Shishir Agarwal - W271 Assignment 2

Due Sunday 7 March 2021 11:59pm

```r
rm(list = ls())
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
# Load Libraries
library(car)
library(Hmisc)
library(skimr)
library(ggplot2)
library(stargazer)
library(tidyverse)
library(GGally)
library(patchwork)
library(MASS)
library(mcprofile)
library(vcd)
library(nnet)
```

```r
setwd("/home/jovyan/r_bridge/student_work/shagarwa/Assignment#2")
```

1. Strategic Placement of Products in Grocery Stores (5 points)

These questions are taken from Question 12 of chapter of the textbook.

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item—breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular boxes of cereals on specific shelves, grocery stores may better attract customers to them. To investigate this further, a random sample of size 10 was taken from each of four shelves at a Dillons grocery store in Manhattan, KS. These data are given in the cereal_dillons.csv *file. The response variable is the shelf number, which is numbered from bottom (1) to top (4), and the explanatory variables are the sugar, fat, and sodium content of the cereals.*

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

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.

From the Box Plot we observe

- Sodium is highest in cereals on Shelf 1 and lower on Shelf 2,3,4
- Sugar is highest in cereals on Shelf 2 and lowest on Shelf 3, 4
- Fat is highest in cereals on Shelf 2 and lowest on Shelf 1,3

From the Parallel Coordinate Plot we observe

- Shelf 1 generally has cereal highest in sodium content and generally low in fat
- Shelf 2 generally has cereal with highest in sugar content with mixed bag of sodium and fat
- Shelf 3 and Shelf 4 has cereal with mixed bag of sodium, sugar, and fat

```
#rescale variables between 0 and 1
stand01 <- function(x) {
  (x-min(x))/(max(x)-min(x))
}

#create new dataframe with rescaled variables
cereal.data <- 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)
  )

#conduct basic EDA
str(cereal.data)
```

```
## '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
summary(cereal.data)
```

```
##      Shelf          sugar              fat               sodium
##  Min.   :1.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.75   1st Qu.:0.3339   1st Qu.:0.1582   1st Qu.:0.4200
##  Median :2.50   Median :0.6000   Median :0.3542   Median :0.5354
##  Mean   :2.50   Mean   :0.5209   Mean   :0.3476   Mean   :0.5240
##  3rd Qu.:3.25   3rd Qu.:0.7200   3rd Qu.:0.5400   3rd Qu.:0.6696
##  Max.   :4.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```r
#skim(cereal.data)
describe(cereal.data)
```

```
## cereal.data
##
##  4  Variables      40  Observations
## --------------------------------------------------------------------------------
## Shelf
##        n  missing distinct     Info     Mean      Gmd
##       40        0        4    0.938      2.5    1.282
##
## Value          1    2    3    4
## Frequency     10   10   10   10
## Proportion 0.25 0.25 0.25 0.25
## --------------------------------------------------------------------------------
## sugar
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       32    0.999   0.5209   0.3062   0.1054   0.1158
##      .25      .50      .75      .90      .95
##   0.3339   0.6000   0.7200   0.8075   0.8496
##
## lowest : 0.0000000 0.0360000 0.1090909 0.1125000 0.1161290
## highest: 0.8068966 0.8129032 0.8437500 0.9600000 1.0000000
## --------------------------------------------------------------------------------
## fat
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       20    0.985   0.3476   0.3319   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.1582   0.3542   0.5400   0.7075   1.0000
##
## lowest : 0.0000000 0.1102041 0.1741935 0.1800000 0.1830508
## highest: 0.5400000 0.5890909 0.6000000 0.6750000 1.0000000
## --------------------------------------------------------------------------------
## sodium
```

```
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       35    0.999    0.524   0.2583   0.1612   0.1934
##      .25      .50      .75      .90      .95
##   0.4200   0.5354   0.6696   0.8223   0.9017
##
## lowest : 0.0000000 0.1696970 0.1728395 0.1956989 0.2765432
## highest: 0.8166667 0.8731183 0.9000000 0.9333333 1.0000000
## --------------------------------------------------------------------------------
```

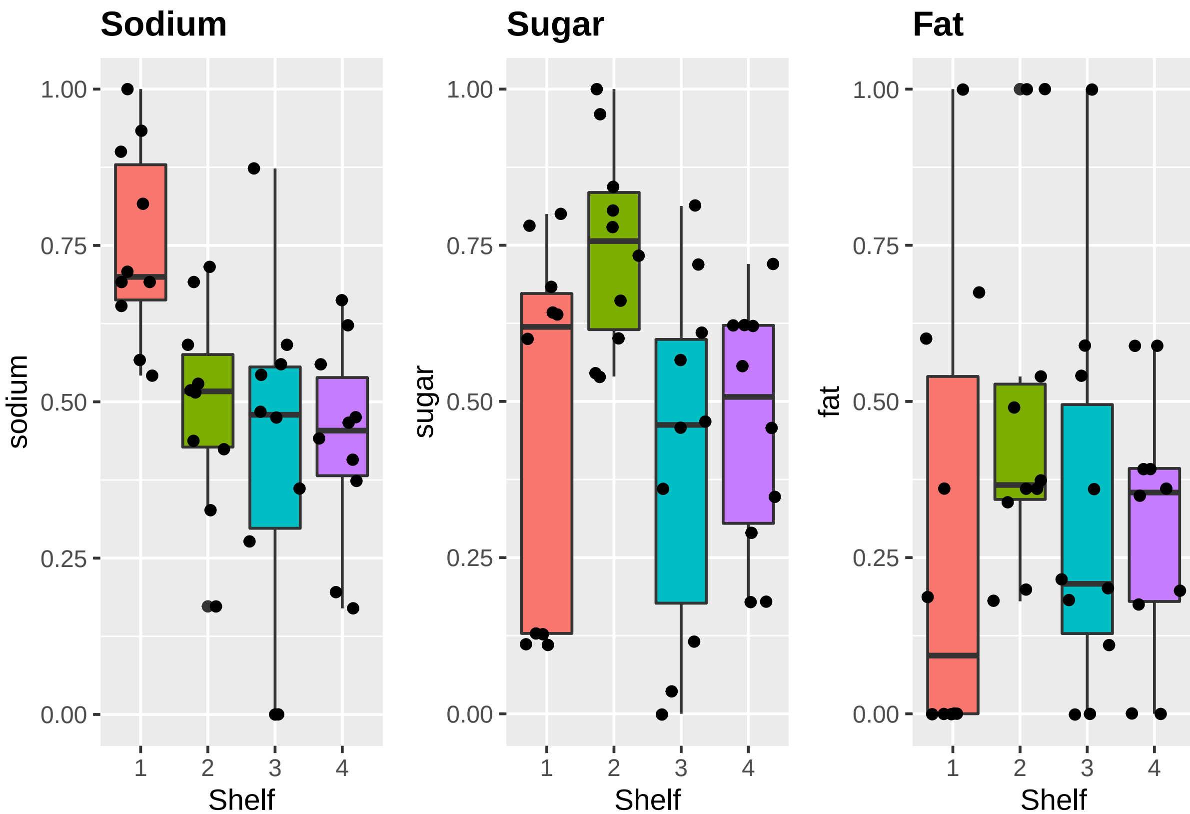
```
cereal[!complete.cases(cereal),]
```

```
## [1] ID         Shelf      Cereal     size_g     sugar_g    fat_g      sodium_mg
## <0 rows> (or 0-length row.names)
```

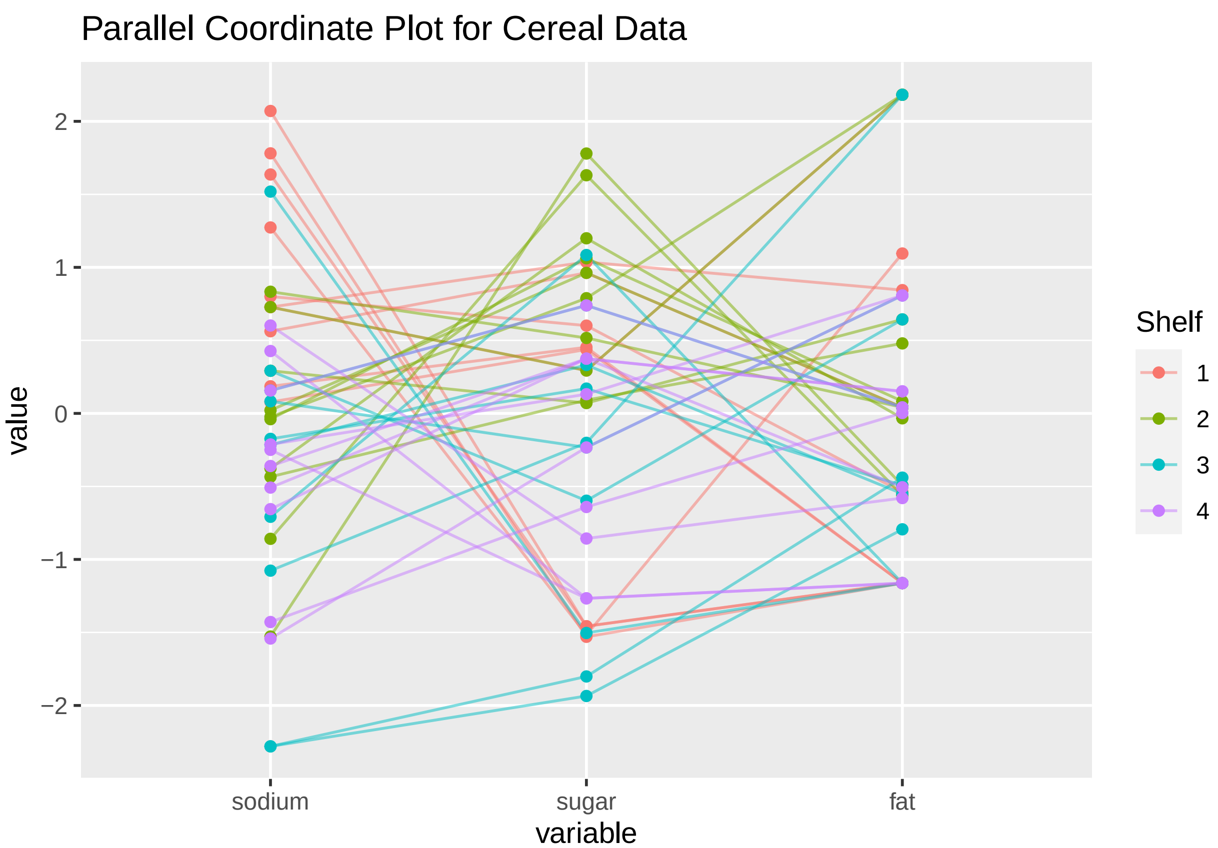
```
sapply(cereal, function(x) sum(is.na(x)))
```

```
##        ID     Shelf    Cereal    size_g   sugar_g     fat_g sodium_mg
##         0         0         0         0         0         0         0
```

```
#box plots
sugar_plot <- ggplot(data = cereal.data) +
  aes(x = factor(Shelf), y = sugar) +
  geom_boxplot(aes(fill = factor(Shelf)), show.legend = FALSE) +
  geom_jitter() +
  ggtitle("Sugar") +
  xlab("Shelf") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

fat_plot <- ggplot(data = cereal.data) +
  aes(x = factor(Shelf), y = fat) +
  geom_boxplot(aes(fill = factor(Shelf)), show.legend = FALSE) +
  geom_jitter() +
  ggtitle("Fat") +
  xlab("Shelf") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

sodium_plot <- ggplot(data = cereal.data) +
  aes(x = factor(Shelf), y = sodium) +
  geom_boxplot(aes(fill = factor(Shelf)), show.legend = FALSE) +
  geom_jitter() +
  ggtitle("Sodium") +
  xlab("Shelf") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

library(patchwork)
sodium_plot + sugar_plot + fat_plot
```




```r
#Parallel Coordinate Plot
cereal.data$Shelf <- factor(cereal.data$Shelf)
library(GGally)
ggparcoord(cereal.data,
           columns = 2:4,
           groupColumn = "Shelf",
           order = "anyClass",
           showPoints = TRUE,
           title = "Parallel Coordinate Plot for Cereal Data",
           alphaLines = 0.5
           )
```



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).

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. Though there is a physical order to these shelves, with 1 being at the bottom and 4 being at the top, we do not know if one shelf is better than another for drawing customer attention. Depending on the height of the shelve, depending on the height of customers, preference for one shelve over another may not follow the physical order. For example, we do not know if there is much of a difference between shelf 2 and shelf 3 when it comes to attracting customer attention. Also, we do not know if shelf 1 and shelf 4 are worse than shelf 2 and shelf 3. Also, we do not know if shelf 1 is preferred to shelf 4 or is it other way around. Thus, because we cannot assume an order among shelves for the purposes of drawing customer attention, we do not consider shelf as the ordinal variable. Instead we assume it to be a multinomial categorical variable.

The estimated regressions are

Equation 1: Shelf2 vs. Shelf1

$$log\left(\frac{\widehat{\pi}_{Shelf2}}{\widehat{\pi}_{Shelf1}}\right) = 6.9 - 17.5sodium + 2.7sugar + 4.1fat$$

Equation 2: Shelf3 vs. Shelf1

$$log\left(\frac{\widehat{\pi}_{Shelf3}}{\widehat{\pi}_{Shelf1}}\right) = 21.7 - 25sodium - 12.2sugar - 0.6fat$$

Equation 3: Shelf4 vs. Shelf1

$$log\left(\frac{\widehat{\pi}_{Shelf4}}{\widehat{\pi}_{Shelf1}}\right) = 21.3 - 24.7sodium - 11.4sugar - 0.9fat$$

```
# We look at sodium as the only dependent variable
cereal.multinom <- multinom(formula = Shelf ~ sodium, data = cereal.data)
```

```
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 46.094750
## final  value 46.089554
## converged
```

```
summary(cereal.multinom)
```

```
## Call:
## multinom(formula = Shelf ~ sodium, data = cereal.data)
##
## Coefficients:
##   (Intercept)     sodium
## 2    6.216387  -9.981426
## 3    7.192076 -12.122536
## 4    6.961633 -11.582694
##
## Std. Errors:
##   (Intercept)   sodium
## 2    2.586333 4.032968
## 3    2.629050 4.208826
## 4    2.622990 4.173343
##
## Residual Deviance: 92.17911
## AIC: 104.1791
```

```
#Calculate significance of sodium to all the categories
Anova(cereal.multinom)
```

```
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 55.451774
## converged
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##        LR Chisq Df Pr(>Chisq)
## sodium   18.724  3  0.0003117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Calculate significance of sodium for each individual category
(z_stat <- as.numeric(coef(cereal.multinom)[1,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[2]))
```

```
## [1] -2.474958
```

```
(z_stat <- as.numeric(coef(cereal.multinom)[2,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[4]))
```

```
## [1] -2.880265
```

```
(z_stat <- as.numeric(coef(cereal.multinom)[3,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[6]))
```

```
## [1] -2.775399
```

```
# We conclude sodium is important as explanatory variable

# We look at sugar as the only dependent variable
cereal.multinom <- multinom(formula = Shelf ~ sugar, data = cereal.data)
```

```
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 48.933328
## final  value 48.916047
## converged
```

```
summary(cereal.multinom)
```

```
## Call:
## multinom(formula = Shelf ~ sugar, data = cereal.data)
##
## Coefficients:
##   (Intercept)       sugar
## 2 -4.77496346  7.56100551
## 3  0.32695579 -0.74496920
## 4  0.01817729 -0.03944922
##
## Std. Errors:
##   (Intercept)    sugar
## 2   2.2421770 3.276525
## 3   0.8986483 1.773178
## 4   0.9369203 1.785385
```

```
##
## Residual Deviance: 97.83209
## AIC: 109.8321
```

```r
#Calculate significance of sugar to all the categories
Anova(cereal.multinom)
```

```
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 55.451774
## converged
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##       LR Chisq Df Pr(>Chisq)
## sugar   13.072  3   0.004485 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Calculate significance of sugar for each individual category
(z_stat <- as.numeric(coef(cereal.multinom)[1,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[2]))
```

```
## [1] 2.30763
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[2,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[4]))
```

```
## [1] -0.4201323
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[3,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[6]))
```

```
## [1] -0.02209564
```

```r
# We conclude sugar is marginally important as explanatory variable
# We conclude sugar is significantly important in explaining Shelf 2 over Shelf 1

# We look at fat as the only dependent variable
cereal.multinom <- multinom(formula = Shelf ~ fat, data = cereal.data)
```

```
## # weights:  12 (6 variable)
## initial  value 55.451774
## iter  10 value 54.029369
## iter  10 value 54.029369
## iter  10 value 54.029369
## final  value 54.029369
## converged
```

```r
summary(cereal.multinom)
```

```
## Call:
```

```
## multinom(formula = Shelf ~ fat, data = cereal.data)
##
## Coefficients:
##   (Intercept)       fat
## 2  -0.8647901 2.3032252
## 3  -0.1573904 0.5236279
## 4  -0.0923381 0.3151565
##
## Std. Errors:
##   (Intercept)      fat
## 2   0.7518341 1.616542
## 3   0.6725137 1.672102
## 4   0.6671339 1.689748
##
## Residual Deviance: 108.0587
## AIC: 120.0587
```

```
#Calculate significance of fat to all the categories
Anova(cereal.multinom)
```

```
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 55.451774
## converged
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##     LR Chisq Df Pr(>Chisq)
## fat   2.8448  3     0.4162
```

```
#Calculate significance of fat for each individual category
(z_stat <- as.numeric(coef(cereal.multinom)[1,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[2]))
```

```
## [1] 1.424785
```

```
(z_stat <- as.numeric(coef(cereal.multinom)[2,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[4]))
```

```
## [1] 0.3131555
```

```
(z_stat <- as.numeric(coef(cereal.multinom)[3,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[6]))
```

```
## [1] 0.186511
```

```
# We conclude fat is not important as explanatory variable

# We look at sugar, sodium, fat and their interaction as dependent variable
cereal.multinom <- multinom(formula = Shelf ~ sodium + sugar + fat +
                              sodium:sugar + sodium:fat + sugar:fat +
```

```r
                              sodium:sugar:fat, data = cereal.data,
                            maxit = 7500, trace = FALSE)
summary(cereal.multinom)
```

```
## Call:
## multinom(formula = Shelf ~ sodium + sugar + fat + sodium:sugar +
##     sodium:fat + sugar:fat + sodium:sugar:fat, data = cereal.data,
##     maxit = 7500, trace = FALSE)
##
## Coefficients:
##   (Intercept)     sodium      sugar       fat sodium:sugar sodium:fat
## 2   -7.807847   3.465455   7.444662  86.36575    -6.543132  -100.7529
## 3   19.959063 -22.504897 -15.962058 154.11602     6.226306  -236.7778
## 4   25.514998 -27.027213 -17.656176 101.13031    -5.285209  -169.8817
##    sugar:fat sodium:sugar:fat
## 2  -14.45414        -9.678778
## 3 -178.46624       282.856975
## 4  -73.72122       152.328329
##
## Std. Errors:
##   (Intercept)   sodium    sugar      fat sodium:sugar sodium:fat sugar:fat
## 2    27.40532 29.68536 32.17191 150.9823     36.05472   183.0248  192.6290
## 3    22.50951 24.11450 25.10943 165.6750     25.49028   222.5891  224.2506
## 4    22.43787 24.12360 25.79090 163.5723     30.00455   221.7291  222.0643
##   sodium:sugar:fat
## 2         224.0570
## 3         301.7841
## 4         306.9750
##
## Residual Deviance: 51.11665
## AIC: 99.11665
```

```r
Anova(cereal.multinom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                  LR Chisq Df Pr(>Chisq)
## sodium            30.8407  3  9.183e-07 ***
## sugar             19.2525  3  0.0002424 ***
## fat                6.1167  3  0.1060686
## sodium:sugar       3.0185  3  0.3887844
## sodium:fat         3.1586  3  0.3678151
## sugar:fat          3.2309  3  0.3573733
## sodium:sugar:fat   5.0167  3  0.1705789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# We conclude interactions is not important as explanatory variable

# We look at sodium, sugar, fat as the only dependent variable as the final model.
cereal.multinom <- multinom(formula = Shelf ~ sodium + sugar +
                              fat, data = cereal.data)
```

```
## # 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(cereal.multinom)
```

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

```r
# We notice fat does not play a significant role
Anova(cereal.multinom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##        LR Chisq Df Pr(>Chisq)
## sodium  26.6197  3  7.073e-06 ***
## sugar   22.7648  3  4.521e-05 ***
## fat      5.2836  3     0.1522
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Sodium
(z_stat <- as.numeric(coef(cereal.multinom)[1,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[2]))
```

```
## [1] -2.464913
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[2,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[6]))
```

```
## [1] -3.091298
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[3,2])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[10]))
```

```
## [1] -3.0604
```

```r
#Sugar
(z_stat <- as.numeric(coef(cereal.multinom)[1,3])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[3]))
```

```
## [1] 0.533103
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[2,3])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[7]))
```

```
## [1] -2.499296
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[3,3])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[11]))
```

```
## [1] -2.338928
```

```r
#Fat
(z_stat <- as.numeric(coef(cereal.multinom)[1,4])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[4]))
```

```
## [1] 1.761712
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[2,4])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[8]))
```

```
## [1] -0.2306981
```

```r
(z_stat <- as.numeric(coef(cereal.multinom)[3,4])/
    as.numeric(sqrt(diag(vcov(cereal.multinom)))[12]))
```

```
## [1] -0.3616886
```

```r
# We notice sodium to be most important, followed by sugar, followed by fat
```

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.

We predict Shelf 2 for this Kellogg's Apple Jacks Cereal.

```
#Create Test Data by binding it to the original data
test.cereal <- rbind(cereal, data.frame(ID = 0, Shelf = 0,
                     Cereal = "Apple Jacks", size_g = 28,
                      sugar_g = 12, fat_g = 0.5, sodium_mg = 130))

#Pre-process Test Data alongside the original data
test.cereal.data <- data.frame(Shelf = test.cereal$Shelf,
                sugar = stand01(x = test.cereal$sugar_g/test.cereal$size_g),
                fat = stand01(x = test.cereal$fat_g/test.cereal$size_g),
                sodium = stand01(x = test.cereal$sodium_mg/test.cereal$size_g)
                              )

#Train the model on the original cereal data
cereal.multinom <- multinom(formula = Shelf ~ sodium + sugar + fat,
                            data = cereal.data, trace = FALSE)

#Predict the model for the new data
test.cereal[41,] #Raw Data
```

```
##    ID Shelf      Cereal size_g sugar_g fat_g sodium_mg
## 41  0     0 Apple Jacks     28      12   0.5       130
```

```
test.cereal.data[41,] #Pre-processed data
```

```
##    Shelf     sugar       fat    sodium
## 41     0 0.7714286 0.1928571 0.4333333
```

```
round(predict(object = cereal.multinom, newdata = test.cereal.data[41,2:4],
              type = "probs"),3)
```

```
##     1     2     3     4
## 0.053 0.472 0.200 0.274
```

```
predict(object = cereal.multinom, newdata = test.cereal.data[41,2:4],
        type = "class")
```

```
## [1] 2
## Levels: 1 2 3 4
```

```
#We predict Shelf 2
```

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.

From the plot we can see if the sugar is low within a cereal then the ceral is typically placed on Shelf3 and Shelf4 instead of Shelf1 and Shelf2. However as the sugar content

increases, the cereal is placed on Shelf1 and Shelf2 with Shelf2 dominating. From this plot we can see how Shelf2 is the preferred shelf for high sugar content. Specifically for sugar less than 0.5mg/serving we find the cereal on Shelf3 and Shelf4 provided sodium and fat stays constant. For sugar more than 0.75mg/serving we find the cereal on Shelf2 provided sodium and fat stays constant

```
cereal.multinom <- multinom(formula = Shelf ~ sodium + sugar + fat,
                            data = cereal.data)
```

```
## # 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
```

```
summary(cereal.multinom)
```

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

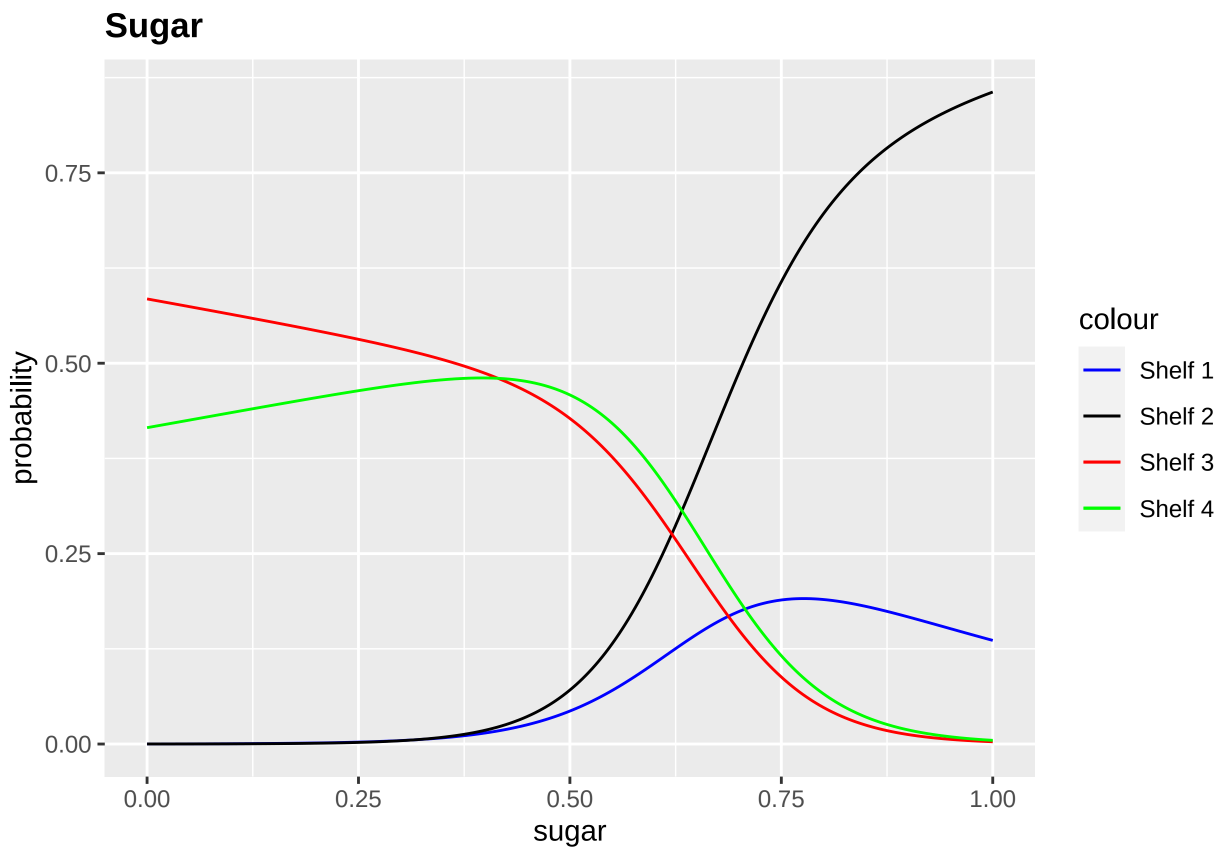
```
Anova(cereal.multinom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##        LR Chisq Df Pr(>Chisq)
## sodium  26.6197  3  7.073e-06 ***
## sugar   22.7648  3  4.521e-05 ***
```

```
## fat       5.2836  3     0.1522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sodium.mean <- mean(cereal.data$sodium)
fat.mean <- mean(cereal.data$fat)
beta.hat<-coefficients(cereal.multinom)
a <- seq(0,1,length = 1000)
b0 <- 1/(1 + exp(beta.hat[1,1] + beta.hat[1,2]*sodium.mean + beta.hat[1,3]*a + beta.hat[1,4]*fa
            exp(beta.hat[2,1] + beta.hat[2,2]*sodium.mean + beta.hat[2,3]*a + beta.hat[2,4]*fat
            exp(beta.hat[3,1] + beta.hat[3,2]*sodium.mean + beta.hat[3,3]*a + beta.hat[3,4]*fat
b1 <- (exp(beta.hat[1,1] + beta.hat[1,2]*sodium.mean + beta.hat[1,3]*a +
             beta.hat[2,4]*fat.mean))/
      (1+ exp(beta.hat[1,1] + beta.hat[1,2]*sodium.mean + beta.hat[1,3]*a + beta.hat[1,4]*fat.m
          exp(beta.hat[2,1] + beta.hat[2,2]*sodium.mean + beta.hat[2,3]*a + beta.hat[2,4]*fat.m
          exp(beta.hat[3,1] + beta.hat[3,2]*sodium.mean + beta.hat[3,3]*a + beta.hat[3,4]*fat.m
b2 <- (exp(beta.hat[2,1] + beta.hat[2,2]*sodium.mean + beta.hat[2,3]*a +
             beta.hat[2,4]*fat.mean))/
  (1+ exp(beta.hat[1,1] + beta.hat[1,2]*sodium.mean + beta.hat[1,3]*a +
            beta.hat[1,4]*fat.mean) +
     exp(beta.hat[2,1] + beta.hat[2,2]*sodium.mean + beta.hat[2,3]*a +
           beta.hat[2,4]*fat.mean) +
     exp(beta.hat[3,1] + beta.hat[3,2]*sodium.mean + beta.hat[3,3]*a +
           beta.hat[3,4]*fat.mean))
b3 <- (exp(beta.hat[3,1] + beta.hat[3,2]*sodium.mean + beta.hat[3,3]*a +
             beta.hat[3,4]*fat.mean))/
  (1+ exp(beta.hat[1,1] + beta.hat[1,2]*sodium.mean + beta.hat[1,3]*a +
            beta.hat[1,4]*fat.mean) +
     exp(beta.hat[2,1] + beta.hat[2,2]*sodium.mean + beta.hat[2,3]*a +
           beta.hat[2,4]*fat.mean) +
     exp(beta.hat[3,1] + beta.hat[3,2]*sodium.mean + beta.hat[3,3]*a +
           beta.hat[3,4]*fat.mean))
#plot(a,b0)
#plot(a,b1)
#plot(a,b2)
#plot(a,b3)
sodium <- rep(mean(cereal.data$sodium),1000)
sugar <- seq(0,1,length = 1000)
fat <- rep(mean(cereal.data$fat),1000)
test.data <- data.frame(sodium, sugar, fat)
predict.data <- predict(object = cereal.multinom, newdata = test.data, type = "probs")
sugar.data <- data.frame(sugar = sugar, predict.data)
#plot(sugar,sugar.data[,2])
#plot(sugar,sugar.data[,3])
#plot(sugar,sugar.data[,4])
#plot(sugar,sugar.data[,5])
ggplot(data = sugar.data) +
  aes(x = sugar) +
```

```
geom_line(aes(y = X1, color="Shelf 1"), linetype="solid") +
geom_line(aes(y = X2, color="Shelf 2"), linetype="solid") +
geom_line(aes(y = X3, color="Shelf 3"), linetype="solid") +
geom_line(aes(y = X4, color="Shelf 4"), linetype="solid") +
scale_color_manual(values = c(
  'Shelf 1' = 'blue',
  'Shelf 2' = 'black',
  'Shelf 3' = 'red',
  'Shelf 4' = 'green')) +
ggtitle("Sugar") +
xlab("sugar") +
ylab("probability") +
theme(plot.title = element_text(lineheight=1, face="bold"))
```



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. **The estimated odd of Shelf2 over Shelf1 change by 0.01 times for a 0.27 increase in sugar holding other variables constant. The estimated odd of Shelf3 over Shelf1 change by 0.0 times for a 0.27 increase in sugar holding other variables constant. The estimated odd of Shelf4 over Shelf1 change by 0.01 times for a 0.27 increase in sugar holding other variables constant**

With 95% confidence, the odds of Shelf2 over Shelf1 changes by (0.12, 43.21) when

sugar increase by 0.27. With 95% confidence, the odds of Shelf3 over Shelf1 changes by (0.0, 0.45) when sugar increase by 0.27. With 95% confidence, the odds of Shelf4 over Shelf1 changes by (0.0, 0.58) when sugar increase by 0.27.

```
sd.cereal <- apply(X = cereal.data[,c(2:4)], MARGIN = 2, FUN = sd)
c.value <- c(1, sd.cereal)
round(c.value,2)
```

```
##          sugar    fat  sodium
##   1.00    0.27   0.30    0.23
```

```
beta2 <- coef(cereal.multinom)[1,1:4]
beta3 <- coef(cereal.multinom)[2,1:4]
beta4 <- coef(cereal.multinom)[3,1:4]

round(exp(c.value*beta2),2)
```

```
##          sugar    fat  sodium
## 992.98    0.01   2.24    2.55
```

```
#round(1/exp(c.value*beta2),2)

round(exp(c.value*beta3),2)
```

```
##                     sugar          fat       sodium
## 2.604952e+09 0.000000e+00 3.000000e-02 8.800000e-01
```

```
#round(1/exp(c.value*beta3),2)

round(exp(c.value*beta4),2)
```

```
##                     sugar          fat       sodium
## 1.759584e+09 0.000000e+00 3.000000e-02 8.200000e-01
```

```
#round(1/exp(c.value*beta4),2)

conf.beta <- confint(object = cereal.multinom, level = 0.95)

ci.OR2 <- exp(c.value * conf.beta[1:4,1:2,1])
round(ci.OR2,2)
```

```
##              2.5 %       97.5 %
## (Intercept)  0.00 330392859.41
## sodium       0.00         0.38
## sugar        0.12        43.21
## fat          0.90         7.20
```

```
ci.OR3 <- exp(c.value * conf.beta[1:4,1:2,2])
round(ci.OR3,2)
```

```
##               2.5 %       97.5 %
## (Intercept) 1184.66 5.728017e+15
```

```
## sodium         0.00 9.000000e-02
## sugar          0.00 4.500000e-01
## fat            0.30 2.610000e+00
```

```r
ci.OR4 <- exp(c.value * conf.beta[1:4,1:2,3])
round(ci.OR4,2)
```

```
##              2.5 %       97.5 %
## (Intercept) 825.32 3.751459e+15
## sodium        0.00 9.000000e-02
## sugar         0.00 5.800000e-01
## fat           0.28 2.420000e+00
```

2. Alcohol, self-esteem and negative relationship interactions (5 points)

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 definitions of its variables: the number of drinks consumed (`numall`), positive romantic-relationship events (`prel`), negative romantic-relationship events (`nrel`), age (`age`), trait (long-term) self-esteem (`rosn`), state (short-term) self-esteem (`state`).

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.

```
dehart <- read.table(file = "DeHartSimplified.csv", header=TRUE, sep=",")
```

2.1 (2 points): Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. Address the reasons for limiting the study to observations from only one day.

We choose one dayof the week for our study because the response variable we are interested is *numall* which is a number between 0 and n. Thus, we can analyze *numall* using Poisson distribution. To analyze using Poisson distribution we want to assume the same intensity from period to period and the period remains constant from one observation to observation. If we were not to keep our unit of observation to one day of the week than the desire to drink (intensity) will vary from observation to observation and our assumptions for Poisson distribution will be violated. In our analysis, we notice *Saturdays* is when the data is most rich and there are least number of 0 drinks on *Saturday*. Thus our unit observation for this analysis is number of drinks consumed by each individual on *Saturdays* and we assume the desire to drink on Saturday (intensity) is constant from Saturday to Saturday which is a reasonabale assumption.

We also perform EDA to understand the data. The response variable can be modeled using Poisson distribution however we see compared to a theoretical poisoon distribution we see fewer data points with 3 or 4 drinks. Also, we note most of the explanatory variables are skewed. Lastly, when we analyze the scatter plots of *numall* against *nrel* for low, medium, high self-esteem individual we see a pattern emerge which shows for individuals with low self-esteem there is a strong relationship between *numall* and *nrel*

```
# We want to first check if there are missing values
dehart[!complete.cases(dehart),]
```

```
##      id studyday dayweek numall nrel prel  negevent posevent gender rosn
## 12    2        5       7      7 0.00    0 0.0000000     0.00      2  3.9
## 17    4        3       5      3 0.25    6 0.5716667     1.42      2  3.7
## 214  42        4       7     NA 0.00    3 0.0000000     1.80      2  4.0
## 402 110        3       1      1 0.00    0 0.1000000     0.70      2  3.6
## 448 116        7       3      2 0.00    2 0.2000000     1.30      2  3.4
##          age  desired    state
## 12  38.00137       NA       NA
## 17  30.04791 5.666667       NA
## 214 35.15674 3.666667 4.555556
## 402 40.82957       NA       NA
## 448 37.38809       NA 4.000000
```

```r
# We notice there are missing values for 1,3,5,7 but not for 6

# We subset the data to variables that are important to the researcher
dehart.data <- dehart[dehart$dayweek == 6, c(3,4,5,7,10,12)]

# We ensure there are no missing values for the subset data
dehart.data[!complete.cases(dehart.data),]
```

```
## [1] dayweek  numall   nrel     negevent rosn     desired
## <0 rows> (or 0-length row.names)
```

```r
# We chedk the data structure for the data
str(dehart.data)
```

```
## 'data.frame':    89 obs. of  6 variables:
##  $ dayweek : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ numall  : int  9 4 1 0 2 7 2 5 0 0 ...
##  $ nrel    : num  1 5.833 0.333 0 0 ...
##  $ negevent: num  0.4 2.377 0.233 0.2 0 ...
##  $ rosn    : num  3.3 3.9 3.7 3 3.3 3.5 3.5 3.1 3.7 3.5 ...
##  $ desired : num  5.67 5.67 5 1.67 4 ...
```

```r
# We summarize the data
summary(dehart.data)
```

```
##     dayweek      numall            nrel           negevent           rosn
##  Min.   :6   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :2.100
##  1st Qu.:6   1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:0.1500   1st Qu.:3.200
##  Median :6   Median : 4.000   Median :0.0000   Median :0.3500   Median :3.500
##  Mean   :6   Mean   : 4.101   Mean   :0.4034   Mean   :0.4404   Mean   :3.436
##  3rd Qu.:6   3rd Qu.: 5.000   3rd Qu.:0.3333   3rd Qu.:0.6000   3rd Qu.:3.800
##  Max.   :6   Max.   :21.000   Max.   :5.8333   Max.   :2.3767   Max.   :4.000
##     desired
##  Min.   :1.000
##  1st Qu.:4.000
##  Median :5.000
##  Mean   :4.846
```

```
##  3rd Qu.:6.000
##  Max.   :8.000
```

```r
#We analyze the response variable and key explanatory variables
describe(dehart.data$numall)
```

```
## dehart.data$numall
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       89        0       15     0.98    4.101     3.67      0.0      1.0
##      .25      .50      .75      .90      .95
##      2.0      4.0      5.0      9.0     10.6
##
## lowest :  0  1  2  3  4, highest: 10 11 12 13 21
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency      7    14    18     5    10    16     3     3     2     3     3
## Proportion 0.079 0.157 0.202 0.056 0.112 0.180 0.034 0.034 0.022 0.034 0.034
##
## Value         11    12    13    21
## Frequency      2     1     1     1
## Proportion 0.022 0.011 0.011 0.011
```

```r
describe(dehart.data$nrel)
```

```
## dehart.data$nrel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       89        0       15    0.592   0.4034   0.6916   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.0000   0.0000   0.3333   1.1000   2.1500
##
## lowest : 0.0000000 0.3333333 0.4000000 0.5000000 0.6000000
## highest: 2.0000000 2.2500000 3.0000000 4.0000000 5.8333333
##
## 0 (66, 0.742), 0.333333333 (1, 0.011), 0.4 (1, 0.011), 0.5 (2, 0.022), 0.6 (1,
## 0.011), 0.65 (1, 0.011), 0.666666667 (1, 0.011), 1 (7, 0.079), 1.5 (1, 0.011),
## 1.666666667 (1, 0.011), 2 (2, 0.022), 2.25 (1, 0.011), 3 (2, 0.022), 4 (1,
## 0.011), 5.833333333 (1, 0.011)
```

```r
describe(dehart.data$negevent)
```

```
## dehart.data$negevent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       89        0       34    0.993   0.4404   0.4328    0.000    0.000
##      .25      .50      .75      .90      .95
##    0.150    0.350    0.600    0.900    1.235
##
## lowest : 0.0000000 0.1000000 0.1333333 0.1400000 0.1500000
## highest: 1.3250000 1.4000000 1.5000000 1.9500000 2.3766667
```

```r
describe(dehart.data$rosn)
```

```
## dehart.data$rosn
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       89        0       17    0.993    3.436   0.4709     2.74     2.90
##      .25      .50      .75      .90      .95
##     3.20     3.50     3.80     3.90     4.00
##
## lowest : 2.1 2.4 2.5 2.7 2.8, highest: 3.6 3.7 3.8 3.9 4.0
##
## Value        2.1   2.4   2.5   2.7   2.8   2.9   3.0   3.1   3.2   3.3   3.4
## Frequency      1     1     2     1     3     5     6     3     4     6     5
## Proportion 0.011 0.011 0.022 0.011 0.034 0.056 0.067 0.034 0.045 0.067 0.056
##
## Value        3.5   3.6   3.7   3.8   3.9   4.0
## Frequency     12     9     7     9     7     8
## Proportion 0.135 0.101 0.079 0.101 0.079 0.090
```

```r
#We notice frequency for 3 or 4 drinks is low comparatively but could be due to chance
table(dehart.data$numall)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 21
##  7 14 18  5 10 16  3  3  2  3  3  2  1  1  1
```

```r
head(dehart.data)
```

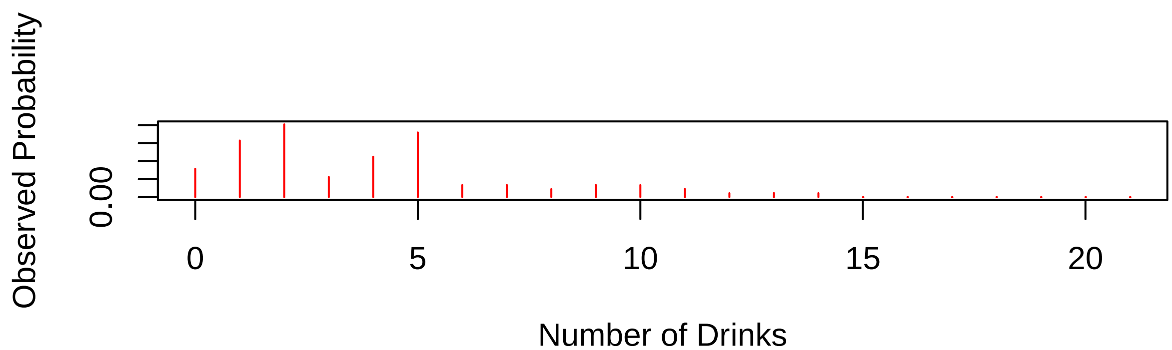
```
##    dayweek numall      nrel  negevent rosn  desired
## 1        6      9 1.0000000 0.4000000  3.3 5.666667
## 11       6      4 5.8333333 2.3766667  3.9 5.666667
## 18       6      1 0.3333333 0.2333333  3.7 5.000000
## 24       6      0 0.0000000 0.2000000  3.0 1.666667
## 35       6      2 0.0000000 0.0000000  3.3 4.000000
## 39       6      7 1.0000000 0.5500000  3.5 7.333333
```

```r
tail(dehart.data)
```

```
##     dayweek numall nrel  negevent rosn  desired
## 584       6      1    0 0.8000000  2.9 1.333333
## 593       6      4    2 1.4000000  3.6 6.000000
## 601       6      6    0 0.5666667  3.6 5.333333
## 603       6      5    0 0.0000000  3.8 5.000000
## 614       6     13    0 0.5000000  3.1 6.000000
## 619       6      5    0 0.5000000  3.5 6.000000
```

```r
# We want to analyze the data against theoritical Poisson distribution
mu.hat <- mean(dehart.data$numall)
mu.var <- var(dehart.data$numall)
alpha <- 0.05
n <-  length(dehart.data$numall)
x <- seq(0,21, by = 1)
```

```r
rel.freq <- table(dehart.data$numall)/length(dehart.data$numall)
rel.freq <- c(rel.freq, rep(0, times = 7))
theory.prob <- dpois(x = x, lambda = mean(dehart.data$numall))
dehart.prob <- data.frame(x, theory.prob, rel.freq)
par(mfrow = c(2,1))
plot(dehart.prob$x, dehart.prob$theory.prob, type = "h",
     ylab = "Theoretical Probability", xlab = "Number of Drinks", col = "black")
plot(dehart.prob$x, dehart.prob$rel.freq, type = "h",
     ylab = "Observed Probability", xlab = "Number of Drinks", col = "red")
```




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

#We calculate the confidence interval for mean and variance
(wald.int <- mu.hat + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(mu.hat/n))
```

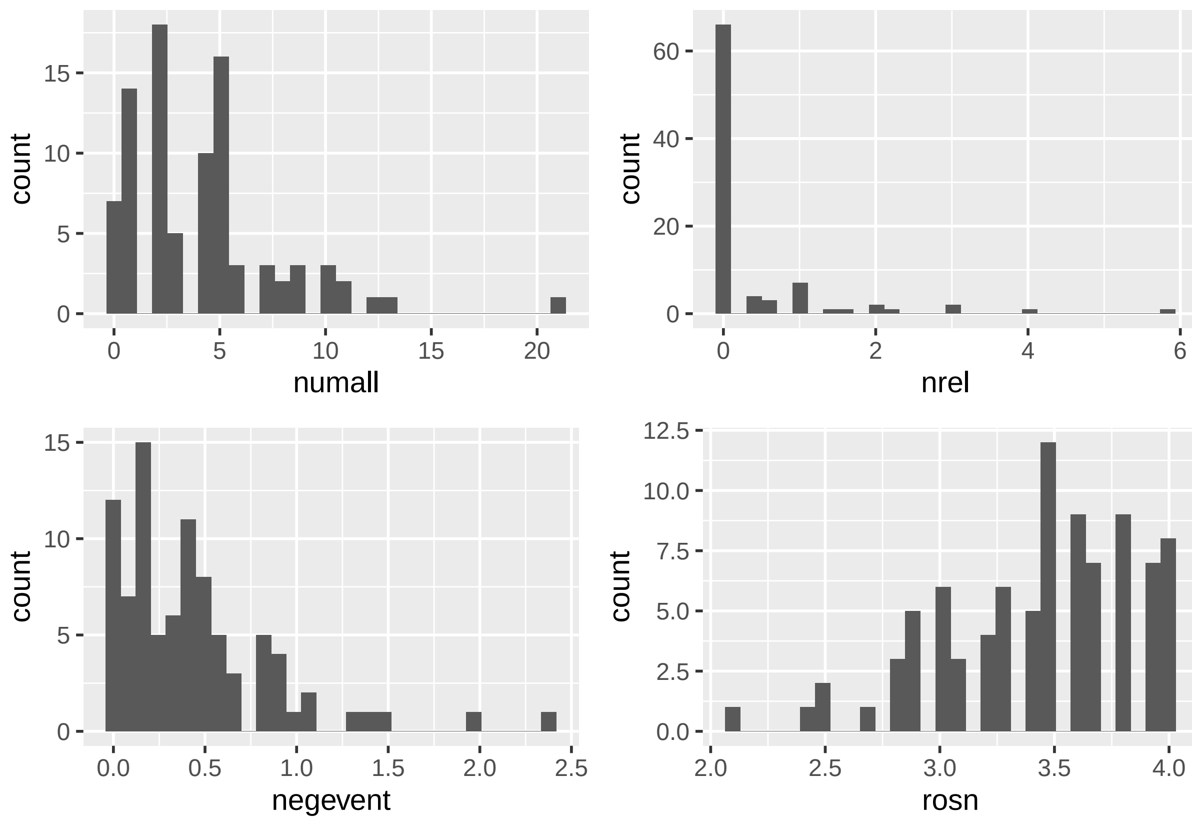
```
## [1] 3.680393 4.521855
```

```r
as.numeric(t.test(dehart.data$numall, conf.level = 0.95)$conf.int)
```

```
## [1] 3.350928 4.851319
```

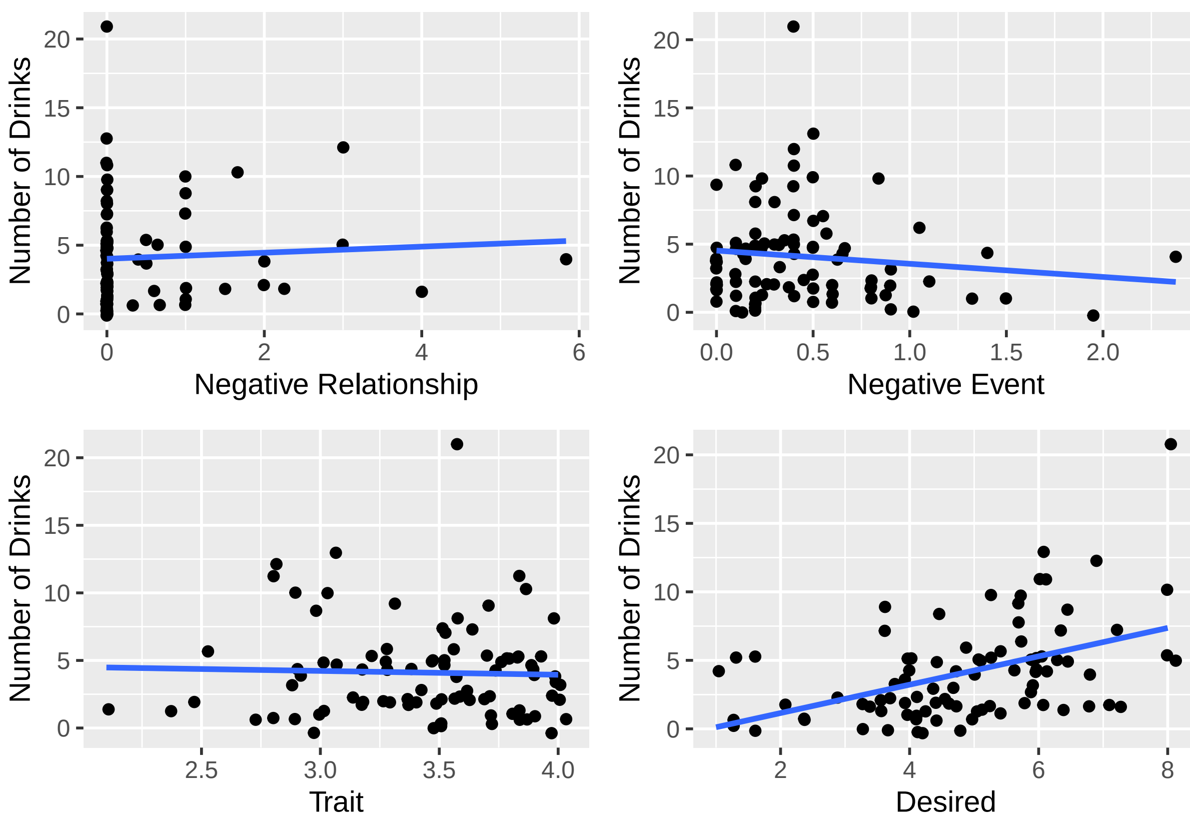
```r
#We analyze the histogram of response and key explanatory variables
#We notice few data points with low self esteem trait
#We notice most of the negative romantic relationship data points are zero
```

```
numall_hist <- dehart.data %>%
  ggplot(aes(numall)) +
  geom_histogram()
nrel_hist <- dehart.data %>%
  ggplot(aes(nrel)) +
  geom_histogram()
negevent_hist <- dehart.data %>%
  ggplot(aes(negevent)) +
  geom_histogram()
rosn_hist <- dehart.data %>%
  ggplot(aes(rosn)) +
  geom_histogram()
library(patchwork)
(numall_hist + nrel_hist) / (negevent_hist + rosn_hist)
```




```
#We analyze relationship between response and explanatory variable
#We see a positive relationship between numall and desired as expected
#We see a surprising negative relationships between numall and negevent
#We see a positive relationship between numall and nrel as expected
#We do not see any relationship between self esteem trait and nrel
nrel_numall <- dehart.data %>%
  ggplot(aes(x = nrel, y = numall)) +
  geom_jitter() +
```

```
  geom_smooth(method = "lm", se = FALSE) +
  labs(y = "Number of Drinks", x = "Negative Relationship")

negevent_numall <- dehart.data %>%
  ggplot(aes(x = negevent, y = numall)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(y = "Number of Drinks", x = "Negative Event")

rosn_numall <- dehart.data %>%
  ggplot(aes(x = rosn, y = numall)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(y = "Number of Drinks", x = "Trait")

desired_numall <- dehart.data %>%
  ggplot(aes(x = desired, y = numall)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(y = "Number of Drinks", x = "Desired")

(nrel_numall + negevent_numall) / (rosn_numall + desired_numall)
```



```r
#Because of research question we further break down the self esteem data
#We look at the different quartile for rosn and accordingly bin the data
#We create a categorical variable trait
summary(dehart.data$rosn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.100   3.200   3.500   3.436   3.800   4.000
```

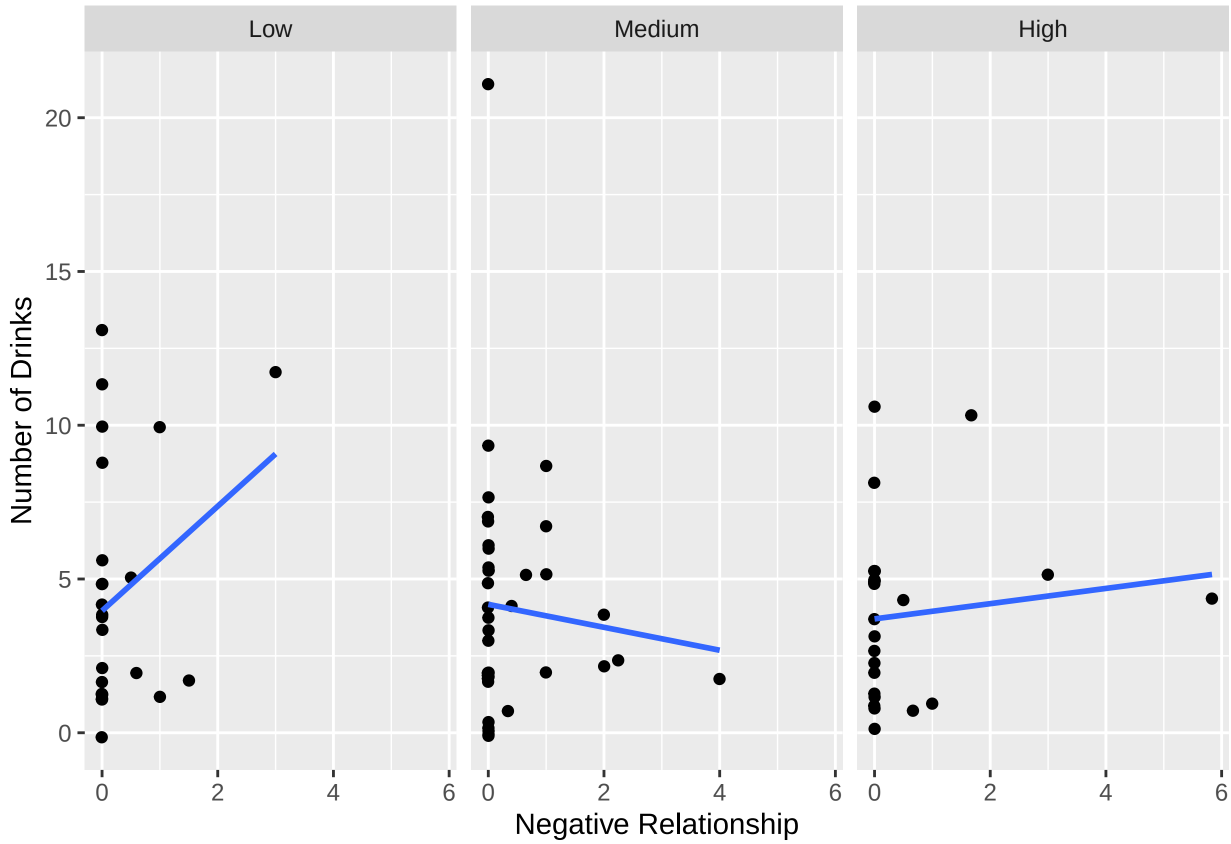
```r
dehart.data <- dehart.data %>%
  mutate(trait = case_when(
    rosn <= 3.2 ~ "Low",
    rosn > 3.2 & rosn < 3.8 ~ "Medium",
    rosn >= 3.8 ~ "High"))
dehart.data$trait <- factor(dehart.data$trait, level = c("Low", "Medium", "High"))
head(dehart.data)
```

```
##    dayweek numall      nrel  negevent rosn  desired  trait
## 1        6      9 1.0000000 0.4000000  3.3 5.666667 Medium
## 11       6      4 5.8333333 2.3766667  3.9 5.666667   High
## 18       6      1 0.3333333 0.2333333  3.7 5.000000 Medium
## 24       6      0 0.0000000 0.2000000  3.0 1.666667    Low
## 35       6      2 0.0000000 0.0000000  3.3 4.000000 Medium
## 39       6      7 1.0000000 0.5500000  3.5 7.333333 Medium
```

```r
tail(dehart.data)
```

```
##     dayweek numall nrel  negevent rosn  desired  trait
## 584       6      1    0 0.8000000  2.9 1.333333    Low
## 593       6      4    2 1.4000000  3.6 6.000000 Medium
## 601       6      6    0 0.5666667  3.6 5.333333 Medium
## 603       6      5    0 0.0000000  3.8 5.000000   High
## 614       6     13    0 0.5000000  3.1 6.000000    Low
## 619       6      5    0 0.5000000  3.5 6.000000 Medium
```

```r
#Because of research question we look at relationship between numall and nrel
#For each bin we notice a different relation between numall and nrel
#We notice a strong positive relationship between numall and nrel for low esteem
#We notice a slight positive relationship between numall and nrel for high esteem
dehart.data %>%
  ggplot(aes(x = nrel, y = numall)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(y = "Number of Drinks", x = "Negative Relationship") +
  facet_wrap(~ trait)
```



2.2 (2 points): The researchers hypothesize that negative interactions with romantic partners would be associated with alcohol consumption and an increased desire to drink. Using appropriate models, evaluate the evidence that negative relationship interactions are associated with higher alcohol consumption and an increased desire to drink.

We find there is no significant relationship between *numall* and *nrel* when we run regression. If any relationship exists it is by chance. We then add self esteem *rosn* to the model and find no significant relationship between *numall* and *nrel* when controlling for *rosn*. However when we add *negevent* along with *nrel* we find both *nrel* to be marginally significant and *negevent* to be strongly significant. Thus, we use this model to explore further. Using this model we find the following relationship:

$$log\left(numall\right) = 1.52221 + 0.12815nrel - 0.39634negevent$$

This leads to 13.67% percent change in *numall* from a unit change in *nrel* while controlling for *negevent*. And the 95% confidence interval for this change is (0.8%, 27.4%). We notice zero is excluded from the confidence interval. We plot this relationship between *numall* and *nrel* for three different values (min, max, mean) of *negevent*. From the plot, we notice as the *numevent* increases the relationship between *numall* and *nrel* becomes more significant. As Without *negevent* the relationship between *numall* and *nrel* is not significant.

We also explore the relationship between *numall* and *desirded* and plot this relationship. This is a strongly significant relationship.

```r
#We see the relationship between nrel and numall is not significant
#Thus this relationship can be due to chance
dehart.poisson.model <- glm(numall ~ nrel, family = poisson(link = "log"),
                            data = dehart.data)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = dehart.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8337  -1.3211  -0.5305   0.4733   5.9597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39003    0.05715  24.320   <2e-16 ***
## nrel         0.04971    0.05076   0.979    0.328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.43  on 87  degrees of freedom
## AIC: 508.83
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(dehart.poisson.model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  0.90934  1     0.3403
```

```r
#When we control for rosn, we still see no relationship between nrel and numall
dehart.poisson.model <- glm(numall ~ nrel + rosn, family = poisson(link = "log"),
                            data = dehart.data)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
##     data = dehart.data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.8809  -1.3074  -0.4411   0.5377   6.0026
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.66338    0.42696   3.896 9.79e-05 ***
## nrel         0.05303    0.05113   1.037    0.300
## rosn        -0.08011    0.12428  -0.645    0.519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.02  on 86  degrees of freedom
## AIC: 510.42
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(dehart.poisson.model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  1.01875  1     0.3128
## rosn  0.41219  1     0.5209
```

```r
#When we control for negevent
#We see a marginal relationship between nrel and numall
dehart.poisson.model <- glm(numall ~ nrel + negevent, family = poisson(link = "log"),
                            data = dehart.data)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ nrel + negevent, family = poisson(link = "log"),
##     data = dehart.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9679  -1.3596  -0.2781   0.5279   6.0346
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.52221    0.07445  20.447  < 2e-16 ***
## nrel         0.12815    0.05960   2.150  0.03155 *
## negevent    -0.39634    0.15132  -2.619  0.00881 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 242.06  on 86  degrees of freedom
## AIC: 503.46
##
## Number of Fisher Scoring iterations: 5
```

```
Anova(dehart.poisson.model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##          LR Chisq Df Pr(>Chisq)
## nrel       4.3358  1    0.03732 *
## negevent   7.3760  1    0.00661 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
100*(exp(dehart.poisson.model$coefficients[2]) -1)
```

```
##     nrel
## 13.67218
```

```
100*(exp(dehart.poisson.model$coefficients[3]) -1)
```

```
##  negevent
## -32.72203
```

```
beta1.int <- confint(dehart.poisson.model, parm = "nrel", level = 0.95)
beta2.int <- confint(dehart.poisson.model, parm = "negevent", level = 0.95)
100*(exp(beta1.int) -1)
```

```
##      2.5 %     97.5 %
##  0.7795388 27.3745457
```

```
100*(exp(beta2.int) -1)
```

```
##     2.5 %    97.5 %
## -50.39355 -10.18919
```

```
x_nrel <- seq(0,10,0.01)
max_negevent <- rep(max(dehart.data$negevent), 1001)
min_negevent <- rep(min(dehart.data$negevent), 1001)
mean_negevent <- rep(mean(dehart.data$negevent), 1001)
y.max <- exp(dehart.poisson.model$coefficients[1] +
           dehart.poisson.model$coefficients[2] * x_nrel +
           dehart.poisson.model$coefficients[3] * max_negevent)
y.min <- exp(dehart.poisson.model$coefficients[1] +
           dehart.poisson.model$coefficients[2] * x_nrel +
```

```r
          dehart.poisson.model$coefficients[3] * min_negevent)
y.mean <- exp(dehart.poisson.model$coefficients[1] +
          dehart.poisson.model$coefficients[2] * x_nrel +
          dehart.poisson.model$coefficients[3] * mean_negevent)

numall_nrel_df <- data.frame(nrel = x_nrel, numall.max = y.max,
                             numall.min = y.min, numall.mean = y.mean)

numall_nrel_plot <- numall_nrel_df %>%
  ggplot() +
  aes(x = nrel) +
  geom_line(aes(y = numall.max, color="With max(numevent) = 2.38"), linetype="solid") +
  geom_line(aes(y = numall.min, color="With min(numevent) = 0.0"), linetype="solid") +
  geom_line(aes(y = numall.mean,color="With mean(numevent) = 0.44"), linetype="solid") +
  scale_color_manual(values = c(
    'With max(numevent) = 2.38' = 'blue',
    'With min(numevent) = 0.0' = 'red',
    'With mean(numevent) = 0.44' = 'black')) +
  ggtitle("Number of Drinks vs. Negative Relationship") +
  xlab("Negative Relationship") +
  ylab("Number of Drinks") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

#When we control for desired
#We see a marginal relationship between nrel and numall
dehart.poisson.model <- glm(numall ~ desired, family = poisson(link = "log"),
                            data = dehart.data)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ desired, family = poisson(link = "log"),
##     data = dehart.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6749  -1.3361  -0.3239   0.5618   3.4753
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.01148    0.20113   0.057    0.954
## desired      0.27068    0.03543   7.640 2.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 189.11  on 87  degrees of freedom
## AIC: 448.51
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(dehart.poisson.model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##         LR Chisq Df Pr(>Chisq)
## desired   61.236  1  5.063e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

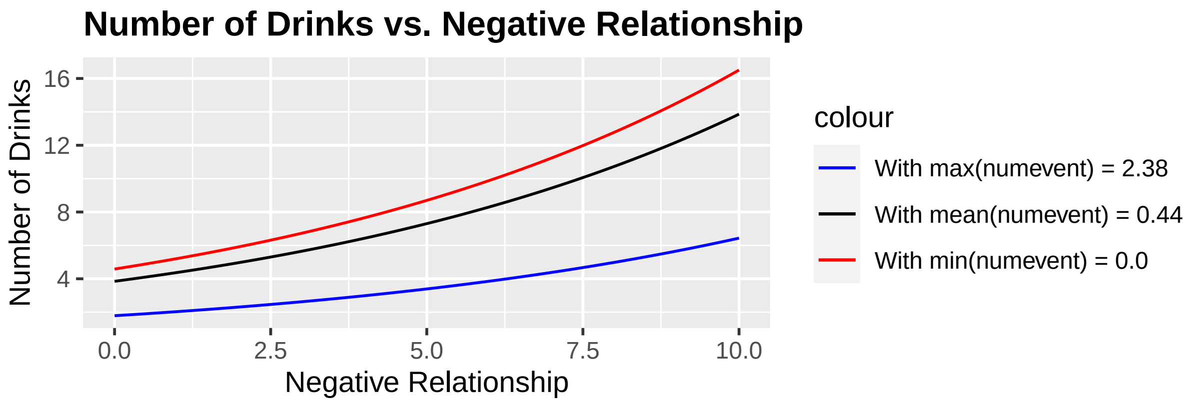
```r
100*(exp(dehart.poisson.model$coefficients[2]) -1)
```

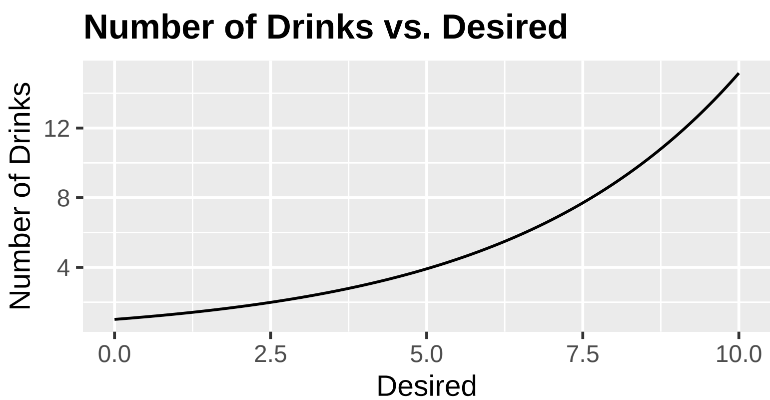
```
##  desired
## 31.08545
```

```r
beta1.int <- confint(dehart.poisson.model, parm = "desired", level = 0.95)
100*(exp(beta1.int) -1)
```

```
##    2.5 %   97.5 %
## 22.34281 40.57103
```

```r
x_desired <- seq(0,10,0.01)
y <- exp(dehart.poisson.model$coefficients[1] +
           dehart.poisson.model$coefficients[2] * x_desired)
numall_desired_df <- data.frame(desired = x_desired, numall = y)

numall_desired_plot <- numall_desired_df %>%
  ggplot() +
  aes(x = desired) +
  geom_line(aes(y = numall), linetype="solid") +
  ggtitle("Number of Drinks vs. Desired") +
  xlab("Desired") +
  ylab("Number of Drinks") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

numall_nrel_plot / numall_desired_plot
```





2.3 (1 points): The researchers hypothesize that the relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem. Conduct an analysis to address this hypothesis.

Because we do not find a significant relationship between $numall$ **and** $rosn$**, I break the data set into 2 parts. One with individuals that have low esteem and another with individual that have high esteem. Then I run regression to see if the relationship between** $numall$ **and** $nrel$ **is significant for the two data sets.**

For the data set that contains individuals with low self esteem, the relationship between $numall$ **and** $nrel$ **is significant. For the data set that contains individuals with high self esteem, the relationship between** $numall$ **and** $nrel$ **is not significant. This is what the researcher expected as well. Thus, we have the following relationship for individuals with low self esteem.**

$$log\,(numall) = 1.3888 + 0.2881nrel$$

This leads to 33.40% percent change in $numall$ **from a unit change in** $nrel$ **for individuals with low self esteem (less than equal to 3.2). And the 95% confidence interval for this change is (6.8%, 62.4%). We notice zero is excluded from the confidence interval. We plot this relationship between** $numall$ **and** $nrel$ **for data set that has low self esteem and compare the plot with data set that has high self esteem. From the plot, we notice for low self esteem, drastic increase in drinking with unit increase in negative relationship.**

```r
dehart.data.low <- dehart.data[dehart.data$trait == "Low",]
dehart.poisson.model <- glm(numall ~ nrel, family = poisson(link = "log"),
                            data = dehart.data.low)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = dehart.data.low)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8320  -1.8008  -0.2667   0.6983   3.5495
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.3888     0.1051  13.213  < 2e-16 ***
## nrel          0.2881     0.1064   2.707  0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 84.516  on 25  degrees of freedom
## Residual deviance: 78.311  on 24  degrees of freedom
## AIC: 159.02
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(dehart.poisson.model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel   6.2058  1    0.01273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
100*(exp(dehart.poisson.model$coefficients[2]) -1)
```

```
##     nrel
## 33.39562
```

```r
beta1.int <- confint(dehart.poisson.model, parm = "nrel", level = 0.95)
100*(exp(beta1.int) -1)
```

```
##     2.5 %    97.5 %
##  6.780151 62.413264
```

```r
x_nrel <- seq(0,10, 0.01)
y_low <- exp(dehart.poisson.model$coefficients[1] +
           dehart.poisson.model$coefficients[2] * x_nrel)

dehart.data.high <- dehart.data[dehart.data$trait == "High",]
dehart.poisson.model <- glm(numall ~ nrel, family = poisson(link = "log"),
                            data = dehart.data.high)
summary(dehart.poisson.model)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = dehart.data.high)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7230  -1.6716  -0.1432   0.6371   3.0563
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.31034    0.11374  11.520   <2e-16 ***
## nrel         0.05751    0.07132   0.806     0.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 49.490  on 23  degrees of freedom
## Residual deviance: 48.886  on 22  degrees of freedom
## AIC: 122.6
##
## Number of Fisher Scoring iterations: 5
```

```r
Anova(dehart.poisson.model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel  0.60356  1     0.4372
```

```r
100*(exp(dehart.poisson.model$coefficients[2]) -1)
```

```
##     nrel
## 5.919096
```

```r
beta1.int <- confint(dehart.poisson.model, parm = "nrel", level = 0.95)
100*(exp(beta1.int) -1)
```

```
##     2.5 %    97.5 %
```

```
## -9.201839 20.416143
```

```r
x_nrel <- seq(0,10, 0.01)
y_high <- exp(dehart.poisson.model$coefficients[1] +
          dehart.poisson.model$coefficients[2] * x_nrel)

numall_nrel_df <- data.frame(nrel = x_nrel, numall.low.rosn = y_low,
                             numall.high.rosn = y_high)

numall_nrel_df %>%
  ggplot() +
  aes(x = nrel) +
  geom_line(aes(y = numall.low.rosn, color="With Low Self Esteem"), linetype="solid") +
  geom_line(aes(y = numall.high.rosn, color="With High Self Esteem"), linetype="solid") +
  scale_color_manual(values = c(
    'With Low Self Esteem' = 'red',
    'With High Self Esteem' = 'black')) +
  ggtitle("Number of Drinks vs. Negative Relationship") +
  xlab("Negative Relationship") +
  ylab("Number of Drinks") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```