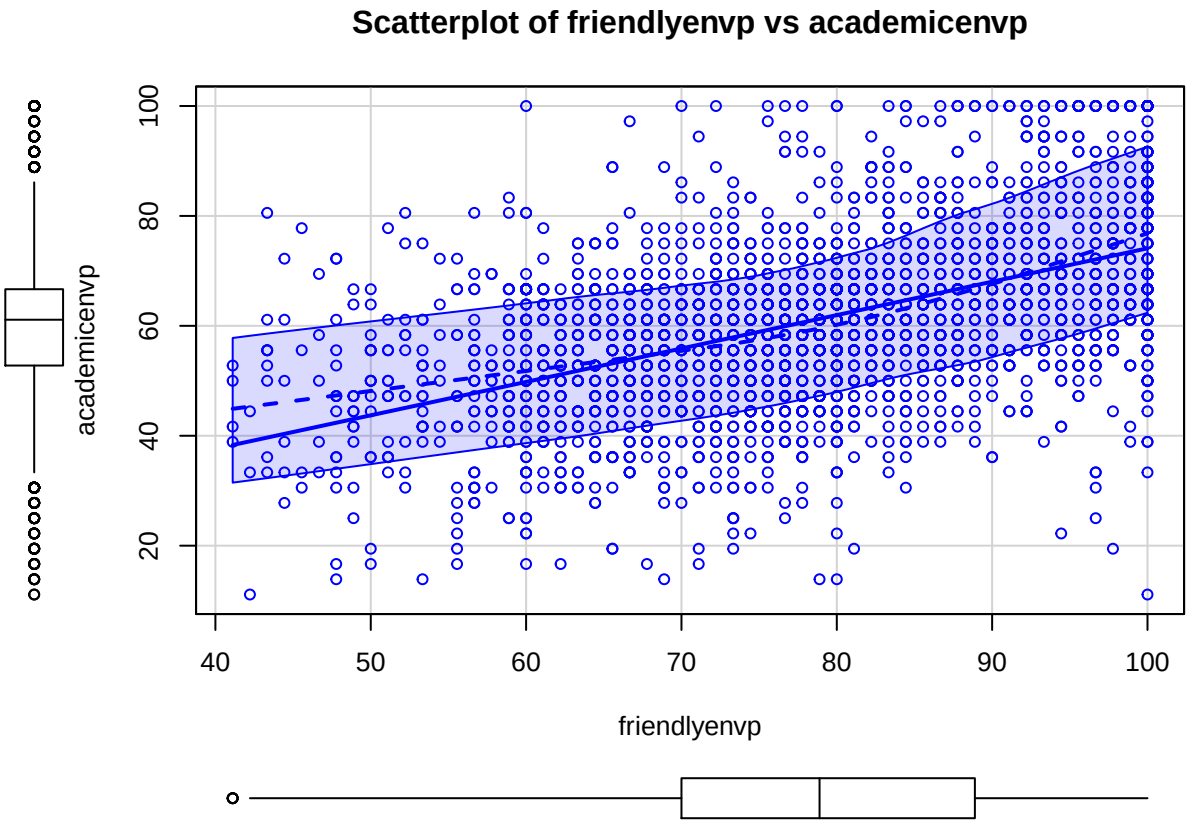
b.

```
## Loading required package: carData
```



From this plot, we can see the solid straight line is the line of best fit from fitting a regression. Meanwhile, the dotted line is the lowess line that is non-parametric and is fitted through the averages of Y and is accompanied by some thin lines that indicate the confidence band. The Boxplot of friendliness has a longer whisker for the left side and a lot of lower outliers suggesting a left skew. Similarly, the boxplot for academic satisfaction is asymetric on the left side indicating a left skew but also notice that there are a couple upper outliers.

c.



Some noticeable changes start with how the regression line's fit has changed and seems to fit the data better and looks closer to the points. The regression line is also a lot closer to the lowess line. There's also a lot less outliers for friendliness.

d.

```
##
## Call:
## lm(formula = academicenvp ~ friendlyenvp, data = clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.491  -8.524  -0.400   7.671  73.589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.42112    1.42961   10.79   <2e-16 ***
## friendlyenvp  0.58181    0.01796   32.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.96 on 2980 degrees of freedom
## Multiple R-squared:  0.2605, Adjusted R-squared:  0.2603
## F-statistic:  1050 on 1 and 2980 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = academicenvp ~ friendlyenvp, data = subs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.001  -8.427  -0.246   7.565  50.200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.3301     1.4965   8.908   <2e-16 ***
## friendlyenvp   0.6078     0.0187  32.504   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.62 on 2940 degrees of freedom
## Multiple R-squared:  0.2644, Adjusted R-squared:  0.2641
## F-statistic:  1057 on 1 and 2940 DF,  p-value: < 2.2e-16
```

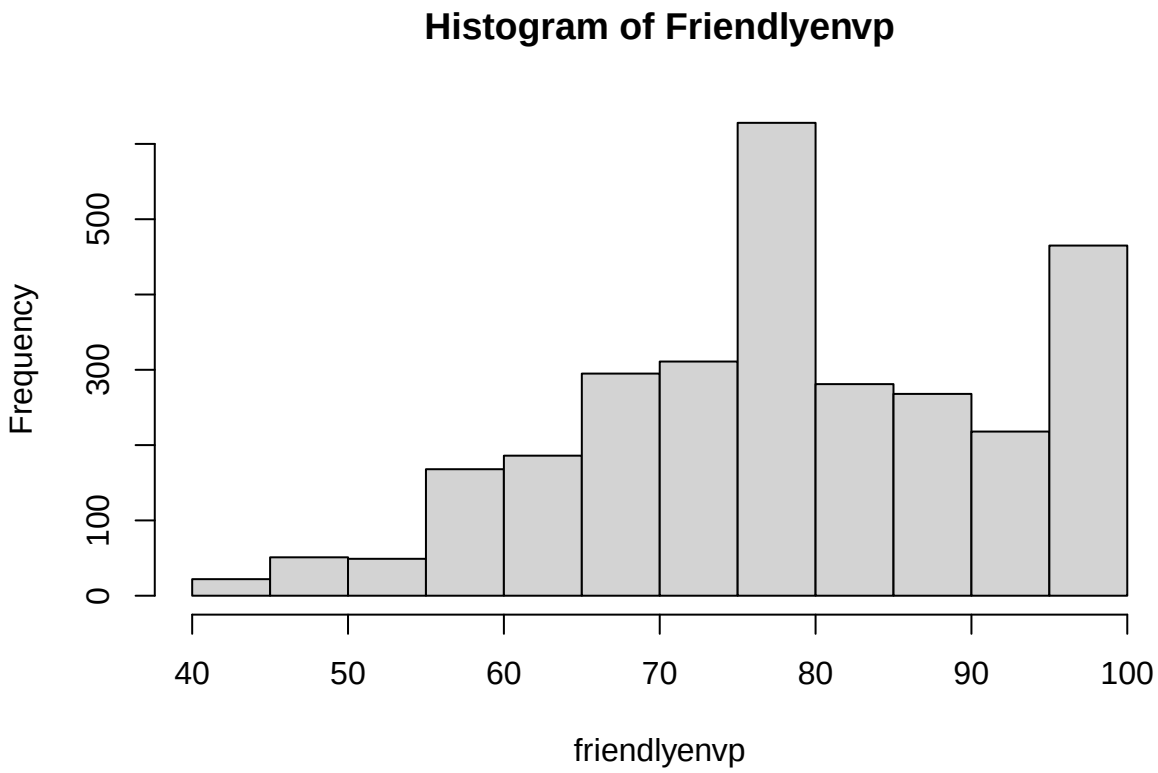
The slope on the model using the subset data is .6078 while it's .5818 for the original data meaning the average rate of change in the academicenvp score is slightly higher for the 2nd model per unit increase in the friendlyenvp score. The R^2 for the subset model .2644 is slightly higher than the model for the original data .2603 indicating a slightly better fit and more variance in academicenvp is explained by friendlyenvp. Also, the intercept and standard error for the residuals is smaller for the model using the subset data compared to the model using the original data.

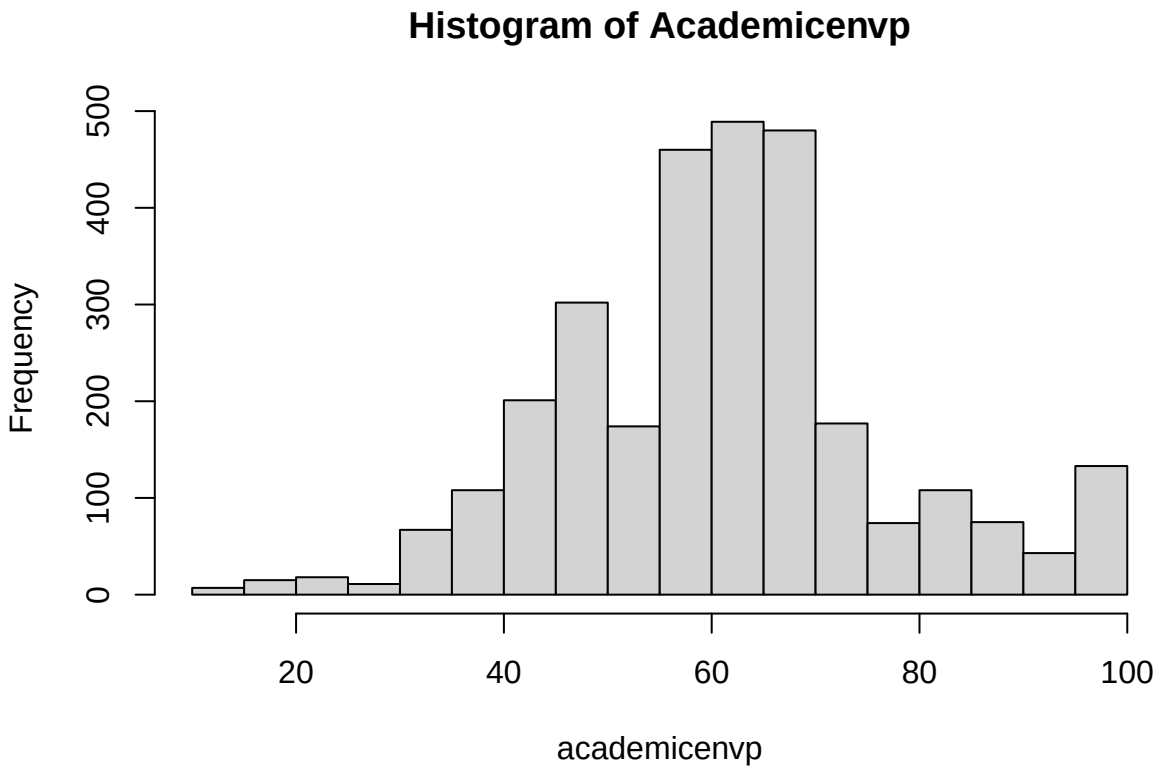
e.

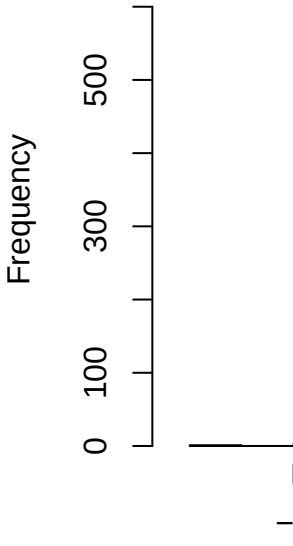
From the results of fitting a regression onto the subset data,

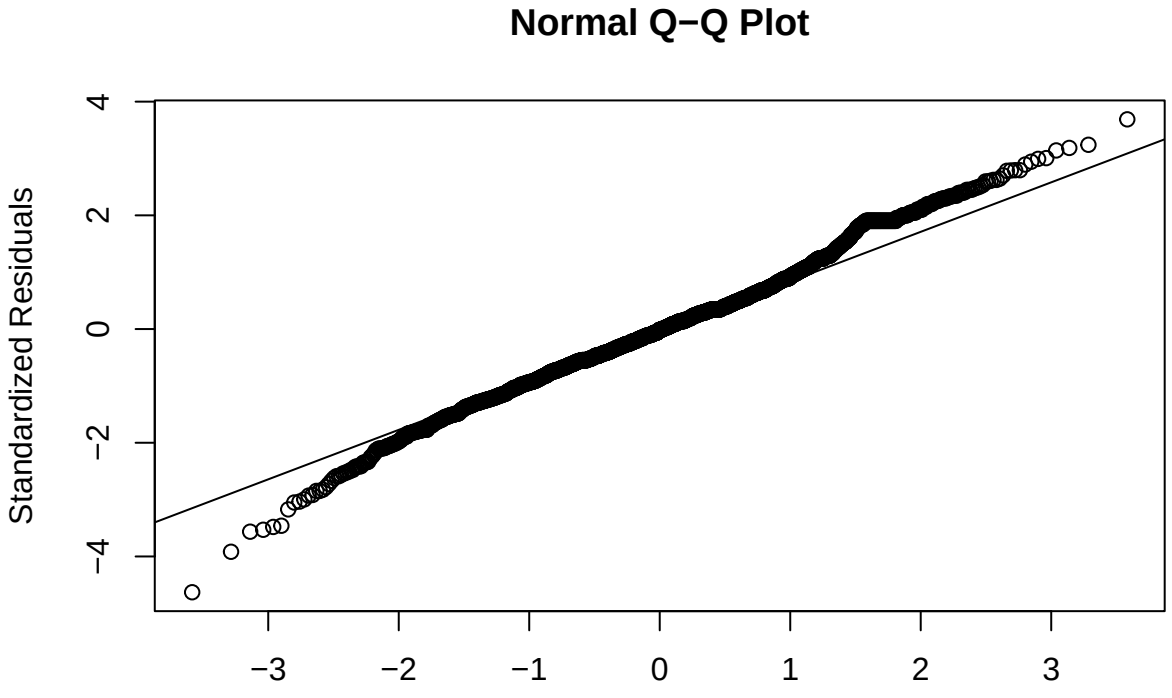
$\hat{\beta}_1$For a one unit increase in the friendlyenvp score, the acamdeicenvp score increases on average by .6078.
$\hat{\beta}_0$When the friendlyenvp score is 0, the average academicenvp score is 13.33.
R^2 26.44% of the variance in academicenvp can be explained by friendlyenvp.

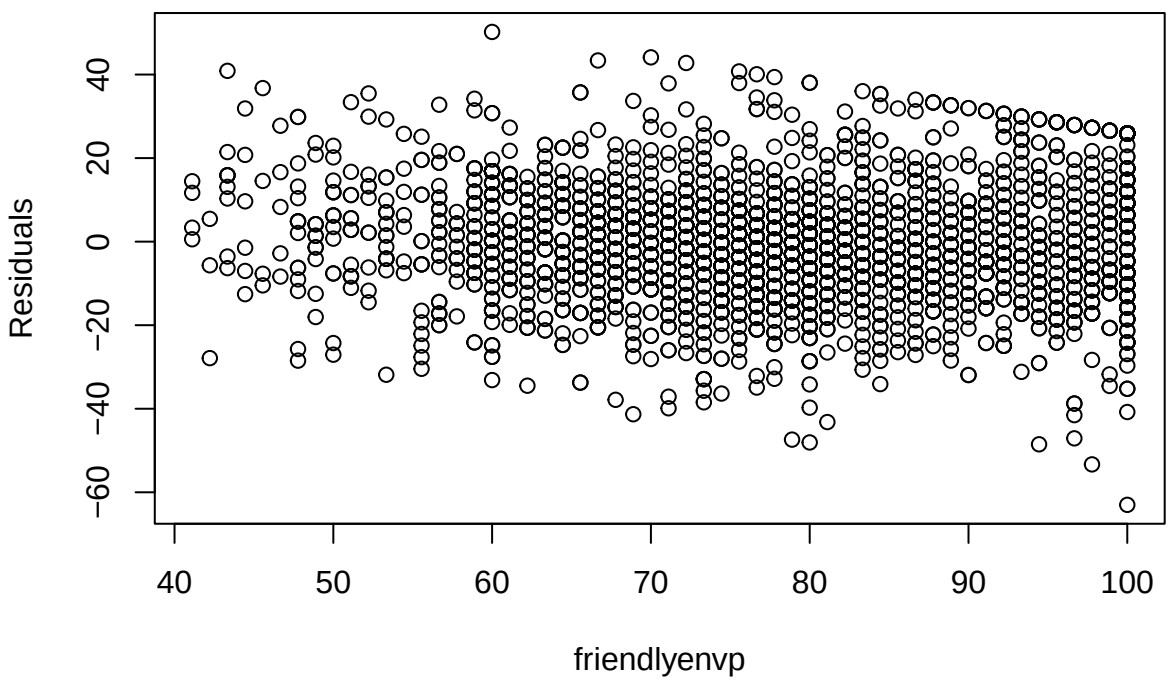
f.


Histogram of Friendlyenvp
Frequency
0
100
300
500
40
50
60
70
80
90
100
friendlyenvp


Histogram of Academicenvp
Frequency
0
100
200
300
400
500
20
40
60
80
100
academicenvp


Frequency
0
100
300
500

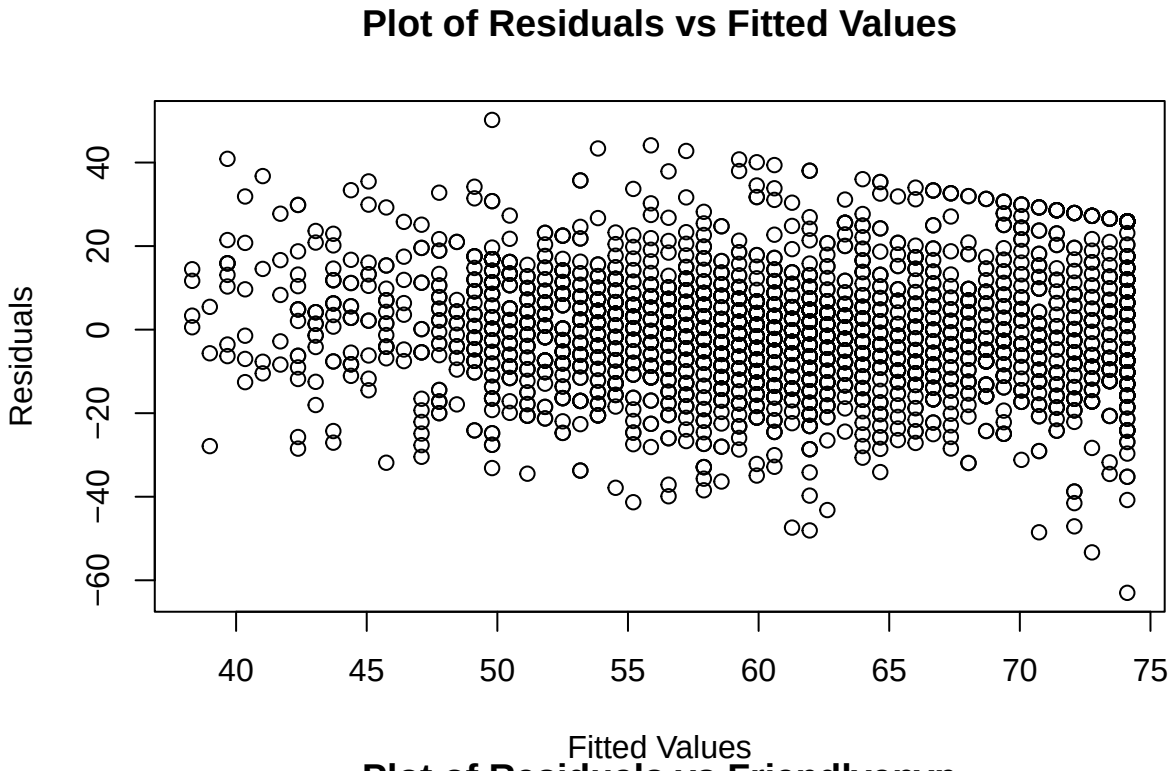


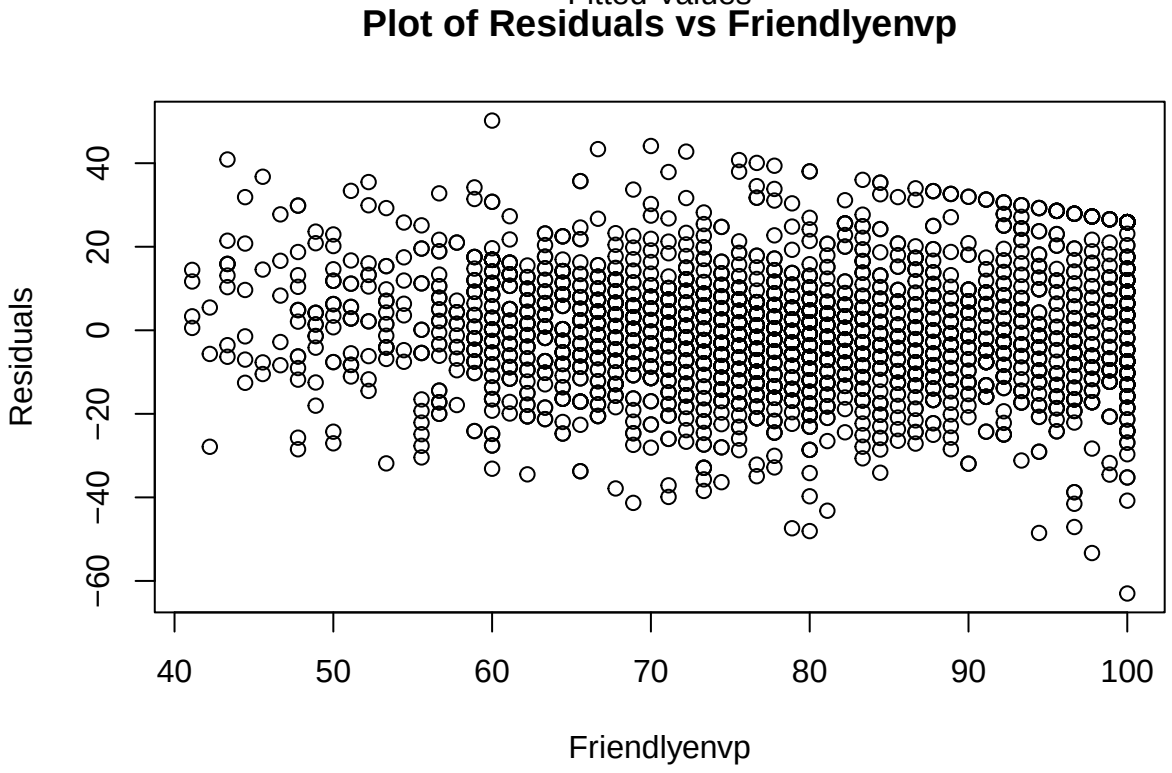


As stated earlier, friendlyenvp has a noticeable left skew while academicenvp has a more moderate left skew and almost looks symmetric. Meanwhile, we can see that the studentized residiuals seem to follow a normal distribution fairly well. Furthermore the qqplot gives better evidence of this and shows that our data roughly follows a normal distribution due to that nearly straight line pattern with the exception of the tiny bump around 1.5 and some points above/below 3 standard deviations. The Residual plot also indicates linearity

due to the lack of pattern and the scatter across all x values are roughly equivalent (although there seems to be less spread past 80 that's not drastically different) implying roughly equal variance.

g.





They look almost identical and it's not exactly a surprise considering the fact that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ the fitted value is simply a linear transformation of the predictor which does not affect the plot.

h.

Utilizing the model fitted on the subset data,
H_0:Homoscedascity is present, H_a:Heteroscedascity is present and Let $\alpha = .05$

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 30.21301, Df = 1, p = 3.871e-08
```

Since our p-value is extremely low and $< .05$, we reject the null hypothesis that homoscedacity is present. However, our residual plots look alright to conclude equal variance as the spread seems roughly equal across all the predictor values. As stated earlier, the spread past 80 seems to be a little bit less which our test may have picked up on but it isn't a drastic violation to conclude unequal variance. Also, this may be due to our large sample size that makes the test more powerful leading to statistical significance which may not always imply practical significance.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.003011394, Df = 1, p = 0.95624
```

Interestingly, if we don't remove the outliers, we fail to reject the null hypothesis and conclude that there is homoscedascity since the p-value is very large and greater than our significance level $0.9562 > 0.05$. This is consistent with what we're seeing in the residual plots of roughly equal variance.

2.1.

What we mean when we say equality of error variance is that the errors/residuals should be spread out the same across all values of the predictor variable. In other words we should not have scatterplots where we see things like a funnel shape.

2.2.

The principle of least squares is where we fit a regression/line of best fit based on selecting the coefficients of the line that minimize (the least part) the sum of squared residuals (the squares part). i.e. $argmin_{\beta_0,\beta_1} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

2.3.

As started in 1g, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ which means that $\hat{Y}$ is really similar to X already which is no surprise they would have almost an identical performance when it comes to evaluating equal error variance. Mathematically speaking, $\hat{Y}$ is a linear transformation of X which will not alter the shape/look of the plot.

2.4.

TSS $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 =$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 =$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$
$\sum_{i=1}^{n}\hat{\epsilon}_i(\hat{Y}_i - \bar{Y}) =$
$\sum_{i=1}^{n}\hat{\epsilon}_i\hat{Y}_i - \sum_{i=1}^{n}\hat{\epsilon}_i\bar{Y} =$
$\sum_{i=1}^{n}\hat{\epsilon}_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}\hat{\epsilon}_i$
recall $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$ so we are left with $\sum_{i=1}^{n}\hat{\epsilon}_i\hat{Y}_i = \sum_{i=1}^{n}\hat{\epsilon}_i(\hat{\beta}_0 + \beta_1\hat{X}_i) =$
$\hat{\beta}_0\sum_{i=1}^{n}\hat{\epsilon}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{\epsilon}_i X_i =$
$0 + \hat{\beta}_1\sum_{i=1}^{n}\hat{\epsilon}_i X_i$
Looking at $\sum_{i=1}^{n}\hat{\epsilon}_i X_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)X_i = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i =$
$\sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\sum_{i=1}^{n}X_i - \hat{\beta}_1\sum_{i=1}^{n}X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - (\bar{Y} - \hat{\beta}_1\bar{X})\sum_{i=1}^{n}X_i - \hat{\beta}_1\sum_{i=1}^{n}X_i^2$
Note $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2}$ which will be proved in 2.5
$\sum_{i=1}^{n}X_iY_i - (\bar{Y} - \hat{\beta}_1\bar{X})n\bar{X} - \hat{\beta}_1\sum_{i=1}^{n}X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} + \hat{\beta}_1 n\bar{X}^2 - \hat{\beta}_1\sum_{i=1}^{n}X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} - \hat{\beta}_1(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2)$
By algebra we get $\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} = \hat{\beta}_1(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2)$
$\hat{\beta}_1(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2) - \hat{\beta}_1(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2) = 0$
Hence we can conclude $2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$
and $TSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ and notice how the formulas on the right is simply RSS and ESS where $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ and $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

Therefore $TSS = RSS + ESS$

2.5

We are interested in deriving $\hat{\beta}_0$ & $\hat{\beta}_1$ or the coefficients of the line of best fit. The line of best fit is derived from minimizing the sum of squared residuals/loss function i.e. $L = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$. In order to minimize such a function, we can take the partial derivatives and set them equal to 0.

First i) $\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0$ and then ii) $\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$

From i, divide both sides by -2 and we get $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = \sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}\beta_0 - \beta_1\sum_{i=1}^{n}X_i = n\bar{Y} - n\beta_0 - \beta_1 n\bar{X}$ Since this is set to 0 solve for β_0 and we get $n\beta_0 = n\bar{Y} - \beta_1 n\bar{X}$
$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

From ii, divide both sides by -2 and we get $\sum_{i=1}^{n}X_iY_i - X_i\beta_0 - \beta_1 X_i^2$. Recall $\beta_0 = \bar{Y} - \beta_1\bar{X}$ so
$\sum_{i=1}^{n}X_iY_i - X_i(\bar{Y} - \beta_1\bar{X}) - \beta_1 X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - X_i\bar{Y} + \beta_1 X_i\bar{X} - \beta_1 X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\bar{Y} + \beta_1\sum_{i=1}^{n}X_i\bar{X} - \beta_1\sum_{i=1}^{n}X_i^2 =$
$\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} + \beta_1 n\bar{X}^2 - \beta_1\sum_{i=1}^{n}X_i^2$
Set equal to 0 and we get $\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} = \beta_1(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2)$ Solving for $\hat{\beta}_1$ we get $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2}$

Notice $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y} = \sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\bar{Y} - \sum_{i=1}^{n}\bar{X}Y_i + \sum_{i=1}^{n}\bar{X}\bar{Y} = \sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = \sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}$

Using the above results, if we replace $Y_i - \bar{Y}$ with another $X_i - \bar{X}$, we get $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}X_i^2 - n\bar{X}^2$

so we can conclude $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

Hence we have derived the coefficients $\hat{\beta}_0, \hat{\beta}_1$ using partial derivatives.

3a.

No because covariance does not tell us about the strength of the relationship directly but rather how two variables vary around their two means. Furthermore regarding the math, there is a lack of standardization to truly be able to compare the strength of the relationships between researchers A and B. To make a meaningful comparison, we would need to calculate the correlations which we would do by dividing each covariance by their respective standard deviations for the predictor and response variable. $r = \frac{S_{xy}}{S_x S_y}$

b.

We call correlation standardized covariance because if we look at the formulas,$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{S_x S_y*(n-1)}$ and $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$, we can rewrite r such that it looks similar to the covariance where the x and y variables are standardized.
$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{S_x S_y*(n-1)} =$
$\frac{1}{n-1}\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{S_x S_y} =$
$\frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{S_x})(\frac{Y_i-\bar{Y}}{S_y})$
$\frac{1}{n-1}\sum_{i=1}^{n} Z_{X_i} Z_{Y_i}$ Note $Z_{X_i} = \frac{X_i-\bar{X}}{S_x}, Z_{Y_i} = \frac{Y_i-\bar{Y}}{S_y}$ notice how the correlation now looks like the covariance formula with the only difference being that the x and y variables have been standardized since we are subtracting each variable by its mean and dividing by its standard deviation. Hence, this is why correlation is called standardized covariance.

4a.

```
xbar <- mean(friendlyenvp)
sx2 <- var(friendlyenvp)
sx <- sd(friendlyenvp)
ybar <- mean(academicenvp)
sy2 <- var(academicenvp)
sy <- sd(academicenvp)
xbar
```

```
## [1] 78.89115
```

```
sx2
```

```
## [1] 180.3267
```

```
sx
```

```
## [1] 13.42858
```

```r
ybar
```

```
## [1] 61.28201
```

```r
sy2
```

```
## [1] 252.0083
```

```r
sy
```

```
## [1] 15.87477
```

$\bar{X} = 78.8912, S_x^2 = 180.3267, S_x = 13.4286$
$\bar{Y} = 61.282, S_y^2 = 252.0083, S_y = 15.8748$

b.

```r
sxy <- cov(friendlyenvp,academicenvp)
rxy <- cor(friendlyenvp,academicenvp)
sxy
```

```
## [1] 109.6069
```

```r
rxy
```

```
## [1] 0.5141625
```

$S_{xy} = 109.6069, r_{xy} = .5142$

c.

```r
TSS <- (nrow(subs)-1)*sy2
TSS
```

```
## [1] 741156.4
```

```r
SSX <- (nrow(subs)-1)*sx2
SSX
```

```
## [1] 530340.7
```

```
RSS <- TSS*(1-rxy^2)
RSS
```

```
## [1] 545222
```

```
Se2 <- RSS/(nrow(subs)-2)
Seb <- sqrt(Se2/SSX)
Seb
```

```
## [1] 0.01869974
```

```
beta1 <- sxy/sx2
tstat <- beta1/Seb
tstat
```

```
## [1] 32.50441
```

```
Fstat <- (TSS-RSS)/Se2
Fstat
```

```
## [1] 1056.537
```

i)$TSS = S_y^2 * (n-1) = 252.0083 * (2942-1) = 741156.4$
ii)$SSX = S_x^2 * (n-1) = 180.3267 * (2942-1) = 530340.7$
iii)$1 - R^2 = 1 - r^2 = \frac{RSS}{TSS}$ thus $RSS = TSS * (1-r^2) = 741156.4 * (1 - .5142^2) = 545222$
iv)$S(e)_{\hat{\beta}_1} = \frac{S_e}{\sqrt{SSX}} = \frac{\sqrt{\frac{RSS}{n-2}}}{\sqrt{SSX}} = \frac{\sqrt{\frac{545222}{2942-2}}}{\sqrt{530340.7}} = 0.0187$

v)$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{109.6069}{180.3267} = 0.6078$
$t = \frac{\hat{\beta}_1 - 0}{S(e)_{\hat{\beta}_1}} = \frac{.6078}{.0187} = 32.5044$

vi)$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} = \frac{\frac{741156.4-545222}{1}}{\frac{545222}{2942-1-1}} = 1056.537$

vii)Notice how $t^2 = 32.5044^2 = 1056.537 = F$ Hence we can conclude $t^2 = F$

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
df <- read.csv('campusclimate.csv')
head(df[,c('friendlyenvp','academicenvp')])
clean <- df[,c('friendlyenvp','academicenvp')]
clean <- clean[complete.cases(clean),] #Drop NA Rows
plot(x = clean$friendlyenvp, y = clean$academicenvp,
     xlab = 'friendlyenvp',ylab = 'academicenvp',
     main = 'Scatterplot of friendlyenvp vs academicenvp')
library(car)
scatterplot(academicenvp ~ friendlyenvp, span=0.6, lwd=3,
 id.n=4, data=clean,
```

```
 main ="Scatterplot of friendlyenvp vs academicenvp",
 xlab = 'friendlyenvp',ylab = 'academicenvp')

subs <- clean[clean$friendlyenvp > 40 & clean$academicenvp > 10,]
#Select rows where friendlyenvp is greater than 40
#AND academicenvp is greater than 10
attach(subs)
scatterplot(academicenvp ~ friendlyenvp, span=0.6, lwd=3,
 id.n=4, data=subs, main ="Scatterplot of friendlyenvp vs academicenvp"
 ,xlab = 'friendlyenvp',ylab = 'academicenvp')
model1 <- lm(academicenvp ~ friendlyenvp, data = clean)
model2 <- lm(academicenvp ~ friendlyenvp, data = subs)
summary(model1)
summary(model2)
hist(friendlyenvp,main = "Histogram of Friendlyenvp")
hist(academicenvp,main = "Histogram of Academicenvp")
hist(rstudent(model2), main = "Histogram of Studentized Residuals",xlab="Studentized Residuals")
qqnorm(rstandard(model2), ylab = 'Standardized Residuals')
qqline(rstandard(model2))
plot(x = friendlyenvp, y = model2$residuals, xlab = "friendlyenvp",
     ylab = 'Residuals',main ='Plot of Residuals vs friendlyenvp')
plot(x = model2$fitted.values, y = model2$residuals,
     xlab = "Fitted Values", ylab = 'Residuals',
     main ='Plot of Residuals vs Fitted Values')
plot(x = friendlyenvp, y = model2$residuals,
     xlab = "Friendlyenvp", ylab = 'Residuals',
     main ='Plot of Residuals vs Friendlyenvp')
ncvTest(model2)
ncvTest(model1)
xbar <- mean(friendlyenvp)
sx2 <- var(friendlyenvp)
sx <- sd(friendlyenvp)
ybar <- mean(academicenvp)
sy2 <- var(academicenvp)
sy <- sd(academicenvp)
xbar
sx2
sx
ybar
sy2
sy
sxy <- cov(friendlyenvp,academicenvp)
rxy <- cor(friendlyenvp,academicenvp)
sxy
rxy
TSS <- (nrow(subs)-1)*sy2
TSS
SSX <- (nrow(subs)-1)*sx2
SSX
RSS <- TSS*(1-rxy^2)
RSS

Se2 <- RSS/(nrow(subs)-2)
```

```
Seb <- sqrt(Se2/SSX)
Seb
beta1 <- sxy/sx2
tstat <- beta1/Seb
tstat

Fstat <- (TSS-RSS)/Se2
Fstat
```