W271 - Assignment2

Chandra Shekar Bikkanur

10/9/2019

```
#Loading some libraries for this assignment
library(car)
library(dplyr)
library(Hmisc)
library(ggplot2)
library(mcprofile)
library(nnet)
library(MASS)
library(GGally)
library("ggpubr")
```

1. Strategic Placement of Products in Grocery Stores

Let us load the data into a data frame and do the initial EDA.

```
cereal <- read.csv("cereal_dillons.csv", header=TRUE, sep=",")
head(cereal, 5)
```

```
##   ID Shelf                                Cereal size_g sugar_g fat_g
## 1  1     1 Kellog's Razzle Dazzle Rice Crispies     28      10     0
## 2  2     1            Post Toasties Corn Flakes     28       2     0
## 3  3     1                Kellogg's Corn Flakes     28       2     0
## 4  4     1               Food Club Toasted Oats     32       2     2
## 5  5     1                     Frosted Cheerios     30      13     1
##   sodium_mg
## 1       170
## 2       270
## 3       300
## 4       280
## 5       210
```

```
str(cereal)
```

```
## 'data.frame':    40 obs. of  7 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Shelf    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Cereal   : Factor w/ 38 levels "Basic 4","Capn Crunch",..: 17 34 19 13 16 9 2 3 30 8 ...
##  $ size_g   : int  28 28 28 32 30 31 27 27 29 33 ...
##  $ sugar_g  : int  10 2 2 2 13 11 12 9 11 2 ...
##  $ fat_g    : num  0 0 0 2 1 0 1.5 2.5 0.5 0 ...
##  $ sodium_mg: int  170 270 300 280 210 180 200 200 220 330 ...
```

```r
summary(cereal)
```

```
##        ID            Shelf                                        Cereal
##  Min.   : 1.00   Min.   :1.00   Capn Crunch's Peanut Butter Crunch: 2
##  1st Qu.:10.75   1st Qu.:1.75   Food Club Toasted Oats            : 2
##  Median :20.50   Median :2.50   Basic 4                           : 1
##  Mean   :20.50   Mean   :2.50   Capn Crunch                       : 1
##  3rd Qu.:30.25   3rd Qu.:3.25   Cinnamon Grahams                  : 1
##  Max.   :40.00   Max.   :4.00   Cocoa Pebbles                     : 1
##                                 (Other)                           :32
##      size_g         sugar_g          fat_g          sodium_mg
##  Min.   :27.00   Min.   : 0.0   Min.   :0.000   Min.   :  0.0
##  1st Qu.:29.75   1st Qu.: 6.0   1st Qu.:0.500   1st Qu.:157.5
##  Median :31.00   Median :11.0   Median :1.000   Median :200.0
##  Mean   :37.20   Mean   :10.4   Mean   :1.200   Mean   :195.5
##  3rd Qu.:51.00   3rd Qu.:14.0   3rd Qu.:1.625   3rd Qu.:262.5
##  Max.   :60.00   Max.   :20.0   Max.   :5.000   Max.   :330.0
##
```

We see that *Shelf* is of integer type. This should be changed to a *factor* data type to do any regressions on the data.

1.1 (1 point): The explanatory variables need to be reformatted before proceeding further (sample code is provided in the textbook). First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Second, rescale each variable to be within 0 and 1. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables. Also, construct a parallel coordinates plot for the explanatory variables and the shelf number. Discuss whether possible content differences exist among the shelves.

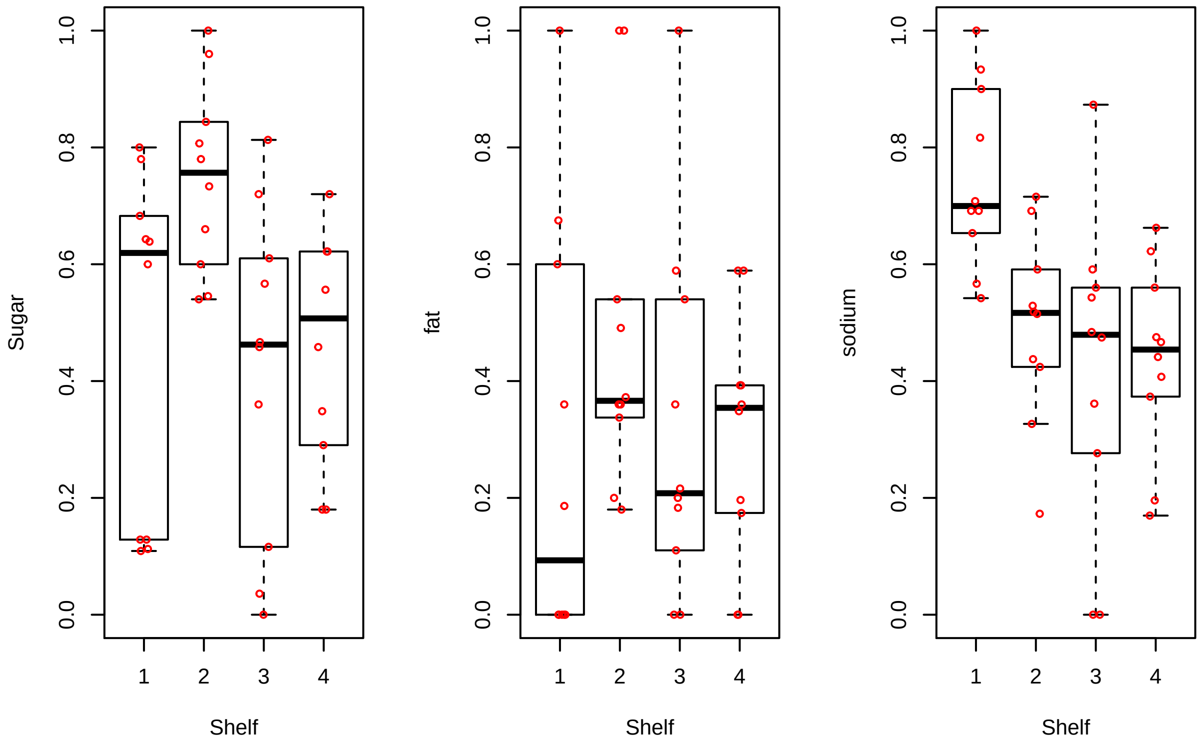
```r
stand01 <- function(x) { (x - min(x))/(max(x) - min(x)) } # function to standardize a dataset
cereal2 <- data.frame(Shelf = cereal$Shelf,
                      sugar = stand01(x = cereal$sugar_g/cereal$size_g),
                      fat = stand01(x = cereal$fat_g/cereal$size_g),
                      sodium = stand01(x = cereal$sodium_mg/cereal$size_g))
                      # new data frame consisting of Shelf, sugar, fat and sodium
str(cereal2)
```

```
## 'data.frame':    40 obs. of  4 variables:
##  $ Shelf : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sugar : num  0.643 0.129 0.129 0.112 0.78 ...
##  $ fat   : num  0 0 0 0.675 0.36 ...
##  $ sodium: num  0.567 0.9 1 0.817 0.653 ...
```

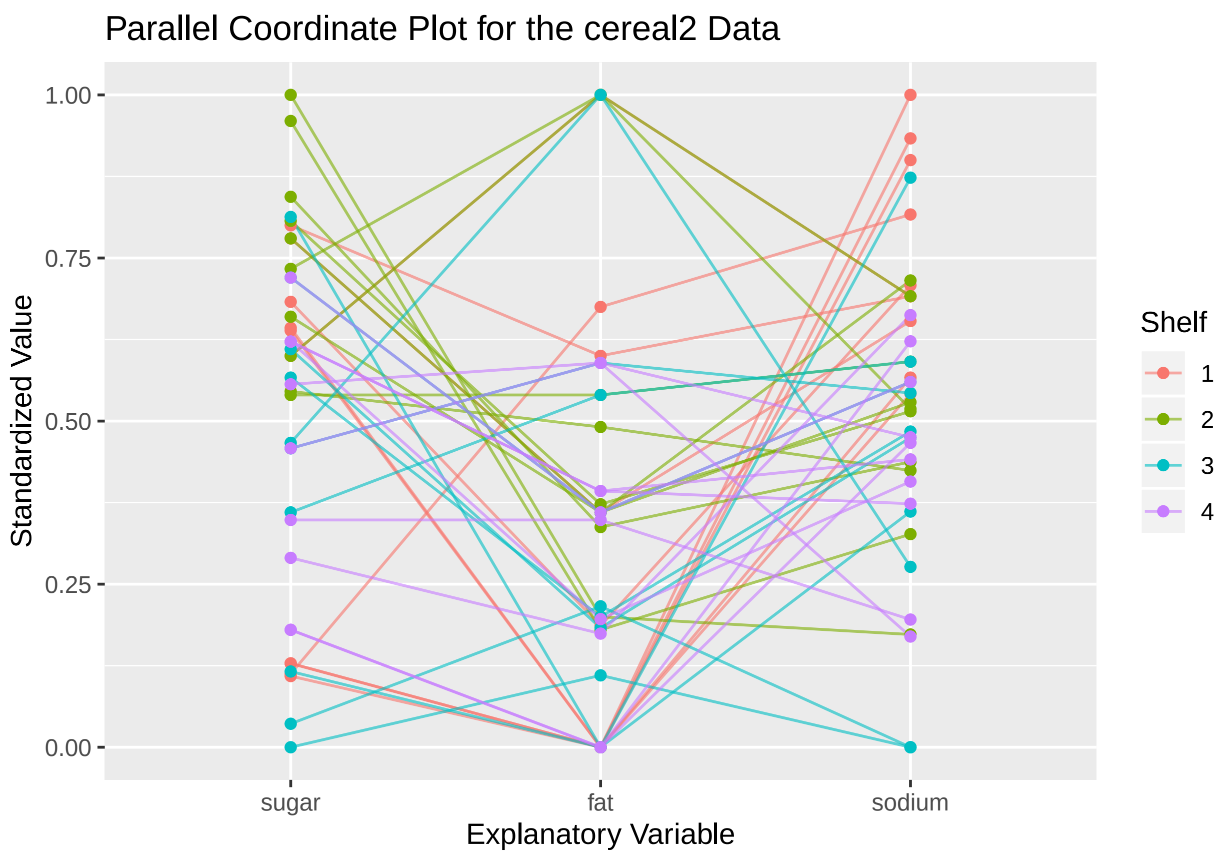
```r
tail(cereal2, 5)
```

```
##    Shelf     sugar       fat    sodium
## 36     4 0.3483871 0.3483871 0.1956989
## 37     4 0.4581818 0.5890909 0.1696970
## 38     4 0.6218182 0.3927273 0.4412121
## 39     4 0.5563636 0.5890909 0.4751515
## 40     4 0.1800000 0.0000000 0.6222222
```

```r
par(mfrow=c(1,3))
boxplot(formula = sugar ~ Shelf, data = cereal2,
        ylab = "Sugar", xlab = "Shelf", pars = list(outpch=NA))
stripchart(x = cereal2$sugar ~cereal2$Shelf, lwd = 1,
           col = "red", method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
boxplot(formula = fat ~ Shelf, data = cereal2,
        ylab = "fat", xlab = "Shelf", pars = list(outpch=NA))
stripchart(x = cereal2$fat ~cereal2$Shelf, lwd = 1,
           col = "red", method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
boxplot(formula = sodium ~ Shelf, data = cereal2,
        ylab = "sodium", xlab = "Shelf", pars = list(outpch=NA))
stripchart(x = cereal2$sodium ~cereal2$Shelf, lwd = 1,
           col = "red", method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```




```r
ggparcoord(cereal2, columns = 2:4, groupColumn = 'Shelf', scale = 'globalminmax',
           showPoints = TRUE, title = "Parallel Coordinate Plot for the cereal2 Data",
           alphaLines = 0.6, mapping=aes(color=as.factor(Shelf))) +
            xlab("Explanatory Variable") + ylab("Standardized Value") +
           scale_color_discrete("Shelf")
```



From above 2 plots, we can see that the *shelf 2* has cereals with relatively more *sugar*, *fat* and *sodium* contents.

1.2 (1 point): The response has values of 1, 2, 3, and 4. Explain under what setting would it be desirable to take into account ordinality, and whether you think that this setting occurs here. Then estimate a suitable multinomial regression model with linear forms of the sugar, fat, and sodium variables. Perform LRTs to examine the importance of each explanatory variable. Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables).

If placing cereals in shelf 4 (top shelf) is more conducive in the likelihood of purchasing cereals from shelf 4 than shelf 1 (bottom), then we can take ordinality into account. Here, we do not have such information about the shelves. So, we can try to fit a nominal response regression model for the *Shelf* with *fat*, *sugar* and *sodium* as explanatory variables.

```
cereal2$Shelf <- as.factor(cereal2$Shelf) # convert int data_type of Shelf to factor
mod.fit.nom <- multinom(Shelf~sugar + fat + sodium, data = cereal2)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 37.329384
## iter  20 value 33.775257
## iter  30 value 33.608495
## iter  40 value 33.596631
## iter  50 value 33.595909
## iter  60 value 33.595564
## iter  70 value 33.595277
## iter  80 value 33.595147
```

```
## final  value 33.595139
## converged
```

```r
summary(mod.fit.nom)
```

```
## Call:
## multinom(formula = Shelf ~ sugar + fat + sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar        fat    sodium
## 2    6.900708   2.693071  4.0647092 -17.49373
## 3   21.680680 -12.216442 -0.5571273 -24.97850
## 4   21.288343 -11.393710 -0.8701180 -24.67385
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 2    6.487408 5.051689 2.307250 7.097098
## 3    7.450885 4.887954 2.414963 8.080261
## 4    7.435125 4.871338 2.405710 8.062295
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

```r
Anova(mod.fit.nom) # performing LR Test
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##        LR Chisq Df Pr(>Chisq)
## sugar   22.7648  3  4.521e-05 ***
## fat      5.2836  3     0.1522
## sodium  26.6197  3  7.073e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above LR test for the *mod.fit.nom* model, we can see that *sugar* and *sodium* have significance in determining the *Shelf* the cereals could be placed in. Let us now check if these explanatory variables have any interactions among them

```r
mod.fit.nom2 <- multinom(Shelf~sugar + fat + sodium + sugar*fat +
                           sugar*sodium + fat*sodium + sugar*fat*sodium,
                         data = cereal2)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 36.170336
## iter  20 value 31.166546
## iter  30 value 29.963705
## iter  40 value 28.414027
## iter  50 value 27.891712
## iter  60 value 27.763967
## iter  70 value 27.622579
```

```
## iter  80 value 27.438263
## iter  90 value 27.015534
## iter 100 value 26.772481
## final  value 26.772481
## stopped after 100 iterations
```

```r
summary(mod.fit.nom2)
```

```
## Call:
## multinom(formula = Shelf ~ sugar + fat + sodium + sugar * fat +
##     sugar * sodium + fat * sodium + sugar * fat * sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar       fat     sodium sugar:fat sugar:sodium
## 2   -4.563627   8.944868 22.063003   1.030077  35.60873   -12.250084
## 3   24.498320 -22.248456 35.981865 -27.899087 -17.12487    13.253103
## 4   27.246742 -21.852777  7.298799 -29.106797  41.08251     2.887805
##   fat:sodium sugar:fat:sodium
## 2  -23.75955        -55.88455
## 3  -59.54150         37.71571
## 4  -30.85250        -22.59552
##
## Std. Errors:
##   (Intercept)    sugar       fat   sodium sugar:fat sugar:sodium
## 2    25.21113 29.72894  96.57821 27.29915  135.1117     31.98647
## 3    22.83750 25.81043 101.17670 24.61166  150.1228     26.89827
## 4    22.80359 26.00692 100.83444 24.51538  150.6750     28.86631
##   fat:sodium sugar:fat:sodium
## 2   116.0776         158.8091
## 3   138.0237         212.2222
## 4   138.5448         217.3953
##
## Residual Deviance: 53.54496
## AIC: 101.545
```

```r
Anova(mod.fit.nom2) # performing LR Test
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                  LR Chisq Df Pr(>Chisq)
## sugar             19.2525  3  0.0002424 ***
## fat                6.1167  3  0.1060686
## sodium            30.8407  3  9.183e-07 ***
## sugar:fat          3.2309  3  0.3573733
## sugar:sodium       3.0185  3  0.3887844
## fat:sodium         3.1586  3  0.3678151
## sugar:fat:sodium   2.5884  3  0.4595299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above LR test for *mod.fit.nom2* model, we can see that there are no explanatory variables' interactions in determining the *Shelf*.

1.3 (1 point): Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed toward children. For a serving size of 28 grams, its sugar content is 12 grams, fat content is 0.5 grams, and sodium content is 130 milligrams. Estimate the shelf probabilities for Apple Jacks.

```
stand01_2 <- function(x, min , max) { (x - min)/(max - min) } # function to standardize a dataset
Kellogs_data <- data.frame(sugar = stand01_2(x = 12/28, min = min(cereal$sugar_g/cereal$size_g),
                           max  =    max(cereal$sugar_g/cereal$size_g)),
                           fat = stand01_2(x = 0.5/28, min = min(cereal$fat_g/cereal$size_g),
                           max =  max(cereal$fat_g/cereal$size_g)),
                           sodium = stand01_2(x = 130/28, min = min(cereal$sodium_mg/cereal$size_g),
                           max =  max(cereal$sodium_mg/cereal$size_g)))
str(Kellogs_data)
```

```
## 'data.frame':    1 obs. of  3 variables:
##  $ sugar : num 0.771
##  $ fat   : num 0.193
##  $ sodium: num 0.433
```

```
pi.hat <- predict(object = mod.fit.nom, newdata = Kellogs_data, type = "probs")
pi.hat
```

```
##          1          2          3          4
## 0.05326849 0.47194264 0.20042742 0.27436145
```

1.4 (1 point): Construct a plot similar to Figure 3.3 where the estimated probability for a shelf is on the *y-axis* and the sugar content is on the *x-axis*. Use the mean overall fat and sodium content as the corresponding variable values in the model. Interpret the plot with respect to sugar content.

```
sugar <- seq(0,1, 0.01)
mean_fat <- mean(cereal2$fat)
mean_sodium <- mean(cereal2$sodium)
new_data = data.frame(sugar=sugar, fat=mean_fat,sodium=mean_sodium)
head(new_data)
```

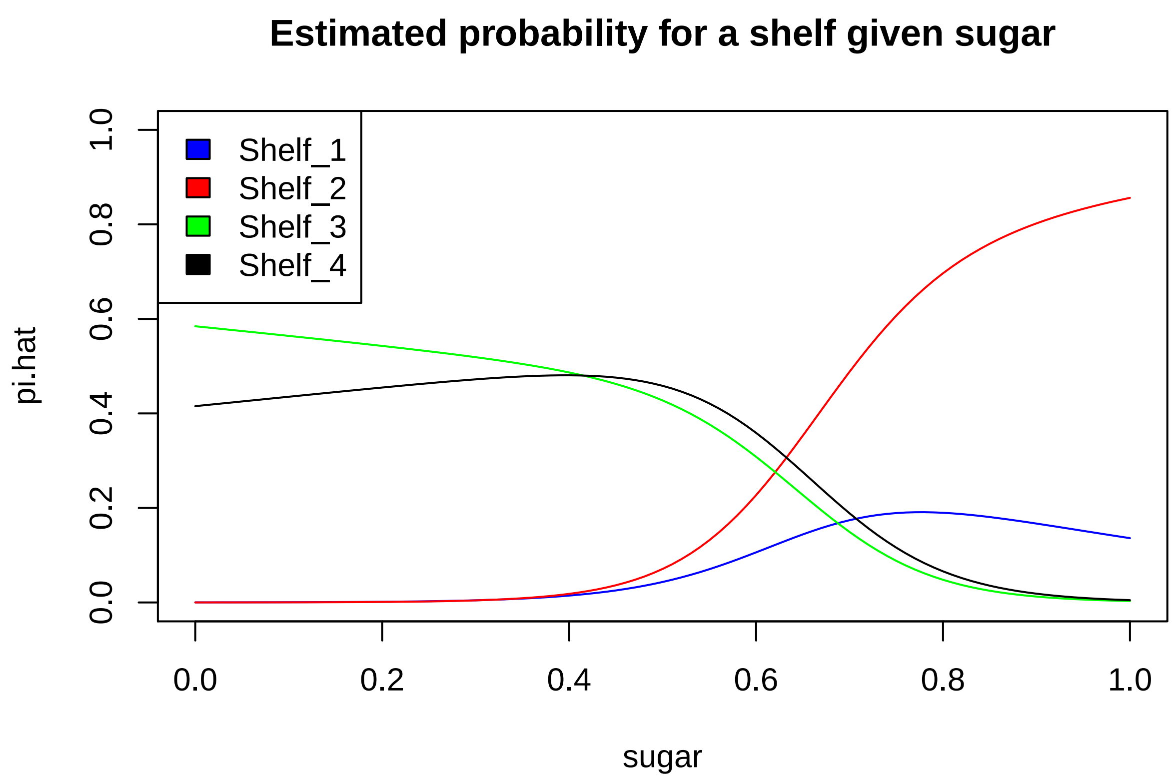
```
##   sugar       fat   sodium
## 1  0.00 0.3475739 0.524039
## 2  0.01 0.3475739 0.524039
## 3  0.02 0.3475739 0.524039
## 4  0.03 0.3475739 0.524039
## 5  0.04 0.3475739 0.524039
## 6  0.05 0.3475739 0.524039
```

```
pi.hat <- predict(object = mod.fit.nom, newdata = new_data, type = "probs")
pi.hat_df <- as.data.frame(pi.hat)
pi.hat_df$sugar <- sugar
names(pi.hat_df) <- c("Shelf_1", "Shelf_2", "Shelf_3", "Shelf_4", "Sugar")
head(pi.hat_df)
```

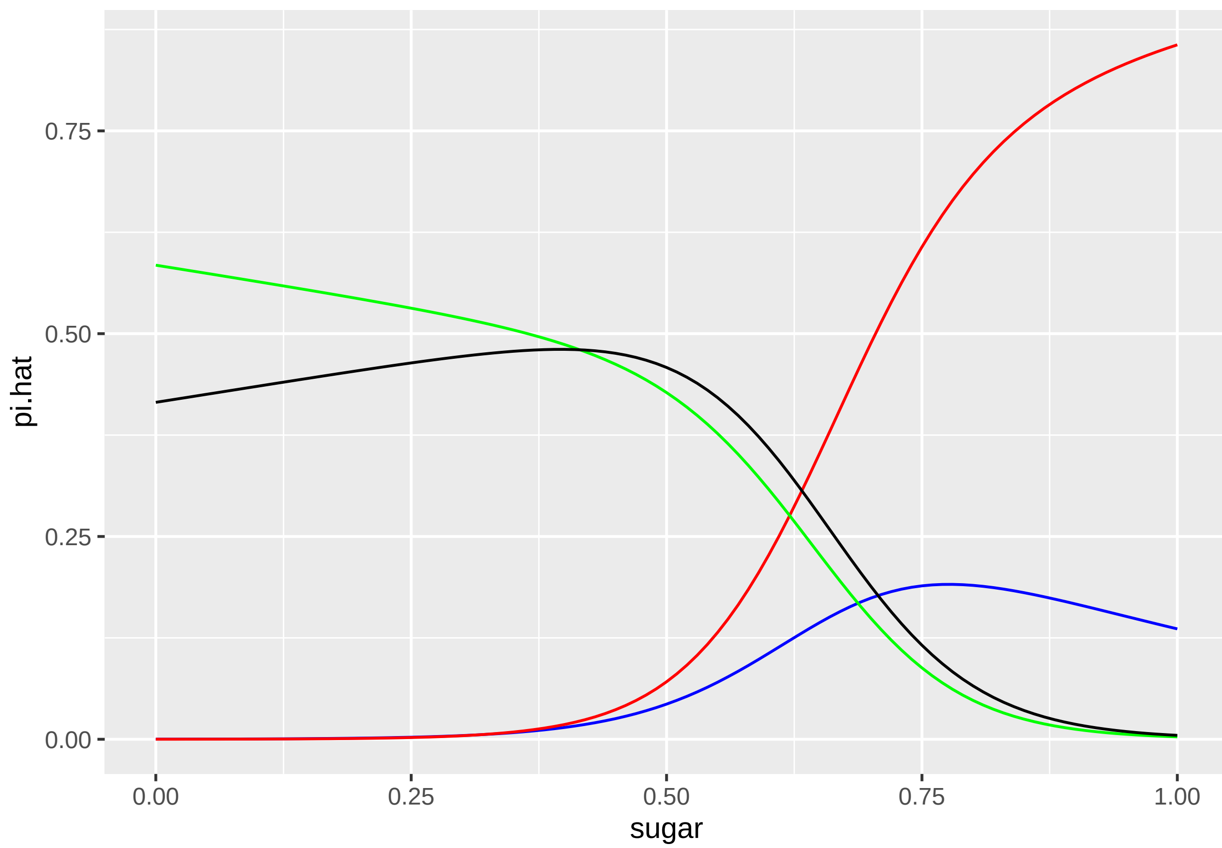
```
##        Shelf_1      Shelf_2   Shelf_3   Shelf_4 Sugar
## 1 0.0001317725 5.610226e-05 0.5844393 0.4153729  0.00
## 2 0.0001483820 6.489817e-05 0.5824254 0.4173613  0.01
```

```
## 3 0.0001670817 7.507165e-05 0.5804070 0.4193508  0.02
## 4 0.0001881341 8.683815e-05 0.5783837 0.4213413  0.03
## 5 0.0002118348 1.004468e-04 0.5763554 0.4233323  0.04
## 6 0.0002385160 1.161856e-04 0.5743218 0.4253235  0.05
```

```
plot(pi.hat_df$Sugar,pi.hat_df$Shelf_1,type="l",col="blue",
     main="Estimated probability for a shelf given sugar",
     ylab="pi.hat", xlab = "sugar", ylim = c(0,1))
lines(pi.hat_df$Sugar,pi.hat_df$Shelf_2, col="red")
lines(pi.hat_df$Sugar,pi.hat_df$Shelf_3, col="green")
lines(pi.hat_df$Sugar,pi.hat_df$Shelf_4, col="black")
legend("topleft",
c("Shelf_1","Shelf_2","Shelf_3","Shelf_4"),
fill=c("blue","red","green","black")
)
```




```
g <- ggplot(pi.hat_df, aes(Sugar))
g <- g + geom_line(aes(y=Shelf_1), colour="blue")
g <- g + geom_line(aes(y=Shelf_2), colour="red")
g <- g + geom_line(aes(y=Shelf_3), colour="green")
g <- g + geom_line(aes(y=Shelf_4), colour="black")
g <- g + ylab("pi.hat") + xlab("sugar")
g
```



1.5 (1 point): Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretations back to the plots constructed for this exercise.

```
c.value <- apply(X = cereal2[c(2:4)], MARGIN = 2, FUN = sd)
beta.hat2 <- coefficients(mod.fit.nom)[1, 2:4]
beta.hat3 <- coefficients(mod.fit.nom)[2, 2:4]
beta.hat4 <- coefficients(mod.fit.nom)[3, 2:4]
```

```
print("-- Odds Ratio (Shelf2 vs. Shelf1) for increase in explanatory variables by one sd --")
```

```
## [1] "-- Odds Ratio (Shelf2 vs. Shelf1) for increase in explanatory variables by one sd --"
```

```
round(exp(c.value*beta.hat2), 2)
```

```
##  sugar    fat sodium
##   2.06   3.37   0.02
```

```
print("--Odds Ratio (Shelf2 vs. Shelf1) for decrease in explanatory variables by one sd --")
```

```
## [1] "--Odds Ratio (Shelf2 vs. Shelf1) for decrease in explanatory variables by one sd --"
```

```
round(1/exp(c.value*beta.hat2), 2)
```

```
##  sugar    fat sodium
##   0.48   0.30  55.74
```

```r
print("--Odds Ratio (Shelf3 vs. Shelf1) for increase in explanatory variables by one sd --")
```

```
## [1] "--Odds Ratio (Shelf3 vs. Shelf1) for increase in explanatory variables by one sd --"
```

```r
round(exp(c.value*beta.hat3), 2)
```

```
##  sugar    fat sodium
##   0.04   0.85   0.00
```

```r
print("--Odds Ratio (Shelf3 vs. Shelf1) for decrease in explanatory variables by one sd --")
```

```
## [1] "--Odds Ratio (Shelf3 vs. Shelf1) for decrease in explanatory variables by one sd --"
```

```r
round(1/exp(c.value*beta.hat3), 2)
```

```
##  sugar    fat sodium
##  26.81   1.18 311.36
```

```r
print("--Odds Ratio (Shelf4 vs. Shelf1) for increase in explanatory variables by one sd --")
```

```
## [1] "--Odds Ratio (Shelf4 vs. Shelf1) for increase in explanatory variables by one sd --"
```

```r
round(exp(c.value*beta.hat4), 2)
```

```
##  sugar    fat sodium
##   0.05   0.77   0.00
```

```r
print("--Odds Ratio (Shelf4 vs. Shelf1) for decrease in explanatory variables by one sd --")
```

```
## [1] "--Odds Ratio (Shelf4 vs. Shelf1) for decrease in explanatory variables by one sd --"
```

```r
round(1/exp(c.value*beta.hat4), 2)
```

```
##  sugar    fat sodium
##  21.48   1.30 290.31
```

Effect of sugar:

From above Odds Ratios for shelf_2, shelf_3, shelf_4 against shelf_1 for one standard deviation change in sugar, we can see that, as the sugar of the cereal increases, it is more likely (2.06 times) to be in Shelf_2; the same is already proven from boxplot. Also,if the sugar decreases, the cereal could end up in shelf_3 (26.81 times likely) or shelf_4 (21.48 times likely); we can inspect the initial boxplot for the same.

Effect of fat:

From above Odds Ratios for shelf_2, shelf_3, shelf_4 against shelf_1 for one standard deviation change in fat, we can see that, as the fat of the cereal increases, it is more likely (3.37 times) to be in Shelf_2; the same is already proven from boxplot. Also,if the fat decreases, the cereal could end up in shelf_3 (1.18 times likely) or shelf_4 (1.30); we can inspect the initial boxplot for the same.

Effect of sodium:

From above Odds Ratios for shelf_2, shelf_3, shelf_4 against shelf_1 for one standard deviation change in sodium, we can see that, as the sugar of the cereal increases, it is more likely to be in Shelf_1; as high sodium cereals are in shelf_1 which is already proven from boxplot. Also,if the sodium decreases, the cereal could end up in shelf_3 (311.36 times likely) or shelf_4 (290.31 times likely); we can inspect the initial boxplot for the same.

2. Alcohol, self-esteem and negative relationship interactions

Read the example **'Alcohol Consumption'** in chapter 4.2.2 of the textbook. This is based on a study in which moderate-to-heavy drinkers (defined as at least 12 alcoholic drinks/week for women, 15 for men) were recruited to keep a daily record of each drink that they consumed over a 30-day study period. Participants also completed a variety of rating scales covering daily events in their lives and items related to self-esteem. The data are given in the *DeHartSimplified.csv* data set. Questions 24-26 of chapter 3 of the textbook also relate to this data set and give more explanation of its variables.

The researchers stated the following hypothesis:

We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem.

2.1 (2 points): Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. You will use this to guide the model specification in the following questions.

```
dehart <- read.csv("DeHartSimplified.csv", header=TRUE, sep=",", na.strings = " ") # Load the data
head(dehart, 5)
```

```
##   id studyday dayweek numall     nrel      prel  negevent  posevent gender
## 1  1        1       6      9 1.000000 0.0000000 0.4000000 0.5250000      2
## 2  1        2       7      1 0.000000 0.0000000 0.2500000 0.7000000      2
## 3  1        3       1      1 1.000000 0.0000000 0.2666667 1.0000000      2
## 4  1        4       2      2 0.000000 1.0000000 0.5333333 0.6083333      2
## 5  1        5       3      2 1.333333 0.3333333 0.6633333 0.6933333      2
##   rosn      age  desired    state
## 1  3.3 39.48528 5.666667 4.000000
## 2  3.3 39.48528 2.000000 2.777778
## 3  3.3 39.48528 3.000000 4.222222
## 4  3.3 39.48528 3.666667 4.111111
## 5  3.3 39.48528 3.000000 4.222222
```

```
str(dehart)
```

```
## 'data.frame':    623 obs. of  13 variables:
##  $ id      : int  1 1 1 1 1 1 1 2 2 2 ...
##  $ studyday: int  1 2 3 4 5 6 7 1 2 3 ...
##  $ dayweek : int  6 7 1 2 3 4 5 3 4 5 ...
##  $ numall  : int  9 1 1 2 2 1 4 3 4 0 ...
##  $ nrel    : num  1 0 1 0 1.33 ...
##  $ prel    : num  0 0 0 1 0.333 ...
##  $ negevent: num  0.4 0.25 0.267 0.533 0.663 ...
##  $ posevent: num  0.525 0.7 1 0.608 0.693 ...
##  $ gender  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ rosn    : num  3.3 3.3 3.3 3.3 3.3 3.3 3.3 3.9 3.9 3.9 ...
##  $ age     : num  39.5 39.5 39.5 39.5 39.5 ...
##  $ desired : num  5.67 2 3 3.67 3 ...
##  $ state   : num  4 2.78 4.22 4.11 4.22 ...
```

```r
summary(dehart)
```

```
##        id            studyday     dayweek       numall           nrel      
##  Min.   :  1.00   Min.   :1   Min.   :1   Min.   : 0.000   Min.   :0.000  
##  1st Qu.: 33.00   1st Qu.:2   1st Qu.:2   1st Qu.: 1.000   1st Qu.:0.000  
##  Median : 60.00   Median :4   Median :4   Median : 2.000   Median :0.000  
##  Mean   : 75.89   Mean   :4   Mean   :4   Mean   : 2.524   Mean   :0.359  
##  3rd Qu.:123.00   3rd Qu.:6   3rd Qu.:6   3rd Qu.: 3.750   3rd Qu.:0.000  
##  Max.   :160.00   Max.   :7   Max.   :7   Max.   :21.000   Max.   :9.000  
##                                           NA's   :1                       
##       prel           negevent         posevent         gender     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :1.000  
##  1st Qu.:0.4167   1st Qu.:0.1583   1st Qu.:0.600   1st Qu.:1.000  
##  Median :2.0000   Median :0.3500   Median :0.950   Median :2.000  
##  Mean   :2.5830   Mean   :0.4414   Mean   :1.048   Mean   :1.562  
##  3rd Qu.:4.0000   3rd Qu.:0.6292   3rd Qu.:1.378   3rd Qu.:2.000  
##  Max.   :9.0000   Max.   :2.3767   Max.   :3.883   Max.   :2.000  
##                                                                   
##       rosn            age           desired          state      
##  Min.   :2.100   Min.   :24.43   Min.   :1.000   Min.   :2.333  
##  1st Qu.:3.200   1st Qu.:30.53   1st Qu.:3.333   1st Qu.:3.667  
##  Median :3.500   Median :34.57   Median :4.667   Median :4.000  
##  Mean   :3.436   Mean   :34.29   Mean   :4.465   Mean   :3.966  
##  3rd Qu.:3.800   3rd Qu.:38.19   3rd Qu.:5.667   3rd Qu.:4.222  
##  Max.   :4.000   Max.   :42.28   Max.   :8.000   Max.   :5.000  
##                                  NA's   :3       NA's   :3      
```

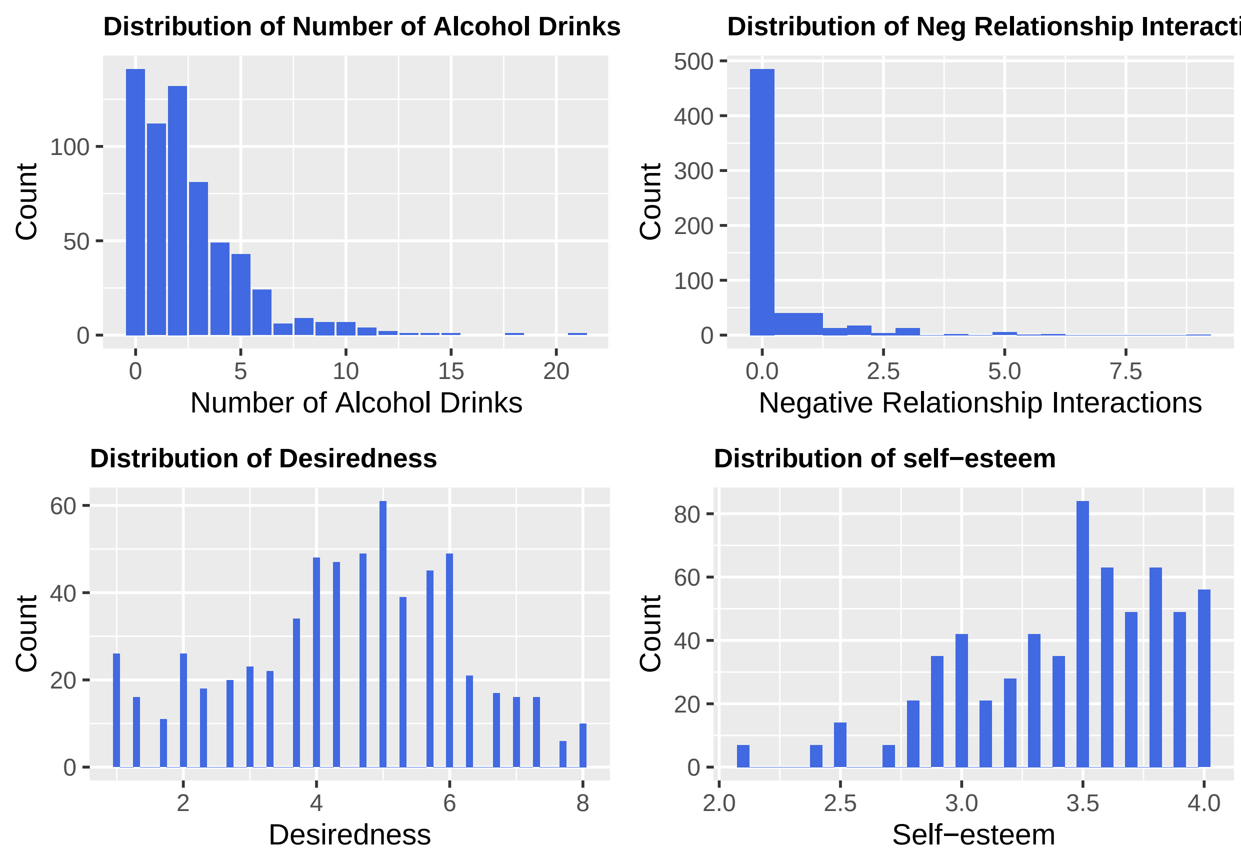
From above **structure** and **summary** of **dehart** dataframe, we can see that total number of alcohol consumed(**numall**)ranges from o to 21. Also, Negative Relationship Interactions range from 0 to 9 and of **num** data type.

Univariate Analysis:

Let us now plot some distribution plots for explanatory variables of interest.

```r
theme_set(theme_gray())
dist_drinks <- ggplot(dehart, aes(numall)) + geom_bar(fill="royalblue") +
                         ggtitle("Distribution of Number of Alcohol Drinks") +
                         xlab("Number of Alcohol Drinks") + ylab("Count")+
                         theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

dist_nrel <- ggplot(dehart, aes(nrel)) + geom_histogram(binwidth = 0.5, fill="royalblue") +
                        ggtitle("Distribution of Neg Relationship Interactions") +
                        xlab("Negative Relationship Interactions") +  ylab("Count") +
                        theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

dist_desired <- ggplot(dehart, aes(desired)) + geom_histogram(binwidth = 0.1, fill="royalblue") +
                        ggtitle("Distribution of Desiredness") +
                        xlab("Desiredness") +  ylab("Count") +
                       theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

dist_rosn <- ggplot(dehart, aes(rosn)) + geom_histogram(binwidth = 0.05, fill="royalblue") +
```

```
                        ggtitle("Distribution of self-esteem") +
                        xlab("Self-esteem") +  ylab("Count") +
                       theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))


ggarrange(dist_drinks, dist_nrel, dist_desired, dist_rosn, ncol = 2, nrow = 2)
```



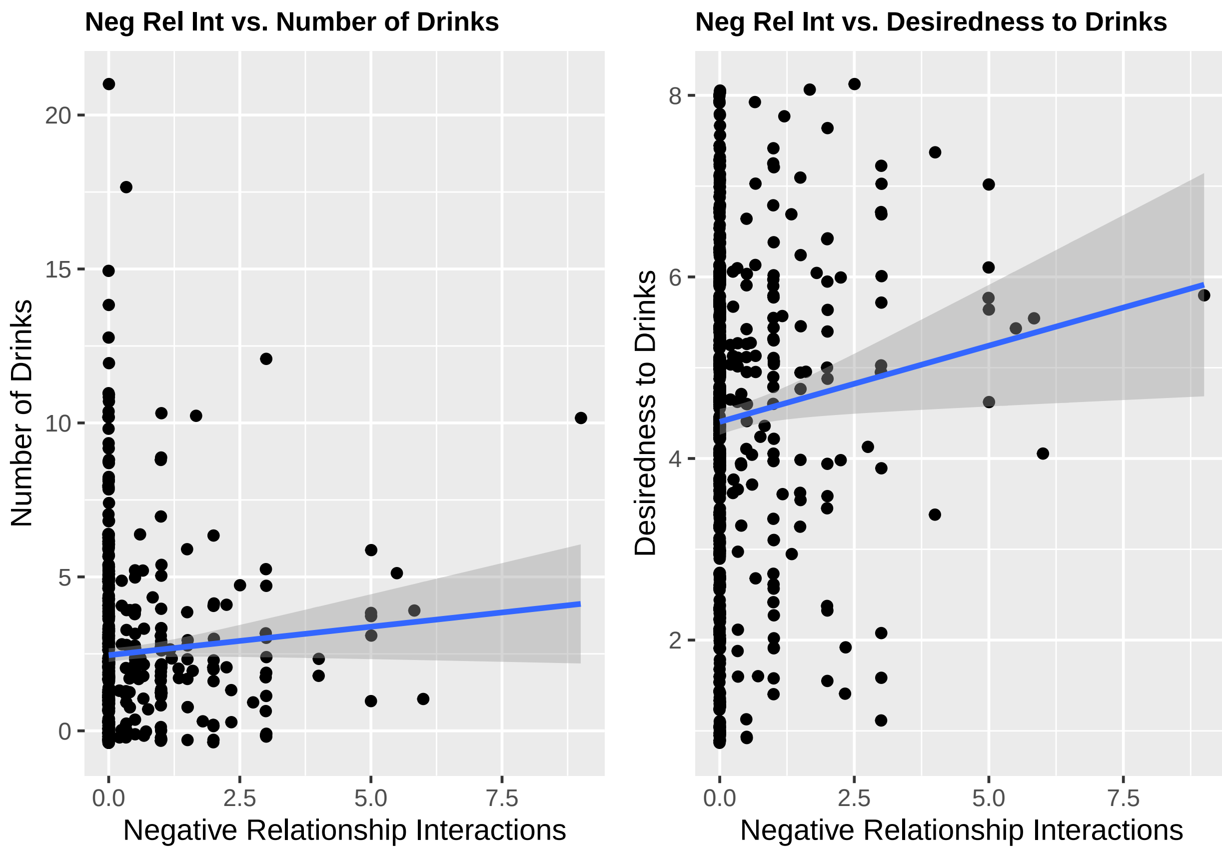
We can see from above histogram/distribution plot, total number of alcohol drinks consumed (**numall**) has median around 2 drinks. Also, desiredness and self-esteem are slightly right-skewed.

Bivariate Analysis:

Let us now see the effects of Negative Relationship Interactions(**nrel**)on total number of drinks consumed (**numall**) and desiredness to drink alcohol(**desired**)

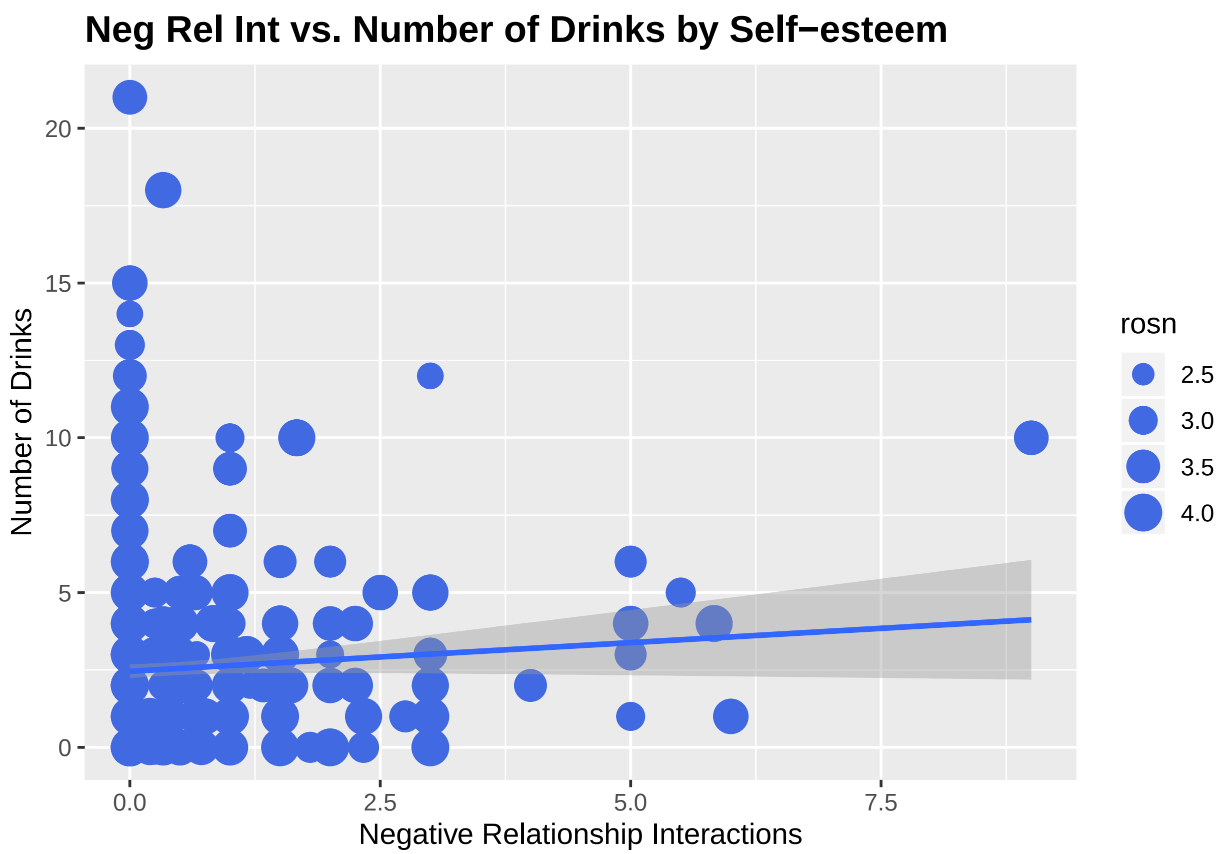
```
theme_set(theme_gray())
lm_nrel_numall <- ggplot(dehart, aes(x = nrel, y=numall)) + geom_point(position = "jitter") +
                  geom_smooth(method = 'lm') +
                  ggtitle('Neg Rel Int vs. Number of Drinks') +
                  xlab("Negative Relationship Interactions") + ylab("Number of Drinks") +
                 theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

lm_desired_numall <- ggplot(dehart, aes(x = nrel, y=desired)) + geom_point(position = "jitter") +
                  geom_smooth(method = 'lm') +
                  ggtitle('Neg Rel Int vs. Desiredness to Drinks') +
```

```r
                    xlab("Negative Relationship Interactions") + ylab("Desiredness to Drinks") +
                   theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))
ggarrange(lm_nrel_numall, lm_desired_numall, ncol = 2, nrow = 1)
```



We can see from above 2 plots that the total number of drinks consumed (**numall**) and desiredness to drink alcohol(**desired**) increase with Negative Relationship Interactions(**nrel**) encountered on that day.

```r
ggplot(dehart, aes(x = nrel, y=numall)) + geom_point(aes(size=rosn), color="royalblue") +
                    geom_smooth(method = 'lm') +
                    ggtitle('Neg Rel Int vs. Number of Drinks by Self-esteem') +
                    xlab("Negative Relationship Interactions") + ylab("Number of Drinks") +
                   theme(plot.title = element_text(lineheight=1, size = 14, face = "bold"))
```



From above plot, we can see that, relatively high self-esteem people had less alcohol consumption and less negative relationship interactions.

2.2 (2 points): Using an appropriate model (or models), evaluate the evidence that negative relationship interactions are associated with higher alcohol consumption and/or an increased desire to drink.

Let us now create a Poisson regression model for mean number of alcohol drinks consumed based on negative relationship interactions and self-esteem.

```
mod.numall.neg <- glm(numall~nrel +rosn , family = poisson(link="log"), data = dehart)
summary(mod.numall.neg)
```

```
##
## Call:
## glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
##     data = dehart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4439  -1.1474  -0.3049   0.3341   7.2786
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.909141   0.209022   4.349 1.36e-05 ***
## nrel         0.064503   0.023671   2.725  0.00643 **
## rosn        -0.002459   0.060439  -0.041  0.96755
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1590.3  on 621  degrees of freedom
## Residual deviance: 1583.4  on 619  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 2962.8
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(mod.numall.neg)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel   6.8258  1   0.008985 **
## rosn   0.0017  1   0.967556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above Poisson regression model's summary and LRT test, we can see that there is a positive effect of negative relationship interactions on mean number of alcohol drinks consumed. However, there is no significant evidence that self-esteem has any effect on alcohol consumption.

Let us now create a linear regression model for desiredness to drink alcohol based on negative relationship interactions and self-esteem.

```r
mod.desired.neg <- glm(desired ~nrel + rosn, data = dehart)
summary(mod.desired.neg)
```

```
##
## Call:
## glm(formula = desired ~ nrel + rosn, data = dehart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8988  -1.0222   0.1498   1.2397   3.6872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.2305     0.5543  11.240  < 2e-16 ***
## nrel          0.1769     0.0715   2.474 0.013621 *
## rosn         -0.5327     0.1603  -3.323 0.000942 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.8032)
##
##     Null deviance: 1776.0  on 619  degrees of freedom
## Residual deviance: 1729.6  on 617  degrees of freedom
##   (3 observations deleted due to missingness)
```

```
## AIC: 2403.5
##
## Number of Fisher Scoring iterations: 2
```

```r
Anova(mod.desired.neg)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired
##      LR Chisq Df Pr(>Chisq)
## nrel   6.1217  1  0.0133534 *
## rosn  11.0448  1  0.0008894 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above linear model, we can see that both Negative Relationship Interactions(**nrel**) and self-esteem(**rosn**) has significant effect on desiedness to consume alcohol(**desired**). Less self-esteem persons had higher desire to drink alcohol and high-esteem persons had lower desire to drink alcohol(as the coefficient of **rosn** has negative magnitude). Both p-value and Chi-square value suggest that self-esteem(**rosn**) is a significant indicator of desiredness to drink alcohol.

2.3 (1 points): Discuss whether the relationship between drinking and negative relationship interactions differs according to individuals' levels of trait self-esteem.

Let us now segregate the data into **lowesteem** and **highesteem** data based on **rosn** values $3rd$ quantile.

```r
lowesteem <- dehart[dehart$rosn <= quantile(dehart$rosn)["75%"], c(1:13)]# < 3rd quantile rosn into low
highesteem <- dehart[dehart$rosn > quantile(dehart$rosn)["75%"], c(1:13)]# > 3rd quantile rosn into hig
```

Let us now create a Poisson regression model on mean number of alcohol drinks consumed(**numall**) by **lowesteem** data and see the findings.

```r
mod.numall.neg_lowesteem <- glm(numall~nrel +rosn , family = poisson(link="log"), data = lowesteem)
summary(mod.numall.neg_lowesteem)
```

```
##
## Call:
## glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
##     data = lowesteem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5052  -1.1200  -0.2944   0.3443   7.2937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.888198   0.244072   3.639 0.000274 ***
## nrel        0.083085   0.024542   3.385 0.000711 ***
## rosn        0.001713   0.072863   0.024 0.981248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 1325.7  on 517  degrees of freedom
## Residual deviance: 1315.5  on 515  degrees of freedom
## AIC: 2467.2
##
## Number of Fisher Scoring iterations: 5
```

```
Anova(mod.numall.neg_lowesteem)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  10.2326  1    0.00138 **
## rosn   0.0006  1    0.98125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above p-value and Chi-square values for **nrel**, we can say that for lowesteem data, number of alcohol consumption(**numall**) are affected by negative relationship interactions(**nrel**). And this is statistically significant.

Let us now create a Poisson regression model on mean number of alcohol drinks consumed(**numall**) by **highesteem** data and see the findings.

```
mod.numall.neg_highesteem <- glm(numall~nrel +rosn , family = poisson(link="log"), data = highesteem)
summary(mod.numall.neg_highesteem)
```

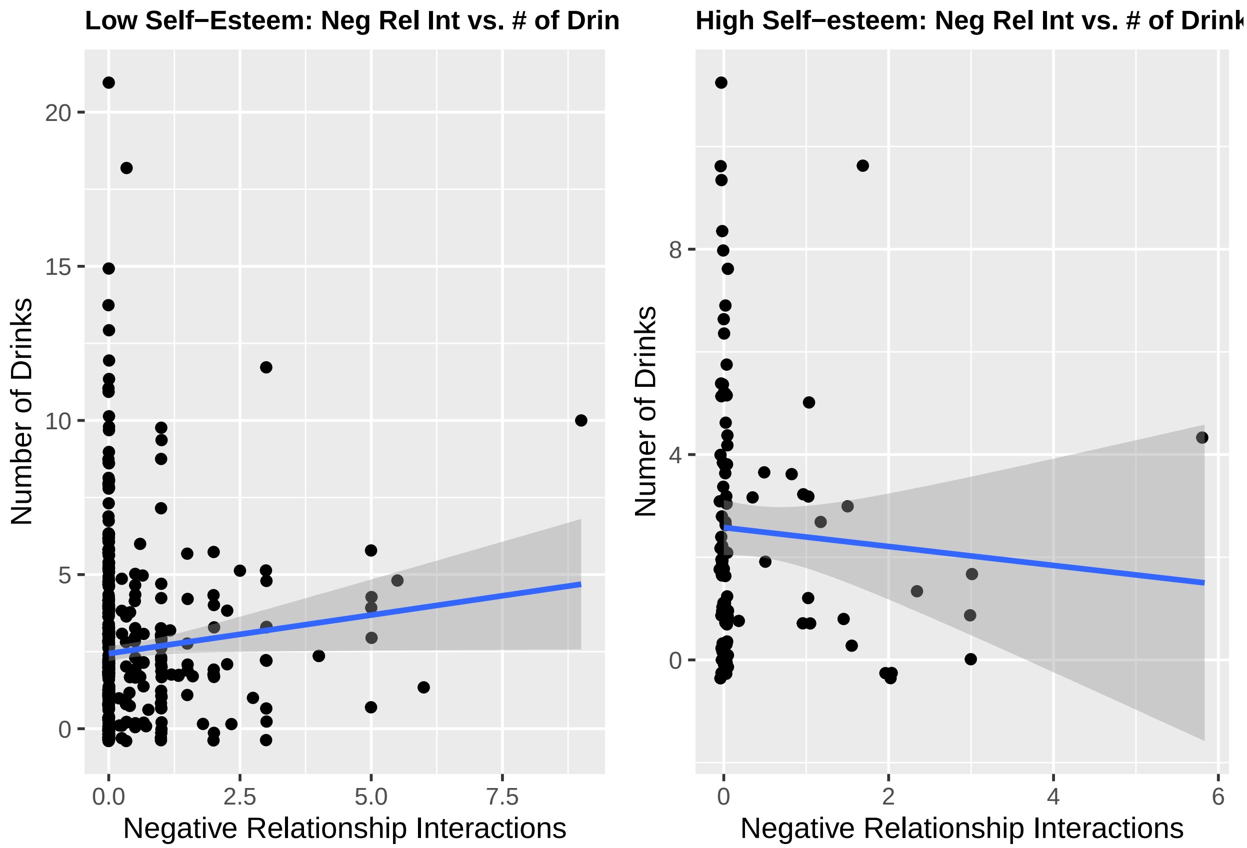
```
##
## Call:
## glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
##     data = highesteem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3408  -1.2098  -0.4697   0.7449   3.9880
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.28448    4.92229   1.074    0.283
## nrel        -0.09017    0.07890  -1.143    0.253
## rosn        -1.09657    1.24518  -0.881    0.379
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 264.52  on 103  degrees of freedom
## Residual deviance: 262.54  on 101  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 496.13
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(mod.numall.neg_highesteem)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  1.42214  1     0.2331
## rosn  0.77475  1     0.3788
```

It is evident from above p-values and Chi-square values for **nrel**, there is no effect of negative relationship interactions on alcohol consumption(**numall**) for persons with high self-esteem (**highesteem** data).

```r
theme_set(theme_gray())
lm_nrel_numall_lowesteem <- ggplot(lowesteem, aes(x = nrel, y=numall)) + geom_point(position = "jitter")
                  geom_smooth(method = 'lm') +
                  ggtitle('Low Self-Esteem: Neg Rel Int vs. # of Drinks') +
                  xlab("Negative Relationship Interactions") + ylab("Number of Drinks") +
                 theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

lm_nrel_numall_highesteem <- ggplot(highesteem, aes(x = nrel, y=numall)) + geom_point(position = "jitte
                  geom_smooth(method = 'lm') +
                  ggtitle('High Self-esteem: Neg Rel Int vs. # of Drinks') +
                  xlab("Negative Relationship Interactions") + ylab("Numer of Drinks") +
                 theme(plot.title = element_text(lineheight=1, size = 10, face = "bold"))

ggarrange(lm_nrel_numall_lowesteem, lm_nrel_numall_highesteem, ncol = 2, nrow = 1)
```



We can visually see the same evidence in above plots. For people with low self-esteem, the negative relationship interactions lead to more number of alcohol drinks consumed. Whereas for people with high self-esteem, the negative relationship interactions do not increase (rather decreases)the number of alcohol drinks consumed.