Stats 402 ~ Homework 1

Group 4 (Britney Brown, Harrison DiStefano, Jaehui(Jaehee) Jeong, Lisa Kaunitz, Tianyang Liu, Yuandong (David) Sun, Jeremy Weidner)

10/19/2020

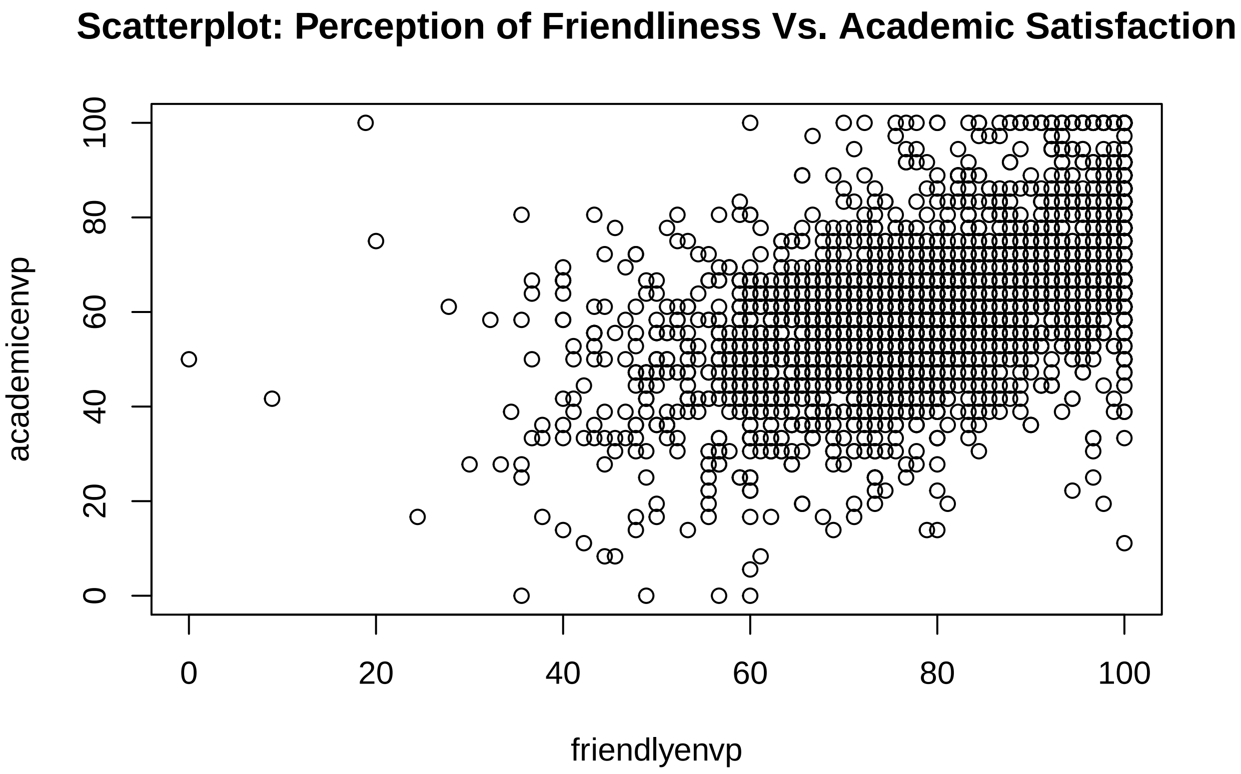
Problem One

Using the following campus climate data set (data folder week one), answer the following questions:

```r
campusclimate <- read.csv("campusclimate.csv")
attach(campusclimate)
```

a) Create a scatterplot for showing UCLA students' perception of friendliness on our campus (friendlyenvp), predictor, as a function of their perception of academic satisfaction on our campus (academicenvp), outcome.

```r
#notice a slight positive trend
plot(academicenvp ~ friendlyenvp,
     main = "Scatterplot: Perception of Friendliness Vs. Academic Satisfaction")
```



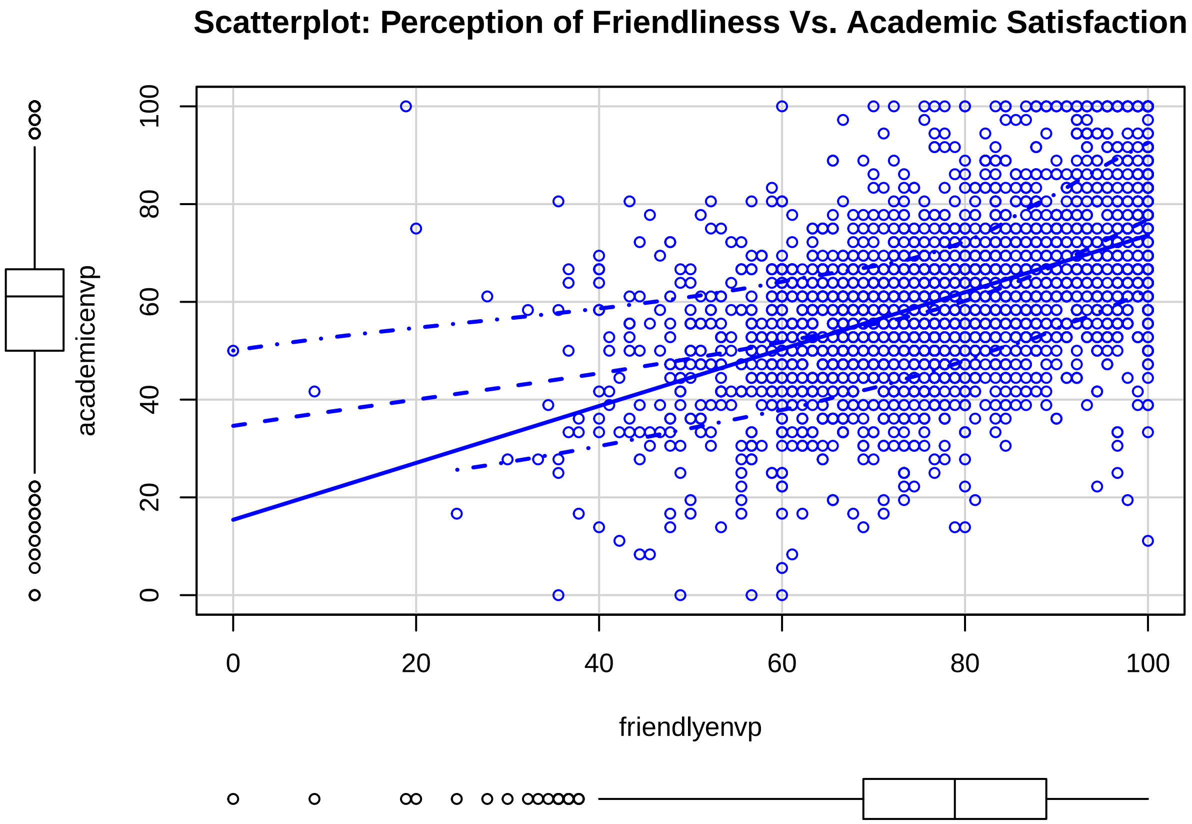
b) Now create a scatterplot like the one shown on (same plot as page 21 of lecture one). Remember to install the "car" package. Explain what the lines in this plot show? What do you conclude from the two boxplots?

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

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

```r
scatterplot(academicenvp ~ friendlyenvp,
            main = "Scatterplot: Perception of Friendliness Vs. Academic Satisfaction")
```

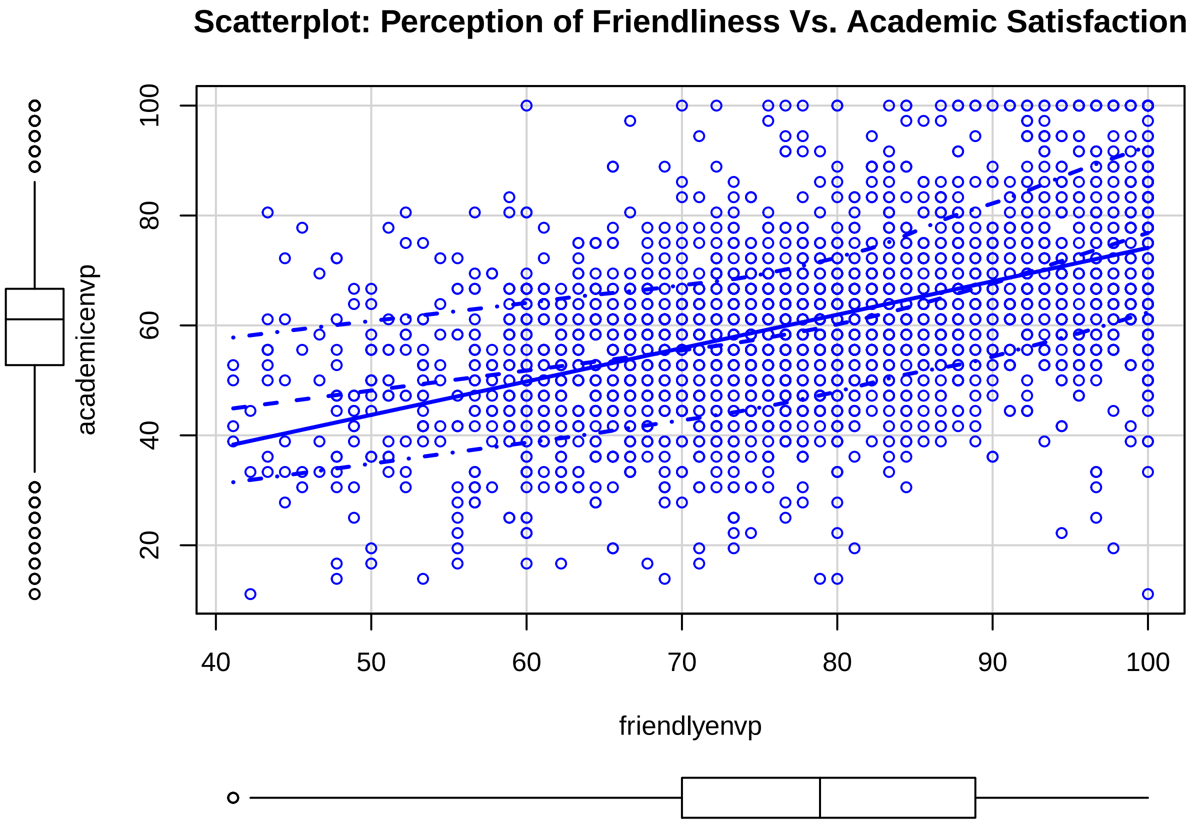


The straight line is the least square regression line where residuals are minimized. The central dashed line is a non-parametric lowess line that passes through the average y value for any given x value while the upper and lower broken lines smooth the positive and negative residuals. While both lines model closely to the right of the graph, they vary greatly on the low outliers.

The boxplots show the conditional distributions for the perception of friendliness and academic satisfaction. The horizontal boxplot represents the relative frequency of academic satisfaction given a specific value of friendliness while the vertical boxplot represents the relative frequency of friendliness given a specific value of academic satisfaction. Like the difference between the least square regression and lowess lines, these boxplots show the variation in the outliers on the lower end of each variable.

c) As you see from the two plots you created in parts a and b, there is very little data below friendlyenvp = 40 and below academicenvp = 10. Create a new subset by deleting the above range of data, attach this new data set and draw the scatter plot you drew in part b. Use the following reference as guideline for sub-setting your data. I have also included the hsb2 data in the data folder of week one so you can recreate what you find in the following reference. Comment how everything changed. https://stats.idre.ucla.edu/r/modules/subsetting-data/

```
#subset original dataset by deleting below academic = 10 and friendly = 40
newdata <- campusclimate[academicenvp > 10 & friendlyenvp > 40,]
attach(newdata) #newdata will mask objects in original climate data
scatterplot(academicenvp ~ friendlyenvp,
            main = "Scatterplot: Perception of Friendliness Vs. Academic Satisfaction")
```



With the new subset of data, the least square regression line appears closer to the lowess line throughout the graph, no longer varying greatly near the lower end of the variables. We also see a lot less outliers in the horizontal boxplot (representing the conditional distribution of academic satisfaction given friendliness). In theory, this means that the subset will be able to create an accurate model.

d) Now create two linear models for the prediction of UCLA students' perception of academics from the friendliness by using: 1) The original campus climate data, and 2) the subset you created. Compare the two models in terms of the slope and R-squared and comment on any changes that happened.

```r
#linear model for original data
lm.original <- lm(academicenvp ~ friendlyenvp, data=campusclimate)
summary(lm.original)
```

```
##
## Call:
## lm(formula = academicenvp ~ friendlyenvp, data = campusclimate)
##
## 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
##   (2400 observations deleted due to missingness)
## Multiple R-squared:  0.2605, Adjusted R-squared:  0.2603
## F-statistic:  1050 on 1 and 2980 DF,  p-value: < 2.2e-16
```

```r
#linear model for subset data
lm.new <- lm(academicenvp ~ friendlyenvp, data=newdata)
summary(lm.new)
```

```
##
## Call:
## lm(formula = academicenvp ~ friendlyenvp, data = newdata)
##
## 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
##   (2373 observations deleted due to missingness)
## 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 is statistically significant in both models: 0.58 for the original and slightly sleeper at 0.61 for the subset. The intercept also changes from 15.42 to 13.33 so both $\hat{\beta}_0$ and $\hat{\beta}_1$ are different in these linear models. However, the models do not vary in their ability to explain the variation in the outcome since the R^2 values show little difference (26.44% > 26.05%).

e) Interpret the slope, intercepts, and R-squared for model resulting from the subset of the data you created within context

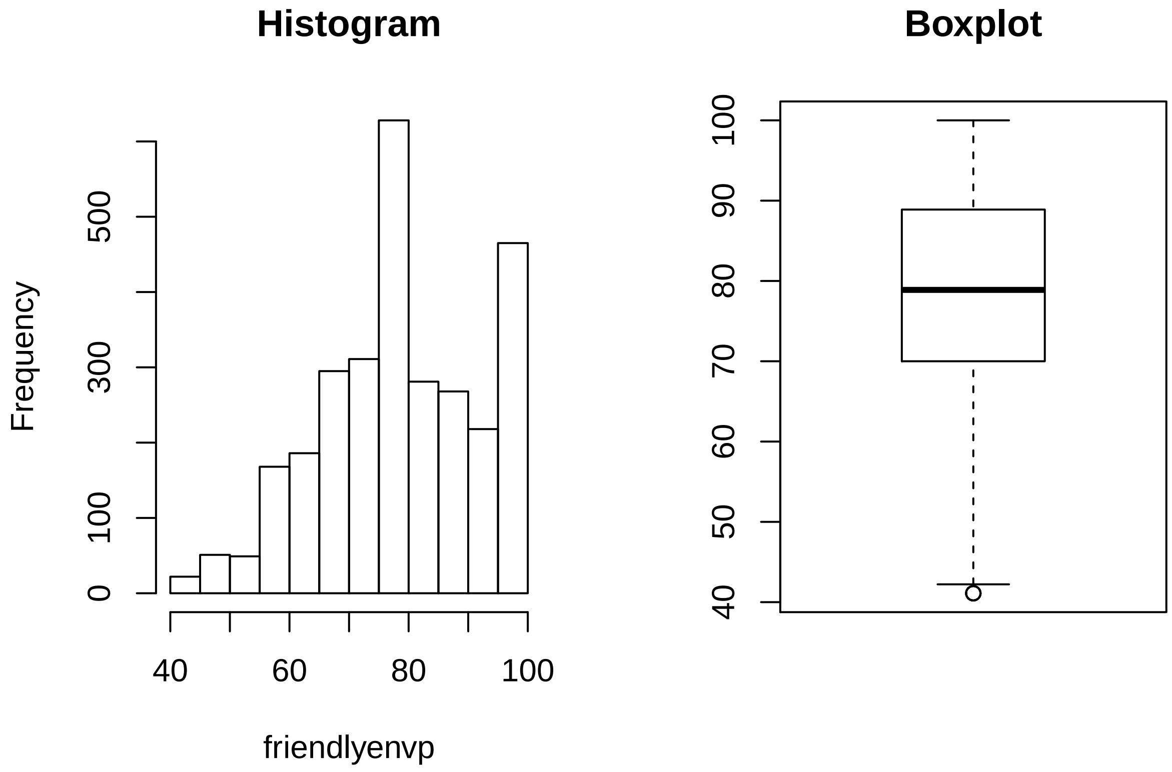
```r
#check range of predictor, the minimum in subset is not 0
summary(newdata$friendlyenvp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   41.11   70.00   78.89   78.89   88.89  100.00    2373
```

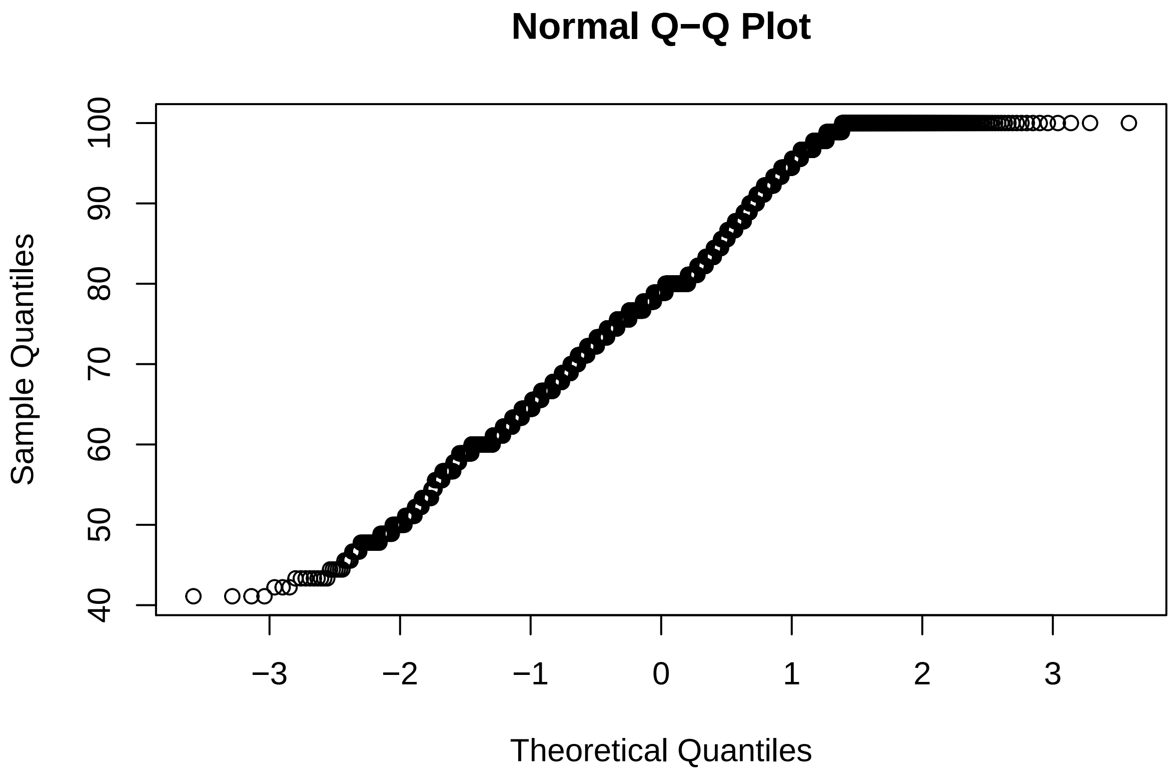
For every unit increase of the perception of friendliness, we except the average of the perception of academic satisfaction to increase by 0.61 units. Since the minimum score for friendlyenvp is not 0, it makes no sense to interpret the intercept. Finally, 26.44% of the variance in the perception of academic satisfaction is explained by the perception of friendliness on campus.

f/g) Conduct exploratory data analysis by creating the histograms, qqplot, plot of residuals vs. predictor. One quick way to perform exploratory data analysis is to use the common plot(name of the model) function. This will provide us with the majority of the plots we need.

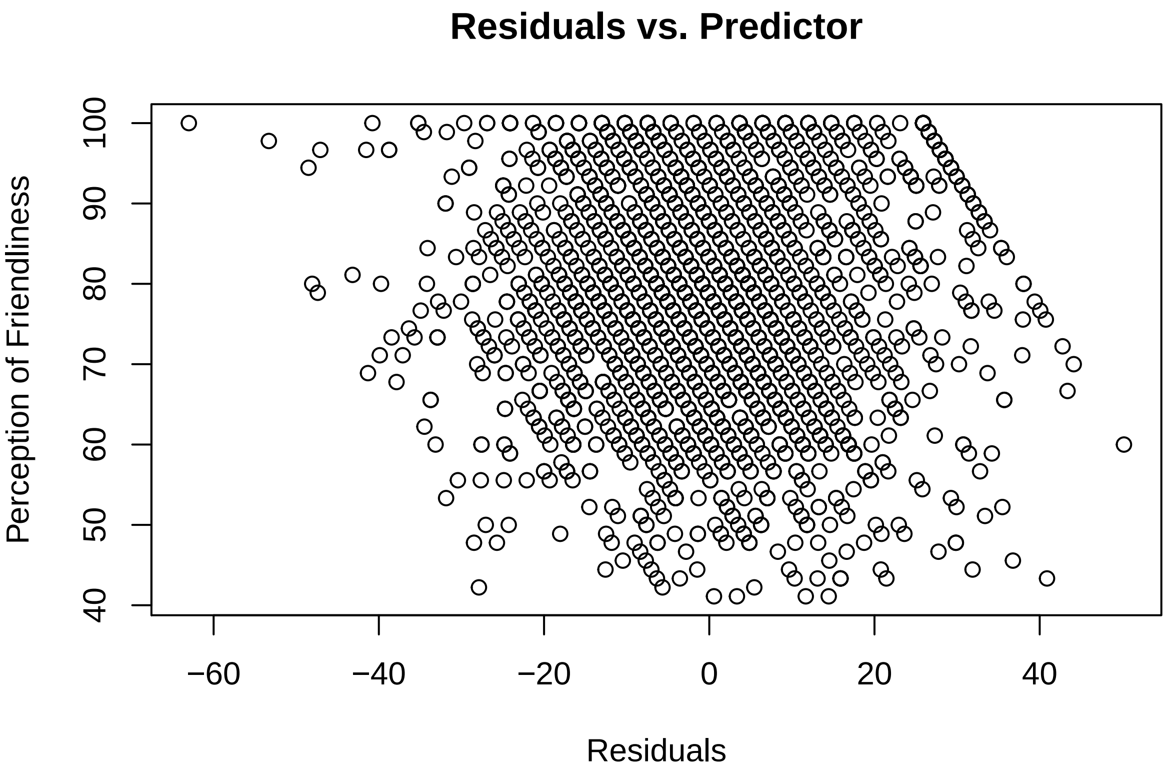
```r
#explore the predictor: Perception of Friendliness
par(mfrow=c(1,2))
hist(newdata$friendlyenvp, main = "Histogram", xlab = "friendlyenvp")
boxplot(newdata$friendlyenvp,main = "Boxplot")
```



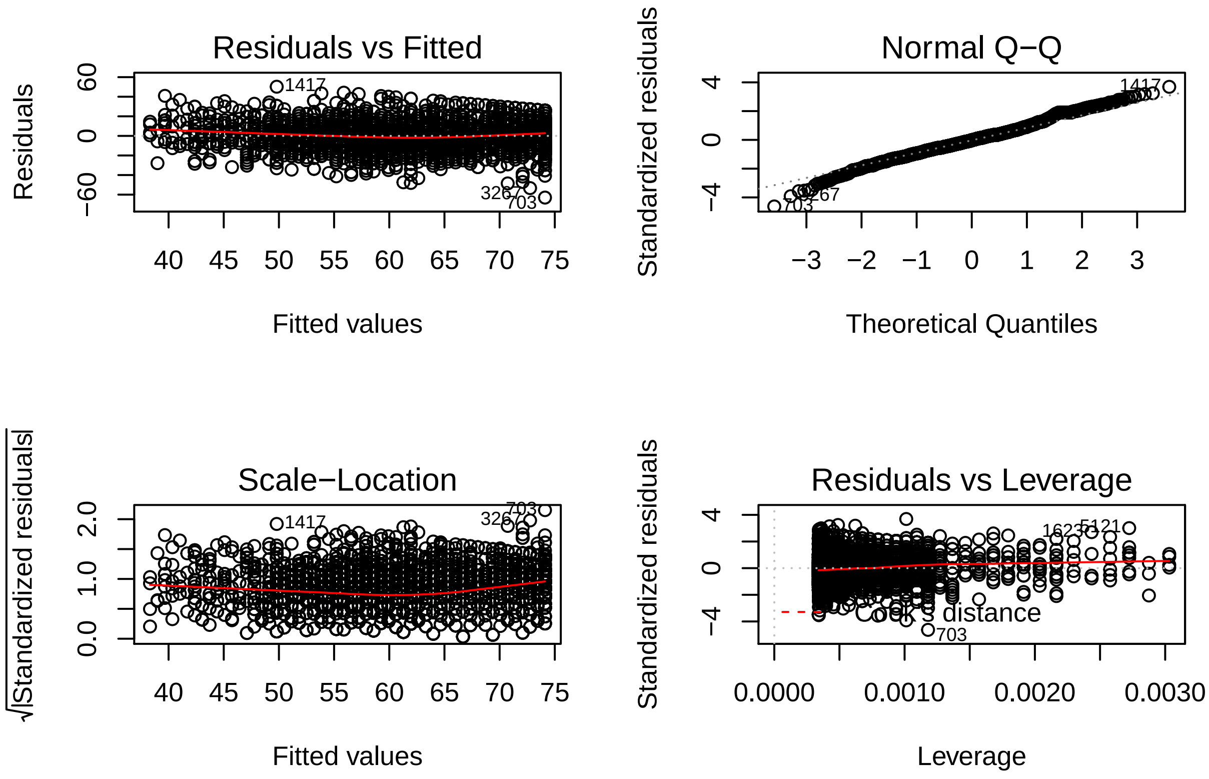
```
#check normality of predictor
qqnorm(newdata$friendlyenvp) #notice lots of 100 values
```



```r
#residual vs predictor
plot(x=lm.new$residuals, y=friendlyenvp[!is.na(academicenvp)],
     xlab = "Residuals", ylab = "Perception of Friendliness",
     main = "Residuals vs. Predictor")
```



```r
par(mfrow=c(2,2))
plot(lm.new)
```

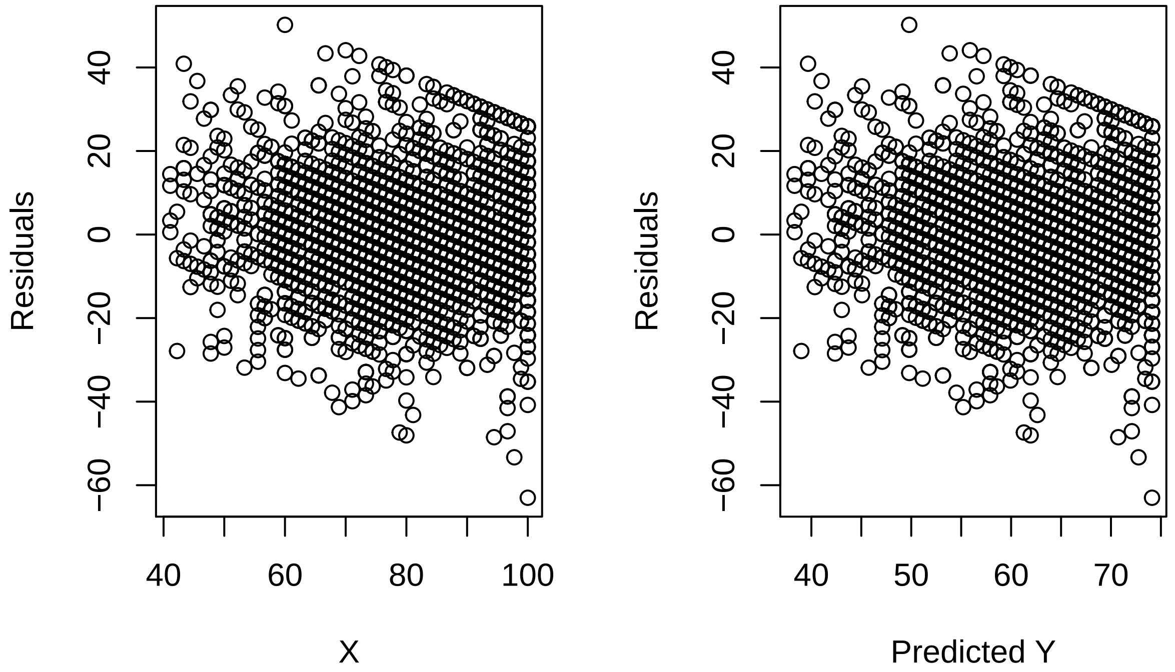


h) [adjusted typo] Draw the plot of academicenvpˆ (Yˆ) vs residual and vs. friendlyenvp. Are the different or the same? Explain why they are the same or different?

```r
par(mfrow=c(1,2))

#residual vs x (friendly)
plot(friendlyenvp[!is.na(friendlyenvp)],lm.new$residuals,
     xlab = "X", ylab = "Residuals")

#residual vs y_hat (predicted academic)
plot(lm.new$fitted.values,lm.new$residuals,
     xlab = "Predicted Y", ylab = "Residuals")
```



These plots look identical (except for the ranges on the horizontal axis). This is because x is used to predict y. In the simple linear regression case, our predictor x is multiplied by the estimated slope and added to the estimated intercept to create the predicted values for y. Therefore, since $\hat{y}$ is a function x, they have the same relationship with residuals, just on a different scale.

i) Conduct the ncv test to show that the principle of equality of error variance holds.

```
ncvTest(lm.new)
```

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

The results of the non-constant variance test has a p-value of approximately 0. Therefore, we have sufficient evidence to reject the null hypothesis of homoscedasticity of residuals and conclude that we are not meeting the assumption of equality of variance in the residuals.

Problem Two.

a) In regression, conceptually speaking what do we mean by the principle of equality of error variance.

Conceptually, the equality of error variance means as the value of predicted y increases, the residuals will not be effected (stay constant) since there is no correlation between the residuals and the predicted y.

b) In regression, mathematically what do we mean by the principle of least squares?

In regression, the goal is to find the line that minimizes the amount of error between the actual and predicted values. Since the sum of errors will always equal 0 (because of positive and negative values), we look at the square of errors: $e_1^2 + e_2^2 + e_3^2 + ... + e_n^2 = \sum_{i=1}^N e_i^2$ where e_i's are the residuals

Therefore, the best fit line is the one that minimizes the sum of residuals squared shown above.

c) In regression plot of residuals vs. X or residuals vs. Y^ serve equally well for checking the principle of equality of error variance. Why is this the case? Explain conceptually and mathematically.

$\hat{y}$ is a linear function of x found by multiplying x by the estimated slope and adding the intercept. Since $\hat{y}$ is just a scaled version of x, they serve the same purpose when checking for a relationship with residuals.

Mathematically, suppose $r = b_1\hat{Y} + b_0$ and $\hat{Y} = a_1X + a_0$, then $r = b_1a_1X + b_1a_0 + b_0$.

d) Prove that sum of square of total = Sum of square of regression + sum of square of residual. (see answers to review exercise one)

Sum of square of total:
$\sum_{i=1}^N (Y_i - \bar{Y})^2$
$= \sum_{i=1}^N (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$
$= \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^N 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$
$= SSE + SSR + \sum_{i=1}^N 2e_i(\hat{Y}_i - \bar{Y})$

Since $\sum_{i=1}^N 2e_i(\hat{Y}_i - \bar{Y}) = 2\sum_{i=1}^N e_i(b_0 + b_1x_i - \bar{y}) = 2b_0\sum_{i=1}^N e_i + 2b_1\sum_{i=1}^N x_ie_i + 2\bar{y}\sum_{i=1}^N e_i = 0$
because $\sum_{i=1}^N e_i = 0$, SST = SSE + SSR

e) Prove that slope and intercept result from placing the derivative of the sum of square of residuals equal to zero.

SSE $= \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N (Y_i - b_0 - b_1X_i)^2$

We want to find b_0 and b_1 such that SSE is minimized. Let's take the derivative with respect to b_0 and b_1 for SSE:

$\frac{\partial SSE}{\partial b_0} = \sum_{i=1}^N -2(Y_i - b_0 - b_1X_i)$
$\frac{\partial SSE}{\partial b_1} = \sum_{i=1}^N -2(Y_i - b_0 - b_1X_i)X_i$

In order to minimize SSE, we set the first derivatives of SSE to 0:
$-2\sum_{i=1}^N (Y_i - b_0 - b_1X_i) = 0$
$-2\sum_{i=1}^N (Y_i - b_0 - b_1X_i)X_i = 0$

Solve for b_0:
$\sum_{i=1}^{N} b_0 = \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} b_1 X_i$
$N b_0 = \sum_{i=1}^{N} Y_i - b_1 \sum_{i=1}^{N} X_i$
$b_0 = \bar{Y} - b_1 \bar{X}$

Solve for b_1 (plug b_0 into $\frac{\partial SSE}{\partial b_1}$):

$-2 \sum_{i=1}^{N} X_i(Y_i - b_0 - b_1 X_i) = 0$
$\sum_{i=1}^{N} X_i(Y_i - \bar{Y} + b_1 \bar{X} - b_1 X_i) = 0$
$b_1 \sum_{i=1}^{N} X_i(\bar{X} - X_i) = \sum_{i=1}^{N} X_i(\bar{Y} - Y_i)$
$b_1 = \frac{\sum_{i=1}^{N} X_i(\bar{Y} - Y_i)}{\sum_{i=1}^{N} X_i(\bar{X} - X_i)} = \frac{\frac{\sum_{i=1}^{N} (X_i - \bar{X}) * (Y_i - \bar{Y})}{N-1}}{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}}$
Therefore, $b_1 = \frac{S_{XY}}{S_X^2}$

Note:
$\sum_{i=1}^{N} \left(X_i - \bar{X}\right) * \left(Y_i - \bar{Y}\right) = \sum_{i=1}^{N} X_i(\bar{Y} - Y_i)$ because $\bar{X} \sum_{i=1}^{N} (Y_i - \bar{Y}) = 0$
$\sum_{i=1}^{N} \left(X_i - \bar{X}\right) * \left(X_i - \bar{X}\right) = \sum_{i=1}^{N} X_i(\bar{X} - X_i)$ because $\bar{X} \sum_{i=1}^{N} (X_i - \bar{X}) = 0$

Problem Three

Two researchers are studying the relationship between math and physics scores. Researcher A finds the covariance to be 700, Researchers B finds covariance to be 900. They both use a sample size of 100.

a) Can we say that researcher B showed a stronger relationship between math and physics scores, yes or no and why? Make your point mathematically and conceptually.

Both covariances show a positive relationship between math and physics scores. However, we cannot say researcher B showed a stronger relationship because covariance is not standardized and therefore does not provide any information on the strength of the relationship.

Mathematically, covariance is $S_{XY} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right) * \left(Y_i - \bar{Y}\right)}{N-1}$. According to the formula, there is no consideration for the standard deviation of each scores. If $\left(X_i - \bar{X}\right) >> 0$ and $\left(Y_i - \bar{Y}\right) >> 0$, the corvariance can be big while the actual relation between the two variables can be small. In such a case, corvariance fails to show the strength of relations between two values.

b) Why do they call correlation standardized covariance?

$r_{XY} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right) * \left(Y_i - \bar{Y}\right)}{S_X * S_Y * N-1}$.

From the formula, we see that the coefficient of correlation takes the standard deviation of each scores into consideration by dividing the covariance by the product of standard deviation of each variable. This standardizes the covariance. Since covariance indicates the direction (positive or negative) of the relationship between variables, correlation is also able to indicate direction as well as compare the strength of variable relationships because it is unitless.

Problem Four

a/b) Using R, calculate the following, using the subset that you created.

```
#remove NA's for analysis
x <- friendlyenvp[!is.na(friendlyenvp)]
y <- academicenvp[!is.na(academicenvp)]

#set N
if(length(x) ==length(y)){
  N <- length(x)
}
N
```

```
## [1] 2942
```

```
#friendlyenvp summary stats
x.mean <- mean(x)
x.sd <- sd(x)
x.var <- var(x)

#academicenvp summary stats
y.mean <- mean(y)
y.sd <- sd(y)
y.var <- var(y)

a <- rbind(c(x.mean, x.sd, x.var),
      c(y.mean, y.sd, y.var))
row.names(a) <- c("friendlyenvp", "academicenvp")
colnames(a) <- c("mean", "sd", "variance")
a
```

```
##                  mean       sd variance
## friendlyenvp 78.89115 13.42858 180.3267
## academicenvp 61.28201 15.87477 252.0083
```

c) Using R, calculate the coefficient of correlation and covariance between UCLA students' perception of academics and friendliness of our environment at UCLA. Use parts a and b to calculate…

```
#covariance
covar <- (sum((x - x.mean)*(y - y.mean)))/(N-1)
covar
```

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

```
#check: covariance function
covar == cov(x, y)
```

```
## [1] TRUE
```

```
#correlation
corr <- covar / (x.sd*y.sd)
corr
```

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

```
#check: correlation function
corr == cor(x, y)
```

```
## [1] TRUE
```

```
k <- 1 #number of predictors
b1 <- covar / (x.sd^2) #slope
b0 <- y.mean - b1*x.mean #intercept
y.hat <- b0 + b1*x #predicted y
se <- sqrt((sum((y - y.hat)^2))/(N-2)) #sd of residuals

# TSS (Total Sum of Square)
tss <- sum((y -y.mean)^2)
tss
```

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

```
# RSS (Residual Sum of Square)
rss <- sum((y - y.hat)^2)
rss
```

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

```
# SSR (Sum of Squares Regression)
ssr <- tss - rss
ssr
```

```
## [1] 195934.4
```

```
# SSX
ssx <- sum((x -x.mean)^2)
ssx
```

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

```
# Standard Error of the Slope
se.b <- se/sqrt(ssx)
se.b
```

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

```
# t-test of the slope (under the null B1 = 0)
t.val <- b1/se.b
t.val
```

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

```
# F-test of R-squared
F.val <- (ssr/k)/(rss/(N-k-1))
F.val
```

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

```
# Show that F = t^2
F.val == t.val^2
```

```
## [1] TRUE
```