W271 Assignment 2

Erin Werner

October 18th, 2020

Contents

```
library(ggplot2)
library(dplyr)
library(car)
library(readr)
library(expss)
library(mcprofile)
library(gridExtra)
library(grid)
library(GGally)
library(nnet)
```

1 Strategic Placement of Products in Grocery Stores

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.

In this context, 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(paste0(here::here(),"/assignments/assignment_2/cereal_dillons.csv"))
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Shelf = col_double(),
##   Cereal = col_character(),
##   size_g = col_double(),
##   sugar_g = col_double(),
##   fat_g = col_double(),
##   sodium_mg = col_double()
## )
```

```
head(cereal)
```

```
## # A tibble: 6 x 7
##      ID Shelf Cereal                                size_g sugar_g fat_g sodium_mg
##   <dbl> <dbl> <chr>                                  <dbl>   <dbl> <dbl>     <dbl>
## 1     1     1 Kellog's Razzle Dazzle Rice Crispi~     28      10     0       170
## 2     2     1 Post Toasties Corn Flakes               28       2     0       270
## 3     3     1 Kellogg's Corn Flakes                   28       2     0       300
## 4     4     1 Food Club Toasted Oats                  32       2     2       280
## 5     5     1 Frosted Cheerios                        30      13     1       210
## 6     6     1 Food Club Frosted Flakes                31      11     0       180
```

```
paste("Sample Size: ", nrow(cereal))
```

```
## [1] "Sample Size:  40"
```

1.1 Question 1.1

Discuss whether possible content differences exist among the shelves.

The explanatory variables needs to be reformatted before proceeding further. As `Shelf` is the dependent categorical variable, I can transform it into a factor for the data set.

```
cereal$Shelf <- factor(cereal$Shelf)
```

Now, my focus is figuring out if there is a possible content difference that exists among the different shelves. First, I will divide each explanatory variable by its serving size to account for the different serving sizes among the cereals.

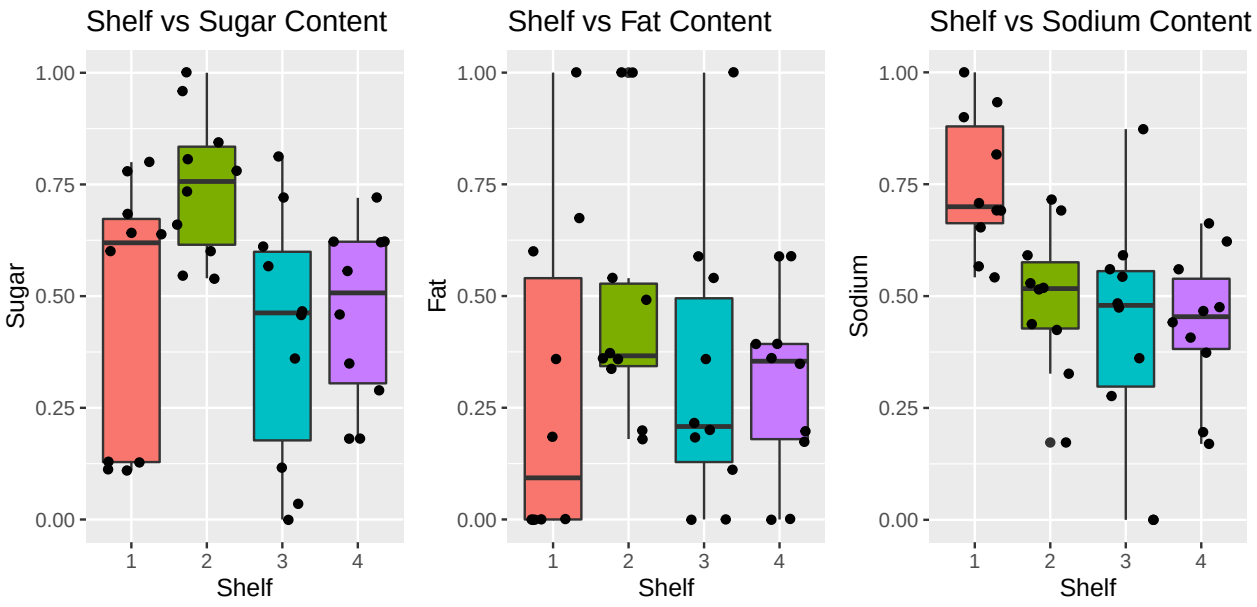
```
cereal$sugar <- cereal$sugar_g/cereal$size_g
cereal$fat <- cereal$fat_g/cereal$size_g
cereal$sodium <- cereal$sodium_mg/cereal$size_g
```

Second, I will re-scale each variable to be within 0 and 1.

```
cereal$sugar <- (cereal$sugar-min(cereal$sugar))/(max(cereal$sugar)-min(cereal$sugar))
cereal$fat <- (cereal$fat-min(cereal$fat))/(max(cereal$fat)-min(cereal$fat))
cereal$sodium<-(cereal$sodium-min(cereal$sodium))/(max(cereal$sodium)-min(cereal$sodium))
```

Then, I can construct side-by-side box plots with dot plots overlaid for each of the explanatory variables.

```
b1 <- ggplot(cereal, aes(Shelf, sugar)) + theme(legend.position="none") +
    geom_boxplot(varwidth=T, aes(fill = Shelf)) + geom_jitter() +
    labs(title="Shelf vs Sugar Content",x="Shelf",y="Sugar")
b2 <- ggplot(cereal, aes(Shelf, fat)) + theme(legend.position="none") +
    geom_boxplot(varwidth=T, aes(fill = Shelf)) + geom_jitter() +
    labs(title="Shelf vs Fat Content",x="Shelf",y="Fat")
b3 <- ggplot(cereal, aes(Shelf, sodium)) + theme(legend.position="none") +
    geom_boxplot(varwidth=T, aes(fill = Shelf)) + geom_jitter() +
    labs(title="Shelf vs Sodium Content",x="Shelf",y="Sodium")
egg::ggarrange(b1, b2, b3, nrow = 1)
```



From the box-plots, I can see:

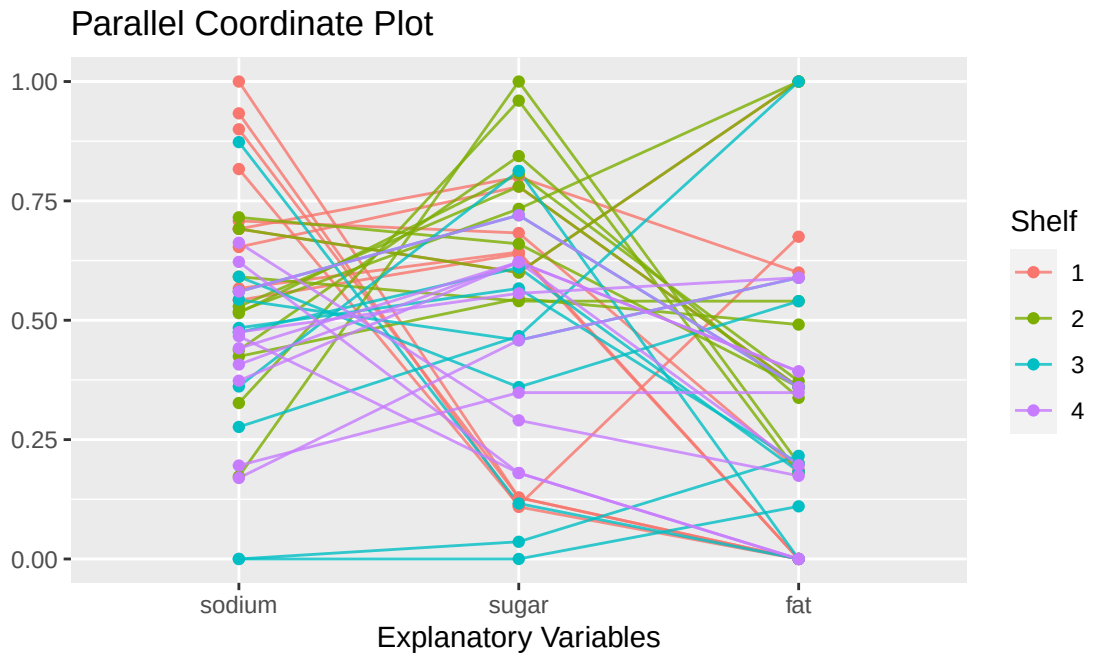
- `Shelf 1` contains a wide distribution of low-to-medium levels of Sugar, with a medium mean value. It also contains a wide distribution of low-to-medium levels of Fat, but with a low mean value. In regards to Sodium, `Shelf 1` has a smaller distribution, with distinctly higher levels of sodium content compared to the other shelves.
- `Shelf 2` contains a small distribution of distinctly high levels of Sugar, compared to the other shelves.

It also has small distributions of medium levels of both Fat and Sodium, each with medium mean values.

- `Shelf 3` contains a wide distribution of low-to-medium levels for Sugar, Fat, and Sodium. Yet, the mean values vary as Sugar and Sodium have medium mean values and Fat has a low mean value.
- `Shelf 4` contains a small distribution of low-to-medium levels for Sugar, Fat, and Sodium, with medium mean values for each content category.

I can also construct a parallel coordinates plot for the explanatory variables and the shelf number.

```
ggparcoord(cereal, columns = 8:10, groupColumn = 2, order = "anyClass",
    showPoints = TRUE, alphaLines = 0.8, scale="uniminmax") +
  labs(title="Parallel Coordinate Plot",x="Explanatory Variables",y="")
```



From the parallel coordinate plot, I can see:

- `Shelf 1` tends to have high levels of Sodium and then a variety of low-to-high levels for Sugar and Fat.
- `Shelf 2` tends to have high levels of Sugar with reasonably high levels for both Sugar and Fat.
- `Shelf 3` tends to have a wide variety of low-to-high levels for Sodium, Sugar, and Fat.
- `Shelf 4` tends to have medium levels for Sodium, Sugar, and Fat.

Now that I have an overall understanding of the data, I can see that it is possible for some content differences to exist among the shelves. Although there is no clear difference in Fat Content distributions between the different shelf levels, the Sugar and Sodium content distributions do appear to have some shelf distinctions. From both the box plots and the parallel coordinate plot, I can see that `Shelf 1` tends to have higher levels of Sodium and `Shelf 2` has higher levels of Sugar. However, `Shelves 3` and `4` tend to have similar distributions across all three content categories, with `Shelf 3` having a slightly wider distribution of values than `Shelf 4`. As a result, some shelves may contain different content.

1.2 Question 1.2

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.

Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories is not known. The difference between ordinal and categorical variables is that there is a clear ordering of the variables. As a result, ordinal is best applied to variables that have a categorical scale such as low, medium, and high.

In this context, the response has values of $1, 2, 3$, and 4. Although it is logical to order the shelves from bottom-to-top, the same scale could have been applied the reverse way and still have the same effect. The shelves could technically even be labeled by random letters and the groupings would still work. Therefore, ordinality does not apply here as there is not a clear ordering among the shelves.

Now, I estimate a simple and suitable multinomial regression model with the following format:

$$log\left(\frac{\widehat{\pi}_j}{\widehat{\pi}_1}\right) = \beta_0 + \beta_1 * sugar + \beta_2 * fat + \beta_3 * sodium$$

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

```
## # 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(mod.nominal)
```

```
## Call:
## multinom(formula = Shelf ~ sugar + fat + sodium, data = cereal)
##
## 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
```

Thus, the estimated regressions for the model are:

Equation 1: Shelf 2 vs. Shelf 1

$$log\left(\frac{\widehat{\pi}_{Shelf\_2}}{\widehat{\pi}_{Shelf\_1}}\right) = 6.9007 + 2.6931sugar + 4.0647fat - 17.494sodium$$

Equation 2: Shelf 3 vs. Shelf 1

$$log\left(\frac{\widehat{\pi}_{Shelf\_3}}{\widehat{\pi}_{Shelf\_1}}\right) = 21.6807 - 12.2164sugar - 0.55717fat - 24.9785sodium$$

Equation 3: Shelf 4 vs. Shelf 1

$$log\left(\frac{\widehat{\pi}_{Shelf\_4}}{\widehat{\pi}_{Shelf\_1}}\right) = 21.2883 - 11.3937sugar - 0.8701fat - 24.6739sodium$$

Next, I can perform LRTs to examine the importance of each explanatory variable in this model. To test the existence of effect of an explanatory variable on all response categories, I set the hypotheses as follows:

$$H_0 : \beta_{jr} = 0, \quad j = 2, \ldots, J \quad \text{assuming j=1 is the base category} H_a : \beta_{jr} \neq 0, \quad \text{for some } j$$

```
Anova(mod.nominal)
```

```
## 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
```

The numbers in the column *LR Chisq* is $-2log(\Lambda)$. For instance, the transformed test statistic is 5.2836 for the explanatory `fat`, and its corresponding p-value ($Pr(> Chisq)$) is greater than $\alpha = 0.05$. However, both `sugar` and `sodium` have corresponding p-values that are less than 0.001. This means that both `sugar` and `sodium` are statistically significant while `fat` is not.

However, it is important to consider the possible effects of the interactions among the three explanatory variables. So, I can build an additional model that includes all possible interactions, such that:

$log\left(\frac{\widehat{\pi}_j}{\widehat{\pi}_1}\right) = \beta_0 + \beta_1 * sugar + \beta_2 * fat + \beta_3 * sodium + \beta_4 * sugar : fat + \beta_5 * sugar : sodium + \beta_6 * fat : sodium + \beta_7 * sugar : fat : sodium$

This will help me to determine if the interactions have any statistical significance and, thus, should be included in the model analysis.

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

```
## # 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
```

```
Anova(mod.nom_inter)
```

```
## 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
```

As a result, no interactions have statistically significant p-values for $\alpha = 0.05$. So, I can see that there are no significant interactions among the explanatory variables (including an interaction among all three variables). Therefore, I will use the simple model, with no interactions, for my analysis.

1.3 Question 1.3

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.

I can calculate π_j, given one explanatory variable and J categories, as follows:

$$\pi_1 = \frac{1}{1 + \Sigma_{j=2}^{J} exp(\beta_{j0} + \beta_{j1} * x_1)}$$

and

$$\pi_j = \frac{exp(\beta_{j0} + \beta_{j1} * x_1)}{1 + \Sigma_{j=2}^{J} exp(\beta_{j0} + \beta_{j1} * x_1)}$$

for j = 2,. . . , J.

In order to estimate the shelf probabilities, I will first set the specified values of the variables.

```
my_size_o <- 28
my_sugar_o <- 12
my_fat_o <- 0.5
my_sodium_o <- 130
```

Next, I need to transform the variables, as I did before, to fit the model data. So, I will divide each explanatory variable by its serving size to account for the different serving sizes among the cereals.

```
my_sugar <- my_sugar_o/my_size_o
my_fat <- my_fat_o/my_size_o
my_sodium <- my_sodium_o/my_size_o
```

I will also rescale each variable to be within 0 and 1.

```
cereal$sugar_t <- cereal$sugar_g/cereal$size_g
cereal$fat_t <- cereal$fat_g/cereal$size_g
cereal$sodium_t <- cereal$sodium_mg/cereal$size_g

my_sugar <- (my_sugar-min(cereal$sugar_t))/(max(cereal$sugar_t)-min(cereal$sugar_t))
```

```
my_fat <- (my_fat-min(cereal$fat_t))/(max(cereal$fat_t)-min(cereal$fat_t))
my_sodium <- (my_sodium-min(cereal$sodium_t))/(max(cereal$sodium_t)-min(cereal$sodium_t))
```

Now, I can estimate the shelf probabilities given the new data. The `predict()` function can compute the estimated shelf probabilities using the `type = "probs"` argument value.

```
newdata <- with(cereal, data.frame(sugar=my_sugar, fat=my_fat, sodium=my_sodium))
pi.hat <- round(predict(object=mod.nominal, newdata=newdata, type="probs"), 4)
head(pi.hat)
```

```
##      1      2      3      4
## 0.0533 0.4719 0.2004 0.2744
```

As a result, $\hat{\pi}_{Shelf\_2} = [1 + exp(6.9007 - 2.6931 * my\_sugar + 4.0647 * my\_fat - 17.4937 * my\_sodium) + exp(21.6807 - 12.2164 * my\_sugar - 0.5571 * my\_fat - 24.9790 * my\_sodium) + exp(21.2883 - 11.3937 * my\_sugar - 0.8701 * my\_fat - 24.6739 * my\_sodium)]^{-1} = 0.472$. This estimation is most likely influenced by a high sugar content in the cereal, which makes sense as Apple Jacks is a cereal marketed to children. Overall, the estimated shelf probabilities are as follows:

- `Shelf 1` $= 5.3\%$
- `Shelf 2` $= 47.2\%$
- `Shelf 3` $= 20.0\%$
- `Shelf 4` $= 27.4\%$

Therefore, `Shelf 2` has the largest probability, making it the most likely shelf location for Kellogg's Apple Jacks, based on its sugar, fat, and sodium content.

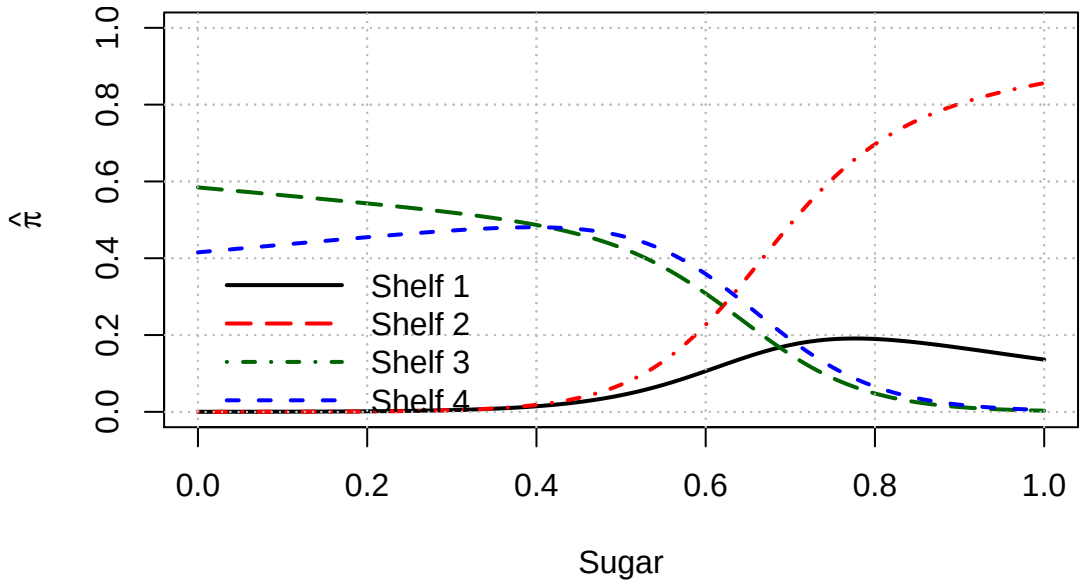
1.4 Question 1.4

Construct a plot where the estimated probability for a shelf is on the *y-axis* and the sugar content is on the *x-axis*. Interpret the plot with respect to sugar content.

The estimated model can be plotted by including the estimated probability for each shelf on the y-axis and one of the explanatory variables on the x-axis, while holding the other variables within the model constant. The explanatory variable I will focus on is `sugar` and I will use the mean overall `fat` and `sodium` content as the corresponding variable values in the model.

```
curve(expr = predict(object = mod.nominal,
  newdata = data.frame(sugar = x,fat = mean(cereal$fat),sodium = mean(cereal$sodium)),
  type = "probs")[,1], ylab = expression(hat(pi)),xlab = "Sugar", ylim = c(0,1),
  xlim=c(min(cereal$sugar),max(cereal$sugar)),col="black",lty="solid",lwd=2, n=1000,
  panel.first = grid(col = "gray", lty = "dotted"))
curve(expr = predict(object = mod.nominal,
  newdata = data.frame(sugar = x,fat = mean(cereal$fat),sodium = mean(cereal$sodium)),
  type = "probs")[,2], ylab = expression(hat(pi)), xlab = "Sugar",
  xlim=c(min(cereal$sugar),max(cereal$sugar)),col="red",lty="dotdash",lwd=2,n=1000,
  add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))
curve(expr = predict(object = mod.nominal,
  newdata = data.frame(sugar = x,fat = mean(cereal$fat),sodium = mean(cereal$sodium)),
  type = "probs")[,3], ylab = expression(hat(pi)), xlab = "Sugar",
  xlim=c(min(cereal$sugar),max(cereal$sugar)),col="darkgreen",lty="longdash",lwd=2,n=1000,
  add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))
curve(expr = predict(object = mod.nominal,
  newdata = data.frame(sugar = x,fat = mean(cereal$fat),sodium = mean(cereal$sodium)),
  type = "probs")[,4], ylab = expression(hat(pi)), xlab = "Sugar",
```

```
  xlim=c(min(cereal$sugar),max(cereal$sugar)),col="blue",lty="dashed",lwd=2,n=1000,
  add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))
legend(x = 0, y = 0.43, legend=c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
       lty=c("solid","longdash","dotdash","dashed"),
       col=c("black","red","darkgreen","blue"), bty="n", lwd = c(2,2,2), seg.len = 4)
```



Now, I can interpret the plot with respect to sugar content. The model is plotted above, where the estimated probabilities for each shelf are drawn between the smallest and the largest observed sugar values for that shelf (0 and 1 in this context).

From the plot, I can see that for the lower sugar levels, the estimated `Shelf 3` probability is the largest, but `Shelf 4` has just a slightly lower probability where `Shelves 1` and `2` have very low (approximately zero) probabilities. The estimated `Shelf 2` probability is the largest for the high sugar levels. The remaining shelves have low probabilities, with `Shelves 3` and `4` having very low (almost zero) probability values. For sugar levels in the middle, `Shelf 4` has the largest estimated probability, although `Shelf 3` has a similar probability distribution. `Shelves 1` and `2` have lower probabilities.

Overall, `Shelf 1` maintains generally low probability values for all sugar levels, although the probability increases slightly for higher levels of sugar. `Shelf 2` has very low probabilities for low sugar levels and very high probabilities for high sugar levels. `Shelves 3` and `4` have higher probabilities for low sugar values and low probabilities for high sugar values.

The parallel coordinates plot above displays similar findings where the sugar levels tend to follow the Shelf 3 $<$ Shelf 4 $<$ Shelf 1 $<$ Shelf 2 ordering in regards to `sugar`.

1.5 Question 1.5

Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretations back to the plots constructed for this exercise.

The odds ratios and corresponding confidence intervals depend on the base categorical value. So, I will start by comparing all shelves to `Shelf 1`, as that is the default base shelf for the model. Then, I will calculate the remaining odds ratios and corresponding confidence intervals by updating the base categorical value of the model accordingly.

Base: Shelf 1

The odds in this problem are constructed as $P(Y = j)/P(Y = 1)$, $j = 2, 3, 4$ where category 1 = `Shelf 1`, 2 = `Shelf 2`, 3 = `Shelf 3`, and 4 = `Shelf 4`.

The estimated odds ratios for `Shelf 2` vs. `Shelf 1` or `Shelf 3` vs. `Shelf 1` or `Shelf 4` vs. `Shelf 1` are calculated for each explanatory variable as $\hat{OR} = \exp(c\hat{\beta}_{jr})$ for $j = 2, 3, 4$ and $r = 1, \ldots, 12$.

To choose appropriate values of `c`, `c` will be equal to one standard deviation for each continuous explanatory variable in the model.

```
sd.cereal <- apply(X = cereal[,-c(1,2,3,4,5,6,7,11,12,13)], MARGIN = 2, FUN = sd)
c.value <- c(sd.cereal)
c.value <- round(c.value, 2)

beta.hat2 <- coefficients(mod.nominal)[1,2:4]
beta.hat3 <- coefficients(mod.nominal)[2,2:4]
beta.hat4 <- coefficients(mod.nominal)[3,2:4]

#Odds ratios for j = 2 vs. j = 1 (Shelf 2 vs. Shelf 1)
Shelf2_OR <- round(exp(c.value * beta.hat2), 2)
Shelf2_OR1 <- round(1/exp(c.value * beta.hat2), 2)
df2 <- data.frame(Shelf2_OR, Shelf2_OR1)
colnames(df2) <- c("OR", "1/OR")

#Odds ratios for j = 3 vs. j = 1 (Shelf 3 vs. Shelf 1)
Shelf3_OR <- round(exp(c.value * beta.hat3), 2)
Shelf3_OR1 <- round(1/exp(c.value * beta.hat3), 2)
df3 <- data.frame(Shelf3_OR, Shelf3_OR1)
colnames(df3) <- c("OR", "1/OR")

#Odds ratios for j = 4 vs. j = 1 (Shelf 4 vs. Shelf 1)
Shelf4_OR <- round(exp(c.value * beta.hat4), 2)
Shelf4_OR1 <- round(1/exp(c.value * beta.hat4), 2)
df4 <- data.frame(Shelf4_OR, Shelf4_OR1)
colnames(df4) <- c("OR", "1/OR")
```

Then, I can calculate the corresponding confidence intervals.

```
conf.beta <- confint(object = mod.nominal, level = 0.95)
```

```
ci.OR2 <- exp(c.value * conf.beta[2:4 ,1:2 ,1])
ci.OR3 <- exp(c.value * conf.beta[2:4 ,1:2 ,2])
ci.OR4 <- exp(c.value * conf.beta[2:4 ,1:2 ,3])

ci2 <- round(data.frame(low = ci.OR2[,1], up = ci.OR2[,2]), 2)
ci2_1 <- round(data.frame(low = 1/ci.OR2[,2], up = 1/ci.OR2[,1]),2)[c(1,2,3),]

ci3 <- round(data.frame(low = ci.OR3[,1], up = ci.OR3[,2]), 2)
ci3_1 <- round(data.frame(low = 1/ci.OR3[,2], up = 1/ci.OR3[,1]),2)[c(1,2,3),]

ci4 <- round(data.frame(low = ci.OR4[,1], up = ci.OR4[,2]), 2)
ci4_1 <- round(data.frame(low = 1/ci.OR4[,2], up = 1/ci.OR4[,1]),2)[c(1,2,3),]
```

```
print(paste0("The estimated OR for Shelf 2 vs Shelf 1:"))
```

```
## [1] "The estimated OR for Shelf 2 vs Shelf 1:"
```

```
grid.arrange(tableGrob(df2), ncol = 1)
```

	OR	1/OR
sugar	2.07	0.48
fat	3.39	0.3
sodium	0.02	55.9

```
print(paste0("The estimated CI & 1/CI for Shelf 2 vs Shelf 1:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 2 vs Shelf 1:"
```

```
grid.arrange(tableGrob(ci2), tableGrob(ci2_1), ncol = 2)
```

	low	up
sugar	0.14	29.98
fat	0.87	13.15
sodium	0	0.44

	low	up
sugar	0.03	7
fat	0.08	1.15
sodium	2.28	1370.42

```
print(paste0("The estimated OR for Shelf 3 vs Shelf 1:"))
```

```
## [1] "The estimated OR for Shelf 3 vs Shelf 1:"
```

```
grid.arrange(tableGrob(df3), ncol = 1)
```

	OR	1/OR
sugar	0.04	27.07
fat	0.85	1.18
sodium	0	312.64

```
print(paste0("The estimated CI & 1/CI for Shelf 3 vs Shelf 1:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 3 vs Shelf 1:"
```

```
grid.arrange(tableGrob(ci3), tableGrob(ci3_1), ncol = 2)
```

	low	up
sugar	0	0.49
fat	0.2	3.5
sodium	0	0.12

	low	up
sugar	2.04	359.64
fat	0.29	4.89
sodium	8.19	11939.03

```
print(paste0("The estimated OR for Shelf 4 vs Shelf 1:"))
```

```
## [1] "The estimated OR for Shelf 4 vs Shelf 1:"
```

```
grid.arrange(tableGrob(df4), ncol = 1)
```

	OR	1/OR
sugar	0.05	21.68
fat	0.77	1.3
sodium	0	291.48

```
print(paste0("The estimated CI & 1/CI for Shelf 4 vs Shelf 1:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 4 vs Shelf 1:"
```

```
grid.arrange(tableGrob(ci4), tableGrob(ci4_1), ncol = 2)
```

	low	up
sugar	0	0.61
fat	0.19	3.17
sodium	0	0.13

	low	up
sugar	1.65	285.48
fat	0.32	5.34
sodium	7.69	11041.33

With 95% confidence, the odds of a `Shelf 2` instead of a `Shelf 1` are 0.48 and change by 0.03 to 7.00 times when sugar is increased by 0.27 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 1` are 20.07 and change by 2.04 and 359.64 times when sugar is increased by 0.27 holding the other variables constant. Furthermore, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 1` are 21.68 and change by 1.65 and 285.48 times when sugar is increased by 0.27 holding the other variables constant.

These results are reinforced by the plots above. From the plot in **Part 1.4**, I know that as the sugar content increases, the probabilities for `Shelf 1` and `Shelf 2` increase, while `Shelf 3` and `Shelf 4` decrease. I can see that `Shelf 2` has the highest probability for high levels of sugar, which results in a low odds ratio as well as a conservative confidence interval. However, I can also see that `Shelf 3` and `Shelf 4` have probabilities of almost zero and, thus, have larger odds ratios and wider confidence intervals. I can also see these results in the plots from **Part 1.1**. The box plot reveals that `Shelf 2` has the most significant levels of sugar. The remaining shelves all have similar, lower levels. Therefore, when `Shelf 2` is compared to `Shelf 1`, the odds ratio is small and the confidence interval is somewhat conservative. Yet, `Shelf 3` and `Shelf 4` have similar distributions to `Shelf 1`, which results in higher odds ratios and wider confidence intervals. The coordinate plot further reveals that while `Shelf 2` has higher levels of sugar, `Shelf 3` and `Shelf 4` have substantially lower levels of sugar. This demonstrates why the odds ratios and confidence intervals for `Shelf 3` or `Shelf 4` instead of `Shelf 1` are so large comparatively.

Additionally, with 95% confidence, the odds of a `Shelf 2` instead of a `Shelf 1` are 0.30 and change by 0.08 to 1.15 times when fat is increased by 0.30 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 1` are 1.18 and change by 0.29 and 4.89 times when fat is increased by 0.30 holding the other variables constant. Furthermore, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 1` are 1.30 and change by 0.32 and 5.34 times when fat is increased by 0.30 holding the other variables constant.

These results are evident in the plots from **Part 1.1**. The box plot reveals that all of the shelves have similar, lower levels of fat. As a result, we see reasonable odds ratios and conservative confidence intervals when each of the shelves are compared to `Shelf 1`. The coordinate plot further demonstrates why the odds ratios and confidence intervals are so reasonable, as there is no clear distinction among the levels of fat across each of the shelves.

With 95% confidence, the odds of a `Shelf 2` instead of a `Shelf 1` are 55.90 and change by 2.28 to 1370.42 times when sodium is increased by 0.23 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 1` are 312.64 and change by 8.18 and 11939.03 times when sodium is

increased by 0.23 holding the other variables constant. Furthermore, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 1` are 291.48 and change by 7.69 and 11041.33 times when sodium is increased by 0.23 holding the other variables constant.

These results are also evident in the plots from **Part 1.1**. The box plot reveals that `Shelf 1` has the most significant levels of sodium. The remaining shelves all have similar, lower levels of sodium. As a result, we see high odds ratios and wide confidence intervals when those shelves are compared to `Shelf 1`. The coordinate plot further reveals that while `Shelf 1` has higher levels of sodium, `Shelf 3` has substantially lower levels of sodium. This demonstrates why the odds ratio and confidence interval for `Shelf 3` instead of `Shelf 1` are so large.

Base: Shelf 2

Next, I need to calculate the odds ratios and corresponding confidence intervals for `Shelves 3` and `4` compared to `Shelf 2`. As a result, the odds in this problem are constructed as $\mathrm{P(Y = j)/P(Y = 2)}$, $\mathrm{j = 3, 4}$ where category 2 = `Shelf 2`, 3 = `Shelf 3`, and 4 = `Shelf 4`.

The estimated odds ratios for `Shelf 3` vs. `Shelf 2` or `Shelf 4` vs. `Shelf 2` are calculated for each explanatory variable as $\hat{OR} = \exp(\mathrm{c}\hat{\beta}_{jr})$ for $\mathrm{j = 3, 4}$ and $\mathrm{r = 1,\ldots, 9}$.

By subtracting the appropriate log-odds, I can rewrite the equation from **Part 1.2** to compare any pair of response categories. So, to find $\log(\pi_j/\pi_{j'})$ where $j' \neq 1$ and $j' \neq \mathrm{j}$, one can compute

$$log\left(\frac{\widehat{\pi}_j}{\widehat{\pi}_{j'}}\right) = log(\pi_j) - log(\pi_{j'}) = log(\frac{\pi_j}{\pi_1}) - log(\frac{\pi_{j'}}{\pi_1})$$

I can refactor the model in order to accomplish this. As I have already computed the comparisons with `Shelf 1`, I will focus on the remaining combination of shelves.

```
cereal$Shelf2 <- factor(cereal$Shelf, levels = c(2,3,4,1))
mod.nominal2 <- multinom(Shelf2 ~  sugar + fat + sodium, data = cereal)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 33.794856
## iter  20 value 33.616990
## iter  30 value 33.595713
## iter  40 value 33.595185
## iter  50 value 33.595142
## final  value 33.595141
## converged
```

```
summary(mod.nominal2)
```

```
## Call:
## multinom(formula = Shelf2 ~ sugar + fat + sodium, data = cereal)
##
## Coefficients:
##   (Intercept)      sugar       fat    sodium
## 3    14.78681 -14.912714 -4.621863 -7.495409
## 4    14.39434 -14.090538 -4.934381 -7.189731
## 1    -6.89779  -2.692069 -4.063019 17.486515
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 3    5.513107 5.059891 2.752076 5.558649
## 4    5.496190 4.989776 2.745046 5.522355
## 1    6.486685 5.051034 2.306998 7.095999
```

```
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

Thus, the estimated regressions for the model are:

Equation 1: Shelf 3 vs. Shelf 2

$$log\left(\frac{\widehat{\pi}_{Shelf\_3}}{\widehat{\pi}_{Shelf\_2}}\right) = 14.7868 - 14.9127sugar - 4.6219fat - 7.4954sodium$$

Equation 2: Shelf 4 vs. Shelf 2

$$log\left(\frac{\widehat{\pi}_{Shelf\_4}}{\widehat{\pi}_{Shelf\_2}}\right) = 14.3943 - 14.0905sugar - 4.9344fat - 7.1897sodium$$

```
beta.hat2_3 <- coefficients(mod.nominal2)[1,2:4]
beta.hat2_4 <- coefficients(mod.nominal2)[2,2:4]

#Odds ratios for j = 3 vs. j = 2 (Shelf 3 vs. Shelf 2)
Shelf2_3_OR <- round(exp(c.value * beta.hat2_3), 2)
Shelf2_3_OR1 <- round(1/exp(c.value * beta.hat2_3), 2)
df2_3 <- data.frame(Shelf2_3_OR, Shelf2_3_OR1)
colnames(df2_3) <- c("OR", "1/OR")

#Odds ratios for j = 4 vs. j = 2 (Shelf 4 vs. Shelf 2)
Shelf2_4_OR <- round(exp(c.value * beta.hat2_4), 2)
Shelf2_4_OR1 <- round(1/exp(c.value * beta.hat2_4), 2)
df2_4 <- data.frame(Shelf2_4_OR, Shelf2_4_OR1)
colnames(df2_4) <- c("OR", "1/OR")

conf.beta2 <- confint(object = mod.nominal2, level = 0.95)

ci.OR2_3 <- exp(c.value * conf.beta2[2:4 ,1:2 ,1])
ci.OR2_4 <- exp(c.value * conf.beta2[2:4 ,1:2 ,2])

ci2_3 <- round(data.frame(low = ci.OR2_3[,1], up = ci.OR2_3[,2]), 2)
ci2_3_1 <- round(data.frame(low = 1/ci.OR2_3[,2], up = 1/ci.OR2_3[,1]),2)[c(1,2,3),]

ci2_4 <- round(data.frame(low = ci.OR2_4[,1], up = ci.OR2_4[,2]), 2)
ci2_4_1 <- round(data.frame(low = 1/ci.OR2_4[,2], up = 1/ci.OR2_4[,1]),2)[c(1,2,3),]
```

```
print(paste0("The estimated OR for Shelf 3 vs Shelf 2:"))
```

```
## [1] "The estimated OR for Shelf 3 vs Shelf 2:"
```

```
grid.arrange(tableGrob(df2_3), ncol = 1)
```

	OR	1/OR
sugar	0.02	56.06
fat	0.25	4
sodium	0.18	5.61

```
print(paste0("The estimated CI & 1/CI for Shelf 3 vs Shelf 2:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 3 vs Shelf 2:"
```

```
grid.arrange(tableGrob(ci2_3), tableGrob(ci2_3_1), ncol = 2)
```

	low	up
sugar	0	0.26
fat	0.05	1.26
sodium	0.01	2.19

	low	up
sugar	3.85	815.73
fat	0.79	20.18
sodium	0.46	68.7

```
print(paste0("The estimated OR for Shelf 4 vs Shelf 2:"))
```

```
## [1] "The estimated OR for Shelf 4 vs Shelf 2:"
```

```
grid.arrange(tableGrob(df2_4), ncol = 1)
```

	OR	1/OR
sugar	0.02	44.9
fat	0.23	4.39
sodium	0.19	5.23

```
print(paste0("The estimated CI & 1/CI for Shelf 4 vs Shelf 2:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 4 vs Shelf 2:"
```

```
grid.arrange(tableGrob(ci2_4), tableGrob(ci2_4_1), ncol = 2)
```

	low	up
sugar	0	0.31
fat	0.05	1.14
sodium	0.02	2.31

	low	up
sugar	3.2	629.54
fat	0.87	22.07
sodium	0.43	63

With 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 2` are 56.06 and change by 3.85 to 815.73 times when sugar is increased by 0.27 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 2` are 44.90 and change by 3.2 and 629.54 times when sugar is increased by 0.27 holding the other variables constant.

These results are also reinforced by the plots above. From the plot in **Part 1.4**, I know that `Shelf 2` has the highest probability for high levels of sugar. So, `Shelf 3` and `Shelf 4` have probabilities of almost zero and, thus, have larger odds ratios and wider confidence intervals. I can also see these results in the plots from **Part 1.1**. The box plot reveals that `Shelf 2` has the most significant levels of sugar. As a result when compared to `Shelf 2`, `Shelf 3` and `Shelf 4` have higher odds ratios and wider confidence intervals. The coordinate plot further reveals that while `Shelf 2` has higher levels of sugar, `Shelf 3` and `Shelf 4` have substantially lower levels of sugar. This demonstrates why the odds ratios and confidence intervals for `Shelf 3` or `Shelf 4` instead of `Shelf 2` are so large.

With 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 2` are 4.00 and change by 0.79 to 20.18 times when fat is increased by 0.30 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 2` are 4.39 and change by 0.87 and 22.07 times when fat is increased by 0.30

holding the other variables constant.

These results are evident in the plots from **Part 1.1**. The box plot reveals that all of the shelves have similar, lower levels of fat. As a result, we see reasonable odds ratios and conservative confidence intervals when each of the shelves are compared to `Shelf 2`. The coordinate plot further demonstrates why the odds ratios and confidence intervals are so reasonable, as there is no clear distinction among the levels of fat across each of the shelves.

With 95% confidence, the odds of a `Shelf 3` instead of a `Shelf 2` are 5.61 and change by 0.46 to 68.7 times when sodium is increased by 0.23 holding the other variables constant. Also, with 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 2` are 5.23 and change by 0.43 and 63 times when sodium is increased by 0.23 holding the other variables constant.

These results are also evident in the plots from **Part 1.1**. The box plot reveals that `Shelf 1` has the most significant levels of sodium. The remaining shelves all have similar, lower levels of sodium. As a result, we see very similar, yet somewhat high odds ratios and somewhat wide confidence intervals when the shelves are compared to `Shelf 2`. The coordinate plot further reveals that even though `Shelf 2` has medium levels of sodium, `Shelf 3` and `Shelf 4` have even lower levels of sodium. This demonstrates why the odds ratios and confidence intervals for `Shelf 3` and `Shelf 4` instead of `Shelf 2` are so large.

Base: Shelf 3

Last, I need to calculate the odds ratios and corresponding confidence intervals for `Shelf 4` compared to `Shelf 3`. As a result, the odds in this problem are constructed as $P(Y = j)/P(Y = 3)$, $j = 4$ where category $3 =$ `Shelf 3` and $4 =$ `Shelf 4`.

The estimated odds ratios for `Shelf 4` vs. `Shelf 3` is calculated for each explanatory variable as $\hat{OR} = \exp(c\hat{\beta}_{jr})$ for $j = 4$ and $r = 1, \ldots, 6$.

```
cereal$Shelf3 <- factor(cereal$Shelf, levels = c(3,4,1,2))
mod.nominal3 <- multinom(Shelf3 ~  sugar + fat + sodium, data = cereal)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 35.514143
## iter  20 value 33.667925
## iter  30 value 33.598476
## iter  40 value 33.595194
## iter  50 value 33.595146
## final  value 33.595139
## converged
```

```
summary(mod.nominal3)
```

```
## Call:
## multinom(formula = Shelf3 ~ sugar + fat + sodium, data = cereal)
##
## Coefficients:
##   (Intercept)     sugar        fat     sodium
## 4  -0.3922872  0.823257 -0.3129878  0.3038989
## 1 -21.6902795 12.221728  0.5572494 24.9882173
## 2 -14.7788984 14.911283  4.6234522  7.4801036
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 4    1.348656 1.954287 1.753219 2.155031
## 1    7.454121 4.889438 2.415212 8.083700
## 2    5.510731 5.059640 2.751679 5.556188
```

```
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

Thus, the estimated regression for the model is:

Equation 1: Shelf 4 vs. Shelf 3

$$log\left(\frac{\widehat{\pi}_{Shelf\_4}}{\widehat{\pi}_{Shelf\_3}}\right) = -0.3923 + 0.8233sugar - 0.3129fat + 0.3039sodium$$

```
beta.hat3_4 <- coefficients(mod.nominal3)[1,2:4]

#Odds ratios for j = 4 vs. j = 3 (Shelf 4 vs. Shelf 3)
Shelf3_4_OR <- round(exp(c.value * beta.hat3_4), 2)
Shelf3_4_OR1 <- round(1/exp(c.value * beta.hat3_4), 2)
df3_4 <- data.frame(Shelf3_4_OR, Shelf3_4_OR1)
colnames(df3_4) <- c("OR", "1/OR")

conf.beta3 <- confint(object = mod.nominal3, level = 0.95)

ci.OR3_4 <- exp(c.value * conf.beta3[2:4 ,1:2 ,1])

ci3_4 <- round(data.frame(low = ci.OR3_4[,1], up = ci.OR3_4[,2]), 2)
ci3_4_1 <- round(data.frame(low = 1/ci.OR3_4[,2], up = 1/ci.OR3_4[,1]),2)[c(1,2,3),]
```

```
print(paste0("The estimated OR for Shelf 4 vs Shelf 3:"))
```

```
## [1] "The estimated OR for Shelf 4 vs Shelf 3:"
```

```
grid.arrange(tableGrob(df3_4), ncol = 1)
```

	OR	1/OR
sugar	1.25	0.8
fat	0.91	1.1
sodium	1.07	0.93

```
print(paste0("The estimated CI & 1/CI for Shelf 4 vs Shelf 3:"))
```

```
## [1] "The estimated CI & 1/CI for Shelf 4 vs Shelf 3:"
```

```
grid.arrange(tableGrob(ci3_4), tableGrob(ci3_4_1), ncol = 2)
```

	low	up
sugar	0.44	3.51
fat	0.32	2.55
sodium	0.41	2.83

	low	up
sugar	0.28	2.25
fat	0.39	3.08
sodium	0.35	2.46

With 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 3` are 0.80 and change by 0.28 to 2.25 times when sugar is increased by 0.27 holding the other variables constant.

These results are still reinforced by the plots above. From the plot in **Part 1.4**, I know that `Shelf 3` and `Shelf 4` have similar probabilities of almost zero and, thus, result in a low odds ratio and a conservative

confidence interval when compared to each other. I can also see these results in the plots from **Part 1.1**. The box plot reveals that compared to `Shelf 3`, `Shelf 4` has a similar distribution and, thus, a lower odds ratio and a conservative confidence interval. The coordinate plot further reveals that while `Shelf 2` has higher levels of sugar, `Shelf 3` and `Shelf 4` have substantially lower levels of sugar. This demonstrates why the odds ratios and confidence intervals for `Shelf 4` instead of `Shelf 3` are so small.

With 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 3` are 1.10 and change by 0.39 to 3.08 times when fat is increased by 0.30 holding the other variables constant.

These results are evident in the plots from **Part 1.1**. The box plot reveals that all of the shelves have similar, lower levels of fat. As a result, we see a reasonable odds ratio and a conservative confidence interval when `Shelf 4` is compared to `Shelf 3`. The coordinate plot further demonstrates why the odds ratio and the confidence interval are so reasonable, as there is no clear distinction among the levels of fat across the shelves.

With 95% confidence, the odds of a `Shelf 4` instead of a `Shelf 3` are 0.93 and change by 0.35 to 2.46 times when sodium is increased by 0.23 holding the other variables constant.

These results are also evident in the plots from **Part 1.1**. The box plot reveals that `Shelf 3` and `Shelf 4` have very similar distributions of sodium. As a result, we see a low odds ratio and a conservative confidence interval when `Shelf 4` is compared to `Shelf 3`. The coordinate plot further reveals that `Shelf 3` and `Shelf 4` have similarly low levels of sodium. This demonstrates why the odds ratio and confidence interval for `Shelf 4` instead of `Shelf 3` are so small.

Overall, I am able to calculate the odds ratios and calculate the corresponding confidence intervals for each explanatory variable and relate the results to my previous plots and general analysis.

2 Alcohol, self-esteem and negative relationship interactions

This exercise 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 completed a variety of rating scales covering daily events in their lives and items related to self-esteem.

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_csv(paste0(here::here(),"/assignments/assignment_2/DeHartSimplified.csv"))
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   studyday = col_double(),
##   dayweek = col_double(),
##   numall = col_double(),
##   nrel = col_double(),
##   prel = col_double(),
##   negevent = col_double(),
##   posevent = col_double(),
##   gender = col_double(),
##   rosn = col_double(),
##   age = col_double(),
##   desired = col_double(),
##   state = col_double()
## )
```

```
head(dehart)
```

```
## # A tibble: 6 x 13
##      id studyday dayweek numall  nrel  prel negevent posevent gender  rosn   age
##   <dbl>    <dbl>   <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1     1        1       6      9  1    0        0.4      0.525      2   3.3  39.5
## 2     1        2       7      1  0    0        0.25     0.7        2   3.3  39.5
## 3     1        3       1      1  1    0        0.267    1          2   3.3  39.5
## 4     1        4       2      2  0    1        0.533    0.608      2   3.3  39.5
## 5     1        5       3      2  1.33 0.333    0.663    0.693      2   3.3  39.5
## 6     1        6       4      1  1    0        0.59     0.68       2   3.3  39.5
## # ... with 2 more variables: desired <dbl>, state <dbl>
```

```
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
```

```
paste("Sample Size: ", nrow(dehart))
```

```
## [1] "Sample Size:  623"
```

2.1 Question 2.1

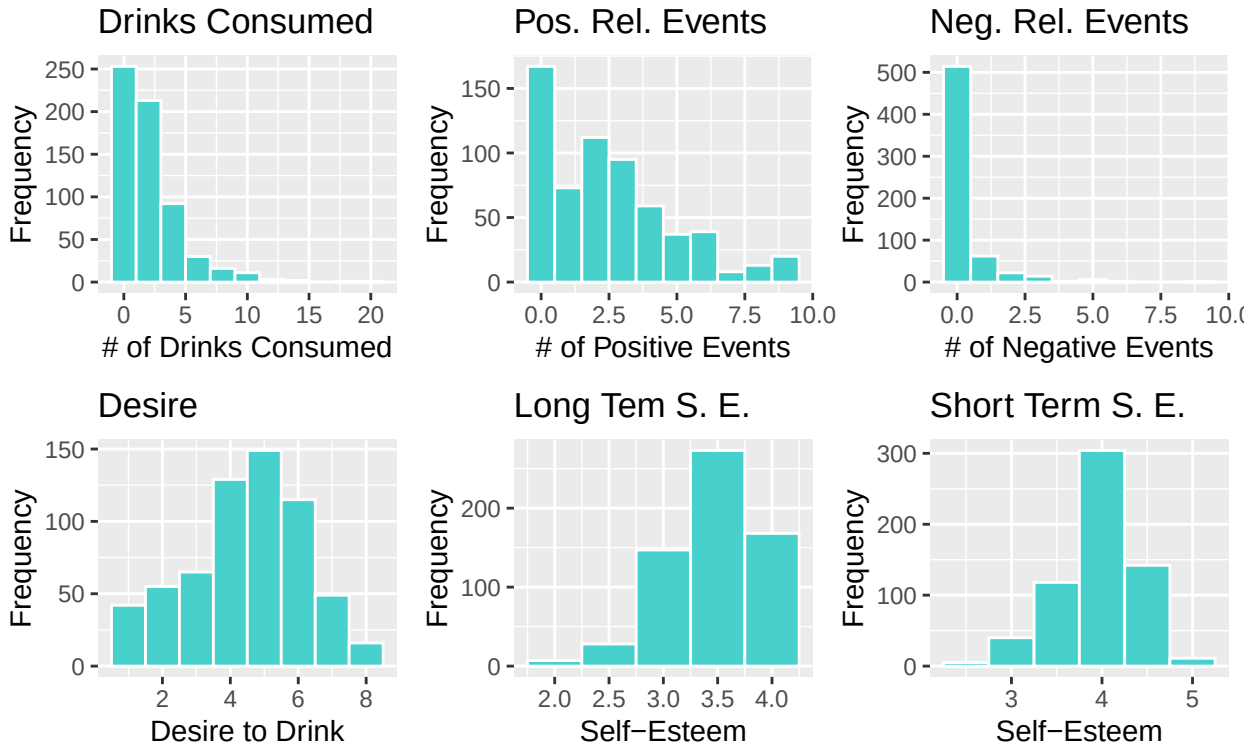
Conduct a thorough EDA of the data set, giving special attention to the relationships relevant to the researchers' hypotheses. Address the reasons for limiting the study to observations from only one day.

The hypothesis that I want to test is focused on how romantic relationship events as well as how self-esteem impacts an individuals' desire to drink alcohol and their overall alcoholic beverage consumption level. So, I will take a closer look at the variables in the data set (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717559/) that relate to these aspects, which include:

- the number of drinks consumed (`numall`) - Participants reported the number of standard alcoholic drinks they consumed the previous evening. They were instructed that "one drink equals one 12-oz. can or bottle of beer, one 4-oz. glass of wine, one 12-oz. wine cooler, or 1-oz. of liquor straight or in a mixed drink."
- the desire to drink alcohol (`desired`) - Participants reported their intent to drink alcohol that evening. This single item ("Do you intend to drink alcohol tonight") was assessed using a 8-point scale (1 = not at all, 8 = definitely).
- the negative romantic-relationship events (`nrel`) - Participants were instructed to check any events that occurred that day and rated how desirable or undesirable the event was on a 7-point scale (1 = extremely undesirable, 7 = extremely desirable). The study categorized 3 events as interpersonal. Negative interpersonal events were those events rated as 1, 2, or 3. The study then computed negative interpersonal events by separately summing the ratings for each event for each day and then averaged the items.
- the positive romantic-relationship events (`prel`) - Participants did the same for positive interpersonal events, which were those events rated as 5, 6, or 7. Positive interpersonal events were recoded so that slightly desirable was rated 1 and extremely desirable was rated 3. The study then computed positive interpersonal events by separately summing the ratings for each event for each day and then averaged the items.
- trait (long-term) self-esteem (`rosn`) - The study used Rosenberg's (1965) 10-item self-esteem scale that taps global self-evaluations (e.g., "I feel that I have a number of good qualities"). Participants responded using a 7-point scale (1 = strongly disagree, 7 = strongly agree). Negative items were reverse-scored, such that higher scores indicated higher self-esteem.

- state (short-term) self-esteem (`state`) - Previous research suggests that name-letter preferences are a valid indicator of self-esteem. So, participants were instructed to "trust your intuitions, work quickly, and report your gut impressions". Participants reported their liking for every letter of the alphabet using a 9-point scale (1 = dislike very much, 9 = like very much). The study then computed a liking score that was the difference between each participant's rating of his or her first and last name initials and the mean liking for these two letters provided by people whose names did not include that letter (thus, more positive numbers would indicate higher name-letter preferences). Participants' name-letter preferences were computed by taking the average liking scores for their first and last name initials, which formed the short-term self-esteem variable where higher scores indicate higher self-esteem.

```
p1 <- ggplot(dehart, aes(numall)) +
  geom_histogram(binwidth = 2, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Drinks Consumed", x = "# of Drinks Consumed",y="Frequency")
p2 <- ggplot(dehart, aes(prel)) +
  geom_histogram(binwidth = 1, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Pos. Rel. Events",x="# of Positive Events",y="Frequency")
p3 <- ggplot(dehart, aes(nrel)) +
  geom_histogram(binwidth = 1, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Neg. Rel. Events", x="# of Negative Events",y = "Frequency")
p4 <- ggplot(dehart, aes(desired)) +
  geom_histogram(binwidth = 1, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Desire", x = "Desire to Drink",y = "Frequency")
p5 <- ggplot(dehart, aes(rosn)) +
  geom_histogram(binwidth = 0.5, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Long Tem S. E.", x = "Self-Esteem",y = "Frequency")
p6 <- ggplot(dehart, aes(state)) +
  geom_histogram(binwidth = 0.5, fill = "mediumturquoise",col="white",size = 0.5)+
  labs(title="Short Term S. E.", x = "Self-Esteem",y = "Frequency")
egg::ggarrange(p1, p2, p3, p4, p5, p6, nrow = 2)
```



From the histograms, I can see:

- `numall` appears to have some outliers as well as a tail leading to the higher end of the distribution. Overall, the distribution appears to be primarily uni-modal with a peak around the mean (2.5) although the peak is at the lower end of the consumed drinks distribution. I can also see a non-normal distribution, with a positive skew towards the lower number of drinks consumed. Yet, since we know that our sample size is 623, which is greater than 30, I can rely on asymptotic assumptions of normality for our data. This will be important for the model assumptions moving forward. As a result, the majority of individuals in the study tend to have fewer drinks, yet there are some that have more.
- `prel` appears to have no major outliers or any fat tails. Overall, the distribution appears to be somewhat uni-modal with a slight peak around the mean (2.6), although there is a dip in frequency from zero to the mean. I can also see a non-normal distribution, with a slight positive skew towards the lower values of positive romantic relationship ratings. Yet, I can still rely on asymptotic assumptions of normality. In general, the majority of individuals in the study tend to give lower positive romantic relationship scores, although a fair bit give higher scores.
- `nrel` appears to have some outliers as well as a tail leading towards higher ratings of the distribution. Overall, the distribution appears to be primarily uni-modal with a peak around the mean (0.4). I can also see a non-normal distribution, with a distinct positive skew towards the lower values of negative romantic romantic relationship ratings. Yet, I can still rely on asymptotic assumptions of normality. As a result, the majority of individuals in the study tend to give lower negative romantic relationship scores, although a few give higher scores.
- `desired` appears to have no major outliers or any fat tails. Overall, the distribution appears to be mostly uni-modal with a peak around the mean (4.5). I can see that the distribution is practically normal with a slight negative skew towards higher desires to drink. Yet, I can still rely on asymptotic assumptions of normality. In general, the individuals in the study tend to have mostly average level desires to drink alcohol.
- `rosn` appears to have no major outliers or any fat tails. Overall, the distribution appears to be uni-modal with a peak around the mean (3.4). I can also see a mostly normal distribution, with a slight negative skew towards the higher long-term self-esteem scores. Yet, I can still rely on asymptotic assumptions of normality. In general, the majority of individuals in the study tend to have somewhat high levels of long-term self-esteem, although a few have lower levels.
- `state` appears to have no major outliers or any fat tails. Overall, the distribution appears to be uni-modal with a peak around the mean (3.9). I can also see a mostly normal distribution, with a very slight negative skew towards the higher short-term self-esteem scores. Yet, I can still rely on asymptotic assumptions of normality. In general, the majority of individuals in the study tend to have somewhat high levels of short-term self-esteem, although a small amount have lower levels.

As there appear to be some outliers in the variable `nrel`, so I will take a closer as it is an important feature in the model.

```
dehart[which(dehart$nrel > 7),]
```

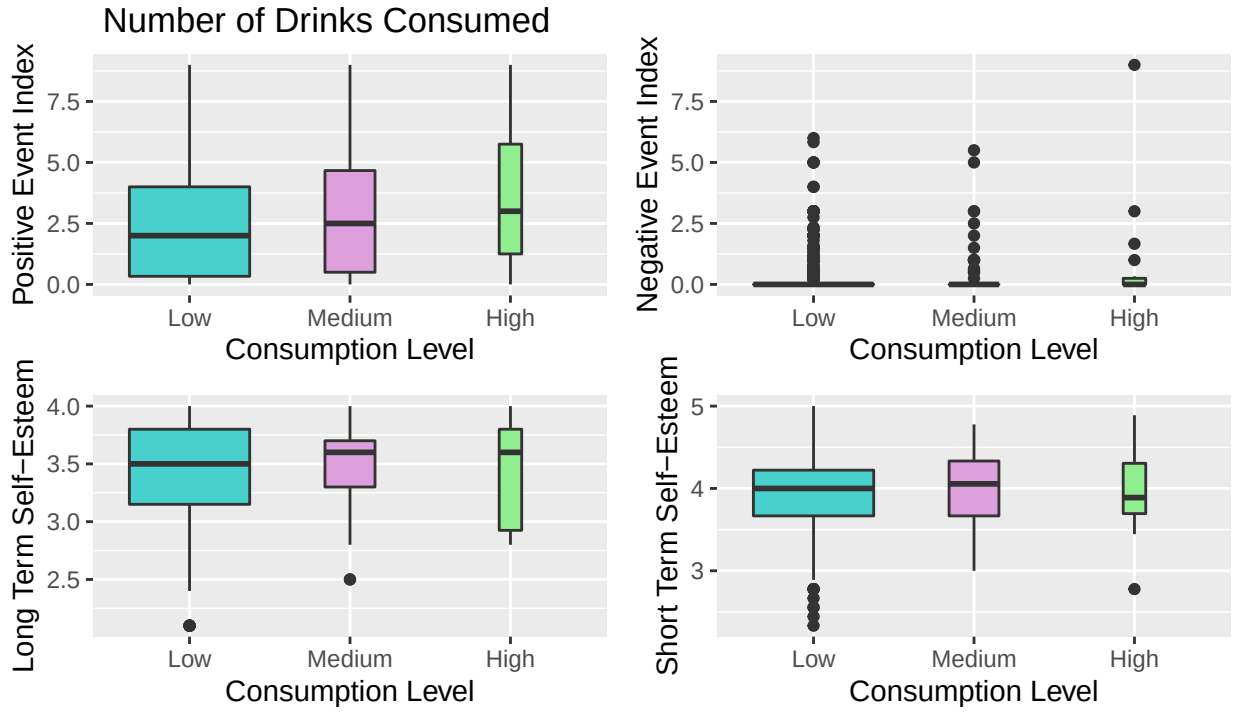
```
## # A tibble: 1 x 13
##      id studyday dayweek numall  nrel  prel negevent posevent gender  rosn   age
##   <dbl>    <dbl>   <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1   153        4       5     10     9     0      1.6      1.4      2   3.6  30.9
## # ... with 2 more variables: desired <dbl>, state <dbl>
```

Although the value for `nrel` is high, "9" is within the range of valid values and the other responses (such as 10 drinks consumed on a Friday night with no positive romantic relationship interactions). Thus, it does not seem to be erroneous and I will not remove the data point.

For the hypothesis I am trying to test, it is important to consider the relationship between these different variables and `numall`. To start, I will create an ordinal categorical variable to represent `numall`. These will consist of "Low" (less than 5), "Medium" (greater than 5 and less than 10), and "High" (greater than 10) values. This will give me an overall, more general understanding of the various interactions.

```
library(forcats)
dehart$drinks_cat <- ifelse(dehart$numall < 5, "Low",
                            ifelse(dehart$numall < 10, "Medium", "High"))
dehart_t <- dehart[which(!is.na(dehart$drinks_cat)),]
```

```
p1 <- ggplot(dehart_t,aes(fct_reorder(drinks_cat,numall),prel))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(drinks_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(title=" Number of Drinks Consumed",x="Consumption Level",
       y="Positive Event Index")
p2 <- ggplot(dehart_t,aes(fct_reorder(drinks_cat,numall),nrel))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(drinks_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Consumption Level",y="Negative Event Index")
p3 <- ggplot(dehart_t,aes(fct_reorder(drinks_cat,numall),rosn))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(drinks_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Consumption Level",y="Long Term Self-Esteem")
p4 <- ggplot(dehart_t,aes(fct_reorder(drinks_cat,numall),state))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(drinks_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Consumption Level",y="Short Term Self-Esteem")
egg::ggarrange(p1, p2, p3, p4, nrow = 2)
```



From the box-plots, I can see:

- `prel` appears to have slightly different distributions for each of the consumption levels. Although they are similar, individuals that have "Low" consumption levels tend to have slightly lower ratings of

positive romantic relationship events. Additionally, the ratings of positive romantic relationship events seem to increase slightly with each increased consumption level. However, the "Low" category has the greatest density of individuals, with density decreasing with each increased consumption level. As a result, this could be an indication that the number of alcoholic drinks consumed does not necessarily have a strong impact on the ratings of positive romantic relationship events.

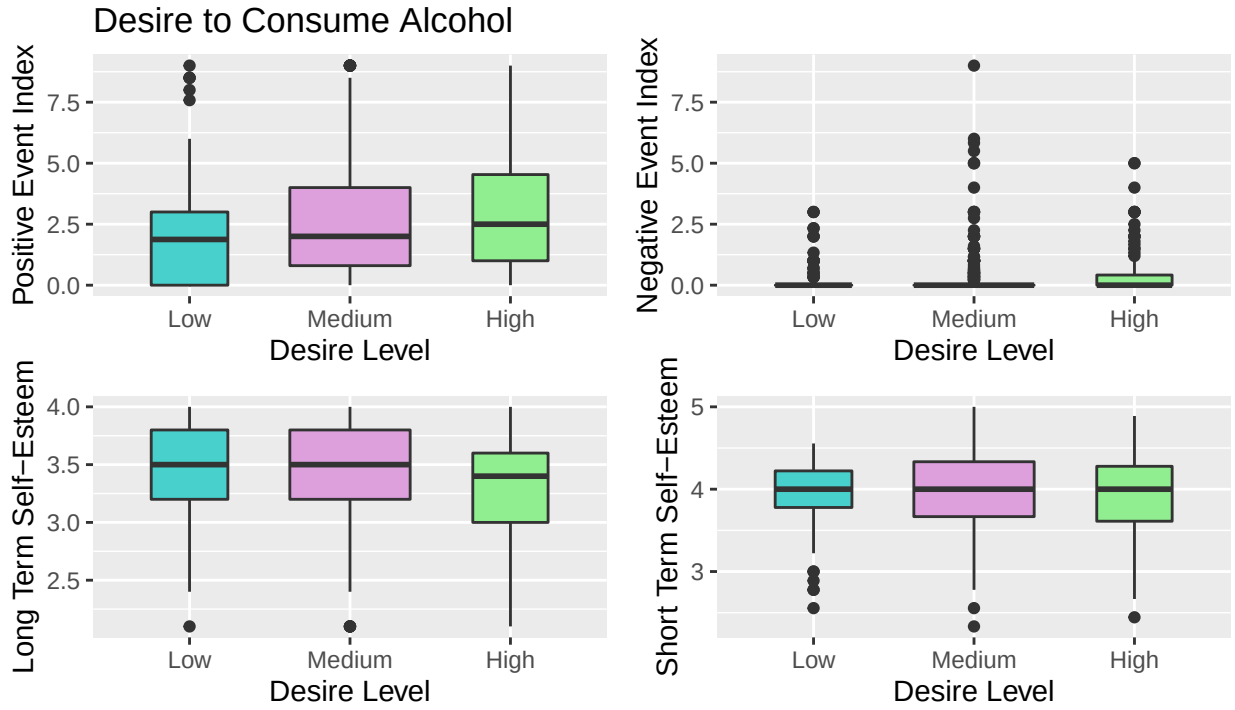
- `nrel` appears to have similarly low, yet slightly different distributions for each of the consumption levels. Although they are similar, individuals that have "High" consumption levels tend to have slightly higher ratings of negative romantic relationship events. Yet, each category has generally low ratings, with outliers representing higher ratings. However, the "Low" category still has the greatest density of individuals, with density decreasing with each increased consumption level. Overall, this could be an indication that the number of alcoholic drinks consumed might actually have an impact on the ratings of negative romantic relationship events.
- `rosn` appears to have similarly high, yet slightly different distributions for each of the consumption levels. Although they are similar, individuals that have "Low" consumption levels tend to have slightly lower levels of long-term self-esteem, whereas the "High" consumption category has a wider distribution of self-esteem levels. However, the "Low" category has the greatest density of individuals, with density decreasing with each increased consumption level. As a result, this could be an indication that the number of alcoholic drinks consumed does not necessarily have a strong relationship with long-term self-esteem.
- `state` appears to have quite similar distributions for each of the consumption levels. Although they are similar, individuals that have "High" consumption levels tend to have slightly lower levels of short-term self-esteem, whereas the "Low" consumption category has more outliers ranging towards the lower end of self-esteem levels. However, the "Low" category has the greatest density of individuals, with density decreasing with each increased consumption level. As a result, this could be an indication that the number of alcoholic drinks consumed does not necessarily have a strong relationship with short-term self-esteem.

For the hypothesis I am trying to test, it is also important to consider the relationship between these different variables and `desired`. So, I will set create a different ordinal categorical variable to represent `desired`. These will consist of "Low" (less than 3), "Medium" (greater than 3 and less than 6), and "High" (greater than 6) values. This will give me an overall, more general understanding of the various interactions.

```
dehart$des_cat <- ifelse(dehart$desired <= 3, "Low",
                         ifelse(dehart$desired < 6, "Medium", "High"))
dehart_t <- dehart[which(!is.na(dehart$des_cat)),]
```

```
p1 <- ggplot(dehart_t,aes(fct_reorder(des_cat,desired),prel))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(des_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(title="Desire to Consume Alcohol",
       x="Desire Level",y="Positive Event Index")
p2 <- ggplot(dehart_t,aes(fct_reorder(des_cat,desired),nrel))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(des_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Desire Level",y="Negative Event Index")
p3 <- ggplot(dehart_t,aes(fct_reorder(des_cat,desired),rosn))+
  theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(des_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Desire Level",y="Long Term Self-Esteem")
p4 <- ggplot(dehart_t,aes(fct_reorder(des_cat,desired),state))+
  theme(legend.position="none") +
```

```
  geom_boxplot(varwidth=T, aes(fill = factor(des_cat))) +
  scale_fill_manual(values=c("lightgreen","mediumturquoise","plum")) +
  labs(x="Desire Level",y="Short Term Self-Esteem")
egg::ggarrange(p1, p2, p3, p4, nrow = 2)
```



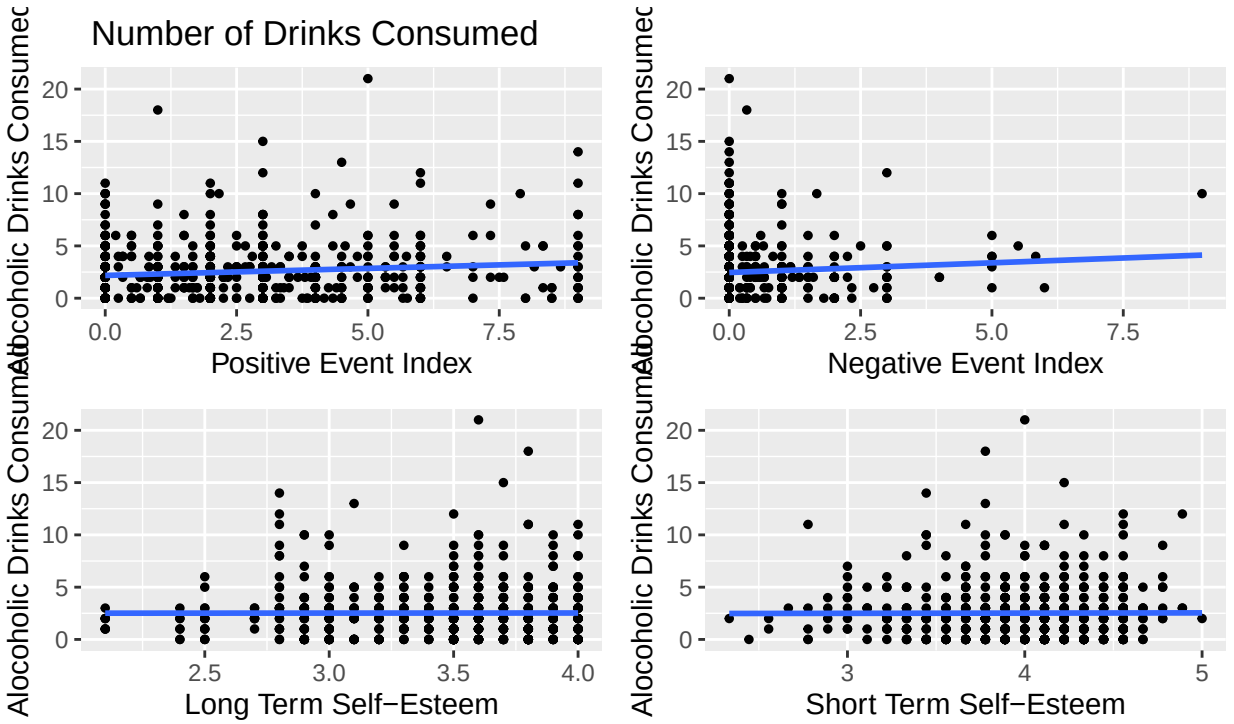
From the box-plots, I can see:

- `prel` appears to have slightly different distributions for each of the desire levels. Although they are similar, individuals that have "Low" desire levels tend to have slightly lower ratings of positive romantic relationship events. Additionally, the ratings of positive romantic relationship events seem to increase slightly with each increased desire level. As a result, this could be an indication that the desire for alcoholic drinks does not necessarily have a strong impact on the ratings of positive romantic relationship events.
- `nrel` appears to have similarly low, yet slightly different distributions for each of the desire levels. Although they are similar, individuals that have "High" desire levels tend to have slightly higher ratings of negative romantic relationship events. Yet, each category has generally low ratings, with outliers representing higher ratings. Overall, this could be an indication that the desire for alcoholic drinks might actually have an impact on the ratings of negative romantic relationship events.
- `rosn` appears to have similarly high, yet slightly different distributions for each of the desire levels. Although they are similar, individuals that have "Low" desire levels tend to have slightly higher levels of long-term self-esteem, whereas the "High" desire category has a wider distribution of self-esteem levels. As a result, this could be an indication that the number of alcoholic drinks does not necessarily have a strong relationship with long-term self-esteem.
- `state` appears to have quite similar distributions for each of the desire levels. The main distinction between each distribution is the density of points that belong to each category. Otherwise, the width of their distributions vary slightly in relation to their outliers. As a result, this could be an indication that the desire for alcoholic drinks does not necessarily have a strong relationship with short-term self-esteem.

However, these box plots only provide a broad understanding of the relationships. So now, I can take a look

at the more detailed relationship through my scatter plots, as well as the predicted linear regression lines.

```
p1 <- ggplot(dehart, aes(x=prel, y=numall)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(title=" Number of Drinks Consumed",y="Alocoholic Drinks Consumed",
       x="Positive Event Index")
p2 <- ggplot(dehart, aes(x=nrel, y=numall)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Alocoholic Drinks Consumed",x="Negative Event Index")
p3 <- ggplot(dehart, aes(x=rosn, y=numall)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Alocoholic Drinks Consumed",x="Long Term Self-Esteem")
p4 <- ggplot(dehart, aes(x=state, y=numall)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Alocoholic Drinks Consumed",x="Short Term Self-Esteem")
egg::ggarrange(p1, p2, p3, p4, nrow = 2)
```

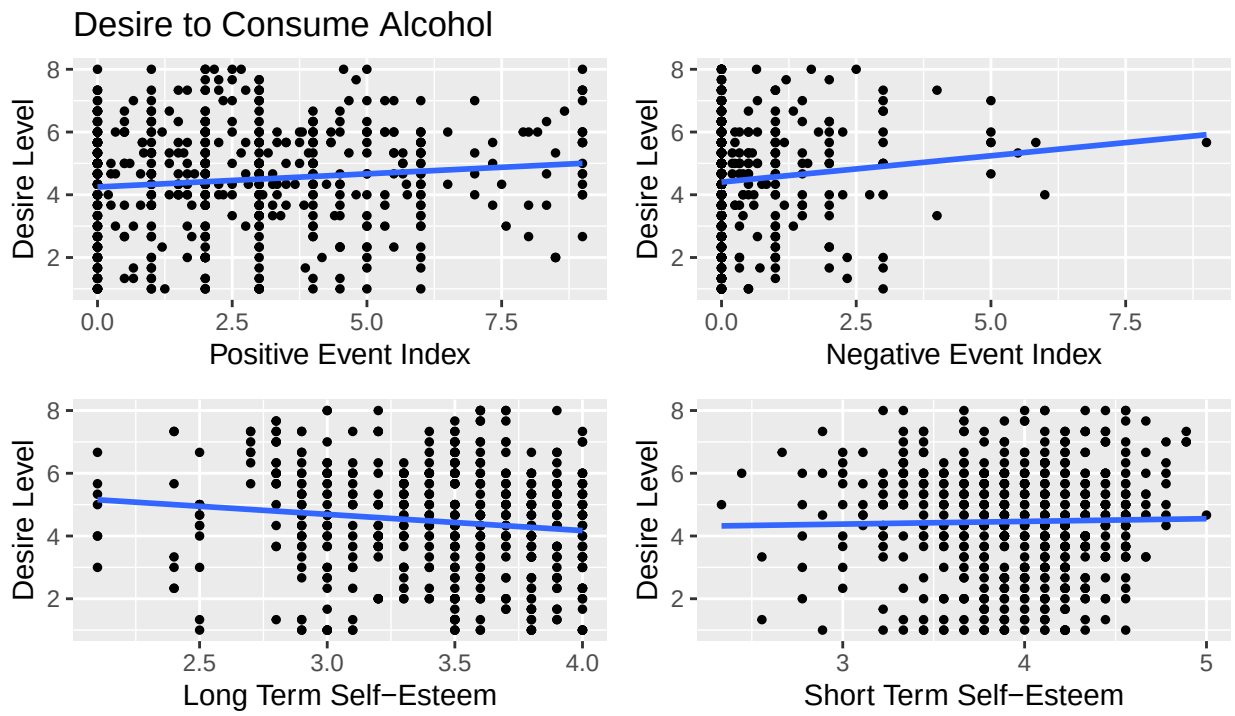


From the scatter plots, I can see:

- `prel` appears to have a slight positive relationship between the number of alcoholic drinks consumed and the ratings of positive romantic relationship events. This means that as the ratings increase, the number of consumed drinks is also likely to increase a small amount.
- `nrel` appears to also have a slight positive relationship between the number of alcoholic drinks consumed and the ratings of negative romantic relationship events. This means that as the ratings increase, the number of consumed drinks is also likely to increase a small amount.
- `rosn` appears to have a practically neutral relationship between the number of alcoholic drinks consumed and the level of long-term self-esteem. This means that as the level of self-esteem increases, the number of consumed drinks is not likely to change very much.
- `state` appears to also have a practically neutral relationship between the number of alcoholic drinks consumed and the level of short-term self-esteem. This means that as the level of self-esteem increases,

the number of consumed drinks is not likely to change very much.

```
p1 <- ggplot(dehart, aes(x=prel, y=desired)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(title="Desire to Consume Alcohol",y="Desire Level",
       x="Positive Event Index")
p2 <- ggplot(dehart, aes(x=nrel, y=desired)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Desire Level",x="Negative Event Index")
p3 <- ggplot(dehart, aes(x=rosn, y=desired)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Desire Level",x="Long Term Self-Esteem")
p4 <- ggplot(dehart, aes(x=state, y=desired)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Desire Level",x="Short Term Self-Esteem")
egg::ggarrange(p1, p2, p3, p4, nrow = 2)
```

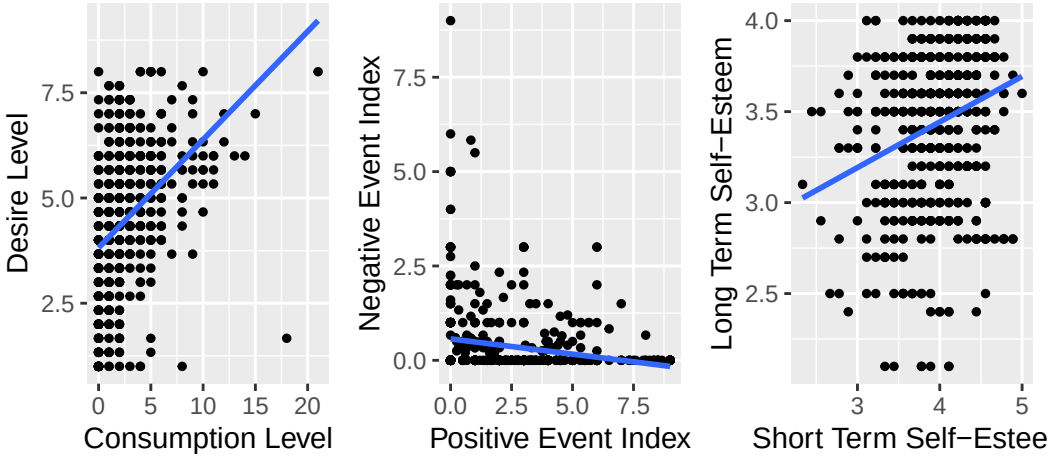


From the scatter plots, I can see:

- `prel` appears to have a slight positive relationship between the desire for alcoholic drinks and the ratings of positive romantic relationship events. This means that as the ratings increase, the desire for drinks is also likely to increase a small amount.
- `nrel` appears to also have a slight positive relationship between the desire for alcoholic drinks and the ratings of negative romantic relationship events. This means that as the ratings increase, the desire for drinks is also likely to increase a small amount.
- `rosn` appears to have a slightly negative relationship between the desire for alcoholic drinks and the level of long-term self-esteem. This means that as the level of self-esteem increases, the desire for drinks is likely to decrease a small amount.
- `state` appears to have an almost neutral, slightly positive relationship between the desire for alcoholic drinks and the level of short-term self-esteem. This means that as the level of self-esteem increases, the number of consumed drinks is likely to increase a small amount.

I can also examine the scatter plots for the desire for drinks versus the number of drinks consumed, the negative and positive romantic relationship ratings, as well as long and short term self-esteems. This will provide a more nuanced look into how these variables might relate to each other.

```
p1 <- ggplot(dehart, aes(x=numall, y=desired)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Desire Level",x="Consumption Level")
p2 <- ggplot(dehart, aes(x=prel, y=nrel)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Negative Event Index",x="Positive Event Index")
p3 <- ggplot(dehart, aes(x=state, y=rosn)) +
  geom_point(fill="gray", size = 1) + geom_smooth(method="lm", se=F) +
  labs(y="Long Term Self-Esteem",x="Short Term Self-Esteem")
egg::ggarrange(p1, p2, p3, nrow = 1)
```



From the scatter plots, I can see:

- `numall` & `desired` appear to have a very positive relationship between the desire to drink and the number of drinks consumed. This makes sense in this context as it means that as an individuals' desire to drink increases, the number of drinks they consume also increases.
- `prel` & `nrel` appear to have a slightly negative relationship between the negative and positive relationship ratings. This makes sense in this context as it means that as the ratings for positive romantic relationship events increase, the ratings for negative romantic relationship events decrease.
- `rosn` & `state` appear to have a positive relationship between the long and short term self-esteem levels. This makes sense in this context as it means that as the short-term self-esteem levels increase, the long-term self-esteem levels also increase.

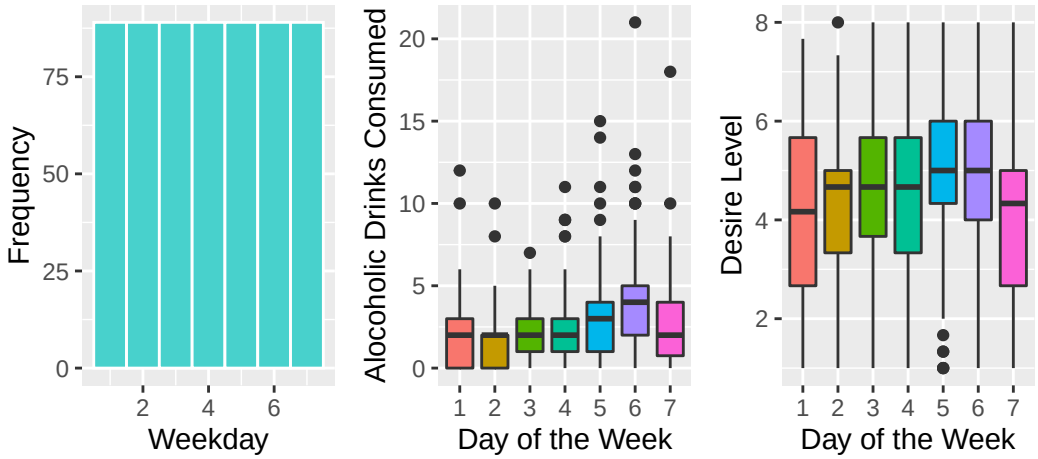
However, the repeated-measures aspect of this study (the 7 days of measurements on each participant) is beyond the current goal of this exercise. The repeated days would relate more to a time-series problem, which would not be useful in a ordinal or count response model. It would also violate the different models assumption of independence. So, for the sake of this exercise, I will need to subset to data to a single day in order to conduct my analysis.

I will conduct further diagnostics in order to determine which day I should use for my analysis. The variable that I will focus on is:

- the day of the week (`dayweek`) - Each day is represented by a number for its given day of the week. For example, "Monday" is considered as "1" and "Sunday" is considered as "7".

```
p1 <- ggplot(dehart, aes(dayweek)) + labs(x = "Weekday",y="Frequency") +
  geom_histogram(binwidth = 1, fill = "mediumturquoise",col="white",size = 0.5)
```

```
p2 <- ggplot(dehart,aes(factor(dayweek),numall))+theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(dayweek))) +
  labs(x="Day of the Week",y="Alocoholic Drinks Consumed")
p3 <- ggplot(dehart,aes(factor(dayweek),desired))+theme(legend.position="none") +
  geom_boxplot(varwidth=T, aes(fill = factor(dayweek))) +
  labs(x="Day of the Week",y="Desire Level")
egg::ggarrange(p1, p2, p3, nrow = 1)
```



From the histogram, I can see that there is a uniform distribution of days within the study. However, the middle box plot reveals the amount of drinking that takes place on each day. As "5", "6", and "7" represent "Friday", "Saturday", and "Sunday", I can see that there tends to be higher amounts of drinking that takes place on the weekends. Therefore, drinking is less common during the week (days "1" through "4"). This makes sense as people typically are busy during the week and are more social on the weekends. Increased social activity, which includes romantic relationship events, can likely be associated with consuming alcohol. Yet, the box plot on the right reflects the desire to drink alcohol on each day. Although the desire is slightly higher on the weekends, there is still a decent desire to drink throughout the week. Therefore, I want to choose a day that both desire and actual consumption are relatively high.

Because part of my goal is to measure a count response, I want to consider the Poisson assumptions with the day of the week that I choose for my subset. The assumptions to consider are:

- *Random Poisson Response* - The response variable is a count per unit of time or space, described by a Poisson distribution.
- *Independence* - The observations must be independent of one another, for example, they are not serially correlated nor do they form clusters.
- *Systematic Linearity* - The log of the mean rate, $\log(\lambda)$, must be a linear function of x such that $\beta_0 + \beta_1 * x_x + ... + \beta_p * x_p$.
- *Log Link Function* - The Poisson distribution requires that $\mu > 0$, and the identity link can lead to a non-positive value of μ for particular values of the explanatory variables. A consequence of the log link function is that the explanatory variables affect the response mean in multiplicative way.

I know that the *Random Poisson Response* and *Independence* assumptions are maintained based on the `numall` variable in the data subset. Yet, the day that I choose should help to maintain the *Systematic Linearity* and *Log Link Function* assumptions. So, for the purpose of my analysis, I want to choose a day in which the desire to drink alcohol, actual alcoholic drink consumption, and romantic social interactions (positive or negative) all take place more frequently. I can see that "Friday" has a decent effect. Friday has higher averages for both the desire to drink and actual consumed beverages. But, these are not the most extreme values, which makes Friday more generally representative of a general sample set. As a result, I will focus on each participant's first Friday in the study. This makes sense in this context because Friday is a day

in which the desire to drink, alcohol consumption, and romantic social interaction might normally be high. Therefore, I will model the desire to drink - the ordinal response - and the number of drinks consumed — the count response — as a function of the variables measuring total positive and negative romantic relationship events, as well as long and short term self-esteem.

```
dehart_fri <- dehart[which(dehart$dayweek == 5),]
head(dehart_fri)
```

```
## # A tibble: 6 x 15
##      id studyday dayweek numall  nrel  prel negevent posevent gender  rosn   age
##   <dbl>    <dbl>   <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1     1        7       5      4  0      3.5    0.633     0.8        2   3.3  39.5
## 2     2        3       5      0  2      3      0.668     1.94       2   3.9  38.0
## 3     4        3       5      3  0.25   6      0.572     1.42       2   3.7  30.0
## 4     5        2       5      3  0      0      0.4       0.9        2   3    27.6
## 5     7        6       5      1  0      2      0         1.1        2   3.3  40.4
## 6     9        3       5      1  0      4.5    1.1       1.73       2   3.5  33.0
## # ... with 4 more variables: desired <dbl>, state <dbl>, drinks_cat <chr>,
## #   des_cat <chr>
```

```
summary(dehart_fri)[,c(4,12,5,10)]
```

```
##      numall          desired           nrel             rosn
##  Min.   : 0.000   Min.   :1.000   Min.   :0.0000   Min.   :2.100
##  1st Qu.: 1.000   1st Qu.:4.333   1st Qu.:0.0000   1st Qu.:3.200
##  Median : 3.000   Median :5.000   Median :0.0000   Median :3.500
##  Mean   : 2.966   Mean   :4.816   Mean   :0.4015   Mean   :3.436
##  3rd Qu.: 4.000   3rd Qu.:6.000   3rd Qu.:0.0000   3rd Qu.:3.800
##  Max.   :15.000   Max.   :8.000   Max.   :9.0000   Max.   :4.000
##
```

```
paste("Sample Size: ", nrow(dehart_fri))
```

```
## [1] "Sample Size:  89"
```

Now, I have a subset of the data that is ready for analysis. There are no missing values and the new sample size is 89.

2.2 Question 2.2

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.

Higher Alcohol Consumption

To test the hypothesis that negative relationship interactions are associated with higher alcohol consumption, I will begin with a simplistic model using the ratings of negative events with romantic partners to estimate the mean number of drinks consumed. So, the first model I fit is $Y_i \sim Po(\mu_i)$ with

$$log(\mu_i) = \beta_0 + \beta_1 * nrel$$

where Y_i is the number of drinks consumed (`numall`) for person i = 1,. . . , 89.

```
mod.neg <- glm(formula = numall ~ nrel, family = poisson(link = "log"),
               data = dehart_fri)
```

```
summary(mod.neg)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = dehart_fri)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7346  -1.2604  -0.2269   0.6499   5.0629
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.04277    0.06545  15.933   <2e-16 ***
## nrel         0.09201    0.03813   2.413   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 234.89  on 88  degrees of freedom
## Residual deviance: 230.04  on 87  degrees of freedom
## AIC: 448.3
##
## Number of Fisher Scoring iterations: 5
```

The estimated model is $log(\hat{\mu}) = 1.04 + 0.09 * nrel$. The parameter estimates are $\beta_0 = 1.04$ and $\beta_1 = 0.09$. The positive "slope" parameter indicates that the number of drinks is increasing as the negative romantic events increase.

Next, I can calculate the odds ratio and the corresponding confidence interval to understand the impact of a change in negative romantic events on the consumption of alcoholic beverages.

```
100*(exp(mod.neg$coefficients[2]) - 1)
```

```
##     nrel
## 9.637751
```

As a result, a 1 unit increase in an individuals negative romantic relationship is associated with a 9.64% increase in the mean number of alcoholic drinks consumed.

```
beta1.int <- confint(mod.neg, parm = "nrel", level = 0.95)
100*(exp(beta1.int) - 1)
```

```
##     2.5 %    97.5 %
##  1.100186 17.497805
```

Thus, a 1-unit increase in negative romantic events leads to an estimated $\hat{PC} = 9.64\%$ increase in the number of alcoholic beverages with 95% profile LR confidence interval $1.10\% < PC < 17.50\%$.

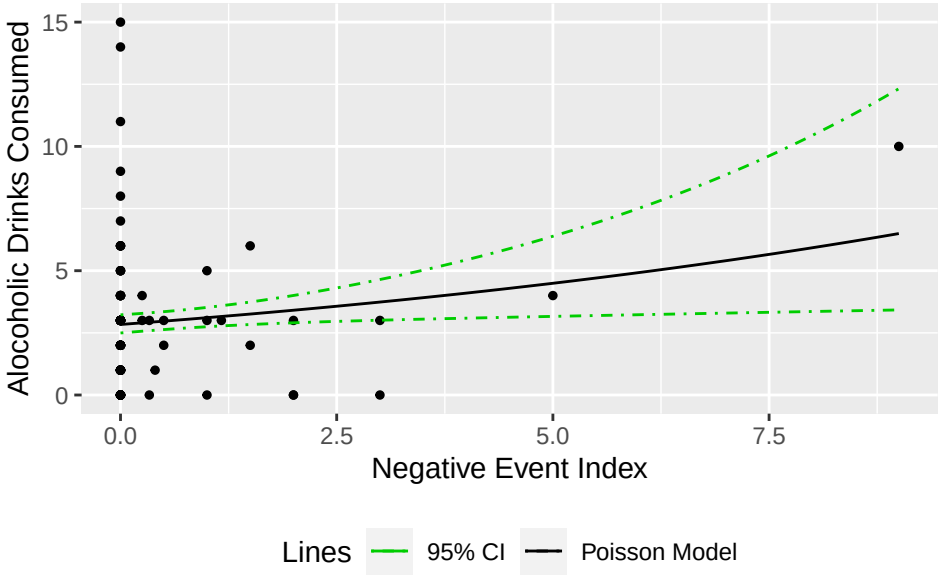
I can add the regression line and the confidence interval bands to my original scatter plot to visualize the model.

```
ci.mu <- function(newdata, mod.fit.obj, alpha) {
  lin.pred.hat<-predict(object=mod.fit.obj,newdata=newdata,type="link",se=TRUE)
  lower <- exp(lin.pred.hat$fit - qnorm(1-alpha/2) * lin.pred.hat$se)
  upper <- exp(lin.pred.hat$fit + qnorm(1-alpha/2) * lin.pred.hat$se)
```

```
  list(lower = lower, upper = upper)
}
```

```
beta0 <- mod.neg$coefficients[1]
beta1 <- mod.neg$coefficients[2]

ggplot(dehart_fri, aes(x=nrel, y=numall)) + theme(legend.position="bottom") +
  geom_point(fill="gray", size = 1) +
  stat_function(aes(colour = "Poisson Model"), show.legend = TRUE,
    fun=function(x) exp(beta0+beta1*x)) +
  stat_function(aes(colour = "95% CI"), show.legend = TRUE,lty = "dotdash",
    fun = function(x) ci.mu(newdata = data.frame(nrel = x),
                        mod.fit.obj = mod.neg, alpha = 0.05)$lower) +
  stat_function(aes(colour = "95% CI"), show.legend = FALSE,lty ="dotdash",
    fun = function(x) ci.mu(newdata = data.frame(nrel = x),
                        mod.fit.obj = mod.neg, alpha = 0.05)$upper) +
  scale_colour_manual("Lines", values = c("green3", "black")) +
  labs(y="Alocoholic Drinks Consumed",x="Negative Event Index")
```



This plot reaffirms the positive association between negative romantic relationship events and the consumption of alcoholic beverages.

Overall, an individual that has had higher negative romantic relationship events is more likely to have consumed a higher number of alcoholic beverages. Thus, there is an evident relation between drinking and negative relationship interactions.

Then, I can test to see if the change is statistically significant.

```
Anova(mod.neg)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel   4.8482  1    0.02767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This change is statistically significant according to both a Wald test (p-value = 0.0158) and the LRT (p-value = 0.0277).

Therefore, there is evidence that negative romantic relationship interactions are associated with higher alcohol consumption.

Increased Desire to Drink

To continue testing the hypothesis that negative relationship interactions are associated with an increased desire to drink, I need to build another simplistic model using the ratings of negative events with romantic partners to estimate the level of desire to drink alcohol. This is an ordinal response and cumulative logits are expressed as

$$logit(P(Y \leq j)) = log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} \ldots \pi_J}\right)$$

So, the next model I fit is:

$$logit(P(Y \leq j)) = \beta_{j0} + \beta_1 * nrel$$

In the study, participants desire to drink was assessed using a 8-point scale (1 = not at all, 8 = definitely). Thus, I will use the ordering of "No Desire" (Y = 1) < ... < "Definitely Desire" (Y = 8) and fit a proportional odds model to these data.

To start, I will create the required ordering of the levels for the `desired` variable.

```
dehart_fri$desire_whole <- round(dehart_fri$desired,0)
dehart_fri$desire.order<-factor(dehart_fri$desire_whole,levels=c(1,2,3,4,5,6,7,8))
levels(dehart_fri$desire.order)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8"
```

The new variable named `desire.order` contains the factor with the proper ordering of its levels. The proportional odds model is then estimated using `polr()`.

```
library(package = MASS)
mod.fit.ord <- polr(desire.order ~ nrel, data = dehart_fri, method = "logistic")
```

```
summary(mod.fit.ord)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = desire.order ~ nrel, data = dehart_fri, method = "logistic")
##
## Coefficients:
##       Value Std. Error t value
## nrel 0.1655     0.1417   1.168
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -2.7719  0.4623    -5.9953
## 2|3 -2.0178  0.3384    -5.9627
## 3|4 -1.5456  0.2866    -5.3920
## 4|5 -0.6702  0.2318    -2.8907
## 5|6  0.7581  0.2371     3.1980
## 6|7  2.2886  0.3675     6.2274
```

```
## 7|8  3.8709  0.7223     5.3591
##
## Residual Deviance: 313.5271
## AIC: 329.5271
```

This results in the model

$$logit(\hat{P}(Y \leq j)) = \hat{\beta_{j0}} - \beta_1 * 0.17$$

where $\hat{\beta_{10}} = -2.77$, $\hat{\beta_{20}} = -2.02$, $\hat{\beta_{30}} = -1.55$, $\hat{\beta_{40}} = -0.67$, $\hat{\beta_{50}} = 0.76$, $\hat{\beta_{60}} = 2.29$, and $\hat{\beta_{70}} = 3.87$. Overall, `nrel` does not seem to have a statistically significant p-value. But, I can observe that the intercept values increase as the desire level increases. As there is a positive "slope" coefficient and incrementally increasing equation intercepts, it is possible that negative relationship interactions are somewhat positively associated with an increased desire to drink.

Next, I can calculate the corresponding confidence interval to understand the impact of a change in negative romantic events on the desire to drink alcoholic beverages.

```
sd.d<-apply(X=dehart_fri[,c(5)],MARGIN=2,FUN=sd)
c.value <- c(sd.d)
conf.beta <- confint(object = mod.fit.ord, level = 0.95)
ci <- exp(c.value*(-conf.beta))
round(data.frame(low = 1/ci[1], up = 1/ci[2], row.names = c("nrel")), 2)
```

```
##       low   up
## nrel 0.86 1.74
```

With 95% confidence, the odds of the desire to drink being below a particular level change by 0.86 to 1.74 times when negative romantic event ratings are increased by 1.22, holding the other variables constant.

Yet, I can test to see if the model is actually statistically significant.

```
Anova(mod.fit.ord)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desire.order
##      LR Chisq Df Pr(>Chisq)
## nrel   1.3164  1     0.2512
```

However, this result is not statistically significant according to the LRT test (p-value = 0.25) for $\alpha = 0.05$. Therefore, there is no evidence to support the hypothesis that negative relationship interactions are associated with an increased desire to drink.

2.3 Question 2.3

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.

To test the hypothesis that the relation between drinking and negative relationship interactions is not evident for individuals with high trait self-esteem, I will build a model using the ratings of negative events with romantic partners and the scores of trait self-esteem to estimate the mean number of drinks consumed. So, the model I fit is $Y_i \sim Po(\mu_i)$ with

$$log(\mu_i) = \beta_0 + \beta_1 * nrel + \beta_2 * rosn$$

where Y_i is the number of drinks consumed (`numall`) for person i = 1,. . . , 89.

```
mod.neg_est <- glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
                   data = dehart_fri)
```

```
summary(mod.neg_est)
```

```
##
## Call:
## glm(formula = numall ~ nrel + rosn, family = poisson(link = "log"),
##     data = dehart_fri)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8798  -1.2047  -0.3725   0.5235   5.3255
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.26858    0.47899   4.736 2.18e-06 ***
## nrel         0.10066    0.03842   2.620  0.00879 **
## rosn        -0.36124    0.14133  -2.556  0.01059 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 234.89  on 88  degrees of freedom
## Residual deviance: 223.72  on 86  degrees of freedom
## AIC: 443.98
##
## Number of Fisher Scoring iterations: 5
```

The estimated model is $log(\hat{\mu}) = 2.27 - 0.10 * nrel + 0.36 * rosn$. The parameter estimates are $\beta_0 = 2.27$, $\beta_1 = 0.10$, and $\beta_2 = -0.36$. The positive and negative "slope" parameters indicate that the number of drinks are increasing as the negative romantic relationship events increase and the trait self-esteem decreases. In this context, this means that people with higher trait self-esteem do not drink more in relation to negative romantic relationship events.

Next, I can calculate the odds ratios and the corresponding confidence intervals to understand the impact of a change in negative romantic events and trait self-esteem on the consumption of alcoholic beverages.

```
100*(exp(mod.neg_est$coefficients) - 1)
```

```
## (Intercept)        nrel        rosn
##   866.56528    10.59059   -30.31849
```

As a result, a 1 unit increase in an individuals negative romantic relationship is associated with a 10.60% increase in the mean number of alcoholic drinks consumed. However, a 1 unit increase in an individuals trait self-esteem is associated with a 30.32% decrease in the mean number of alcoholic drinks consumed.

```
beta1.int <- confint(mod.neg_est, level = 0.95)
100*(exp(beta1.int) - 1)
```

```
##                   2.5 %      97.5 %
## (Intercept) 270.652278 2325.642914
## nrel          1.925524   18.593868
## rosn        -47.002262   -7.747164
```

So, a 1-unit increase in negative romantic events leads to an estimated $\hat{PC} = 10.59\%$ increase in the number

of alcoholic beverages with 95% profile LR confidence interval 1.93% $< PC <$ 18.59%. Also, a 1-unit increase in trait self-esteem leads to an estimated $\hat{PC} = 30.32\%$ decrease in the number of alcoholic beverages with 95% profile LR confidence interval -47.00% $< PC <$ -7.75%.

Overall, an individual who has consumed a high number of drinks and has had negative romantic relationship events is not very likely to have high trait self-esteem. Thus, it is fair to assume that for an individual with high trait self-esteem, there is no evident relation between drinking and negative relationship interactions.

Then, I can test to see if the change the variables is statistically significant.

```
Anova(mod.neg_est)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel   5.6463  1    0.01749 *
## rosn   6.3186  1    0.01195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The change for `nrel` is statistically significant according to both a Wald test (p-value = 0.0088) and the LRT (p-value = 0.0175). The change for `rosn` is also statistically significant according to both a Wald test (p-value = 0.0106) and the LRT (p-value = 0.0120).

Therefore, there is evidence that the relation between drinking and negative relationship interactions are not evident for individuals with high trait self-esteem.