Fall 2019 – Homework Four Stat 112 – Professor Esfandiari

Due Tuesday December 3rd by eleven PM

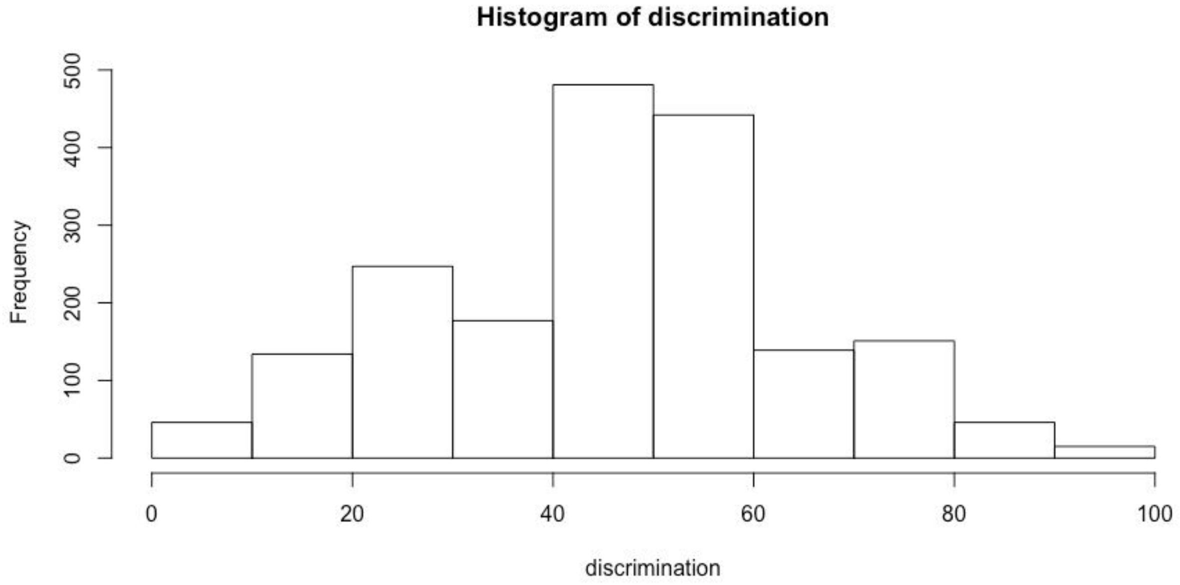
Group Member Names: Yaxin Tan (705121965), Charles Liu (304804942), Kevin Nguyen (305182183), Alexandra Tucker (704813360), Sam Naraghi (305188279), Wan-ling Renee So (105189595)

Tier one:Complete questions one to four

Question one. Using the sex discrimination data set…

1. Make a histogram of scores on discrimination toward women

```
hist(discrimination)
```



2. Compute the summary statistics

```
>Summary(discrimination)

  Min.  1st Qu.  Median    Mean   3rd Qu.    Max.     NA's
  0.00   35.29    47.06    45.99   58.82   100.00     372
```

3. Using the following R command, make discrimination into three factors (below 25% or low, middle 50% or average, top 25% or high)

```
>discut<-cut(discrimination,br=c(0,35.29,58.82,100),lables=c("low","average","high"), right=FALSE)
```

4. find the frequency with the above categories

```
>table<-(discut)
discut
 [0,35.3]      [35.3,58.8]       [58.8,100]


 427           896               551
```

5. Now create a contingency table between discrimination and level of education

```
>table(discut,edur)
 discut              college                 less than college


 [0,35.3]            71                      256


 [35.3,58.8]         373                     520


 [58.8,100]          163                     386
```

The frequencies within the cells look fine.

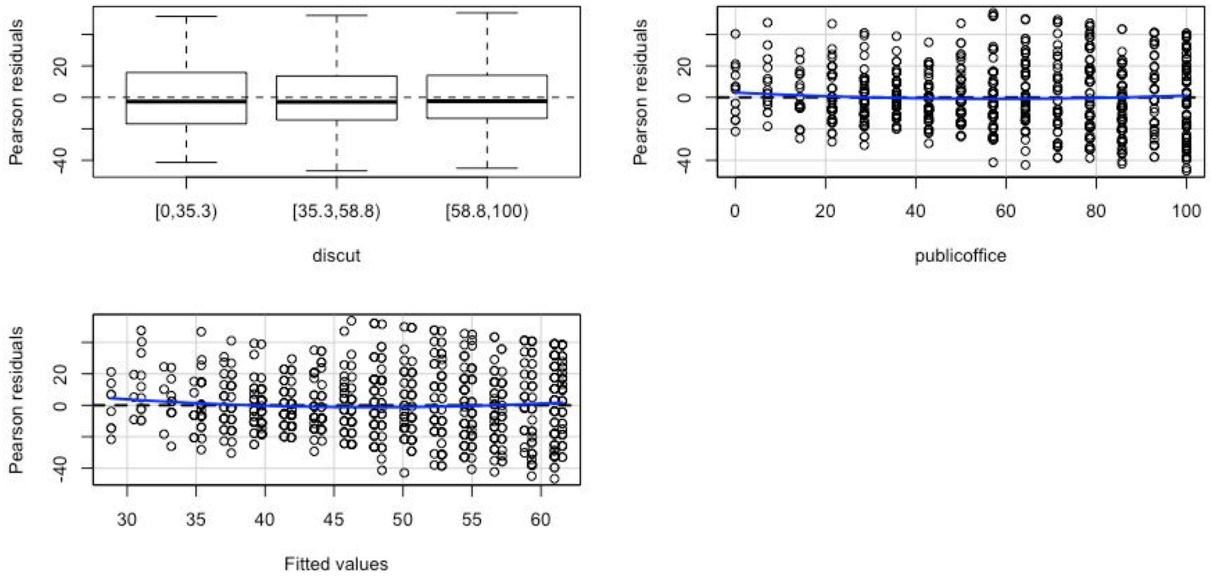
6. Create a linear model for the prediction of attitude toward female leadership, from attitude toward placing women in public office and attitude toward discrimination of females.

```
m1 <- lm(leadership~discut+publicoffice)
```

7. check the relevant assumptions, scatterplots, etc. you will find the following command useful:

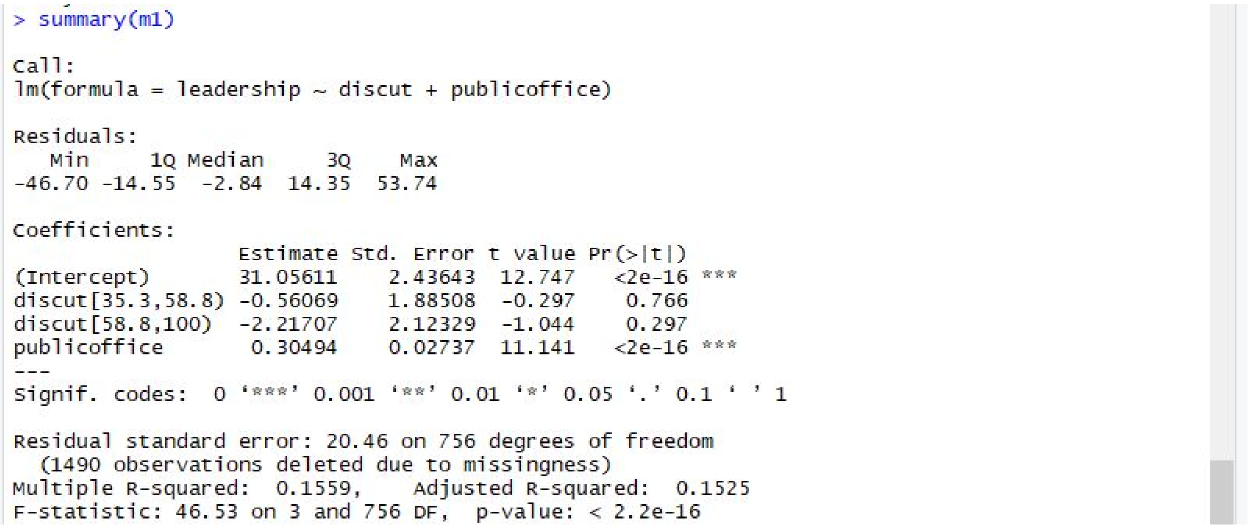
```
>library(car)
```

>residualplots(m1)



- We find that the assumptions are true for Independence and Normality.

summary(m1)



```
> summary(m1)

Call:
lm(formula = leadership ~ discut + publicoffice)

Residuals:
   Min     1Q Median     3Q    Max 
-46.70 -14.55  -2.84  14.35  53.74 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       31.05611    2.43643  12.747   <2e-16 ***
discut[35.3,58.8) -0.56069    1.88508  -0.297    0.766    
discut[58.8,100)  -2.21707    2.12329  -1.044    0.297    
publicoffice       0.30494    0.02737  11.141   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.46 on 756 degrees of freedom
  (1490 observations deleted due to missingness)
Multiple R-squared:  0.1559,	Adjusted R-squared:  0.1525 
F-statistic: 46.53 on 3 and 756 DF,  p-value: < 2.2e-16
```

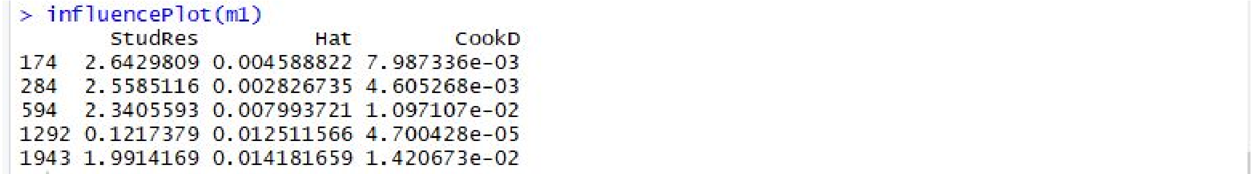
8. Interpret the coefficients of the model within context.
    - For every one unit of increase for “leadership”, ON AVERAGE, “discut[35.3,58.8)” decreases by 0.56069 units (Not significant as p-value is 0.766), “discut[58.8,100)” decreases by 2.21707 units (Not significant as p-value is 0.297), and finally “publicoffice” increases by 0.30494 units (Is significant as p-value < 2e-16).
9. Interpret R-squared within context.
    - The Multiple R-Squared is 15.59%, meaning there is a Variance of the model of 15.59% and fits that much of the model. It is an “okay” R-Squared and acceptable.

10. Check for leverage and influential points. Explain whether there are any points of concern. You will find the following command useful.

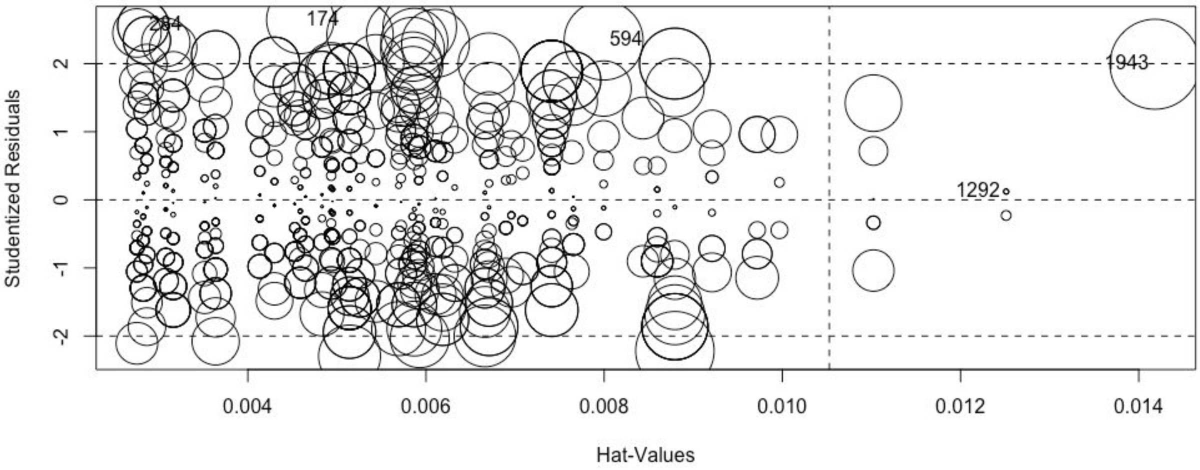
- Using the command function in R known as nrow(...), we see there are 2250 observations. If we consider this a smaller dataset, the observations have standardized residuals larger than +2 and smaller than -2. If we consider this a smaller dataset, the observations have standardized residuals larger than +4 and smaller than -4. If we consider this a smaller dataset, then we say that 174, 284, and 594 are all Bad Leverage values. Meanwhile, 1292 and 1943 are Good Leverage values. R identifies points 174, 284, 594, 1292, and 1943 as points with either high leverage or high standardized residual.

>library(car)

>influenceplots(m1)



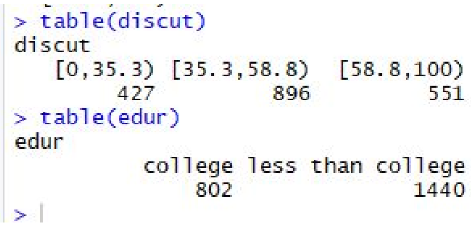
```
> influencePlot(m1)
       StudRes         Hat        CookD
174  2.6429809 0.004588822 7.987336e-03
284  2.5585116 0.002826735 4.605268e-03
594  2.3405593 0.007993721 1.097107e-02
1292 0.1217379 0.012511566 4.700428e-05
1943 1.9914169 0.014181659 1.420673e-02
```



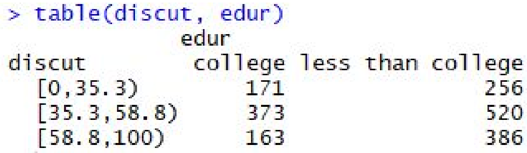
Question two.

Using sex discrimination data set, create a plot for the prediction of leadership (numerical) from discrimination (categorical), level of education (edur), and the interaction between these two factors.

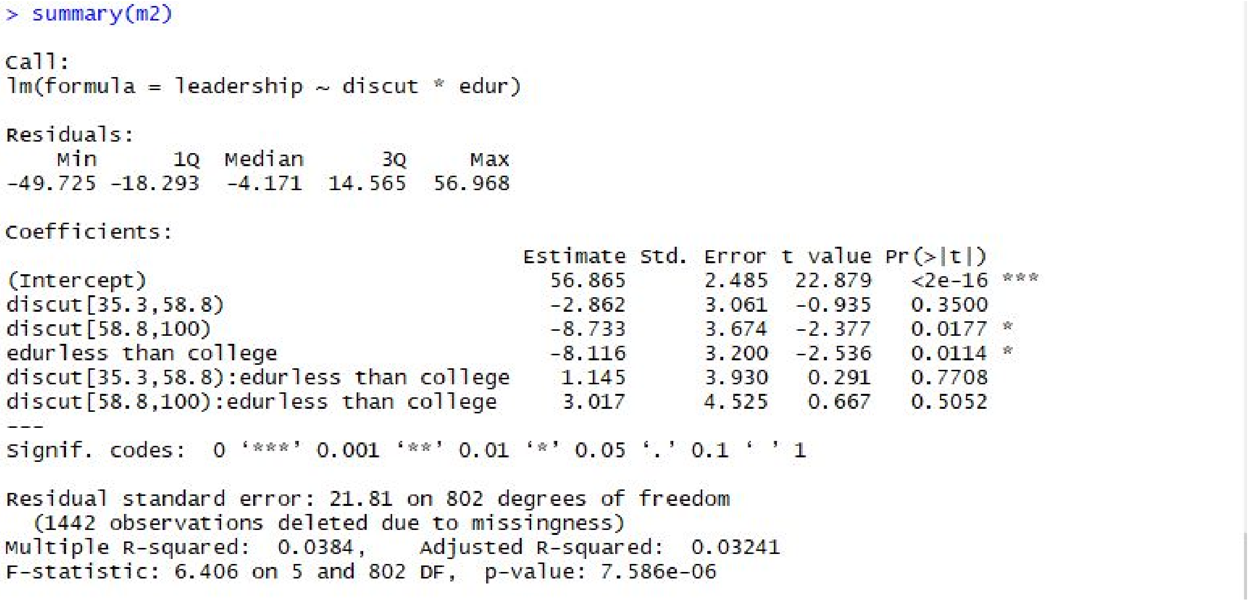
m2<-lm(leadership~discut*edur)



```
> table(discut)
discut
   [0,35.3) [35.3,58.8)  [58.8,100) 
        427         896         551 
> table(edur)
edur
          college less than college 
              802              1440 
> |
```



```
> table(discut, edur)
             edur
discut        college less than college
  [0,35.3)        171               256
  [35.3,58.8)     373               520
  [58.8,100)      163               386
```



```
> summary(m2)

Call:
lm(formula = leadership ~ discut * edur)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.725 -18.293  -4.171  14.565  56.968 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             56.865      2.485  22.879   <2e-16 ***
discut[35.3,58.8)                       -2.862      3.061  -0.935   0.3500    
discut[58.8,100)                        -8.733      3.674  -2.377   0.0177 *  
edurless than college                   -8.116      3.200  -2.536   0.0114 *  
discut[35.3,58.8):edurless than college  1.145      3.930   0.291   0.7708    
discut[58.8,100):edurless than college   3.017      4.525   0.667   0.5052    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.81 on 802 degrees of freedom
  (1442 observations deleted due to missingness)
Multiple R-squared:  0.0384,	Adjusted R-squared:  0.03241 
F-statistic: 6.406 on 5 and 802 DF,  p-value: 7.586e-06
```

1. Interpret the coefficients within context.

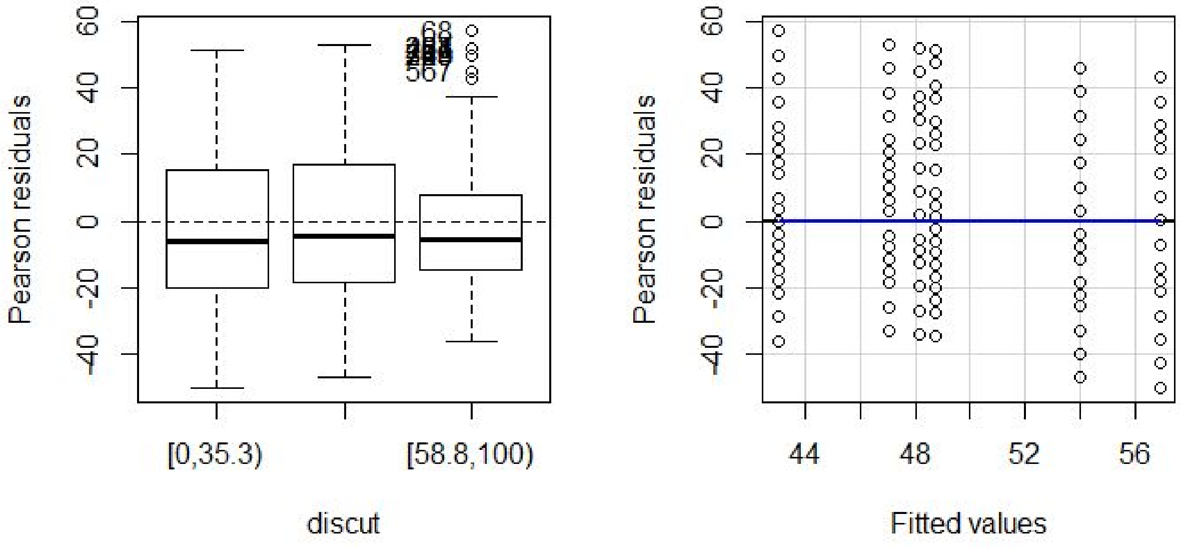
- For discut[35.3,58.8), the base is “[35.3,58.8)”. On average, the respondents who are of discut [35.3,58.8) score 2.862 points lower on their leadership at UCLA. This difference is not statistically significant (p = 0.3500). **Thus, we can say that being discut [35.3,58.8) is not related to leadership at UCLA.**
- For discut [58.8, 100), the base is “[58.8, 100)”. On average, the respondents who are of discut [58.8,100) score 8.733 points lower on their leadership at UCLA. This difference is somewhat statistically significant (p = 0.0177). **Thus, we can say that being discut [58.8,100) is somewhat related to leadership at UCLA.**
- For edurless than college, the base is “less than college”. On average, the respondents who are of edurless than college score 8.116 points lower on their leadership at UCLA. This difference is

somewhat statistically significant (p = 0.0114). **Thus, we can say that being edurless than college is somewhat related to leadership at UCLA.**

2. Interpret R-squared.

- **3.84% of the variance in perception of leadership is explained by edur and discut. (Notice that the insignificant predictors) do not explain any part of the variance in the outcome.**

3. Check the assumptions.



- We can see this is true for Independence and follows Normality.

4. Draw the plot of interaction and interpret it.

Take these steps

· Install the effect package

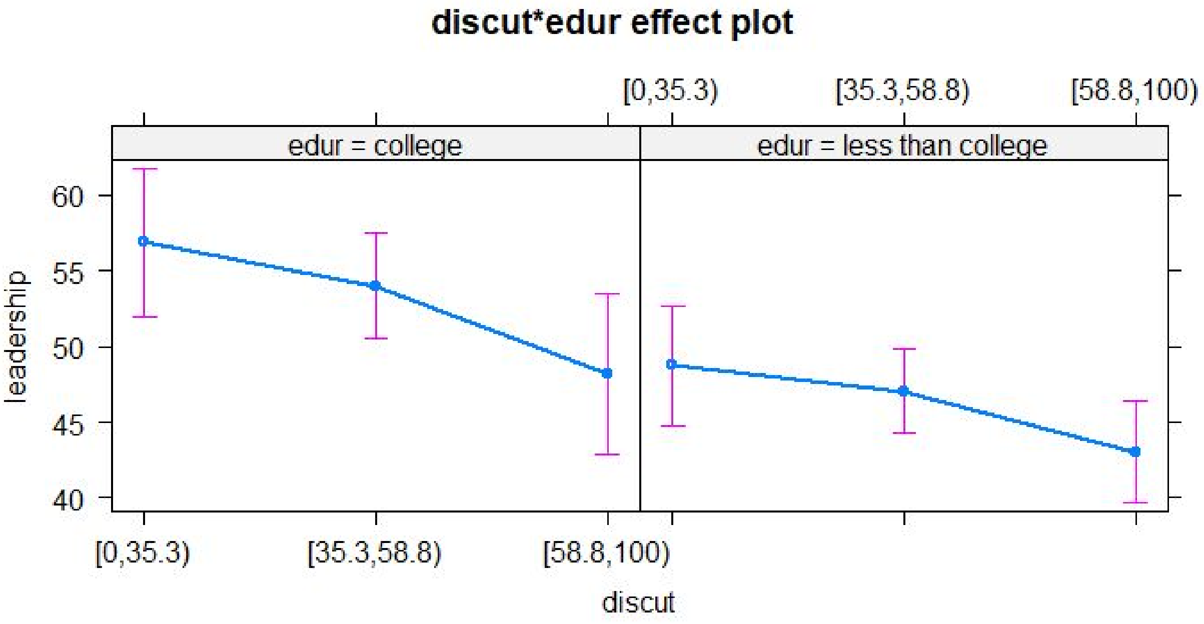
· Tell R that the predictors are factors

· **edur<-factor(edur)**

· **discut<-factor(discut)**

· **library(effects)**

· **plot(allEffects(m2),ask=FALSE)**



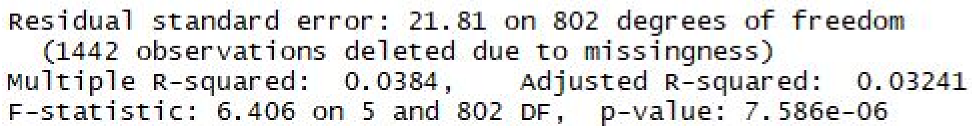
- As it is clear from the above plot, the lines are pretty parallel, showing that the AVERAGE score on leadership is pretty similar for students who have edurless than college and those who do not; regardless of whether they are of discut[35.3,58.8) or discut[58.8,100).

5. Interpret the interaction effect within context.

- The interaction effect for discut [35.3,58.8):edurless than college **AND** discut [58.8,100):edurless than college is not statistically significant (P = 0.7708 **AND** 0.5052, respectively). This means that the effect of edurless than college on students' perception of leadership is similar for students who are of discut[35.3,58.8), discut[58.8,100), and those who are not of these.

6. Show how you can calculate adjusted R-squared.

- $$R^2_{Adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{(N - k - 1)}$$ → nrow(...) = 2250 - 1442 (removed observations due to missingness) = 808 → 1 - [(1-0.0384)*(808-1)]/(808-5-1) = **0.03241**
- Double check using summary(m2)...



```
Residual standard error: 21.81 on 802 degrees of freedom
  (1442 observations deleted due to missingness)
Multiple R-squared:  0.0384,    Adjusted R-squared:  0.03241
F-statistic: 6.406 on 5 and 802 DF,  p-value: 7.586e-06
```

- This is true and Adjusted R-squared is calculated.

Question three. Using satactgpa data posted in the homework four folder in week nine… Create a linear model for the prediction of ACT for SATV, SATM, and gpa.

Check whether there is a multicollinearity problem by using R to calculate variance inflation factor (VIF
R commands

```
m3 <- lm(ACT ~ SATV + SATM + GPA)
library(car)
vif(m3)

   SATV       SATM        GPA
1.773921    1.644818     1.443561
```

Each value does not exceed +5 or -5, therefore, there is not a problematic amount of multicollinearity between the variables.

Based on the guidelines given in the course, discuss whether there is any multicollinearity issue. Show how you can calculate VIF for SATV. Hint you first need to regress SATV on the rest of the predictors in the model. You should find the same answer reported by R.

```
m4 <- lm(SATV ~ ACT + SATM + GPA)
vif(m4)

  ACT        SATM        GPA
2.570421  2.282987    1.446963
```

Question four. Using campus climate data…

1) Table data for overall comfort

```
>table<-(overallcomfort)
comfortable       somewhat uncomfortable  very comfortable        very uncomfortable
 2930                687          220                      1494                      48
```

2) Pool very uncomfortable with uncomfortable by using the following R code

```
>overallcomfort<-recode(overallcomfort,"'very uncomfortable'='uncomfortable'")
> table(overallcomfort)

overallcomfort

    comfortable        somewhat    uncomfortable   very comfortable
         2930               687            268               1494
```

3) Now that you have made sure that the number of frequencies in each cell is OK, create a contingency table for overall comfort in our climate and being first generation vs. not being first generation.

```
Table(overallcomfort,leaveucla)

                 leaveUCLA

overallcomfort             no  yes

  comfortable           2530  397

  somewhat               447  240

  uncomfortable          121  147

  very comfortable     1389  103
```

4) Once you have created this table, **let comfortable be the base**, and calculate the following odds…

a) Odds of being **very comfortable compared to comfortable** for those who plan to leave UCLA.

103/397=0.25945

b) The odds of being **somewhat comfortable** to comfortable for those who plan to leave UCLA.

240/397=0.60453

c) The odds of being **uncomfortable compared to comfortable** for those who plan to UCLA.

147/397=0.37028

d) The odds ratio of being very comfortable compared to comfortable for those who plan to leave UCLA compared to those who do not.

(103/397)/(1389/2530)=0.47257

e) The odds ratio of being somewhat comfortable compared to comfortable for those who plan to leave UCLA compared to those who do not.

(240/397)/(447/2530)=3.42164

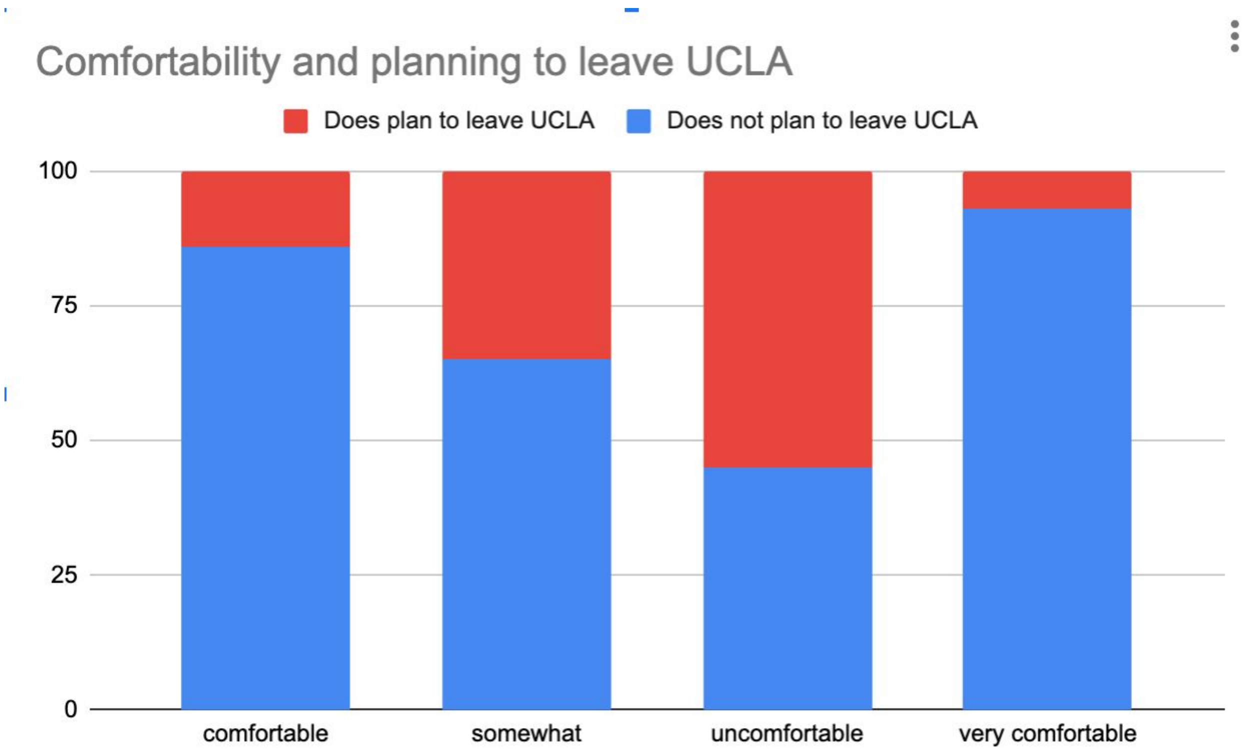
f) The odds ratio of being uncomfortable compared to comfortable for those who plan to leave UCLA compared to those who do not

(147/397)/(121/2530)=7.74216

5) Using R calculate row percentage.

```
> mytable<-table(leaveUCLA,overallcomfort)

> prop.table(mytable,1)

leaveUCLA   comfortable    somewhat      uncomfortable   very uncomfortable

    no       0.56385113    0.09962113     0.02696679       0.30956095

    yes      0.44757610    0.27057497     0.16572717       0.11612176
```

6) Use Excel to make a segmented bar chart based on these row percentages. (directions are given in the odds ratio lecture on week eight of CCLE)



7) Interpret the segmented bar chart that you created within context.

- As the level of comfort increases, the percentage of students planning to leave UCLA tend to decreases.