Generalized Linear Models: Multinomial, Poisson, Proportional Odds

Stone Jiang

This report will explore 3 generalized linear models (GLM) for the analysis of categorical data: multinomial regression for multiclass classification, Poisson regression for analyzing count data, and the proportional odds model for modeling cumulative probability distribution of a multiclass outcome. We will perform exploratory data analysis, extensive model exploration and selection, and inference. We will study cereal placement and its success as a marketing strategy, and predictors for alcohol consumption.

Data is taken from Chris Bilder's Textbook:

Bilder, Christopher R., and Thomas M. Loughin. Analysis of Categorical Data with R. CRC Press, 2015.

Data is provided on his website: http://www.chrisbilder.com/categorical/programs_and_data.html

Strategic Placement of Products in Grocery Stores (Multinomial Regression)

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.*

1.1 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.

For size and rescaling, we will use the code provided in the text. Note that while sodium is in milligrams, we do not need to convert units since we perform min/max scaling later, which makes the resulting value unitless.

```
cereal <- read.csv("data/cereal_dillons.csv")
stand01 <- function(x) {
    (x - min(x))/(max(x) - min(x))
}
cereal2 <- data.frame(Shelf = cereal$Shelf, sugar = stand01(x = cereal$sugar_g/cereal$size_g),
    fat = stand01(x = cereal$fat_g/cereal$size_g), sodium = stand01(x = cereal$sodium_mg/cereal
```

For the boxplots, we will make the plots in ggplot, and stich plots together with patchwork. Note that for the dot plot, the binwidth is chosen to be very narrow, at 0.01, so values that are 0.01 of each other may be grouped into the same bin.

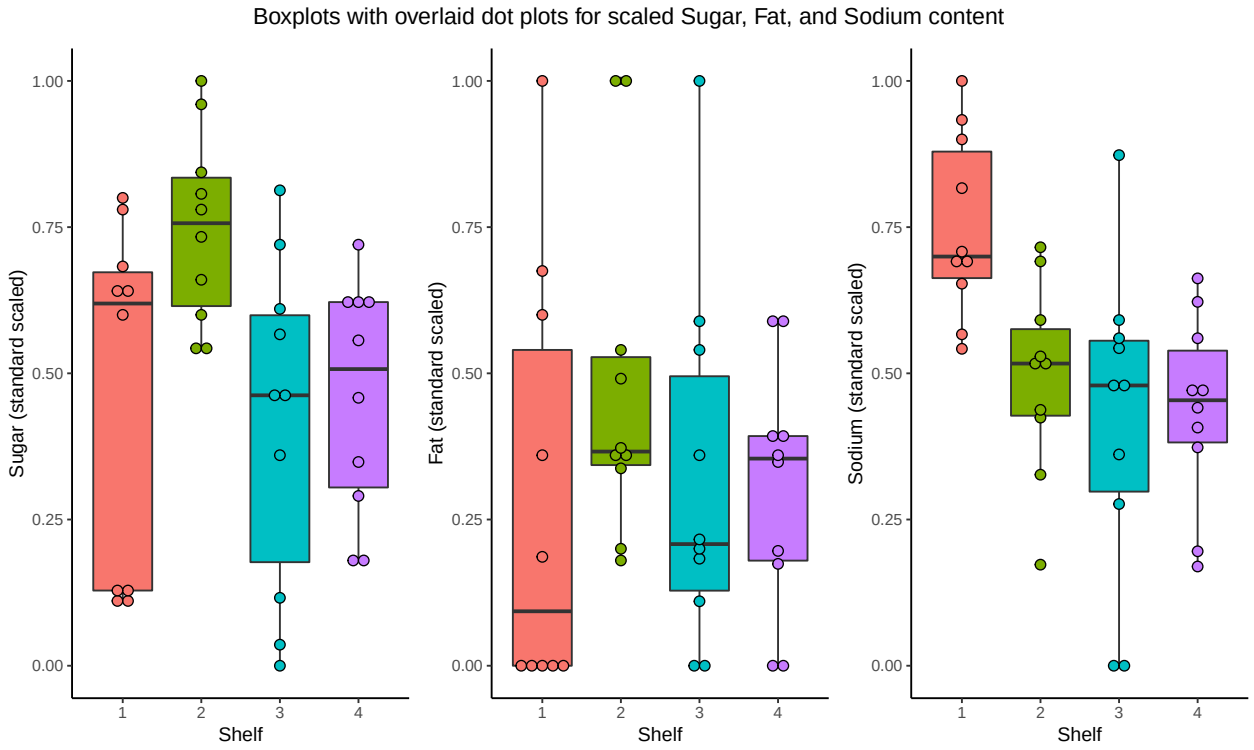
```
library(ggplot2)
library(patchwork)
```

```
##
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```
sugar <- ggplot(data = cereal2, aes(x = Shelf, group = Shelf,
    y = sugar, fill = factor(Shelf))) + geom_boxplot() + geom_dotplot(binaxis = "y",
    stackdir = "center", dotsize = 1.8, binwidth = 0.01) + labs(x = "Shelf",
    y = "Sugar (standard scaled)") + theme_classic() + theme(legend.position = "none")
fat <- ggplot(data = cereal2, aes(x = Shelf, group = Shelf, y = fat,
    fill = factor(Shelf))) + geom_boxplot() + geom_dotplot(binaxis = "y",
    stackdir = "center", dotsize = 1.8, binwidth = 0.01) + labs(x = "Shelf",
```

```
    y = "Fat (standard scaled)") + theme_classic() + theme(legend.position = "none")
sodium <- ggplot(data = cereal2, aes(x = Shelf, group = Shelf,
    y = sodium, fill = factor(Shelf))) + geom_boxplot() + geom_dotplot(binaxis = "y",
    stackdir = "center", dotsize = 1.8, binwidth = 0.01) + labs(x = "Shelf",
    y = "Sodium (standard scaled)") + theme_classic() + theme(legend.position = "none")

sugar | fat | sodium | plot_annotation(theme = theme(plot.title = element_text(hjust = 0.5)),
    title = "Boxplots with overlaid dot plots for scaled Sugar, Fat, and Sodium content")
```



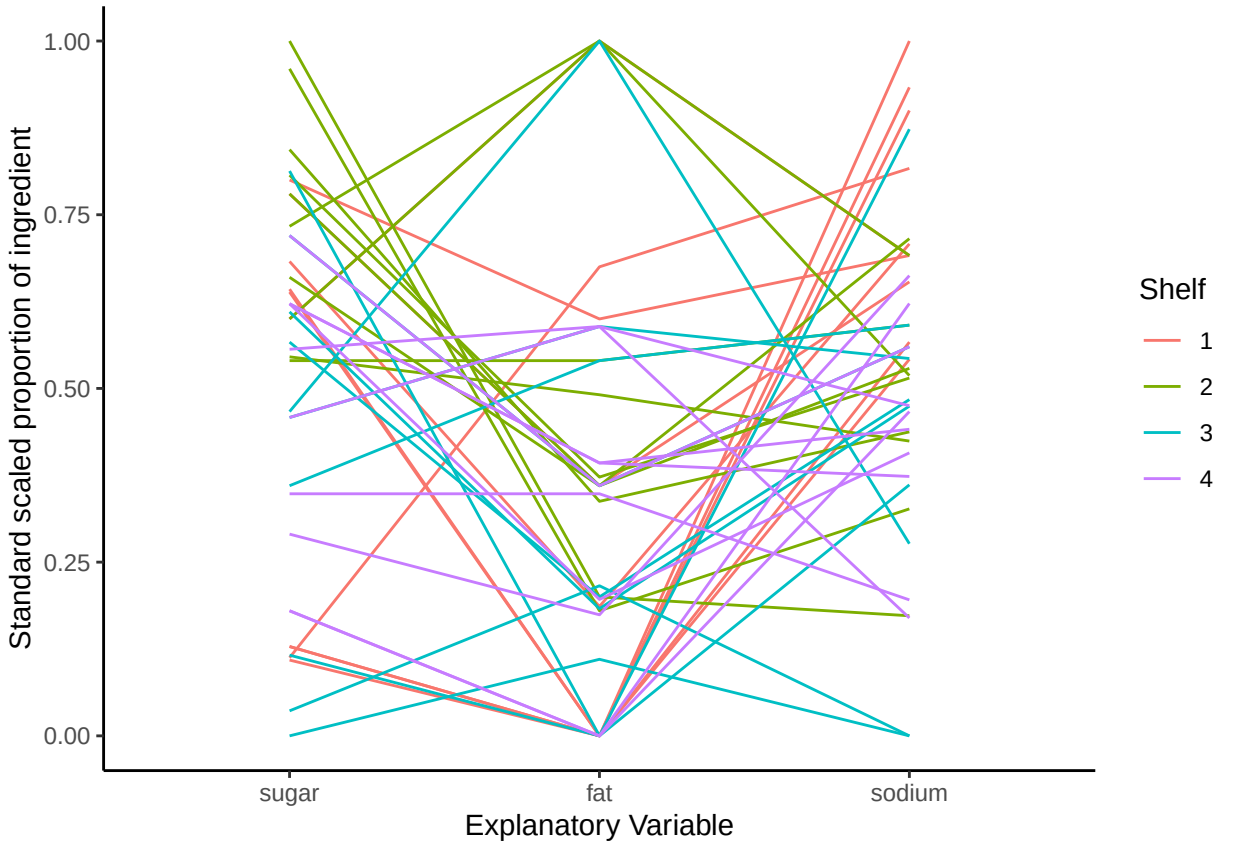
For parallel coordinates plot, we will use ggparcoord function from the GGally library, which utilizes ggplot backend for constructing the plot. Note that Y-axis is grams/milligrams of each ingredient per gram of cereal, and min/max scaled.

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggparcoord(cereal2, columns = 2:4, groupColumn = 1, scale = "globalminmax",
    mapping = aes(color = factor(Shelf))) + labs(x = "Explanatory Variable",
    y = "Standard scaled proportion of ingredient", color = "Shelf") +
    theme_classic()
```



Based on the boxplots, it seems like Shelf 2 cereals typically have higher sugar content, and Shelf 1 cereals typically have higher sodium per serving. This is also represented on the parallel coordinates plot, where Shelf 2 tends to be mostly clustered on the top for Sugar, and Shelf 1 tends to be mostly clustered at the top for Sodium. Shelf 1 has high variability in terms of fat and sugar, so it seems that some type of cereal specifically high in sodium but agnostic to fat and sugar exist on this shelf 1. There is less of a trend in the fat distributions between shelves, although Shelf 4 appears generally lower than other shelves. Shelf 2, even though the top whisker does not extend very far, has 2 outlier points with high fat content. In all plots and ingredient content, Shelf 3 tends to have one of the greatest variability, with a spread ranging from 0 to >0.75 for all three explanatory categories, which could indicate that all types of different cereals exist on this shelf. Shelf 2 tends to have the least variability, which could suggest that cereals sharing a certain trait (particularly high in sugar), tend to exist together on shelf 2. From the parallel coordinates plot, the story appears a bit more complex: two types of cereals tend to exist on shelf 2: 1. high in sugar, and low/medium levels of fat and sodium; 2. medium level of sugar and sodium, but high in fat.

1.2 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).

An ordinal scales makes the most sense when there is a natural ordering to the response variable. There are many situations where different shelves may inheritantly have different values to the store. For example, people typically see items that are at eye level first, and depending on the height of

the shelves, it may be that shelf 3 is closest to eye level for an average adult. It may also be easier for someone to look up rather than down, making shelf 4 more valuable than 2, and 2 more valuable than 1. From a marketing perspective, the store might prefer to place the most attrative looking packaging (to attract adults to that aisle), best selling adult cereals (in order to make these brands easier to find and improve shopping experience), or cereals they are trying to promote, on shelves in that are closer to eye level (in the order 3, 4, 2, 1). In this case, there would be a natural ordering to which shelves the store might prefer to utilize for different cereals. In a completely analogous situation, shelf 2 for example might be eye level with kids. The store might value shelf 2 for kids cereal (perhaps higher in sugar content) more so than shelf 4, in order to attract the attention of kids.

In a different context, the shelves might have a completely different natural ordering. Imagine that a store is color-coding the cereals from darkest boxes on top to lightest on the bottom. This may be a tactic used to fit a store theme or make displayed products more attractive. They could also be arranging the cereal based on any numeric criteria, such as number of calories per serving, etc. In these cases, there would be a natural ordering of Shelf 4, 3, 2, 1 that corresponds to decreasing pixel intensity of the boxes (or any other numeric criteria). We could certainly form a proportional odds model that predicts the odds of a new cereal being on a particular shelf or below it.

These situations are highly context dependent. We do not have further details as to whether these particular stores are trying to achieve any natural ordering. Based on the distribution of the boxplots, it does not appear at least that the stores are arranging their cereals based on our numeric explanatory variables. The spread for content level of each variable is too high for each shelf for there to be a pattern. As a result, we will fit a nominal response model that does not assume a natural ordering to the shelves.

We will also convert Shelf into a factor (since it's categorical and not numeric), and use shelf 1 as the base level. The model we are trying to fit is:

$$\log \frac{\pi_j}{\pi_1} = \beta_{j0} + \beta_{j1} * \text{sugar} + \beta_{j2} * \text{fat} + \beta_{j3} * \text{sodium} \qquad \text{j} = 2,3,4$$

where j ranges from 2 to 4 corresponding to shelves 2 to 4. We have 3 equations that model the log ratio of the probability of being in shelves 2 through 4, versus the probability of being in shelf 1.

```
library(nnet)
mod.nom <- multinom(factor(Shelf) ~ sugar + fat + sodium, data = cereal2)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 37.329384
## iter  20 value 33.775257
## iter  30 value 33.608495
## iter  40 value 33.596631
## iter  50 value 33.595909
## iter  60 value 33.595564
## iter  70 value 33.595277
## iter  80 value 33.595147
## final  value 33.595139
## converged
```

```
mod.nom
```

```
## Call:
## multinom(formula = factor(Shelf) ~ sugar + fat + sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar        fat    sodium
## 2    6.900708   2.693071  4.0647092 -17.49373
## 3   21.680680 -12.216442 -0.5571273 -24.97850
## 4   21.288343 -11.393710 -0.8701180 -24.67385
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

The actual model we have fit is:

$$\log \frac{\pi_2}{\pi_1} = 6.901 + 2.693 * \text{sugar} + 4.065 * \text{fat} - 17.494 * \text{sodium}$$
$$\log \frac{\pi_3}{\pi_1} = 21.681 - 12.216 * \text{sugar} - 0.557 * \text{fat} - 24.979 * \text{sodium}$$
$$\log \frac{\pi_4}{\pi_1} = 21.288 - 11.394 * \text{sugar} - 0.870 * \text{fat} - 24.674 * \text{sodium}$$

To perform LRTs to the examine the importance of each explanatory variable, we perform hypothesis testing at a significance level of 0.05. The null hypothesis is that **all** coefficients corresponding to the explanatory variable of interest at each of the levels is 0.

For example, for fat which corresponds to β_{j2}, the null and alternative hypotheses is as follows:

$$H_0 : \beta_{22} = \beta_{32} = \beta_{42} = 0$$
$$H_a : \text{not all of } \beta_{22}, \beta_{32}, \beta_{42} \text{ are } 0$$

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Anova(mod.nom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(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
```

Based on the LRT, fat has a p-value > 0.05, which means we fail to reject the null hypothesis, and this variable is not statistically significant given that the other two variables are in the model. However, both sugar and sodium have p-value $<< 0.01$, which means in both cases we reject the null hypothesis and state that the variables are both highly statistically significant given the other variables in the model. This is consistent with the EDA, where we showed that fat appeared evenly distributed across the 4 shelves, but shelf 2 was higher in sugar, and shelf 1 higher in fat.

To show that were are no significant interactions among the explanatory variables, we generate another model which includes all pairwise interactions, as well as the three pair interaction. The model we want to fit is (note again there are 3 equations for each level j = 2, 3, 4):

$$\log \frac{\pi_j}{\pi_1} = \beta_{j0} + \beta_{j1} * \text{sugar} + \beta_{j2} * \text{fat} + \beta_{j3} * \text{sodium}$$
$$+ \beta_{j4} * \text{sugar} * \text{fat} + \beta_{j5} * \text{sugar} * \text{sodium} + \beta_{j6} * \text{fat} * \text{sodium} + \beta_{j7} * \text{sugar} * \text{fat} * \text{sodium}$$

```
mod.nom.inter <- multinom(factor(Shelf) ~ sugar + fat + sodium +
    sugar:fat + sugar:sodium + fat:sodium + sugar:fat:sodium,
    data = cereal2)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 36.170336
## iter  20 value 31.166546
## iter  30 value 29.963705
## iter  40 value 28.414027
## iter  50 value 27.891712
## iter  60 value 27.763967
## iter  70 value 27.622579
## iter  80 value 27.438263
## iter  90 value 27.015534
## iter 100 value 26.772481
## final  value 26.772481
## stopped after 100 iterations
```

```
mod.nom.inter
```

```
## Call:
## multinom(formula = factor(Shelf) ~ sugar + fat + sodium + sugar:fat +
##     sugar:sodium + fat:sodium + sugar:fat:sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar       fat     sodium sugar:fat sugar:sodium
## 2   -4.563627   8.944868 22.063003   1.030077  35.60873   -12.250084
## 3   24.498320 -22.248456 35.981865 -27.899087 -17.12487    13.253103
## 4   27.246742 -21.852777  7.298799 -29.106797  41.08251     2.887805
```

```
##   fat:sodium sugar:fat:sodium
## 2  -23.75955        -55.88455
## 3  -59.54150         37.71571
## 4  -30.85250        -22.59552
##
## Residual Deviance: 53.54496
## AIC: 101.545
```

Since this model is constructed merely for testing interactions and not as our final model, we will not re-write the model numerically. To test whether the interactions are significant, we test at a significance level of 0.05, and the null hypothesis is again that all coefficients corresponding to the interaction is 0 for all levels. For example, testing the three variable interaction corresponding to coefficient β_{j6}:

$$H_0 : \beta_{27} = \beta_{37} = \beta_{47} = 0$$
$$H_a : \text{not all of } \beta_{27}, \beta_{37}, \beta_{47} \text{ are } 0$$

```
Anova(mod.nom.inter)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(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
```

The LRT test shows that the interaction terms all have p-values $>>$ 0.05. As a result, we fail to reject the null hypothesis that these coefficients are 0 given that the main order 1 effects and other interactions (besides the one we are testing) are in the model, and state that these interaction terms including the three way interaction are not statistically significant. This test is conducted individually for each interaction given that the other interactions are in the model. For example, the three way interaction term hypothesis test says that the three way interaction is not statistically significant given that the main order 1 effects and two way interactions are included in the model.

We can also test the joint significance of all interaction terms using the profile LR test by comparing the two models we generated at a level of 0.05. H0 states that the model contains no interaction terms (mod.non), and Ha states the interactions do exist (Ha):

```
anova(mod.nom, mod.nom.inter, test = "Chisq")
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: factor(Shelf)
```

```
##                                                                                 Model
## 1                                                                 sugar + fat + sodium
## 2 sugar + fat + sodium + sugar:fat + sugar:sodium + fat:sodium + sugar:fat:sodium
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1       108   67.19028
## 2        96   53.54496 1 vs 2    12 13.64531 0.3239288
```

Since the p-value > 0.05, we fail to reject H0 that the interactions do not exist in the model, and conclude that the 2 and 3-way interaction terms are also not jointly statistically significant.

We can also build models where we include one interaction term at a time in order to test its significance given that only order 1 terms are included in the model, but the two tests conducted should suffice to say that interactions are most likely not needed.

1.3: 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.

The probability of being on shelf 1 (base level) can be derived as:

$$\pi_1 = \frac{1}{1 + \sum_{j=2}^{4} e^{\beta_{j0} + \beta_{j1} * \text{sugar} + \beta_{j2} * \text{fat} + \beta_{j3} * \text{sodium}}}$$

and the probability of being on a non-base level is:

$$\pi_j = \pi_1 * e^{\beta_{j0} + \beta_{j1} * \text{sugar} + \beta_{j2} * \text{fat} + \beta_{j3}}$$

To predict a new probabilities, we can use the predict function with this new data point. However, we must transform our data point in the same way that the traiting data was transformed. Also, it is customary to normalize based on the min/max of the fitted (training) data only, and not include the value to be predicted at part of the determination of min/max. As a result, we will calculate what the min/max pairs are for each explanatory variable based on the training data.

```
apple_jacks <- data.frame(serving_size = 28, sugar = 12, fat = 0.5,
    sodium = 130)

# Get a new dataframe for just the explanatory variables
# scaled by the serving size
cereal3 <- cereal/cereal$size_g

# This function will transform a data frame using serving
# size scaling, and min/max scaling from the fitted data
transform_new_data <- function(data) {

    # define the standardize function that takes min/max from the
    # fitted data set
    standardize <- function(x, min_trait, max_trait) {
        (x - min_trait)/(max_trait - min_trait)
    }
```

```
    # Generate the output dataframe
    data_frame = data.frame(sugar = standardize(data$sugar/data$serving_size,
        min(cereal3$sugar_g), max(cereal3$sugar_g)), fat = standardize(data$fat/data$serving_si
        min(cereal3$fat_g), max(cereal3$fat_g)), sodium = standardize(data$sodium/data$serving_
        min(cereal3$sodium_mg), max(cereal3$sodium_mg)))

    return(data_frame)
}

apple_jacks_transformed <- transform_new_data(apple_jacks)

probs <- predict(mod.nom, newdata = apple_jacks_transformed,
    type = "probs")
pdat <- data.frame(Shelf_1 = probs[1], Shelf_2 = probs[2], Shelf_3 = probs[3],
    Shelf_4 = probs[4])
rownames(pdat) = "Probability"
pdat
```

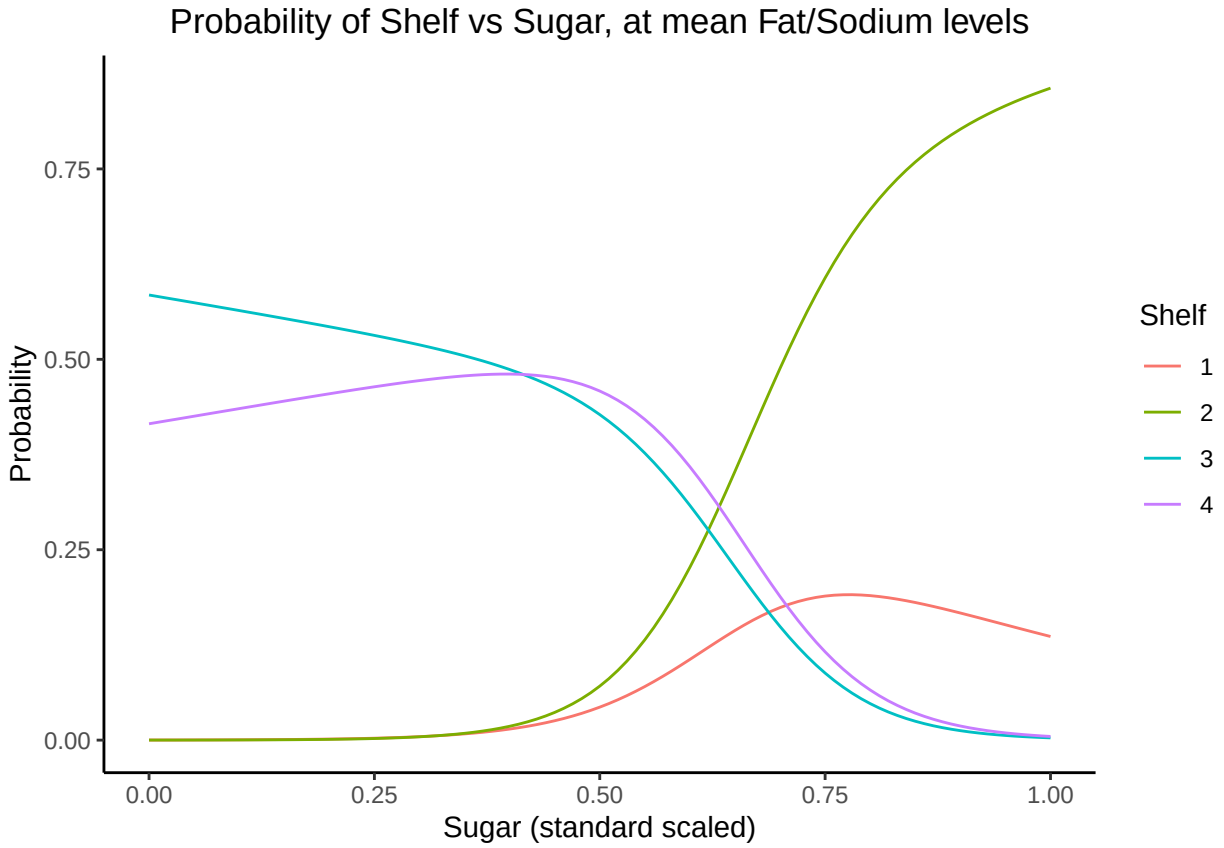
```
##                Shelf_1   Shelf_2   Shelf_3   Shelf_4
## Probability 0.05326849 0.4719426 0.2004274 0.2743615
```

We can see that all of the shelves have some probability > 0, but the largest is Shelf 2 at 0.472. As a result, Shelf 2 would be the predicted shelf. Notice though that it is slightly more likely that this cereal is on shelf 3 or shelf 4 ($0.20042742 + 0.27436145 \approx 0.475$) than it is on shelf 2, and probability of shelf 1 is close to 0.

1.4: 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.

```
library(dplyr)
# First get the mean fat and sodium values from cereal2
# Cereal2 contains values already normalized by serving_size,
# and min/max scaling

mean_fat <- mean(cereal2$fat)
mean_sodium <- mean(cereal2$sodium)

# The range of sugar we will scan is from min(sugar) to
# max(sugar), at intervals of 0.01
sugar_scan <- seq(min(cereal2$sugar), max(cereal2$sugar), 0.01)

# Generate input dataframe
sugar_scan_data <- data.frame(sugar = sugar_scan, fat = mean_fat,
    sodium = mean_sodium)

# Generate prediction probabilities for each Shelf
pi_hat_sugar_scan <- cbind(sugar = sugar_scan, predict(mod.nom,
    newdata = sugar_scan_data, type = "prob")) %>% data.frame()
```

```r
names(pi_hat_sugar_scan) <- c("sugar", "1", "2", "3", "4")

library(tidyr)
# We recast the wide format into a long format for ggplot

pi_hat_plot <- gather(pi_hat_sugar_scan, key = Shelf, value = probs,
    2:5)

# Generate line plot from dataset
ggplot(data = pi_hat_plot, aes(x = sugar, y = probs, color = Shelf)) +
    geom_line() + labs(x = "Sugar (standard scaled)", y = "Probability",
    color = "Shelf", title = "Probability of Shelf vs Sugar, at mean Fat/Sodium levels") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5))
```



From this plot, we see that keeping fat and sodium levels at their mean, the estimated probability of being on shelf 1 and 2 are close to 0 when the sugar content is low (<0.4). In this case, the probability of being on Shelf 3 is slightly higher than 4. The probability of being on shelf 2 drastically increases when the scaled sugar level goes above 0.5, and becomes the most probable shelf after around ~0.65. For scaled sugar content less than 0.65, the probability distribution of Shelf 3 and 4 are similar to each other, although 3 is higher than 4 for very low levels of sugar (<0.4). The probability of being on Shelf 1 is never the greatest, although for values of sugar past ~0.75, probability of Shelf 1 is higher than both 3 and 4, which diminish to 0. The overall trend is that shelf 3 probability monotonically decreases as sugar increases, shelf 2 monotonically increases as sugar increases, while

shelf 1 and 4 both peak in the middle, at sugar levels ~0.75 and ~0.5 respectively.

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 actual model we have fit is:

$$\log \frac{\pi_2}{\pi_1} = 6.901 + 2.693 * \text{sugar} + 4.065 * \text{fat} - 17.494 * \text{sodium}$$
$$\log \frac{\pi_3}{\pi_1} = 21.681 - 12.216 * \text{sugar} - 0.557 * \text{fat} - 24.979 * \text{sodium}$$
$$\log \frac{\pi_4}{\pi_1} = 21.288 - 11.394 * \text{sugar} - 0.870 * \text{fat} - 24.674 * \text{sodium}$$

Therefore, we have modeled the log of the odds of shelves 2 through 4 with respect to shelf 1. To recover the odds ratio with respect to a change in the explanatory variable of interest r, we have:

$$OR = \frac{e^{\beta_{jr}(x_r+c)}}{e^{\beta_{jr}(x_r)}} = e^{c\beta_{jr}}$$

Since we do not have theory to guide reasonable values of changes of each explanatory variable, we will use the standard deviation of each variable. This is a general approach that is amenable to interpretation.

```
# Get standard deviation of each variable
var.sd <- cereal2 %>% summarise_all(sd)

# Get c values corresponding to explanatory variables
c.value <- var.sd[2:4]

# Get coefficients corresponding to each explanatory variable
# for each of shelves 2-4 compared to shelf 1
coefs <- coef(mod.nom)[, 2:4]

# Perform ; c.value expanded into 2 columns for element-wise
# multiplcation with coefs matrix This generates per sd unit
# increase in the explanatory variables
OR <- exp(coefs * c.value[c(1, 1, 1), ])
data.OR <- data.frame(OR, row.names = c("2vs1", "3vs1", "4vs1")) %>%
    round(4)
print("sd increase in explanatory variables")
```

```
## [1] "sd increase in explanatory variables"
```

```
data.OR
```

```
##       sugar    fat sodium
## 2vs1 2.0647 3.3719 0.0179
## 3vs1 0.0373 0.8465 0.0032
## 4vs1 0.0465 0.7709 0.0034
```

```
# This generates per sd unit decrease in the explanatory
# variables
OR <- exp(coefs * -c.value[c(1, 1, 1), ])
data.OR <- data.frame(OR, row.names = c("2vs1", "3vs1", "4vs1")) %>%
    round(4)
print("sd decrease in explanatory variables")
```

```
## [1] "sd decrease in explanatory variables"
```

```
data.OR
```

```
##        sugar    fat   sodium
## 2vs1  0.4843 0.2966  55.7393
## 3vs1 26.8096 1.1813 311.3613
## 4vs1 21.4833 1.2972 290.3058
```

This table says that the odds of being on Shelf 2 versus Shelf 1 changes by 2.065 times for every standard deviation increase in sugar (0.269) while holding other variables constant. This direction of change makes sense because Shelf 2 was the shelf with highest sugar content. Shelf 1 was bimodal in terms of sugar content, with a majority >0.6 with a few observations below .25. Shelf 3 and 4 were more spread out, with a majority of the observations below 0.6. This translates to the interpretation of the model nicely. The odds of being on Shelf 3 versus Shelf 1 changes by 26.810 times, and the odds of being on Shelf 4 versus Shelf 1 changes by 21.483 times, for a standard deviation **decrease** in sugar while holding other variables constant. In other words, the odds increase on being on Shelf 1 versus 3 or 4 as sugar content increases.

Fat has a similar story as sugar. This table says that for every standard deviation increase in fat while holding other variables constant, the odds of being on Shelf 2 versus Shelf 1 changes by 3.372 times, the odds of being on Shelf 3 versus Shelf 1 changes by 0.847 times, and the odds of being on Shelf 4 versus Shelf 1 changes by 0.771 times. As fat goes up, the odds of being on shelf 2 vs 1 increases, while the odds of being on shelf 3 or shelf 4 vs shelf 1 decreases. It is not surprising given the plots earlier that the odds increase between Shelf 2 and Shelf 1, since Shelf 2 was the shelf with large outlier points in fat.

Finally, sodium is the most clearcut in our example, since Shelf 1 has signficantly higher sodium content than any of the other shelves. Therefore, the odds ratios are much less than 1. This table says that for every standard deviation **decrease** in sodium while holding other variables constant, the odds of being on Shelf 2 versus Shelf 1 changes by 55.739 times, the odds of being on Shelf 3 versus Shelf 1 changes by 311.361 times, and the odds of being on Shelf 4 versus Shelf 1 changes by 290.306 times. Odds on being on any other shelf compared to the base shelf 1 increases by a large margin as sodium content decreases. Implicitly, as sodium content increases, the odds of being on Shelf 1 greatly increases.

For confidence intervals, the confint function will return the Wald CI for each parameter. We calculate this at 95%. To get confidence intervals for each explanatory variable coefficient, we calculate the CI for the parameter first, then exponentiate the product with c.value to get the CI for the Odds Ratio.

```
param.CI <- confint(mod.nom, level = 0.95)

level_2vs1 <- exp(param.CI[, , 1][2:4, ] * t(c.value[c(1, 1),
```

```
    ]))
level_3vs1 <- exp(param.CI[, , 2][2:4, ] * t(c.value[c(1, 1),
    ]))
level_4vs1 <- exp(param.CI[, , 3][2:4, ] * t(c.value[c(1, 1),
    ]))
print("For std increase in explanatory variables")
```

```
## [1] "For std increase in explanatory variables"
```

```
print("Shelf 2 vs 1")
```

```
## [1] "Shelf 2 vs 1"
```

```
level_2vs1
```

```
##               2.5 %     97.5 %
## sugar  0.143637126 29.6794905
## fat    0.872160890 13.0360059
## sodium 0.000733476  0.4388243
```

```
print("Shelf 3 vs 1")
```

```
## [1] "Shelf 3 vs 1"
```

```
level_3vs1
```

```
##                2.5 %    97.5 %
## sugar  2.829009e-03 0.4917952
## fat    2.055687e-01 3.4860877
## sodium 8.432188e-05 0.1223293
```

```
print("Shelf 4 vs 1")
```

```
## [1] "Shelf 4 vs 1"
```

```
level_4vs1
```

```
##                2.5 %    97.5 %
## sugar  0.0035614923 0.6083685
## fat    0.1882197197 3.1574471
## sodium 0.0000911725 0.1301442
```

Based on these intervals, we see that the intervals for sodiums for all Shelves 2-4 compared to 1 does not contain 0. This means that it is statistically significant that the odds of being on shelf 1 versus 2, 3 or 4 is higher as sodium level increases. For example, with 95% confidence, the odds of being on shelf 1 compared to shelf 2 changes by 0.00073 to 0.4388 times for a standard deviation increase in sodium, holding other variables constant. The CIs for fat all contain 0, meaning none of the levels are statistically significant. This is consistent with the Anova analysis which found that the fat variable is not statistically significant.

Interestingly, sugar is also statistically significant, but only comparing Shelf 3 to Shelf 1, and Shelf 4 to Shelf 1. We hypothesize that the odds of being on Shelf 2 vs 3 or 4 is also statistically significant for a standard deviation increase in sugar. We can relevel using Shelf 2 as base:

$$\begin{aligned}\frac{\pi_3}{\pi_2} &= \frac{\frac{\pi_3}{\pi_1}}{\frac{\pi_2}{\pi_1}} \\ &= \frac{e^{21.681-12.216*\text{sugar}-0.557*\text{fat}-24.979*\text{sodium}}}{e^{6.901+2.693*\text{sugar}+4.065*\text{fat}-17.494*\text{sodium}}} \\ &= e^{14.78-14.909*\text{sugar}-4.622*\text{fat}-7.485*\text{sodium}}\end{aligned}$$

While we can calculate this, it is easier to just refit using π_2 as base level:

```
mod.nom2 <- multinom(factor(Shelf, levels = c("2", "1", "3",
    "4")) ~ sugar + fat + sodium, data = cereal2)
```

```
## # 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
```

```
mod.nom2
```

```
## Call:
## multinom(formula = factor(Shelf, levels = c("2", "1", "3", "4")) ~
##     sugar + fat + sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar       fat    sodium
## 1    -6.89779  -2.692069 -4.063019 17.486515
## 3    14.78681 -14.912714 -4.621863 -7.495409
## 4    14.39434 -14.090538 -4.934381 -7.189731
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

The model we've fit is now:

$$\begin{aligned}\log\frac{\pi_1}{\pi_2} &= -6.90 - 2.693 * \text{sugar} - 4.063 * \text{fat} + 17.487 * \text{sodium} \\ \log\frac{\pi_3}{\pi_3} &= 14.787 - 14.921 * \text{sugar} - 4.622 * \text{fat} - 7.495 * \text{sodium} \\ \log\frac{\pi_4}{\pi_3} &= 14.394 - 14.091 * \text{sugar} - 4.934 * \text{fat} - 7.189 * \text{sodium}\end{aligned}$$

Re-performing the CI calculation:

```
param.CI <- confint(mod.nom2)

level_1vs2 <- exp(param.CI[, , 1][2:4, ] * t(c.value[c(1, 1),
```

```
    ]))
level_3vs2 <- exp(param.CI[, , 2][2:4, ] * t(c.value[c(1, 1),
    ]))
level_4vs2 <- exp(param.CI[, , 3][2:4, ] * t(c.value[c(1, 1),
    ]))
print("Shelf 1 vs 2")
```

```
## [1] "Shelf 1 vs 2"
```

```
level_1vs2
```

```
##             2.5 %      97.5 %
## sugar  0.03371405    6.961459
## fat    0.07676074    1.146988
## sodium 2.27616841 1360.439155
```

```
print("Shelf 3 vs 2")
```

```
## [1] "Shelf 3 vs 2"
```

```
level_3vs2
```

```
##              2.5 %   97.5 %
## sugar  0.001250256 0.260584
## fat    0.050035280 1.259711
## sodium 0.014600221 2.184290
```

```
print("Shelf 4 vs 2")
```

```
## [1] "Shelf 4 vs 2"
```

```
level_4vs2
```

```
##              2.5 %    97.5 %
## sugar  0.001618792 0.3133329
## fat    0.045759365 1.1426048
## sodium 0.015921047 2.3052692
```

As suspected, the CIs for sugar comparing Shelf 2 to 3 and 4 do not contain 0, meaning it is statistically signficant that a standard deviation increase in sugar increases the odds of being on Shelf 2 vs 3 or 4. This supports the graphical analysis in the EDA. More specifically as an example, with 95% confidence, the odds of being on Shelf 3 versus 2 changes by 0.00125 to 0.2606 times for a standard deviation increase in sugar, holding other variables constant.

2. Alcohol, self-esteem and negative relationship interactions

The data are given in the *DeHartSimplified.csv* data set. 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 researchers stated the following hypothesis:

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

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

We begin by examining each of the variables in tabular form.

```
drink <- read.csv("data/DeHartSimplified.csv")
Hmisc::describe(drink)
```

```
## drink
##
##  13  Variables      623  Observations
## ---------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       89        1    75.89    56.82      7.0     16.2
##      .25      .50      .75      .90      .95
##     33.0     60.0    123.0    147.2    153.0
##
## lowest :   1   2   4   5   7, highest: 153 154 155 156 160
## ---------------------------------------------------------------------------
## studyday
##        n  missing distinct     Info     Mean      Gmd
##      623        0        7     0.98        4    2.289
##
## Value          1     2     3     4     5     6     7
## Frequency     89    89    89    89    89    89    89
## Proportion 0.143 0.143 0.143 0.143 0.143 0.143 0.143
## ---------------------------------------------------------------------------
## dayweek
##        n  missing distinct     Info     Mean      Gmd
##      623        0        7     0.98        4    2.289
##
## Value          1     2     3     4     5     6     7
```

```
## Frequency     89    89    89    89    89    89    89
## Proportion 0.143 0.143 0.143 0.143 0.143 0.143 0.143
## ---------------------------------------------------------------------------
## numall
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      622        1       18     0.97    2.524    2.636     0.00     0.00
##      .25      .50      .75      .90      .95
##     1.00     2.00     3.75     6.00     8.00
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency    141   112   132    81    49    43    24     6     9     7
## Proportion 0.227 0.180 0.212 0.130 0.079 0.069 0.039 0.010 0.014 0.011
##
## Value         10    11    12    13    14    15    18    21
## Frequency      7     4     2     1     1     1     1     1
## Proportion 0.011 0.006 0.003 0.002 0.002 0.002 0.002 0.002
## ---------------------------------------------------------------------------
## nrel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       33    0.551    0.359   0.6252        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        1        2
##
## lowest : 0.0000000 0.2000000 0.2500000 0.3333333 0.4000000
## highest: 5.0000000 5.5000000 5.8333333 6.0000000 9.0000000
## ---------------------------------------------------------------------------
## prel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       68    0.982    2.583    2.613   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.4167   2.0000   4.0000   6.0000   7.8683
##
## lowest : 0.0000000 0.2000000 0.2500000 0.3333333 0.5000000
## highest: 8.1666667 8.3333333 8.5000000 8.6666667 9.0000000
## ---------------------------------------------------------------------------
## negevent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0      131    0.996   0.4414   0.4123   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.1583   0.3500   0.6292   1.0000   1.1500
##
## lowest : 0.00000000 0.02500000 0.03333333 0.05000000 0.07500000
## highest: 1.70000000 1.93000000 1.95000000 2.01666667 2.37666667
## ---------------------------------------------------------------------------
## posevent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0      216        1    1.048   0.7077    0.200    0.300
##      .25      .50      .75      .90      .95
```

```
##    0.600    0.950    1.378    1.938    2.200
##
## lowest : 0.00000000 0.04000000 0.05000000 0.06666667 0.10000000
## highest: 3.23333333 3.25000000 3.30000000 3.40000000 3.88333333
## ---------------------------------------------------------------------------
## gender
##        n  missing distinct     Info     Mean      Gmd
##      623        0        2    0.739    1.562   0.4932
##
## Value          1     2
## Frequency    273   350
## Proportion 0.438 0.562
## ---------------------------------------------------------------------------
## rosn
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       17    0.993    3.436   0.4663      2.7      2.9
##      .25      .50      .75      .90      .95
##      3.2      3.5      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
## Frequency      7     7    14     7    21    35    42    21    28    42
## Proportion 0.011 0.011 0.022 0.011 0.034 0.056 0.067 0.034 0.045 0.067
##
## Value        3.4   3.5   3.6   3.7   3.8   3.9   4.0
## Frequency     35    84    63    49    63    49    56
## Proportion 0.056 0.135 0.101 0.079 0.101 0.079 0.090
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      623        0       89        1    34.29     5.18    26.24    27.82
##      .25      .50      .75      .90      .95
##    30.53    34.57    38.19    40.15    40.56
##
## lowest : 24.43258 25.57700 26.05613 26.14100 26.23682
## highest: 40.56400 40.58864 40.68720 40.82957 42.27789
## ---------------------------------------------------------------------------
## desired
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      620        3       22    0.996    4.465    1.921    1.333    2.000
##      .25      .50      .75      .90      .95
##    3.333    4.667    5.667    6.667    7.333
##
## lowest : 1.000000 1.333333 1.666667 2.000000 2.333333
## highest: 6.666667 7.000000 7.333333 7.666667 8.000000
## ---------------------------------------------------------------------------
## state
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      620        3       25    0.993    3.966   0.4894    3.222    3.333
```

```
##      .25      .50      .75      .90      .95
##    3.667    4.000    4.222    4.556    4.556
##
## lowest : 2.333333 2.444444 2.555556 2.666667 2.777778
## highest: 4.555556 4.666667 4.777778 4.888889 5.000000
## ------------------------------------------------------------------------
```

First, we note that there is one missing data point for number of drinks, and this corresponds to participant 42 on Sunday:

```
drink[drink$id == 42, ]
```

```
##     id studyday dayweek numall nrel prel negevent posevent gender rosn
## 211 42        1       4      6    0    6     0.00     1.30      2    4
## 212 42        2       5      4    0    5     0.50     1.60      2    4
## 213 42        3       6      3    0    6     0.90     2.10      2    4
## 214 42        4       7     NA    0    3     0.00     1.80      2    4
## 215 42        5       1      5    0    3     0.15     1.35      2    4
## 216 42        6       2      0    0    3     0.80     0.60      2    4
## 217 42        7       3      3    0    3     0.60     0.90      2    4
##          age  desired    state
## 211 35.15674 4.666667 4.333333
## 212 35.15674 4.666667 4.444444
## 213 35.15674 3.666667 4.555556
## 214 35.15674 3.666667 4.555556
## 215 35.15674 3.333333 4.555556
## 216 35.15674 1.000000 4.555556
## 217 35.15674 3.666667 4.555556
```

Since this is the response variable, we should not assign a value for this day without further information on why the data point is missing. If for example this was a clerical error, it may be reasonable to impute the value with the median of the other days (since the value is an integer), or the median the number of drinks on Sunday for other participants. If the participant declined to answer, this may indicate for example that they were embarassed by the number of drinks they had that particular day, so the average may not be reasonable. Since we do not have information regarding the nature of the missing value, removing this data point is our best option for EDA as well as regression analysis.

There are also 3 data points missing state (short term self-esteem) and desired. We draw out the rows missing the data below:

```
drink[is.na(drink$desired) | is.na(drink$state), ]
```

```
##      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
## 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
```

```
## 402 40.82957       NA    NA
## 448 37.38809       NA     4
```

Desire will be a response variable, so with the same argument above, missing rows will be eliminated from the regression analysis and EDA. State is a variable that measures short-term self-esteem, which is not a variable of interest to the author's hypothesis. As a result, we will drop this variable from consideration, as well as rows that include NA values for desired and numall.

```
drink_cleaned <- drink %>% dplyr::select(-state) %>% drop_na()
```

For the analysis, we need to determine a reasonable threshold to divide individuals between "high" trait self-esteem and "low" trait self-esteem. To do this, we can perform univariate EDA on the trait self-esteem variable rosn. We first examine a summary of this variable:

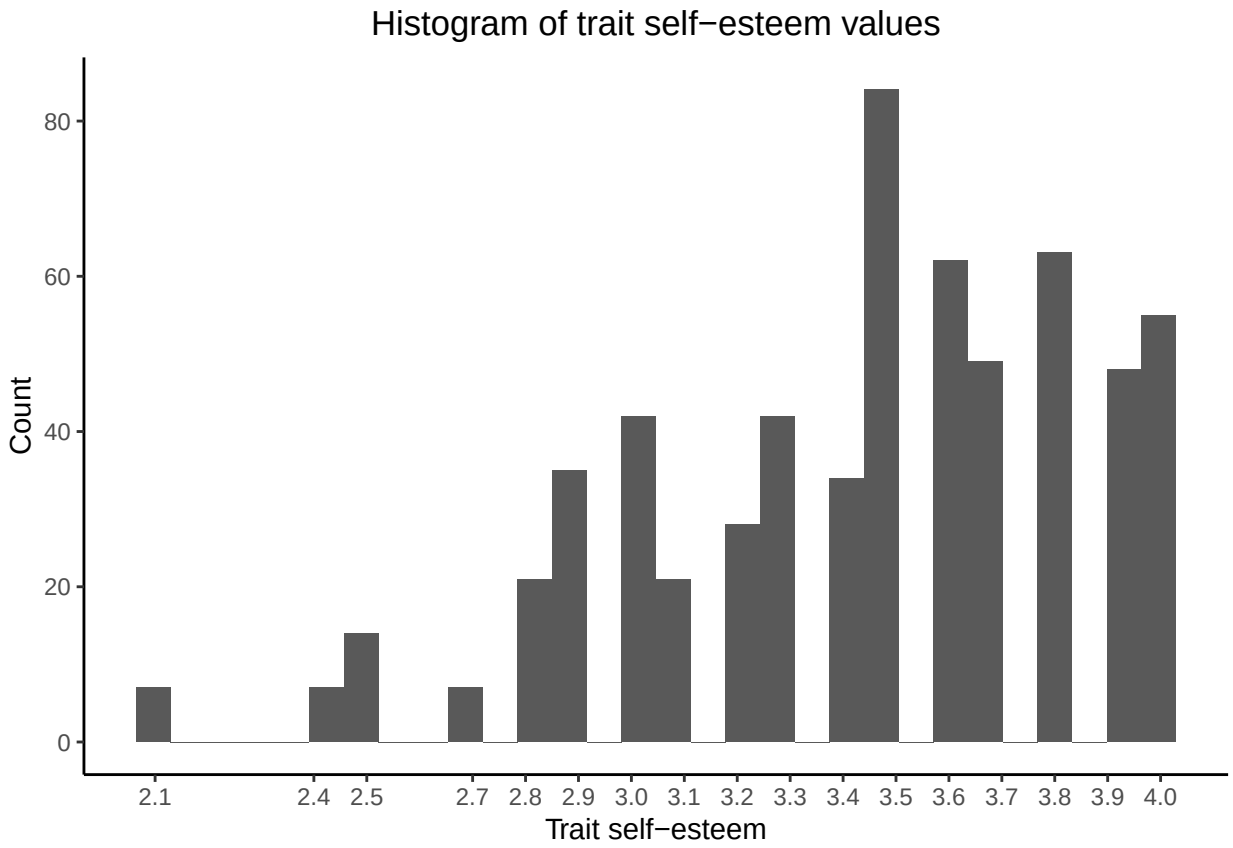
```
summary(drink_cleaned$rosn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.100   3.200   3.500   3.434   3.800   4.000
```

The variable ranges from 2.1 to 4, with a median of 3.5, and a smaller mean at 3.434, meaning the data is slightly negatively skewed. From the tabular analysis above, the variable also takes 17 distinct values. We will then plot a histogram of this variable.

```
ggplot(data = drink_cleaned, aes(x = rosn)) + geom_histogram() +
    labs(x = "Trait self-esteem", y = "Count", title = "Histogram of trait self-esteem values")
    theme_classic() + theme(plot.title = element_text(hjust = 0.5)) +
    scale_x_continuous(breaks = table(drink_cleaned$rosn) %>%
        names() %>% as.numeric())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



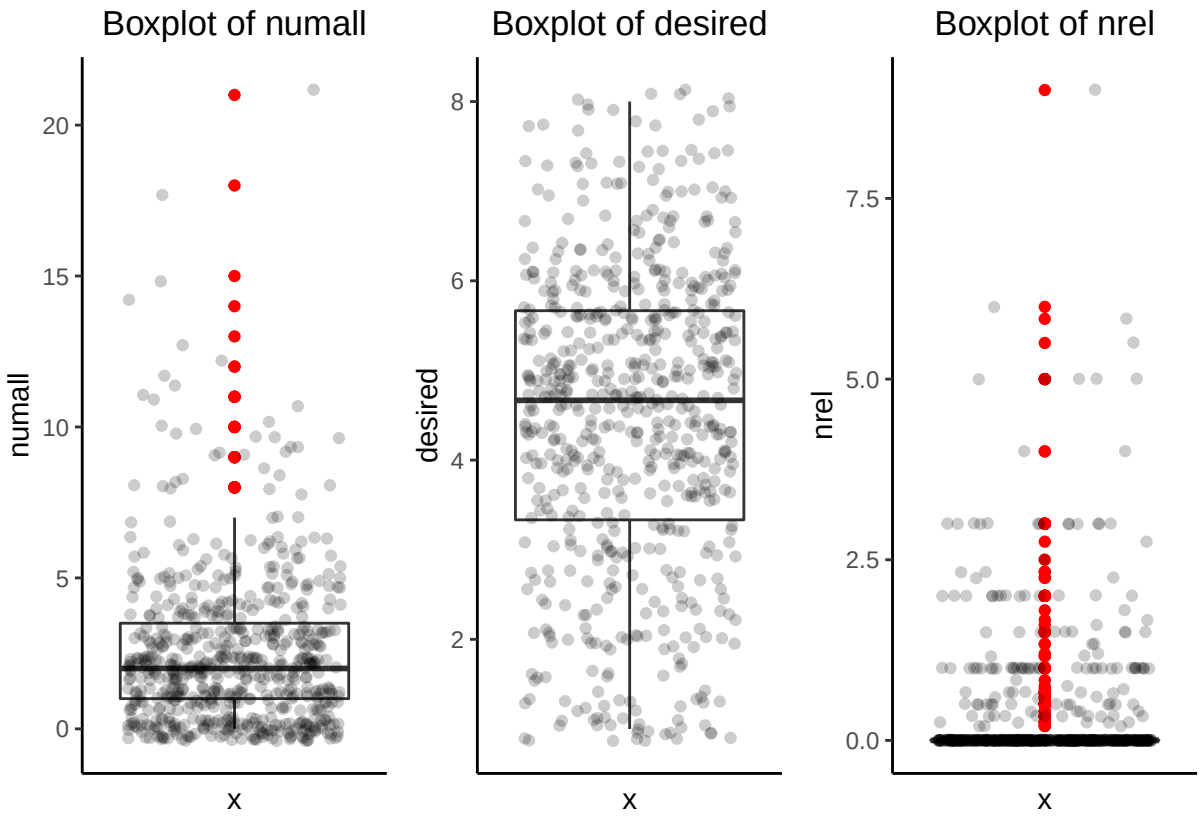
Based on the histogram, there is not an obvious location to draw a cutoff between high and low self esteem. For example, if there were two distinct populations less than or greater than some value, we could draw the cutoff there. As a result, we will use the mean value 3.434 as the cutoff. In other words, individuals with trait self-esteem values greater than 1 standard deviation above average will be considered high, while those with 1 standard deviation below average trait self-esteem values will be considered as having a low self esteem. Everyone else in the middle will be considered medium. The purpose of doing this is for EDA purposes. We can model rosn as a numeric variable, but for visualization, it is convenient to split the variable into three categories. We will generate a new variable factor_rosn to store this.

```
high_cutoff <- mean(drink_cleaned$rosn) + sd(drink_cleaned$rosn)
low_cutoff <- mean(drink_cleaned$rosn) - sd(drink_cleaned$rosn)
drink_cleaned$factor_rosn <- ifelse(drink_cleaned$rosn > high_cutoff,
    yes = "H", no = ifelse(drink_cleaned$rosn < low_cutoff, yes = "L",
        no = "M")) %>% factor()
```

Next, we conduct univariate analysis of our dependent variables, numall and desired, and explanatory variable of interest, nrel, but generating boxplots of their distributions. Note that outliers are colored in red, and a jitter plot with jitter in the x-direction is appended so that all points can be visualized.

```
numall.box <- ggplot(drink_cleaned, aes(x = "", y = numall)) +
    geom_boxplot(outlier.colour = "red") + geom_jitter(width = 0.35,
    alpha = 0.2) + labs(y = "numall", title = "Boxplot of numall") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5),
```

```
    axis.text.x = element_blank(), axis.ticks.x = element_blank())
desired.box <- ggplot(drink_cleaned, aes(x = "", y = desired)) +
    geom_boxplot(outlier.colour = "red") + geom_jitter(width = 0.35,
    alpha = 0.2) + labs(y = "desired", title = "Boxplot of desired") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_blank(), axis.ticks.x = element_blank())
nrel.box <- ggplot(drink_cleaned, aes(x = "", y = nrel)) + geom_boxplot(outlier.colour = "red"
    geom_jitter(width = 0.35, alpha = 0.2) + labs(y = "nrel",
    title = "Boxplot of nrel") + theme_classic() + theme(plot.title = element_text(hjust = 0.5
    axis.text.x = element_blank(), axis.ticks.x = element_blank())
numall.box + desired.box + nrel.box
```



Based on these boxplots, we can see that numall is mostly centered around 0-5, with a countable number of outliers. the distribution of desired is almost entirely homogenous, ranging from 0 - 8 fairly evenly. nrel on the other hand is extremely right skewed. While the majority of points are 0 (no negative relationship interactions), there are a large handful of outliers that range past 5. We look at these points in more detail:

```
drink_cleaned[drink_cleaned$nrel >= 5, ]
```

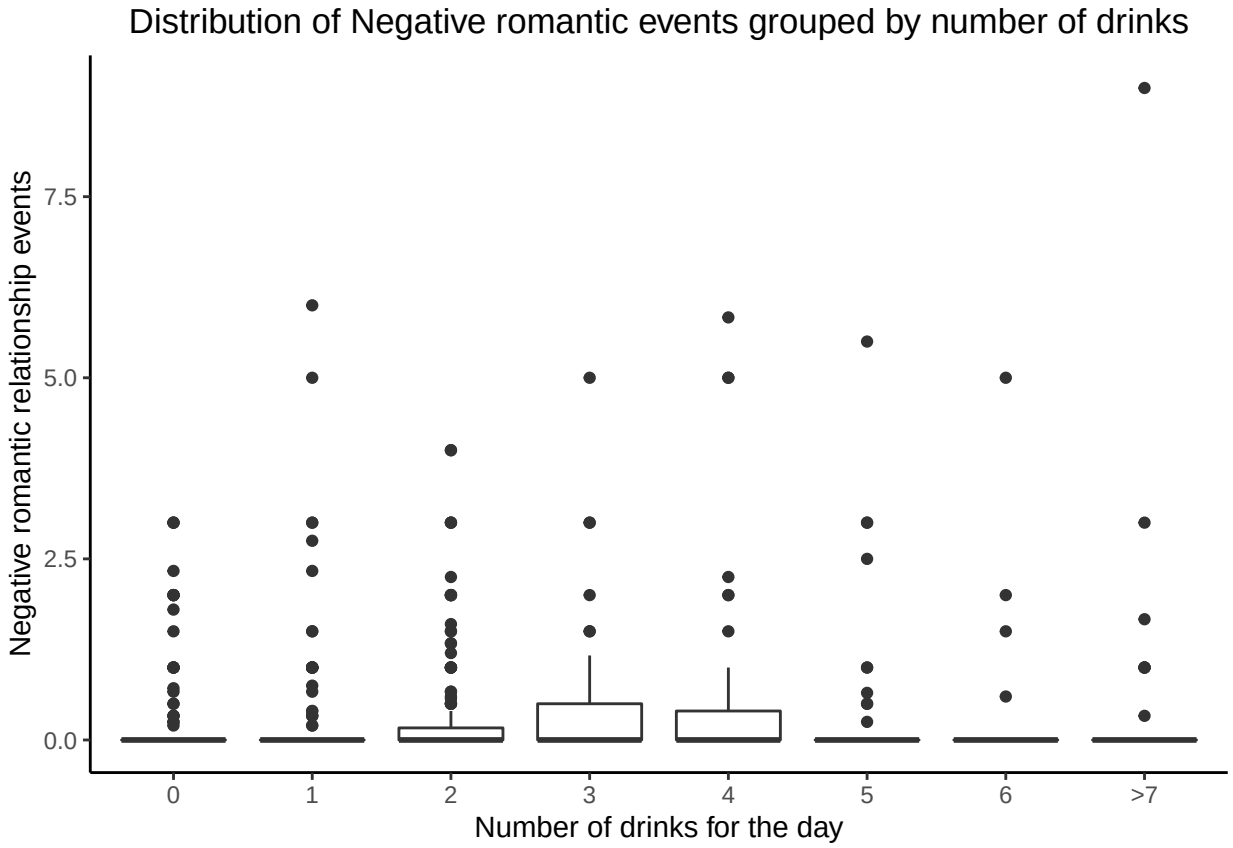
```
##       id studyday dayweek numall     nrel      prel negevent  posevent
## 11     2        4       6      4 5.833333 0.8333333 2.376667 0.9241667
## 155   33        1       3      5 5.500000 1.0000000 1.930000 0.6000000
## 311   60        3       7      4 5.000000 0.0000000 0.500000 0.5000000
## 312   60        4       1      6 5.000000 0.0000000 0.600000 0.1000000
```

```
## 315  60        7       4      3 5.000000 0.0000000 0.600000 0.1000000
## 372 104        1       3      1 5.000000 0.0000000 1.450000 0.5000000
## 533 139        1       5      4 5.000000 0.0000000 1.450000 1.3333333
## 537 139        5       2      1 6.000000 0.0000000 1.600000 0.1500000
## 592 153        4       5     10 9.000000 0.0000000 1.600000 1.4000000
##     gender rosn      age  desired factor_rosn
## 11       2  3.9 38.00137 5.666667           H
## 155      2  3.1 32.47365 5.333333           M
## 311      1  3.3 34.07255 5.666667           M
## 312      1  3.3 34.07255 5.666667           M
## 315      1  3.3 34.07255 4.666667           M
## 372      2  3.0 38.57084 7.000000           L
## 533      2  3.7 39.83025 6.000000           M
## 537      2  3.7 39.83025 4.000000           M
## 592      2  3.6 30.86653 5.666667           M
```

We see that these data points come from a wide range of subjects across many weekdays of study, which is good because that means there's not a systematic trend in where these outliers originated. For example, if all outliers came from a single subject on all 7 days, that individual might not be representative of the entire population. There are 3 data points that come from subject 60, but they are not on consecutive days. While these points are certainly outliers, they should not be removed without further information on the subjects or circumstances. We have no reason to believe that they are not representative of our population of interest (moderate to heavy drinkers) given this EDA. Model diagnostics can tell us whether these outliers are highly influential.

Next, we evaluate the relationship between numall (number of drinks) and nrel (negative romantic relationship events) for all individuals irrespective of their trait self-esteem. This will give us a sense of how these two variables are related for all individuals. To do this, we will generate boxplots of nrel for different values of numall. Note from the tabular analysis above, numall takes on highly positively skewed values. Up until 6 drinks, there are more than 20 observations for each level of numall. However, for numall > 7, we only have single obeservations. For the purpose of EDA, we will numall groups together. As a result, we will have a boxplot for 0 drinks up to 6 drinks, and an 8th boxplot for >7 drinks.

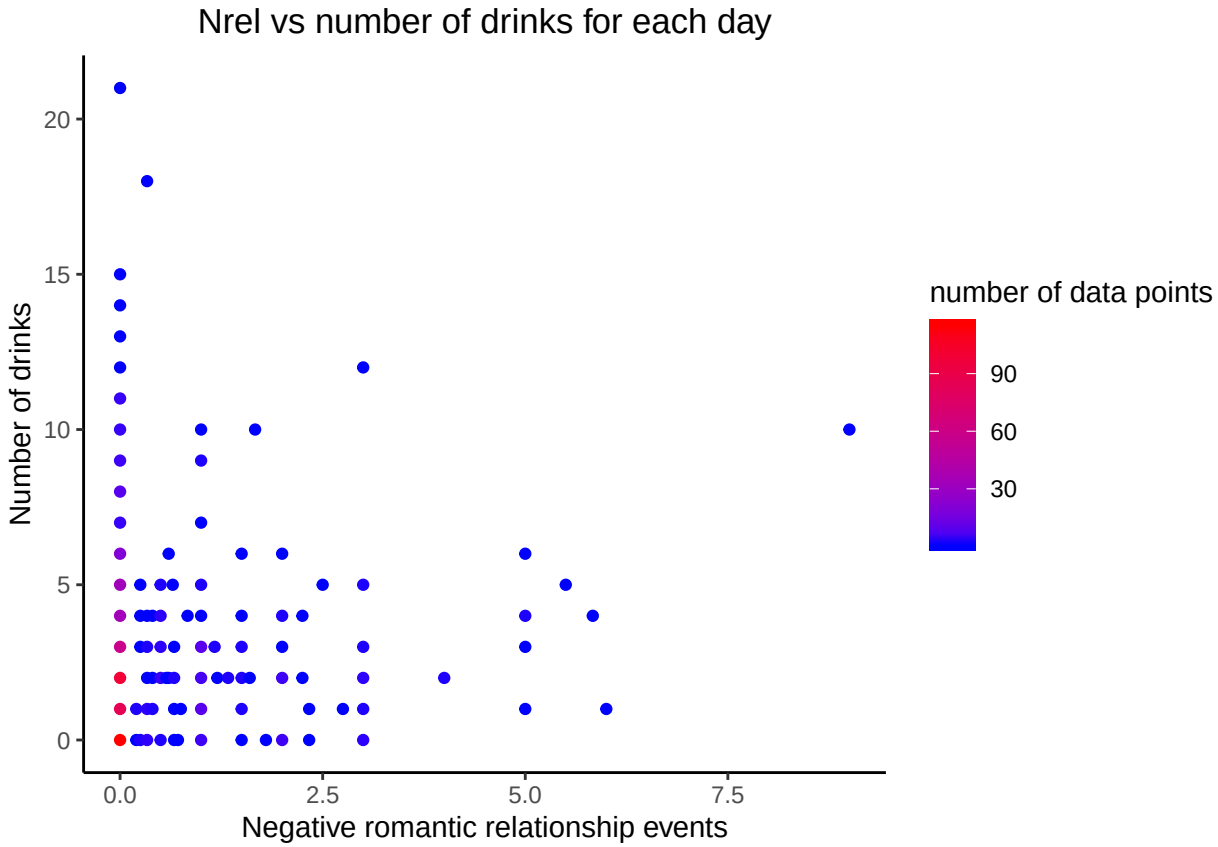
```
# Perform the binning described above
drink_cleaned$numall_binned <- ifelse(drink_cleaned$numall >
    7, yes = 7, no = drink_cleaned$numall) %>% factor()

ggplot(data = drink_cleaned, aes(x = numall_binned, y = nrel)) +
    geom_boxplot() + labs(x = "Number of drinks for the day",
    y = "Negative romantic relationship events", title = "Distribution of Negative romantic eve
    scale_x_discrete(labels = c("0", "1", "2", "3", "4", "5",
        "6", ">7")) + theme_classic() + theme(plot.title = element_text(hjust = 0.5))
```



Based on this boxplot, it does not appear that there are any strong trends in the distribution of negative romantic events grouped by number of drinks per day. Those that had one drink for example had outliers as high as ~6 negative romantic events, which is comparable to outliers observed for those who had 6 drinks.

Now, we will look at number of drinks vs nrel using a scatterplot. The advantage here is that we no longer have to compress the number of drinks to >7 in order to reduce the number of bins. We will simply treat number of drinks as an integral variable. One caveat is that some rows of our table will overlap due to the fact that most of the reported data has values of 0 for number of drinks and nrel. Since there are only 33 distinct values for nrel, for each unique pair of observed nrel/numall, we will **color the point based on how many times that particular value pair was observed. This will be labeled as number of data points.** For example, there are 5 rows where numall = 0, nrel = 1. We will color that point accordingly on the spectrum.

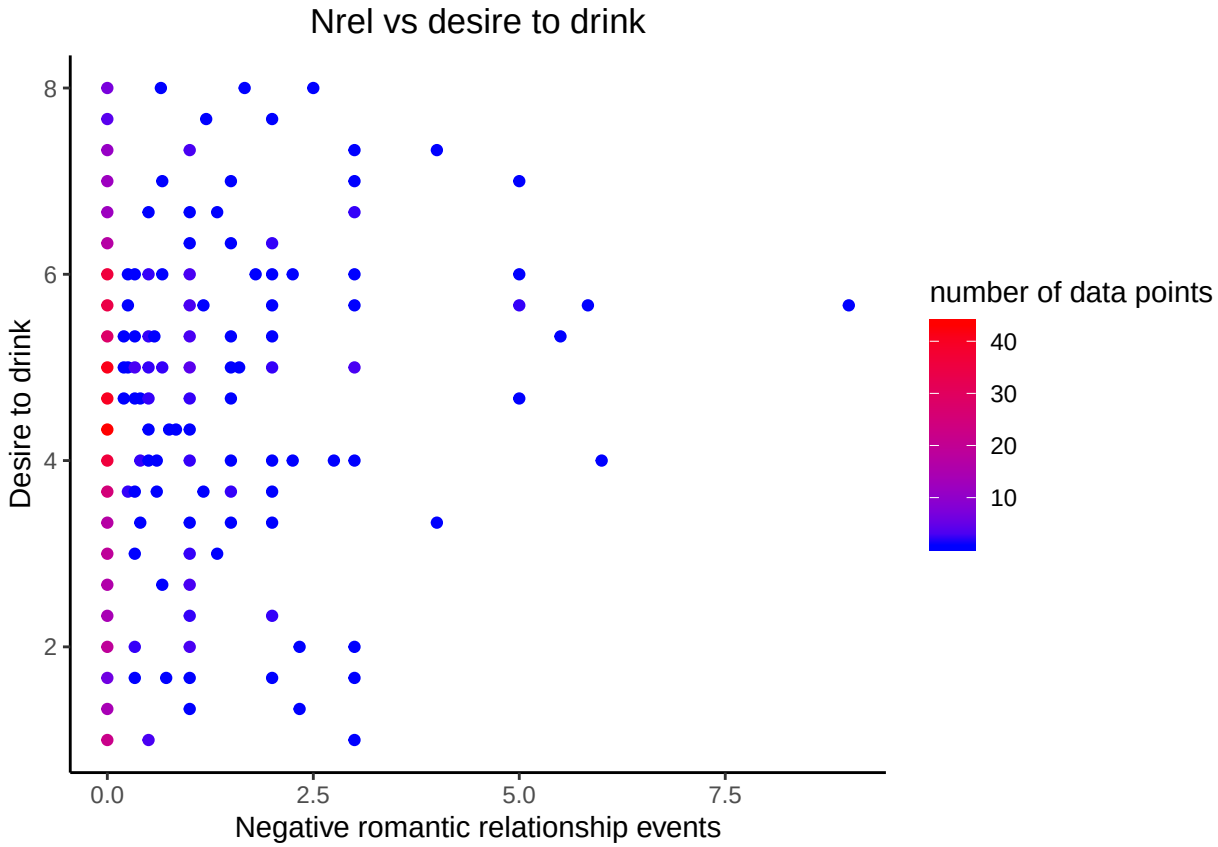
```
counts_1 <- drink_cleaned %>% dplyr::select(nrel, numall) %>%
    group_by_all() %>% summarise(count = n())

ggplot(data = counts_1, aes(x = nrel, y = numall, color = count)) +
    geom_point() + labs(y = "Number of drinks", x = "Negative romantic relationship events",
    title = "Nrel vs number of drinks for each day", color = "number of data points") +
    theme_classic() + scale_color_gradient(low = "blue", high = "red") +
    theme(plot.title = element_text(hjust = 0.5))
```



Based on this plot, we see that some minor patterns emerge. For days in which the nrel variable was >2.5, all participants had some level of alcohol consumption. Ignoring days when nrel was 0 (which was the majority), there might be a very weak upward trend. However, there are two large outliers out of the points where nrel > 0 (as observed by the boxplots as well), which are the points with >7.5 nrel, and >15 drinks.

We now look at relationship between desire to drink and nrel (negative romantic relationship events) for all individuals irrespective of their trait self-esteem. To generate a visual of this relationship, we will generate a scatterplot of these two continous variables using the same counting technique as for nrel and numall.

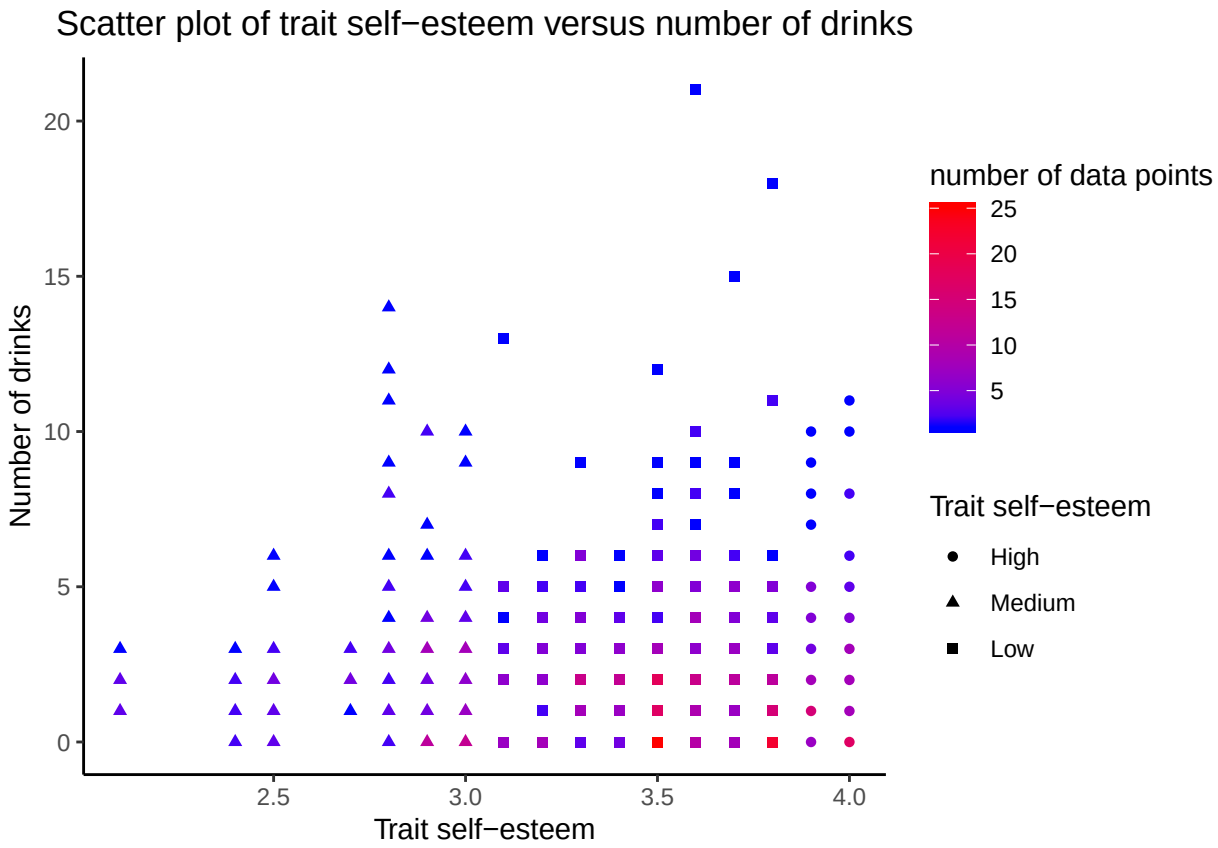
```
counts_2 <- drink_cleaned %>% select(nrel, desired) %>% group_by_all() %>%
    summarise(count = n())

ggplot(data = counts_2, aes(x = nrel, y = desired, color = count)) +
    geom_point() + labs(y = "Desire to drink", x = "Negative romantic relationship events",
    title = "Nrel vs desire to drink", color = "number of data points") +
    theme_classic() + scale_color_gradient(low = "blue", high = "red") +
    theme(plot.title = element_text(hjust = 0.5))
```



Based on this plot, we can see that for those with 0 negative romantic events, there is a large distribution for desire to drink, with most values between 4 and 6. There is not a clear trend between the two variables based on this visualization. The variance of desire to drink as a function of nrel seems relatively constant (except going up to high number of nrel due to small number of data points). Also importantly, the mean desire as a function of nrel appears constant, at around 4.5, which is closer to the global mean of desire (4.46). This indicates that nrel has likely little to no effect on desire.

We will now look at numall and desire as a function of rosn alone, before proceeding to look a trivariate analysis of numall/desire, versus rosn and nrel. For EDA, we will treat rosn as a continuous variable, noting on the plot using shape where we drew the cutoff between high and low self-esteem (which again was based off of the mean of rosn). We will again use color to indicate the number of points at a particular rosn/numall or rosn/desire pair.

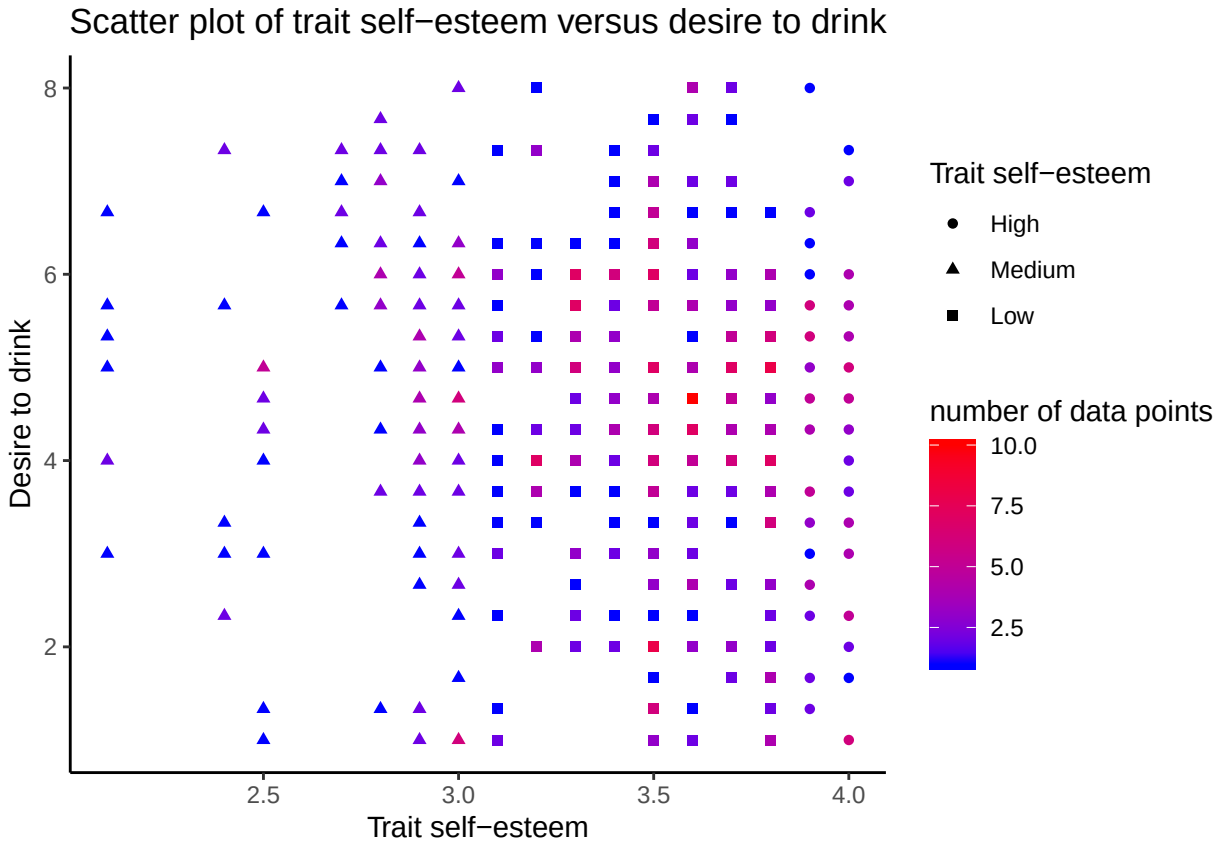
```
counts_3 <- drink_cleaned %>% select(rosn, numall, factor_rosn) %>%
    group_by_all() %>% summarise(count = n())

ggplot(counts_3, aes(x = rosn, y = numall, color = count, shape = factor_rosn)) +
    geom_point() + labs(y = "Number of drinks", x = "Trait self-esteem",
    title = "Scatter plot of trait self-esteem versus number of drinks",
    color = "number of data points", shape = "Trait self-esteem") +
    scale_shape(labels = c("High", "Medium", "Low")) + theme_classic() +
    scale_color_gradient(low = "blue", high = "red") + theme(plot.title = element_text(hjust =
```



Based on this trend, it appears that there is a trend between number of drinks the trait self-esteem: higher esteem individuals tend to drink more on average. The trend is upward, although based on our grouping, individuals at the high medium level tends to drink the most. The graph tends to have an upward to the right tend in the scatter of the dots, and also more events as we move to the right, indicating there might be a relationship between these variables.

We draw the same plot for desire to drink.

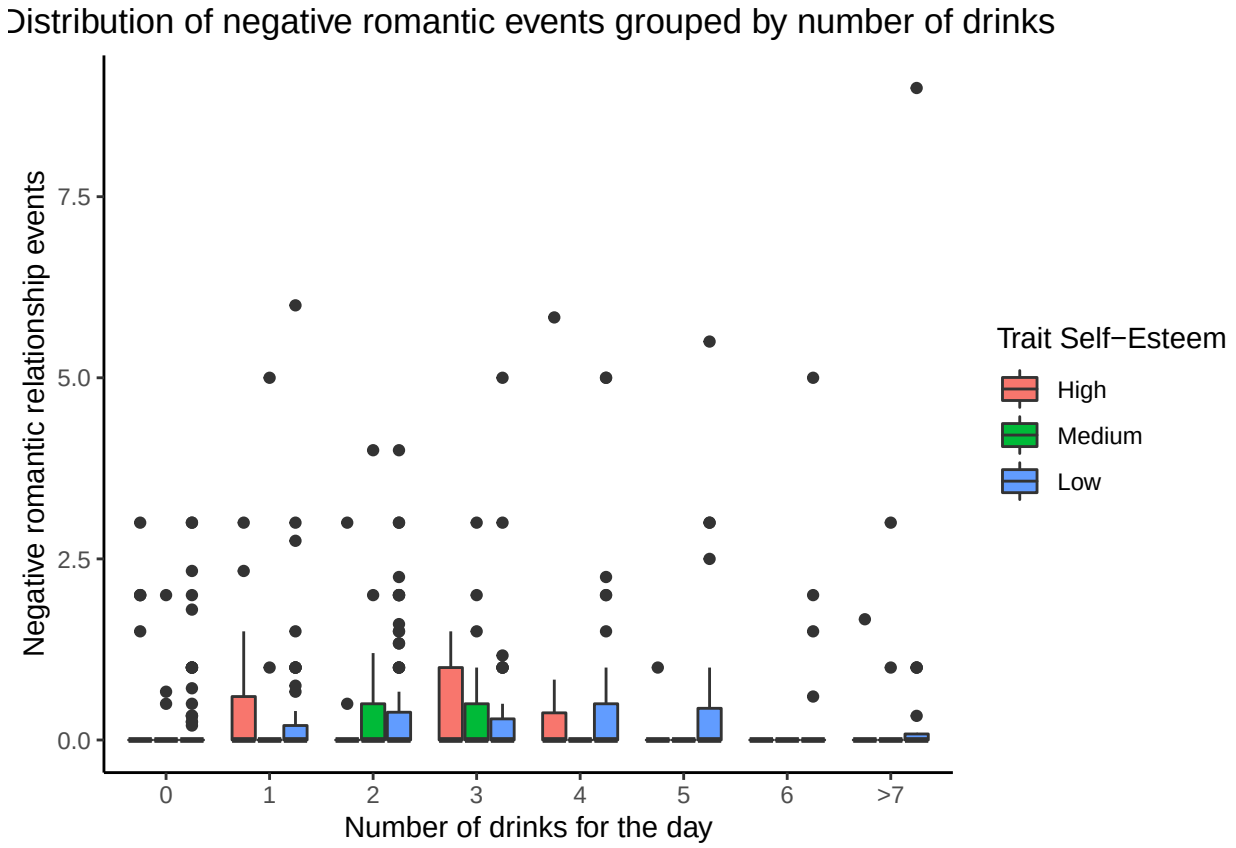
```
counts_4 <- drink_cleaned %>% select(rosn, desired, factor_rosn) %>%
    group_by_all() %>% summarise(count = n())

ggplot(counts_4, aes(x = rosn, y = desired, color = count, shape = factor_rosn)) +
    geom_point() + labs(y = "Desire to drink", x = "Trait self-esteem",
    title = "Scatter plot of trait self-esteem versus desire to drink",
    color = "number of data points", shape = "Trait self-esteem") +
    scale_shape(labels = c("High", "Medium", "Low")) + theme_classic() +
    scale_color_gradient(low = "blue", high = "red") + theme(plot.title = element_text(hjust =
```



The relationship here is less clear. The plot is nearly homogenous in terms of the scatter. Though there are fewer points for very low trait self-esteem, the span is across a similar range of desire to drink. As a result, we expect a weaker, if any, relationship between these two variables.

We now perform EDA for number of drinks as a function of nrel for the three population of individuals: high, medium, and low trait self-esteem. These were defined earlier as individuals with greater than average, and less than average trait self-esteem.

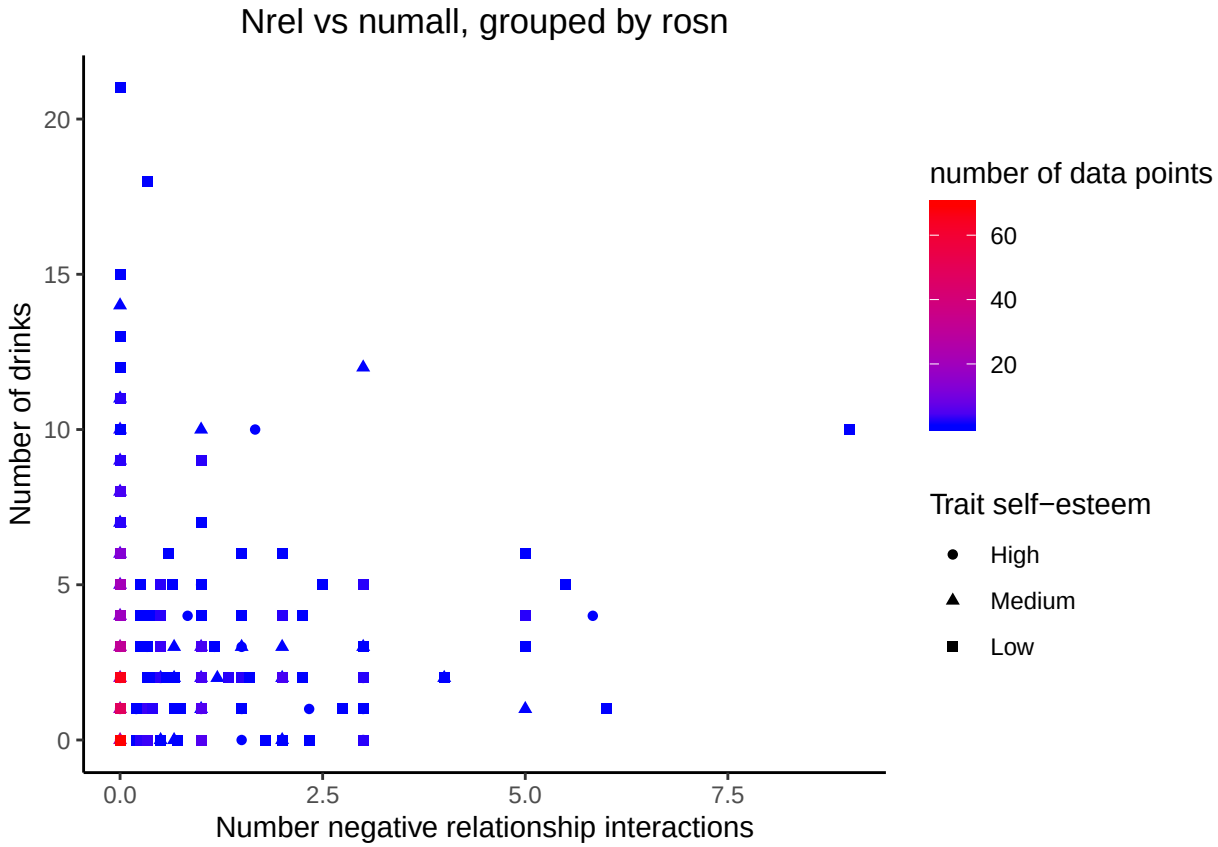
```
ggplot(data = drink_cleaned, aes(x = numall_binned, y = nrel,
    fill = factor_rosn)) + geom_boxplot() + labs(x = "Number of drinks for the day",
    y = "Negative romantic relationship events", title = "Distribution of negative romantic eve
    fill = "Trait Self-Esteem") + scale_fill_discrete(labels = c("High",
    "Medium", "Low")) + scale_x_discrete(labels = c("0", "1",
    "2", "3", "4", "5", "6", ">7")) + theme_classic() + theme(plot.title = element_text(hjust =
```



Those with lower self esteem (blue box) tend to have a slight upward trend, meaning the higher number of negative events, the more they consume. This is especially apparent in the top 75%. From 0 to 5, the 3rd quartile of low trait self-esteem individuals (top bar of the blue boxes), negative romantic events increases with increasing number of drinks. For the bars 6 and >7, this is harder to tell due to small sample size. On the other hand, those with high and medium self esteem have a weaker trend if any, indicating that a relationship between number of drinks and negative relationship events is potentially strong for those with low self-esteem, but weaker for those with medium or high self-esteem.

We explore this 3-way relationship further in a scatterplot, similar to previous, except now to show the level of trait self-esteem for each data point.

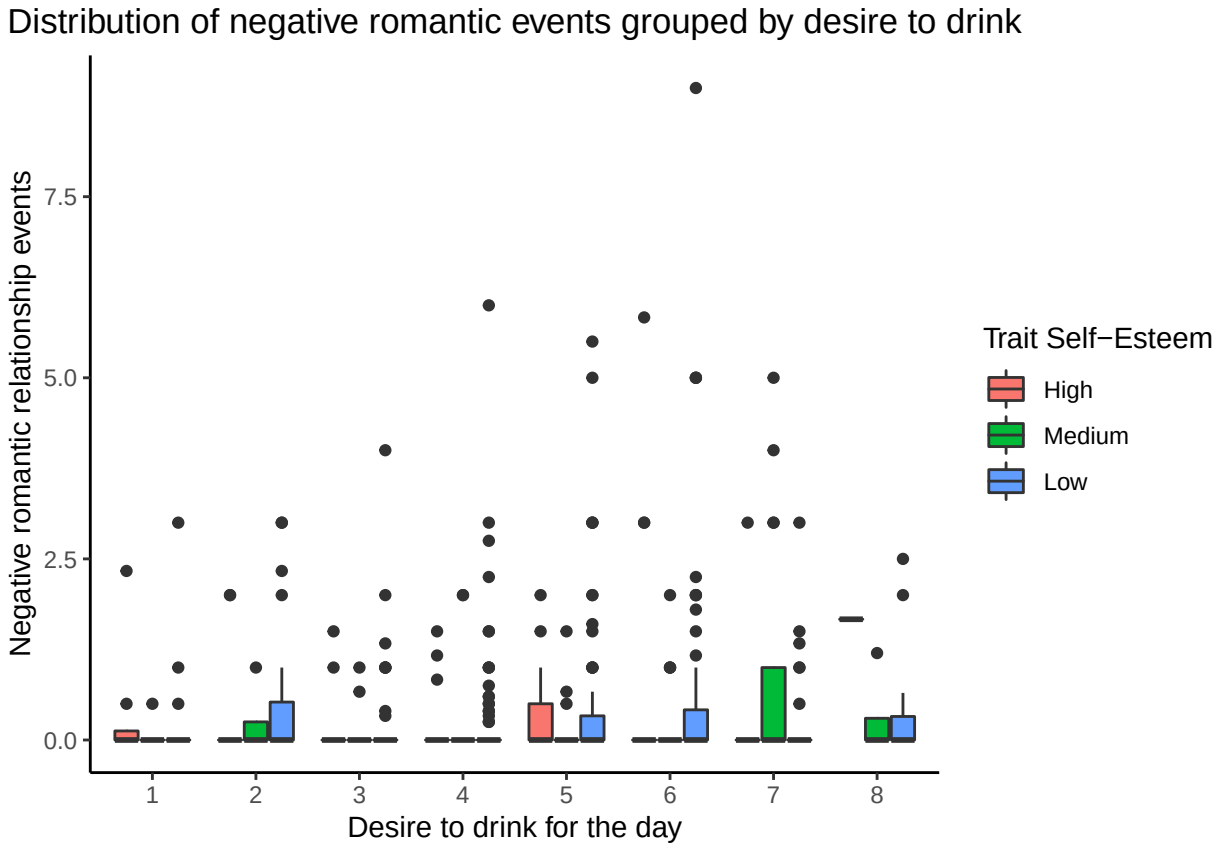
```
counts_5 <- drink_cleaned %>% select(nrel, numall, factor_rosn) %>%
    group_by_all() %>% summarise(count = n())

ggplot(counts_5, aes(x = nrel, y = numall, color = count, shape = factor_rosn)) +
    geom_point() + labs(y = "Number of drinks", x = "Number negative relationship interactions"
    title = "Nrel vs numall, grouped by rosn", color = "number of data points",
    shape = "Trait self-esteem") + scale_shape(labels = c("High",
    "Medium", "Low")) + theme_classic() + scale_color_gradient(low = "blue",
    high = "red") + theme(plot.title = element_text(hjust = 0.5))
```



Ignoring the column where nrel was 0, there is a clear trend in the upward right as seen previously. However, this time, we see that those on the right side of nrel >=5 tend to be mostly individuals with low trait self-esteem. The trend for high trait individuals (circles) is less apparent.

We perform the same EDA with desire to drink. Desired ranges from 1 to 8, with some non-integral values. We will round decimals to the nearest integer in order to get discrete bins:

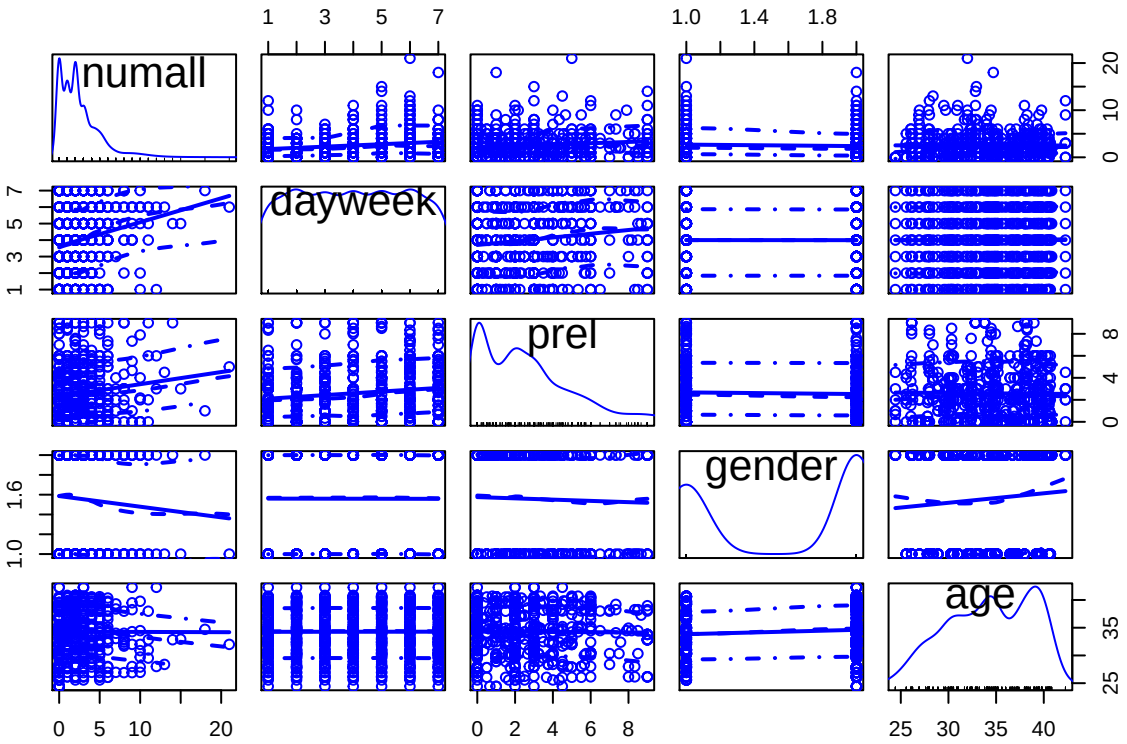
```
drink_cleaned$desired.digit <- round(drink_cleaned$desired)
ggplot(data = drink_cleaned, aes(x = factor(desired.digit), y = nrel,
    fill = factor_rosn)) + geom_boxplot() + labs(x = "Desire to drink for the day",
    y = "Negative romantic relationship events", title = "Distribution of negative romantic eve
    fill = "Trait Self-Esteem") + scale_fill_discrete(labels = c("High",
    "Medium", "Low")) + scale_x_discrete(labels = c("1", "2",
    "3", "4", "5", "6", "7", "8")) + theme_classic() + theme(plot.title = element_text(hjust =
```



Unlike the trend observed previously, this plot is relatively evently scattered from 3 to 7. Bin 8 is difficult to tell due to small sample size. It seems like bin 2 is lower than the average of 3 to 7, and the outliers tend to trend upward to the right, but trend for each level of trait of self-esteem is not clear from this visual analysis. We expect the effect of trait self-esteem on the effect of negative romantic relationship on desire to be weaker than on actual number of drinks.

In order to guide the process of building the best model, we want to capture as many relevant explanatory variables as possible in our model. For example, we expect that numall is highly impacted by the day of the week (dayweek). On average, individuals are more likely to drink on weekends than say a work night. In addition, couples are more likely to get into conflict on weekends than on a work night when they have been apart all day. We would not want to attribute all of the effect on number of drinks consumed to nrel if this is the case. To see which other variables might be relevant, we will look at a scatterplot matrix of dayweek, prel, gender, age, versus numall. We will exclude negevent and posevent because these variables are a sum of negative/positive events, including nrel and prel. We are less interested in other negative/positive events, and will choose to focus on romantic-related matter to address the author's hypothesis.

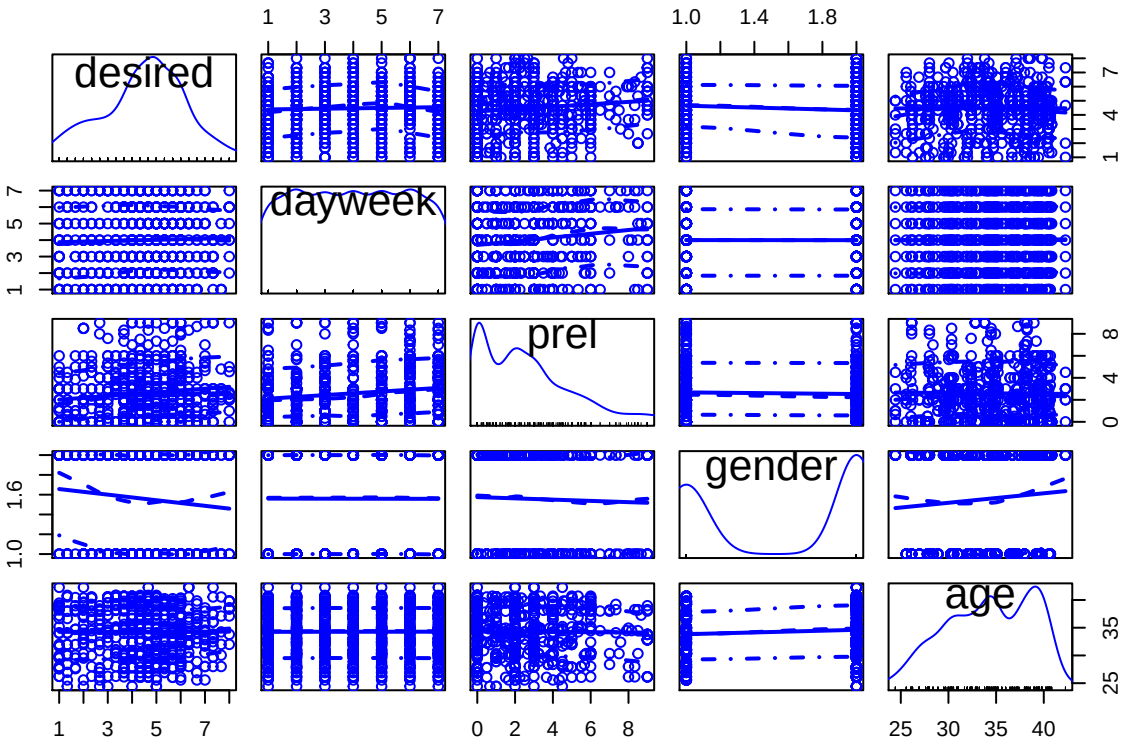
```
data_scatter <- drink_cleaned %>% select(numall, dayweek, prel,
    gender, age)
scatterplotMatrix(data_scatter)
```



Based on the scatter plot, as expected the week is an important factor. Friday, Saturday, Sunday (5,6,7) are all days in which consumption was higher compared to work days Monday-Thursday. We will want to control for day of week in our model. Females tend to drink less on average than male, which is expected, and it may very well be that males and females react differently to romantic difficulties. It will be interesting to include this variable in the model in order to control for gender. The trend with age is relatively flat, so we will not include this variable in our analysis. There also appears to be a positive trend with prel, which we will also consider in our analysis.

We will repeat this exercise for desire to drink.

```
data_scatter <- drink_cleaned %>% select(desired, dayweek, prel,
    gender, age)
scatterplotMatrix(data_scatter)
```



Interestingly, only gender appears to have a trend with desired from this analysis. We will keep this in mind as we build a model for desired.

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

We will begin with Poisson regression. As with all MLE estimation procedures, each observation is assumed to be independent of every other observation so that the likelihood can be formulated as a product of the joint probabilities of each observation. For our dataset, 1. we will assume that each individual is independent of every other individual, and 2. assume for each particular individual, that individual's response is independent for each day. Assumption 1 is easy to fulfill. As long as the participants of the study do not know each other, it is likely that their responses are very close to independent. Assumption 2 is more problematic as we might expect that at least for some participants, the number of drinks, and desire to drink, from consecutive days might be correlated with each other. However, without throwing away most of the data (keeping only a single day per participant), or using time series techniques, we have to proceed assuming Assumption 2 for all participants. If we only kept for example a single day (Saturday) for each person, the dataset would be significantly reduced and the model may not be as useful. Poisson distribution also has the property that Mean == Var, but this is for every set of π values formed by unique combinations of the explanatory variables (explanatory variable pattern, EVP). Since we do not have enough data points for each EVP, we cannot assess this assumption, although examining fitted π values can give a hint of overdispersion.

We will begin with simple models relating negative relationship to number of drinks. The simplest of the model is to fit numall directly to nrel using Poisson regression:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel}$$

```r
mod.n <- glm(numall ~ nrel, data = drink_cleaned, family = poisson(link = "log"))
mod.n
```

```
##
## Call:  glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel
##     0.89877      0.06509
##
## Degrees of Freedom: 618 Total (i.e. Null);  617 Residual
## Null Deviance:       1584
## Residual Deviance: 1577  AIC: 2946
```

The model is:

$$\log \pi_{\text{numall}} = 0.899 + 0.0651 * \text{nrel}$$

We can perform a profile LR test to check the significance of the nrel variable. We test at a significance level of 0.05, under the following hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The test is as follows:

```r
Anova(mod.n)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##      LR Chisq Df Pr(>Chisq)
## nrel   6.9559  1   0.008354 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we only have a single variable, unsurprisingly compared to intercept only model, nrel is highly statistically significant (p-value < 0.01), so we reject H0. This model would suggest that as nrel increases, there is a positive change in the mean number of drinks. However, the AIC is quite high with only a single term, meaning our model has high residual deviance and provides predictions that lie far from the observed data. We can see what happens if we include day of the week as a moderating variable, since we saw from the EDA that day of week has a high effect on number of drinks.

The model is:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek}$$

```r
mod.nd <- glm(numall ~ nrel + dayweek, data = drink_cleaned,
    family = poisson(link = "log"))
mod.nd
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek, family = poisson(link = "log"),
##     data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek
##     0.46420      0.06623      0.10334
##
## Degrees of Freedom: 618 Total (i.e. Null);  616 Residual
## Null Deviance:       1584
## Residual Deviance: 1511  AIC: 2882
```

The model is:

$$\log \pi_{\text{numall}} = 0.464 + 0.066 * \text{nrel} + 0.103 * \text{dayweek}$$

We see that AIC reduced quite a bit with the inclusion of this variable, which makes this a better model. To check the significance of the each variable, we perform Type II test at a significance of 0.05. This tests each variable under the assumption that the other variable is in the model for $j = 1, 2$:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

To test is below:

```r
Anova(mod.nd)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##         LR Chisq Df Pr(>Chisq)
## nrel       7.272  1   0.007005 **
## dayweek   65.337  1  6.312e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for both variables are $<< 0.01$, we reject H0 for both variables, and conclude that both variables are highly statistically significant given that the other variable is in the model.

We now try to improve model further in terms of AIC by adding main effects from rosn, gender, and prel. The reason we select these variables is that based on the EDA, these variables appeared to influence the number of drinks in bi-variate analysis. The model we want is:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{rosn} + \beta_4 * \text{gender} + \beta_5 * \text{prel}$$

```
mod.all <- glm(numall ~ nrel + dayweek + rosn + gender + prel,
    data = drink_cleaned, family = poisson(link = "log"))
mod.all
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek + rosn + gender + prel,
##     family = poisson(link = "log"), data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek         rosn       gender
##    0.539683     0.101849     0.095211     0.008299    -0.141849
##        prel
##    0.049242
##
## Degrees of Freedom: 618 Total (i.e. Null);  613 Residual
## Null Deviance:       1584
## Residual Deviance: 1482  AIC: 2859
```

The model we fit is:

$$\log \pi_{\text{numall}} = 0.5397 + 0.1018 * \text{nrel} + 0.0952 * \text{dayweek} + 0.0083 * \text{rosn} - 0.1418 * \text{gender} + 0.0492 * \text{prel}$$

We see that AIC is further reduced with the inclusion of all variables we believed to be important for the model. To check the significance of the each variable, we perform Type II test at a significance of 0.05 under the same null and alternative hypotheses as previously.

```
Anova(mod.all)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##         LR Chisq Df Pr(>Chisq)
## nrel      15.776  1  7.131e-05 ***
## dayweek   54.382  1  1.651e-13 ***
## rosn       0.019  1   0.891193
## gender     7.463  1   0.006298 **
## prel      21.608  1  3.344e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that nrel, dayweek, prel and gender all have p-value < 0.01, and are highly statistically significant given that the other variables are in the model. Interestingly, rosn, which is trait self-esteem, is not statistically significant, given that the other four variables are in the model. This might be interpreted as the following: Esteem on its own doesn't make a difference in terms of how much individuals drink. People are not expected to drink or less based on their level of trait self-esteem alone.

We will go ahead and remove the main effect of rosn, and evaluate the model:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel}$$

To fit:

```
mod.no.rosn <- glm(numall ~ nrel + dayweek + gender + prel, data = drink_cleaned,
    family = poisson(link = "log"))
mod.no.rosn
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek + gender + prel, family = poisson(link = "log"
##     data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek       gender         prel
##     0.56652      0.10195      0.09520     -0.14082      0.04927
##
## Degrees of Freedom: 618 Total (i.e. Null);  614 Residual
## Null Deviance:       1584
## Residual Deviance: 1482  AIC: 2857
```

The model we fit is:

$$\log \pi_{\text{numall}} = 0.5665 + 0.1020 * \text{nrel} + 0.0952 * \text{dayweek} - 0.1408 * \text{gender} + 0.0493 * \text{prel}$$

The AIC improved slightly with the removal of rosn. To evaluate the significance of other variables, we again perform profile LR:

```
Anova(mod.no.rosn)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##          LR Chisq Df Pr(>Chisq)
## nrel       15.821  1  6.964e-05 ***
## dayweek    54.370  1  1.661e-13 ***
## gender      7.513  1   0.006126 **
## prel       21.664  1  3.248e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are happy to see that with the removal of an insignificant variable, the p-values of the remaining 4 variables have not changed very much, and all four are still highly statistically significant.

One interesting hypothesis regarding trait self-esteem is that though the variable has no effect on its own, rosn might effect how much nrel and prel affect individuals' level of comsumption. In other words, an interaction with negative and positive romantic relationship events might exist. For example, we might expect that people with low self-esteem be more subject to negative influence from negative relationship compared to those with high self-esteem, which in turn leads to larger positive change in alcohol consumption compared to individuals with high self-esteem. In addition,

individuals with high self-esteem may be less positively affected by positive relationship events. We include these terms and evaluate the following model:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel} + \beta_6 * \text{rosn:prel}$$

We now fit:

```
mod.inter <- glm(numall ~ nrel + dayweek + gender + prel + nrel:rosn +
    prel:rosn, data = drink_cleaned, family = poisson(link = "log"))
mod.inter
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek + gender + prel + nrel:rosn +
##     prel:rosn, family = poisson(link = "log"), data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek       gender         prel
##     0.54146      0.49449      0.09607     -0.12578      0.10777
##   nrel:rosn    prel:rosn
##    -0.11329     -0.01728
##
## Degrees of Freedom: 618 Total (i.e. Null);  612 Residual
## Null Deviance:       1584
## Residual Deviance: 1477  AIC: 2856
```

The model we fit is:

$$\begin{aligned}\log \pi_{\text{numall}} = 0.541 + 0.494 * \text{nrel} + 0.096 * \text{dayweek} - 0.126 * \text{gender} \\ + 0.108 * \text{prel} - 0.113 * \text{nrel:rosn} - 0.017 * \text{prel:rosn}\end{aligned}$$

The AIC improved very slightly, and the coefficient (strength of effect) of nrel increased about 5 fold. To perform a Type II test, we evaluate the significance of each coefficient given that the other coefficients are in the model. For example, for nrel, we have the following hypotheses:

$$H_0 : log\pi_{\text{numall}} = \beta_0 + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel} + \beta_6 * \text{rosn:prel}$$
$$H_1 : \log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel} + \beta_6 * \text{rosn:prel}$$

For the interaction between rosn and nrel, we evaluate:

$$H_0 : \log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_6 * \text{rosn:prel}$$
$$H_1 : \log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel} + \beta_6 * \text{rosn:prel}$$

We now evaluate the significance of all variables given that other variables are included in the model:

```
Anova(mod.inter)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##           LR Chisq Df Pr(>Chisq)
## nrel        15.780  1  7.115e-05 ***
## dayweek     55.177  1  1.102e-13 ***
## gender       5.917  1    0.01499 *
## prel        22.148  1  2.524e-06 ***
## nrel:rosn    3.007  1    0.08289 .
## prel:rosn    1.492  1    0.22194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interestingly, the interaction between prel and rosn is not statistically significant, and the interaction between nrel and rosn is marginally statistically significant. All other variables remain statistically significant. This means that trait self-esteem has less of an impact on the effect of positive relationship events on number of drinks. We will remove this interaction and explore the following model:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel}$$

To fit:

```
mod.nrel.inter <- glm(numall ~ nrel + dayweek + gender + prel +
    nrel:rosn, data = drink_cleaned, family = poisson(link = "log"))
mod.nrel.inter
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek + gender + prel + nrel:rosn,
##     family = poisson(link = "log"), data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek       gender         prel
##     0.54511      0.53224      0.09626     -0.13108      0.04982
##   nrel:rosn
##    -0.12410
##
## Degrees of Freedom: 618 Total (i.e. Null);  613 Residual
## Null Deviance:       1584
## Residual Deviance: 1478  AIC: 2855
```

The model we fit is:

$$\begin{aligned}\log \pi_{\text{numall}} = 0.545 + 0.532 * \text{nrel} + 0.096 * \text{dayweek} - 0.131 * \text{gender} \\ + 0.049 * \text{prel} - 0.124 * \text{nrel:rosn}\end{aligned}$$

We will perform Type II test as before for all coefficients:

```
Anova(mod.nrel.inter)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##           LR Chisq Df Pr(>Chisq)
## nrel        15.821  1  6.964e-05 ***
## dayweek     55.391  1  9.878e-14 ***
## gender       6.477  1    0.01093 *
## prel        22.148  1  2.524e-06 ***
## nrel:rosn    3.677  1    0.05515 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variables nrel, dayweek, and prel are all highly statistically significant with p-value < 0.01. Gender is boardline highly statistically significant, and the interaction between nrel:rosn is boardline significant. This is the best AIC we have had so far. As a result, it is reasonable to use this as our best model as the effect of the variables make intuitive sense, and the model is the product of extensive feature engineering.

```
mod.best <- mod.nrel.inter
mod.best
```

```
##
## Call:  glm(formula = numall ~ nrel + dayweek + gender + prel + nrel:rosn,
##     family = poisson(link = "log"), data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek       gender         prel
##     0.54511      0.53224      0.09626     -0.13108      0.04982
##   nrel:rosn
##    -0.12410
##
## Degrees of Freedom: 618 Total (i.e. Null);  613 Residual
## Null Deviance:       1584
## Residual Deviance: 1478  AIC: 2855
```

To evaluate whether nrel has any effect in this model, we will fit a second model without any nrel terms, and perform profile LR test.

The null hypothesis model we fit is:

$$\log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{dayweek} + \beta_2 * \text{gender} + \beta_3 * \text{prel}$$

The fit is:

```
mod.H0 <- glm(numall ~ dayweek + gender + prel, data = drink_cleaned,
    family = poisson(link = "log"))
```

We use this model as purely the null hypothesis, and the test is stated as follows (at alpha = 0.05):

$$H_0 : \log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{dayweek} + \beta_2 * \text{gender} + \beta_3 * \text{prel}$$
$$H_1 : \log \pi_{\text{numall}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn:nrel}$$

The test:

```
anova(mod.H0, mod.best, test = "LR")
```

```
## Analysis of Deviance Table
##
## Model 1: numall ~ dayweek + gender + prel
## Model 2: numall ~ nrel + dayweek + gender + prel + nrel:rosn
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       615     1497.7
## 2       613     1478.2  2   19.498 5.835e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value $<< 0.01$, we reject H0 that the coefficients on any nrel term is not important, and conclude that we have strong evidence for the joint significance of the two terms.

The estimated change in the mean number of drinks for a c increase in nrel is:

$$\frac{e^{0.545+0.532*(\text{nrel}\ +\ \text{c})+0.096*\text{dayweek}-0.131*\text{gender}+0.049*\text{prel}-0.124*\text{rosn}\ (\text{nrel}\ +\ \text{c})}}{e^{0.545+0.532*\text{nrel}+0.096*\text{dayweek}-0.131*\text{gender}+0.049*\text{prel}-0.124*\text{nrel rosn}}} = e^{0.532*c-0.124*\text{rosn}*c}$$

In order to interpret the coefficients for the effect of nrel on number of drinks, we have to look at a change in the variable nrel at a fixed value of rosn. The best value to pick for this is the sample standard deviation of nrel:

```
nrel.sd <- sd(drink_cleaned$nrel)
nrel.sd
```

```
## [1] 0.9425834
```

For the value of rosn, we will use values at the first quartile, median, and third quartile in order to get a sense of how different levels of trait self-esteem influence the effect of nrel.

```
rosn.quart <- quantile(drink_cleaned$rosn) %>% unname()
rosn.quart.first <- rosn.quart[2]
rosn.quart.median <- rosn.quart[3]
rosn.quart.third <- rosn.quart[4]
```

To calculate the 95% confidence interval for this linear combination of parameters, we will use mcprofile, where we can calculate both Wald and LR. We can also extract the estimate from this calculation. In addition, it will make the most sense to convert to a percent change for interpretation.

```
library(mcprofile)

# mcprofile setup; CI for 95% three rows for each level of
# rosn
K <- matrix(data = c(0, nrel.sd, 0, 0, 0, nrel.sd * rosn.quart.first,
    0, nrel.sd, 0, 0, 0, nrel.sd * rosn.quart.median, 0, nrel.sd,
    0, 0, 0, nrel.sd * rosn.quart.third), nrow = 3, byrow = TRUE)
linear.combo <- mcprofile(mod.best, CM = K)
```

```r
# Profile LR
ci.lr.beta <- confint(linear.combo, level = 0.95)

# Wald
wald.lc <- wald(linear.combo)
ci.wald.beta <- confint(wald.lc, level = 0.95)

# Convert to percent change
estimate.pc <- 100 * (exp(ci.lr.beta$estimate) - 1)

# LR CI in percent change
lr.pc <- 100 * (exp(ci.lr.beta$confint) - 1)

# Wald CI in percent change
lr.wald <- 100 * (exp(ci.wald.beta$confint) - 1)

# Compile
headers <- c("Estimate", "lower95%", "upper95%")
rows <- c("q1.rosn", "median.rosn", "q3.rosn")
lr <- data.frame(cbind(estimate.pc, lr.pc))
w <- data.frame(cbind(estimate.pc, lr.wald))

colnames(lr) <- headers
rownames(lr) <- rows

colnames(w) <- headers
rownames(w) <- rows

print("Wald")
```

```
## [1] "Wald"
```

```r
w
```

```
##              Estimate  lower95% upper95%
## q1.rosn     13.581898  6.762982 20.83634
## median.rosn  9.665043  4.006312 15.63165
## q3.rosn      5.883260 -1.474423 13.79040
```

```r
print("LR")
```

```
## [1] "LR"
```

```r
lr
```

```
##              Estimate  lower95% upper95%
## q1.rosn     13.581898  6.486184 20.54800
## median.rosn  9.665043  3.801963 15.41666
## q3.rosn      5.883260 -1.769137 13.46557
```

Since Wald and profile LR are similar, we will use the profile LR CIs. Thus, we can say that with

95% confidence, the estimated number of drinks increase by about 6.48% to 20.55% for individuals with trait self-esteem value of 3.2 (first quartile), for every standard deviation increase in negative relationship interactions, keeping all other variables in the model constant. The estimated increase for mean number of drinks is 13.6%. For every standard deviation increase in negative relationship interactions, keeping all other variables in the model constant, the estimated number of drinks change by about 3.80% to 15.42% for individuals with trait self-esteem value of 3.5 (median), and by -1.78% to 13.47% for individual with self-esteem value of 3.8 (third quartile).

Since the confidence interval does not include 0 for individuals at low (1st quartile) and medium (median) levels of trait self-esteem, we can say that negative relationship interactions is statistically significant given that other variables in the model is included. In other words, for individuals with low and medium levels of trait self-esteem, there is statistical evidence that negative relationship interactions are associated with higher alcohol consumption levels. However, interestingly, we note that since the interval for high trait self-esteem individuals includes 0, the change is not statistically significant for these individuals. Since the interaction between nrel and rosn is statistically significant, the effect of nrel on numall is dependent on the level of rosn.

Next, we look at the effect of negative relationship on desire to drink. We will also use Poisson regression, with the desire to drink rounded to the nearest digit. This was already done as part of the EDA.

Rather than walking through the full exercise, we will start with the model with all variables including interactions, and remove using hypothesis testing and AIC validation (backwards-type selection).

The first model we fit is:

$$\log \pi_{\text{desired}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{dayweek} + \beta_3 * \text{gender} + \beta_4 * \text{prel} + \beta_5 * \text{rosn} + \beta_6 * \text{rosn:nrel} + \beta_7 * \text{rosn:prel}$$

```
mod.d.full <- glm(desired.digit ~ nrel + dayweek + gender + prel +
    rosn + nrel:rosn + prel:rosn, data = drink_cleaned, family = poisson(link = "log"))
mod.d.full
```

```
##
## Call:  glm(formula = desired.digit ~ nrel + dayweek + gender + prel +
##     rosn + nrel:rosn + prel:rosn, family = poisson(link = "log"),
##     data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel      dayweek       gender         prel
##   1.6895535    0.2241568   -0.0002384   -0.0668237    0.0693335
##        rosn    nrel:rosn    prel:rosn
##  -0.0490184   -0.0487332   -0.0137523
##
## Degrees of Freedom: 618 Total (i.e. Null);  611 Residual
## Null Deviance:       461.7
## Residual Deviance: 439.1     AIC: 2484
```

$$\begin{aligned}\log \pi_{\text{desired}} = 1.689 &+ 0.224 * \text{nrel} - 0.0002 * \text{dayweek} - 0.067 * \text{gender} \\ &+ 0.069 * \text{prel} - 0.0490 * \text{rosn} - 0.049 * \text{rosn:nrel} - 0.014 * \text{rosn:prel}\end{aligned}$$

We now perform the Type II profile LR test on each coefficient to determine statistical significance of each coefficient at significance level 0.05.

```
Anova(mod.d.full)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired.digit
##           LR Chisq Df Pr(>Chisq)
## nrel        7.8037  1   0.005214 **
## dayweek     0.0006  1   0.980269
## gender      2.9332  1   0.086773 .
## prel        8.2518  1   0.004071 **
## rosn        5.3508  1   0.020713 *
## nrel:rosn   0.8063  1   0.369220
## prel:rosn   0.6418  1   0.423066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this, we see that prel and nrel are but highly statistically significant. There does not appear to be any significance for dayweek, support by the fact the estimate for the coefficient is basically 0. Gender is only marginally significant. Since we fail to reject our stated H0, we will also remove this variable. The interactions of prel and nrel with rosn do not appear significant from this test, but we will first see whether dayweek and gender are responsible for absoring their effect. We will remove only dayweek and gender, refit, and retest, done in one step below:

```
mod.d.test <- glm(desired.digit ~ nrel + prel + rosn + nrel:rosn +
    prel:rosn, data = drink_cleaned, family = poisson(link = "log"))
mod.d.test
```

```
##
## Call:  glm(formula = desired.digit ~ nrel + prel + rosn + nrel:rosn +
##     prel:rosn, family = poisson(link = "log"), data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel         prel         rosn    nrel:rosn
##     1.62662      0.22978      0.06813     -0.06081     -0.05157
##   prel:rosn
##    -0.01337
##
## Degrees of Freedom: 618 Total (i.e. Null);  613 Residual
## Null Deviance:       461.7
## Residual Deviance: 442   AIC: 2483
```

```
Anova(mod.d.test)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: desired.digit
##           LR Chisq Df Pr(>Chisq)
## nrel        6.7567  1   0.009340 **
## prel        8.4939  1   0.003563 **
## rosn        6.7367  1   0.009445 **
## nrel:rosn   0.9104  1   0.339997
## prel:rosn   0.6041  1   0.437017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on this, we can see that all of the main effects of prel, nrel, rosn are highly statistically significant, while the interaction between rosn of either prel or nrel are not signficant, given the other variables in this model. This means that while negative and positive romantic interactions both increase the desire to drink, the increase is not altered by different levels of trait self-esteem as long as trait self-esteem itself is in the model. Trait self-esteem is also significant with a negative coefficient, meaning as this term increases, the desire to drink decreases.

We will evaluate the following scenario: are interactions important if the main effect from trait self-esteem is left out of the model? We again fit and test in one step:

```
mod.d.test2 <- glm(desired.digit ~ nrel + prel + nrel:rosn +
    prel:rosn, data = drink_cleaned, family = poisson(link = "log"))
Anova(mod.d.test2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired.digit
##           LR Chisq Df Pr(>Chisq)
## nrel        6.4260  1   0.011246 *
## prel        8.2632  1   0.004046 **
## nrel:rosn   1.7623  1   0.184333
## prel:rosn   4.5950  1   0.032066 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, prel:rosn becomes significant, but not nrel:rosn. We will remove nrel:rosn, and ask, given that main effects are all included in the model, is prel:rosn signficant without the interaction between nrel:rosn.

```
mod.d.test3 <- glm(desired.digit ~ nrel + prel + rosn + prel:rosn,
    data = drink_cleaned, family = poisson(link = "log"))
Anova(mod.d.test3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired.digit
##           LR Chisq Df Pr(>Chisq)
## nrel        6.7567  1   0.009340 **
## prel        8.3334  1   0.003892 **
## rosn        6.7367  1   0.009445 **
```

```
## prel:rosn   0.3870  1   0.533882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the interaction increased with the removal of nrel:rosn, so the two interactions do not confound each other's effect. Finally, since authors are interested in negative relationships, we will test to see whether exclusion of all variables except nrel and nrel:rosn produces a significant result for the interaction:

```
mod.d.test4 <- glm(desired.digit ~ nrel + nrel:rosn, data = drink_cleaned,
    family = poisson(link = "log"))
Anova(mod.d.test4)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired.digit
##           LR Chisq Df Pr(>Chisq)
## nrel        3.8290  1    0.05037 .
## nrel:rosn   2.3381  1    0.12625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At a significance level of 0.05, the interaction is not statistically significant, so we fail to reject H0 that the interaction is not significant.

The conclusion from these tests is that with the inclusion of the main effects prel, nrel, and rosn, the interactions between nrel:rosn, and prel:rosn, are statistically insignificant. Additionally, with the removal of any main effect, the statistical significance of the other main effects decrease, signifying potential joint significance. If we remove the main effect rosn, only prel:rosn becomes statistically significant. We generally prefer main effects over interaction terms in the case of confounding effects as observed here because main effects directly account for the effect of the variable, rather than the effect on the effect of another variable. As a result, we will model desire using solely the main effects from prel, nrel, and rosn. The model we will use is:

$$\log \pi_{\text{desired}} = \beta_0 + \beta_1 * \text{nrel} + \beta_2 * \text{prel} + \beta_3 * \text{rosn}$$

```
mod.d.best <- glm(desired.digit ~ nrel + prel + rosn, data = drink_cleaned,
    family = poisson(link = "log"))
mod.d.best
```

```
##
## Call:  glm(formula = desired.digit ~ nrel + prel + rosn, family = poisson(link = "log"),
##     data = drink_cleaned)
##
## Coefficients:
## (Intercept)         nrel         prel         rosn
##     1.81411      0.05181      0.02313     -0.11628
##
## Degrees of Freedom: 618 Total (i.e. Null);  615 Residual
## Null Deviance:       461.7
```

```
## Residual Deviance: 443.3     AIC: 2480
```

The model we've fit is:

$$\log \pi_{\text{desired}} = 1.814 + 0.052 * \text{nrel} + 0.023 * \text{prel} - 0.116 * \text{rosn}$$

In support of a model without interaction terms is the AIC, where we've reduced the AIC from 2484 (original full model) to 2483 (without gender and dayweek) to 2480 (current model without interactions).

To verify the statistical significance of all variables in the model, we perform a Type II test at significance level 0.05:

```
Anova(mod.d.best)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: desired.digit
##      LR Chisq Df Pr(>Chisq)
## nrel   6.8675  1   0.008778 **
## prel   8.3334  1   0.003892 **
## rosn   6.7367  1   0.009445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on this model, the estimated change in the desired number of drinks for a standard deviation increase in nrel is:

$$\frac{e^{1.814+0.052*(\text{nrel}\ +\ \text{c})+0.023*\text{prel}-0.116*\text{rosn}}}{e^{1.814+0.052*\text{nrel}+0.023*\text{prel}-0.116*\text{rosn}}} = e^{0.052*c}$$

We would also like to generate a confidence interval. Since there is only one coefficient of interest β_1, we can directly get the CI for this coefficient using confint, and then exponentiate to get the desired CI corresponding to a sd.nrel change in nrel. We will also calculate the wald and show that they are approximately the same:

```
# Wald
wald.conf <- confint.default(mod.d.best, parm = "nrel", level = 0.95)
wald.ci <- exp(confint.default(mod.d.best, parm = "nrel", level = 0.95) *
    nrel.sd)

# Profile LR
conf.nrel <- confint(mod.d.best, parm = "nrel", level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
ci.nrel <- exp(conf.nrel * nrel.sd)

# wald and profile LR comparison
compare <- rbind(wald = wald.ci, profileLR = ci.nrel)
rownames(compare) <- c("wald", "profileLR")
compare
```

```
##                2.5 %   97.5 %
## wald       1.013515 1.087889
## profileLR 1.012674 1.087013
```

```
# Estimate, and convert to percent change
estimate.pc <- 100 * (exp(nrel.sd * mod.d.best$coefficients[["nrel"]]) -
    1)
ci.nrel.pc <- 100 * (ci.nrel - 1)
t(data.frame(percent_change = c(estimate = estimate.pc, ci.nrel.pc)))
```

```
##                 estimate    2.5 %   97.5 %
## percent_change 5.004382 1.267375 8.701272
```

Since Wald and profile LR are similar, we will use the profile LR. We can say that with 95% confidence, the estimated mean desired number of drinks increase by 1.27% to 8.70% for every standard deviation increase in negative relationship interactions, keeping all other variables in the model constant. The estimated mean change is 5%. Since the confidence interval does not include 0, we can say that negative relationship interactions is statistically significant given that other variables in the model is included, consistent with our Type II hypothesis tests. Since our model does not have interaction terms with rosn, this is the same effect for all levels of rosn. This effect is clearly much smaller than the model for actual number of drinks.

Finally, since desired was a value that ranged from 1-8, we will use a ordinal response model, also known as proportional odds, in order to look at the probability that a respondant would be at least as high of a category. More formally, letting D be the desired level, and j from 1 - 8:

$$\log \frac{P(D < j)}{1 - P(D < j)} = \beta_{j0} + \beta_1 * x_1 + \beta_2 * x_2 + ...$$

In other words, the logit of the CDF is modeled as a linear response with the same coefficients for the explanatory variables, but each discrete level of the CDF with its own intercept. We will use the same variables as in our best model.

```
mod.po <- polr(factor(desired.digit) ~ nrel + prel + rosn, data = drink_cleaned,
    method = "logistic")
mod.po
```

```
## Call:
## polr(formula = factor(desired.digit) ~ nrel + prel + rosn, data = drink_cleaned,
##     method = "logistic")
##
## Coefficients:
##       nrel       prel       rosn
##  0.2664925  0.1062544 -0.5938222
##
## Intercepts:
##        1|2        2|3        3|4        4|5        5|6        6|7
## -4.3482096 -3.4008126 -2.7424973 -1.8097924 -0.7539176  0.5442240
##        7|8
##  2.0495072
##
```

```
## Residual Deviance: 2335.11
## AIC: 2355.11
```

We next check the significance of each explanatory variable using a likelihood ratio test and find that nrel is highly statistically significant.

```
Anova(mod.po)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(desired.digit)
##      LR Chisq Df Pr(>Chisq)
## nrel   12.288  1  0.0004559 ***
## prel   12.318  1  0.0004486 ***
## rosn   11.981  1  0.0005375 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the coefficient on nrel is positive, indicating that higher number of negative relationship events results in a larger category of desired drinking.

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

The relationship between desired number of drinks and negative relationship interactions does not differ according to individuals' level of trait self-esteem. This was evidenced by our mod.d.best, where we extensively showed under multiple situations that the interaction between negative relationship interactions and trait self-esteem was not statistically significant, including in our final model where we account for the main effects of negative relationship interactions, positive relationship interactions, and trait self-esteem. Even with the removal of potential confounding prel and rosn, the interaction was still insigificant. We conclude that the desired number of drinks is affected by negative relationship interactions (an estimated 5% increase for each standard deviation increase in nrel) the same way for all levels of trait self-esteem.

For the effect of trait self-esteem on the relationship between actual number of drinks and negative relationship interactions, we saw that the interaction term is marginally statistically significant, which provides evidence that the trait variable self-esteem has a moderating effect. In other words, the relationship between drinking and negative relationship interactions statistically differs depending on an individual's level of trait self-esteem. Furthermore, the lower the self-esteem, the largest the estimate for the change in the mean number of drinks for a standard deviation change in negative relationship interactions, keeping other variables constant, which supports the author's original hypothesis that lower self-esteem individuals are affected more than high self-esteem individuals by number of negative relationship events.

Our best model estimated the following estimated percent change in the means for a standard deviation change in nrel at three levels of trait self-esteem. The confidence intervals for individuals with first quartile, median, and third quartile rosn:

```
lr
```

```
##              Estimate  lower95% upper95%
## q1.rosn     13.581898  6.486184 20.54800
```

```
## median.rosn  9.665043  3.801963 15.41666
## q3.rosn      5.883260 -1.769137 13.46557
```

The confidence intervals overlap, so it is difficult to determine whether there is statistically significant difference between the changes **at any two levels** of rosn. In particular, we are interested in high trait self-esteem individuals (q3), and low trait individuals (q1). We can see that there is a clear trend in the estimates themselves, decreasing as rosn increases. We see also that between q1 and q3, the CIs for the two levels do not contain the estimate of the other level (for example, 6.48-20.55 does not contain 5.88). Individuals at q3 have an estimated CI that contains 0 (meaning the change in the mean number of drinks is not statistically significant), while those at q1 do not contain 0 indicating statistical significance. Both observations provide concrete evidence that the difference between q3 and q1 individuals is statistically explained by the model. Overall, the increase in drinking as a result of increased negative relationship interactions decreases as trait self-esteem increases.

*This work was done as part of the W271 - Statistical Methods for Discrete Response, Time Series, and Panel Data course under the U.C. Berkeley Master of Information and Data Science program.