Fall 2019 – Homework Four
Stat 112 – Professor Esfandiari
Due Tuesday December 3rd by eleven PM

Tier one: Complete questions one to four
Tier two: Complete questions one to five
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)

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.

model<-lm(leadership~discut+publicoffice)

7. check the relevant assumptions, scatterplots, etc. you will find the following command useful:
>library(car)
>residualplots(m1)

8. Interpret the coefficients of the model within context.
9. Interpret R-squared within context.
10. Check for leverage and influential points. Explain whether there are any points of concern. You will find the following command useful.

>library(car)
>influenceplots(m1)

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)

1. Interpret the coefficients within context.
2. Interpret R-squared.
3. Check the assumptions.
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)**

5. Interpret the interaction effect within context.
6. Show how you can calculate adjusted R-squared.

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)
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.

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)

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.
   b) The odds of being **somewhat comfortable** to comfortable for those who plan to leave UCLA.
   c) The odds of being **uncomfortable compared to comfortable** for those who plan to UCLA.
   d) The odds ratio of being very comfortable compared to comfortable for those who plan to leave UCLA compared to those who do not.
   e) The odds ratio of being somewhat comfortable compared to comfortable for those who plan to leave UCLA compared to those who do not.
   f) The odds ratio of being uncomfortable compared to comfortable for those who plan to leave UCLA compared to those who do not

5)   Using R calculate row percentage.

> mytable<-table(leaveUCLA,overallcomfort)
> prop.table(mytable,1)

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.

Question five: Tier two only

1) Create a logistic regression model for the prediction of being first generation, as a function of climate comfort (overallcomfort). Use the recoded climate comfort that has four levels including very comfortable, comfortable, uncomfortable, and very uncomfortable. Hint: If the glm function did not work recode firstGEN to

Recode firstGEN to zero and one.

```
> firstgencat<-recode(FIRSTGEN,"'no'='0';'yes'='1'")
> table(firstgencat)
firstgencat
  0 (no)      1 (yes)
3678         1689
```

2) Exponentiate the coefficients and interpret the resulting odds within context.
3) State the null hypothesis and the alternative hypotheses in terms of symbols.
4) Calculate R-squared as an estimate of the strength of the model.
5) Create a new model for showing the relationship between being first generation and perception of exclusion (exclusionaryp) on our campus.
6) Now create a model with being first generation as outcome and interaction between perception of exclusion on our campus and overall comfort as predictor.
7) Draw the plot of interaction effects using the "effects" package. For directions on how to do this, see the lecture on interaction effect in regression (week eight). The only difference is that your outcome is binary.
8) Interpret this plot within context.