# Lab 1, Short Questions

## Contents

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

These questions are taken from Question 12 of chapter 3 of the textbook (Bilder and Loughin's "Analysis of Categorical Data with R.

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

```
#cereal <- read_csv('../data/short-questions/cereal_dillons.csv')
cereal <- read_csv('cereal_dillons.csv')
```

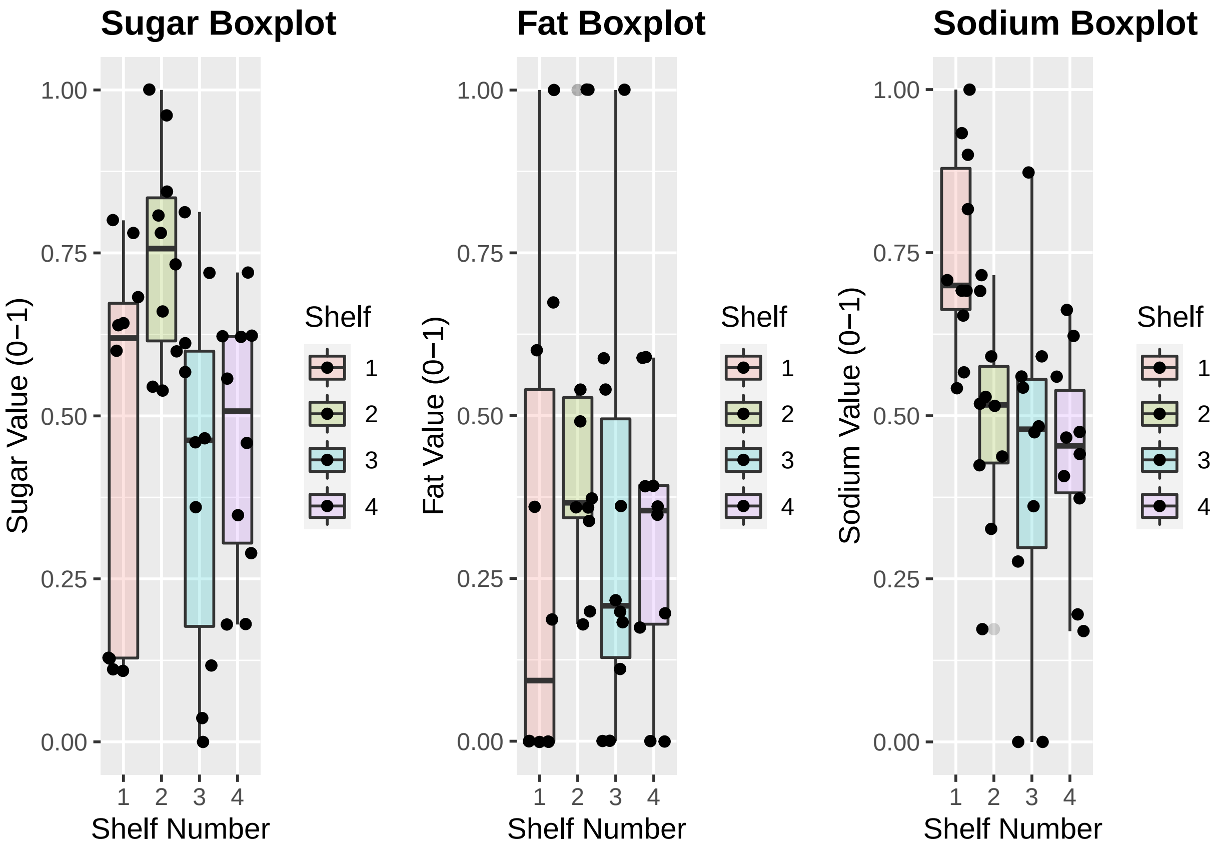
## 1.1 Recode Data

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

---

```
# Set up the function for normalizing data between 0 and 1
stand01 <- function (x) {
  (x - min(x)) / ( max(x) - min(x))
  }

#Establish new dataframe
cereal2 <- data.frame(Shelf = cereal$Shelf,
                      sugar = stand01(x = cereal$sugar_g / cereal$size_g),
                      fat = stand01 (x = cereal$fat_g / cereal$size_g ),
                      sodium = stand01 (x = cereal$sodium_mg / cereal$size_g ))

# Setting shelf up as a factor
cereal2 <- cereal2 %>%
  mutate(
    Shelf = factor(Shelf)
      )
```

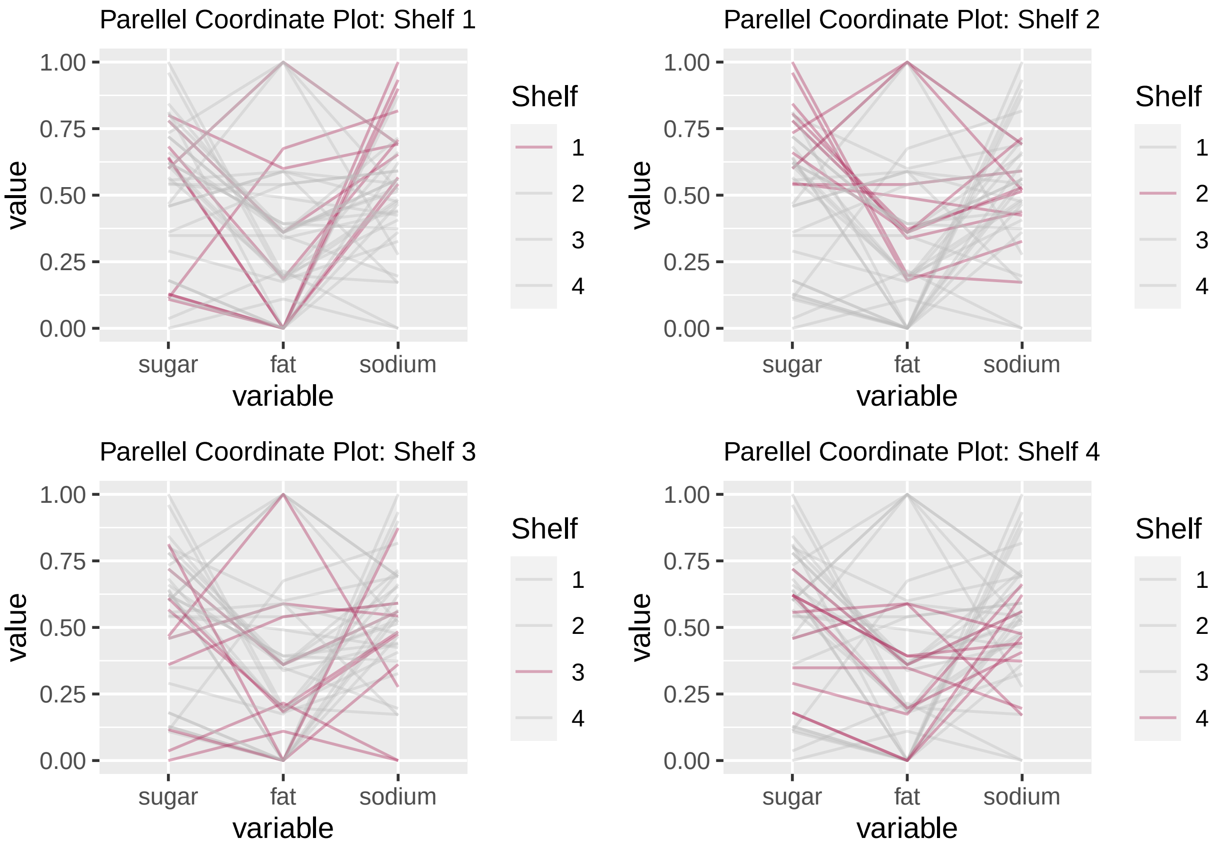
```
# Use box plots to examine response variable with quantitative data
p1 <- cereal2 %>%
  ggplot(aes(x = Shelf, y = sugar, fill = Shelf)) +
  geom_boxplot(varwidth = TRUE, alpha=0.2) +
  geom_jitter() +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ggtitle("Sugar Boxplot") +
  ylab("Sugar Value (0-1)") +
  xlab("Shelf Number")

p2 <- cereal2 %>%
  ggplot(aes(x = Shelf, y = fat, fill = Shelf)) +
  geom_boxplot(varwidth = TRUE, alpha=0.2) +
  geom_jitter() +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ggtitle("Fat Boxplot") +
  ylab("Fat Value (0-1)") +
  xlab("Shelf Number")

p3 <- cereal2 %>%
  ggplot(aes(x = Shelf, y = sodium, fill = Shelf)) +
  geom_boxplot(varwidth = TRUE, alpha=0.2) +
  geom_jitter() +
  theme(plot.title = element_text(lineheight = 1, face = "bold")) +
  ggtitle("Sodium Boxplot") +
  ylab("Sodium Value (0-1)") +
  xlab("Shelf Number")

# Plot on one grid.
grid.arrange(p1, p2, p3, nrow = 1, ncol = 3)
```



Cereal boxes placed on the second shelf appear to have the highest amount of sugar, ranging from 0.6 to 1.0 in standardized terms. The distribution of sugar content is much wider on the other shelves, going as low as 0.1. Shelf 1 has the widest distribution of fat content. Shelves 1, 2, and 3 have a few cereal boxes with fat contents of 1.0 (in standardized terms). Shelf 4 has a maximum fat content of 0.6. Shelf 1 has cereal boxes with the highest sodium content. Shelf 3 has outlier points in both directions (very low and very high sodium content). Shelf 2 and 4 have (relatively) tighter distributions of sodium content.

```r
p4 <- ggparcoord(cereal2, columns = 2:4, scale = "globalminmax",
                 alphaLines = .4, groupColumn = 1, showPoints = FALSE,
                 title = "Parellel Coordinate Plot: Shelf 1") +
  scale_color_manual(values=c("maroon","gray", "gray", "gray")) +
  theme(
    plot.title = element_text(size = 10)
  )

p5 <- ggparcoord(cereal2, columns = 2:4, scale = "globalminmax",
                 alphaLines = .4, groupColumn = 1, showPoints = FALSE,
                 title = "Parellel Coordinate Plot: Shelf 2") +
  scale_color_manual(values=c("gray", "maroon", "gray", "gray")) +
  theme(
    plot.title = element_text(size = 10)
  )
```

```
p6 <- ggparcoord(cereal2, columns = 2:4, scale = "globalminmax",
                 alphaLines = .4, groupColumn = 1, showPoints = FALSE,
                 title = "Parellel Coordinate Plot: Shelf 3") +
  scale_color_manual(values=c("gray", "gray", "maroon", "gray")) +
  theme(
    plot.title = element_text(size = 10)
  )

p7 <- ggparcoord(cereal2, columns = 2:4, scale = "globalminmax",
                 alphaLines = .4, groupColumn = 1, showPoints = FALSE,
                 title = "Parellel Coordinate Plot: Shelf 4") +
  scale_color_manual(values=c("gray", "gray", "gray", "maroon")) +
  theme(
    plot.title = element_text(size = 10)
  )

# Plot on one grid.
grid.arrange(p4, p5, p6, p7, nrow = 2, ncol = 2)
```



Shelf 1 has higher levels of sodium, while shelf 2 appears to have higher levels of sugar and a few observations with the highest fat content. Shelf 3 has a large range of values for sugar, fat, and sodium. Shelf 4 has low-to-medium normalized values for sugar, fat, and sodium.

### 1.2 Evaluate Ordinal vs. Categorical

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

---

The setting under which it would be desirable to take into account ordinality is a situation where 1 represents the lowest physical level of a shelf as well as the lowest preferred level by shoppers. For example, a shelf one would be considered undersirable while a level of 4 would be desirable by all shoppers. Although the dimensions of the shelves matter, it is typically the case that people prefer shelves 2 and 3 because they are easier to access relative to shelves 1 and 4. Under ideal conditions, Dillons would be able to order shelf levels by shopper preference (e.g. level 1 = really not desirable, level 2 = not desirable, level 3 = desirable, level 4 = really desirable). In addition, our team performed a nominal test to check whether the ordinal logistic regression as constructed is valid or not. The nominal test results (displayed below) show that there is no reported nominal effects for sugar indicating that the model is invalid

```r
cereal2.copy <- data.frame(cereal2) # copy dataframe
cereal2.copy$Shelf.order <- factor(cereal2.copy$Shelf, levels = c(1,2,3,4)) # order data
# Set up regression and run nominal test
lrt_cereal_main_effects2 <- ordinal::clm(Shelf.order ~  fat + sodium + sugar,
                                         data = cereal2.copy)
ordinal::nominal_test(lrt_cereal_main_effects2)
```

```
## Tests of nominal effects
##
## formula: Shelf.order ~ fat + sodium + sugar
##          Df  logLik    AIC  LRT Pr(>Chi)
## <none>      -49.265 110.53
## fat      2 -47.465 110.93  3.6 0.165300
## sodium   2 -42.665 101.33 13.2 0.001361 **
## sugar
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model with normalized explanatory variables.
model_cereal_shelves_linear <- multinom(formula = 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
```

```
summary(model_cereal_shelves_linear)
```

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

```
# Model with normalized explanatory variables + interactions terms.
model_cereal_shelves_quadratic <- multinom(formula = Shelf ~ sugar + fat
                                           + sodium + sodium:sugar + sodium:fat
                                           + sugar:fat, data = cereal2)
```

```
## # weights:  32 (21 variable)
## initial  value 55.451774
## iter  10 value 36.108903
## iter  20 value 31.221758
## iter  30 value 30.183143
## iter  40 value 28.725940
## iter  50 value 28.510281
## iter  60 value 28.344031
## iter  70 value 28.286691
## iter  80 value 28.187825
## iter  90 value 28.133526
## iter 100 value 28.066673
## final  value 28.066673
## stopped after 100 iterations
```

```
summary(model_cereal_shelves_quadratic)
```

```
## Call:
## multinom(formula = Shelf ~ sugar + fat + sodium + sodium:sugar +
##     sodium:fat + sugar:fat, data = cereal2)
##
## Coefficients:
##   (Intercept)     sugar       fat       sodium sugar:sodium fat:sodium
## 2   -1.988003  12.63023 25.885108   0.03256223   -25.645040  -41.02996
## 3   26.381935 -22.80111 20.460284 -30.29171281    12.891752  -35.73871
## 4   27.530652 -17.17032  4.312374 -28.60684017    -8.362321  -31.40412
##   sugar:fat
## 2 13.555737
## 3  5.754849
## 4 33.926682
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium sugar:sodium fat:sodium sugar:fat
## 2    24.17632 27.66281 39.72160 25.65306     29.68042   38.88213  35.44831
## 3    21.62372 24.85328 39.71878 23.36754     27.83209   42.84818  33.43987
## 4    21.52055 24.91880 38.89680 23.13444     28.91377   41.87341  33.68466
##
## Residual Deviance: 56.13335
## AIC: 98.13335
```

For the block of coefficients in the models, the first row compares the coefficients for Shelf = 2 to our baseline of Shelf = 1. The second row compares the coefficients for Shelf = 3 to our our baseline of Shelf = 1. Finally, the third row compares the coefficients for Shelf = 4 to our baseline of Shelf = 1.

```
# LRT for model with normalized explanatory variables only.
lrt_cereal_main_effects <- Anova(model_cereal_shelves_linear)
lrt_cereal_main_effects
```

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

```
# LRT for model with normalized explanatory variables + interaction terms.
lrt_cereal_quadratic_effects <- Anova(model_cereal_shelves_quadratic)
lrt_cereal_quadratic_effects
```

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

We performed an LRT using `Anova()` to explore if a given explanatory variable $x_r$ is statistically significant over all response categories in our multinomial model (shelves: 1, 2, 3, 4). The null hypothesis is testing whether all coefficients in the odds ratios for a particular predictor are zero or not.

Our null hypothesis is

$$H_0 : B_{2r} = B_{3r} = 0$$

While our alternative hypothesis is:

$$H_a : B_{2r} \neq 0 or B_{3r} \neq 0$$

In both models the small p-values indicate that sodium and sugar are statistically significant and affect the probability of having the cereal being placed on a shelf level. The p-value for the fat coefficient is quite large, meaning we fail to reject the null hypothesis that the fat coefficient is zero and the insufficient evidence indicates that fat content is not essential to our response variable. Similarly, the p-values for all interaction coefficients is quite large, meaning we also fail to reject the null hypothesis that all interactions are zero, meaning that the interaction terms are not essential to our response variable.

### 1.3 Where do you think Apple Jacks will be placed?

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

---

```r
# Set Apple Jack Variables:
size <- 28
sugar <- 12/size
fat <- 0.5/size
sodium <- 130/size

# Re-scaling sugar to be within 0 and 1
sugar_all <- cereal$sugar_g / cereal$size_g
sugar <- (sugar - min(sugar_all)) / (max(sugar_all) - min(sugar_all))

# Re-scaling fat to be within 0 and 1
fat_all <- cereal$fat_g / cereal$size_g
fat <- (fat - min(fat_all)) / (max(fat_all) - min(fat_all))
# Re-scaling sodium to be within 0 and 1
```

```r
sodium_all <- cereal$sodium_mg/cereal$size_g
sodium <- (sodium - min(sodium_all)) / (max(sodium_all) - min(sodium_all))

# Prediction dataframe
predict.data <- data.frame(sugar = sugar, fat = fat, sodium = sodium)
aj_shelf_probs <- predict(object = model_cereal_shelves_linear,
                          newdata = predict.data, type = "probs")
# Print results
aj_shelf_probs
```

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

Observation: The probability of Apple Jacks being placed on each shelf is:

- Shelf 1: 0.0532685%
- Shelf 2: 0.4719426%
- Shelf 3: 0.2004274%
- Shelf 4: 0.2743615%

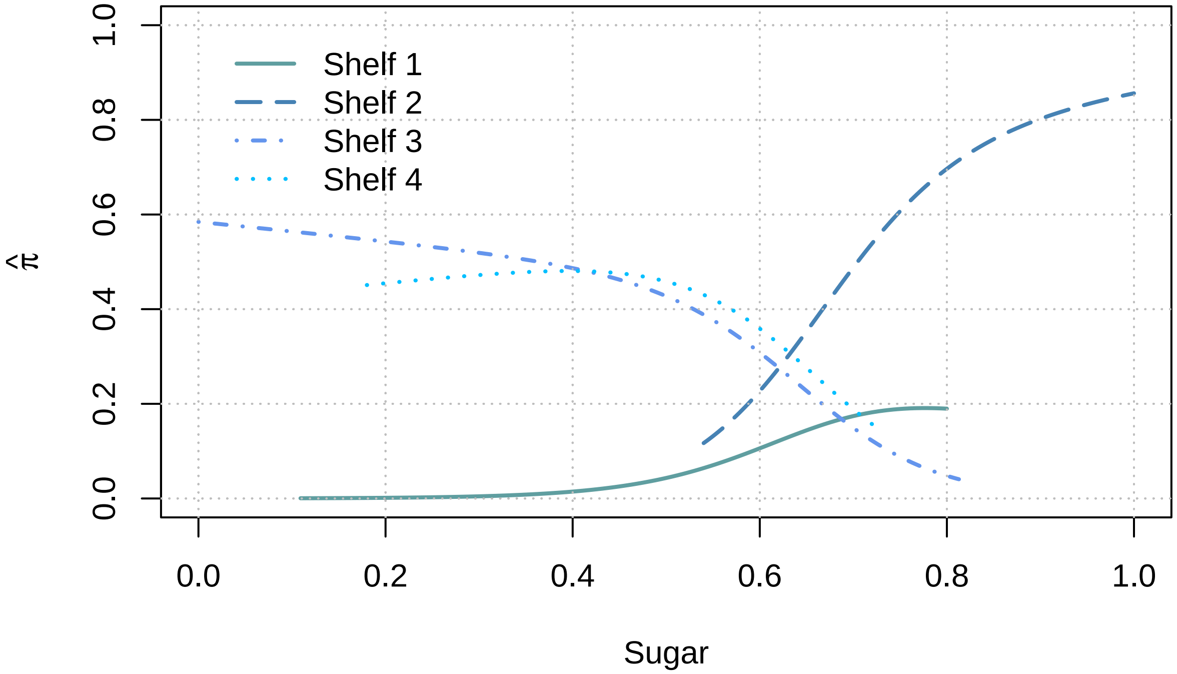
Therefore, Apple Jacks has the highest probability of being placed on shelf 2, followed by shelf 4, shelf 3 and shelf 1.

### 1.4 Figure 3.3

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

---

```r
# Create prediction model to ensure probabilities are estiamted for all four shelves
predict.d <- data.frame(sodium = mean(cereal2$sodium), sugar = c(mean(cereal2$sugar), mean(cere
pi.hat <- predict(object = model_cereal_shelves_linear, newdata = predict.d, type = "probs")
#head(pi.hat)
#pi.hat[,1]

# Create plotting area first to make sure get the whole region with respect to x-axis
curve(expr = predict(object = model_cereal_shelves_linear, newdata = data.frame(sodium = mean(c
                                                                                sugar = x,
                                                                                fat = mean(cere

      ylab = expression(hat(pi)), xlab = "Sugar", type = "n",
      xlim = c(min(cereal2$sugar), max(cereal2$sugar)),
      ylim = c(0,1), col = "#5F9EA0", lty = "solid", lwd = 2, n = 1000,
      panel.first = grid(col = "gray", lty = "dotted"))

# Create plots for all shelves
curve(expr = predict(object = model_cereal_shelves_linear, newdata = data.frame(sodium = mean(c
                                                                                sugar = x,
                                                                                fat = mean(cere
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(cereal2$sugar[cereal2$Shelf == 1]), max(cereal2$sugar[cereal2$Shelf == 1]))
      ylim = c(0,1), col = "#5F9EA0", lty = "solid", lwd = 2, n = 1000,
      add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))

curve(expr = predict(object = model_cereal_shelves_linear, newdata = data.frame(sodium = mean(c
                                                                                sugar = x,
                                                                                fat =
                                                                                  mean(cereal2$
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(cereal2$sugar[cereal2$Shelf == 2]), max(cereal2$sugar[cereal2$Shelf == 2]))
      ylim = c(0,1), col = "#4682B4", lty = "longdash", lwd = 2, n = 1000,
      add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))

curve(expr = predict(object = model_cereal_shelves_linear, newdata = data.frame(sodium = mean(c
                                                                                sugar = x,
                                                                                fat = mean(cere
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(cereal2$sugar[cereal2$Shelf == 3]), max(cereal2$sugar[cereal2$Shelf == 3]))
      ylim = c(0,1), col = "#6495ED", lty = "dotdash", lwd = 2, n = 1000,
      add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))

curve(expr = predict(object = model_cereal_shelves_linear, newdata = data.frame(sodium = mean(c
                                                                                sugar = x,
                                                                                fat = mean(cere
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(cereal2$sugar[cereal2$Shelf == 4]), max(cereal2$sugar[cereal2$Shelf == 4]))
      ylim = c(0,1),col = "#00BFFF", lty = "dotted", lwd = 2, n = 1000,
      add = TRUE, panel.first = grid(col = "gray", lty = "dotted"))

legend(x = .01, y = 1, legend=c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"), lty=c("solid","lon
       col=c("#5F9EA0","#4682B4","#6495ED", "#00BFFF"), bty="n", lwd = c(2,2,2,2))
```



The plot shows that the at low levels of sugar, there is higher probability that cereals will be placed on shelves 3 and 4. However, at higher levels of sugar, there is a higher probability of the cereal being placed on shelf 2. Finally, with increases to the the level of sugar, there is a slight increase in the probability that the cereal will be placed on shelf 1, although the probability starts to plateau at higher levels of sugar.

### 1.5 Odds ratios

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

---

```
# Find the standard deviation for each of the coefficients: sodium, sugar and fat:
sd.cereal <- apply(X = cereal2[, -c(1)], MARGIN = 2, FUN = sd)
print("Standard Deviations of the Model Coefficients")
```

```
## [1] "Standard Deviations of the Model Coefficients"
```

```
sd.cereal
```

```
##     sugar       fat    sodium
## 0.2692078 0.2990292 0.2298359
```

```
# Compute confint for the coefficients for use in CI estimates
conf.beta <- confint(object = model_cereal_shelves_linear, level = 0.95)

# Odds ratios and CI
beta.hat2 <- coefficients(model_cereal_shelves_linear)[1, 2:4] # Shelf 2 Vs. Shelf 1
print("Shelf = 2 vs. Shelf = 1: Odds Ratios and CI")
```

```
## [1] "Shelf = 2 vs. Shelf = 1: Odds Ratios and CI"
```

```
round(exp(sd.cereal * beta.hat2), 4)
```

```
##  sugar    fat sodium
## 2.0647 3.3719 0.0179
```

```
round(exp(sd.cereal * conf.beta [2:4 ,1:2 ,1]),5)
```

```
##          2.5 %   97.5 %
## sugar  0.14364 29.67949
## fat    0.87216 13.03601
## sodium 0.00073  0.43882
```

```
beta.hat3 <- coefficients(model_cereal_shelves_linear)[2, 2:4] # Shelf 1 Vs. Shelf 3
print("Shelf = 3 vs. Shelf = 1: Odds Ratio and CI")
```

```
## [1] "Shelf = 3 vs. Shelf = 1: Odds Ratio and CI"
```

```
round(exp(sd.cereal * beta.hat3 ), 4)
```

```
##  sugar    fat sodium
## 0.0373 0.8465 0.0032
```

```
round(exp(sd.cereal * conf.beta [2:4 ,1:2 ,2]), 5)
```

```
##          2.5 %  97.5 %
## sugar  0.00283 0.49180
## fat    0.20557 3.48609
## sodium 0.00008 0.12233
```

```
beta.hat4 <- coefficients(model_cereal_shelves_linear)[3, 2:4] # Shelf 1 Vs. Shelf 4
print("Shelf = 4 vs. Shelf = 1: Odds Ratio and CI")
```

```
## [1] "Shelf = 4 vs. Shelf = 1: Odds Ratio and CI"
```

```
round(exp(sd.cereal * beta.hat4 ), 4)
```

```
##  sugar    fat sodium
## 0.0465 0.7709 0.0034
```

```
round(exp(sd.cereal * conf.beta [2:4 ,1:2 ,3]), 5)
```

```
##          2.5 %  97.5 %
## sugar  0.00356 0.60837
## fat    0.18822 3.15745
## sodium 0.00009 0.13014
```

Observations:

**Shelf 2 Vs. Shelf 1:** The estimated odds of a cereal being on shelf 2 vs. shelf 1 change by 2.06 times for a 0.27 increase in sugar holding the other variables constant. With 95% confidence, the odds of level 2 instead of level 1 change by 0.14 to 29.68 times when sugar changes by 0.27, while holding the other variables constant. The estimated odds of a cereal being on shelf 2 vs. shelf 1 change by 3.37 times for a 0.30 (rounded) increase in fat holding the other variables constant. With 95% confidence, the odds of level 2 instead of level 1 change by 0.87 to 13.04 times when fat changes by 0.30, while holding the other variables constant. The estimated odds of a cereal being on shelf 2 vs. shelf 1 change by 0.02 times for a 0.23 increase in sodium holding the other variables constant. With 95% confidence, the odds of shelf 2 instead of shelf 1 change by 0.001 to 0.439 times when sodium changes by 0.23, while holding the other variables constant.

Relationship to previous plots: Figure 3 shows that probability of being on shelf 2 increases with higher levels of sugar, which supports the odds ratio of 2.06. Both sugar and fat have odds ratios greater then 1, which makes sense when assessing the parallel coordinate plots for shelf 1 and shelf 2 as it shows that cereals on shelf 2 generally have higher sugar values and mid-to-higher fat values relative to shelf 1. Similarly, the the boxplots for sugar and fat demonstrate higher values relative to shelf 1. An odds ratio less than 1 for sodium (0.02) reflects the findings in the parallel coordinate plot and the box plots as shelf 1 has higher sodium values relative to shelf 2.

**Shelf 3 Vs. Shelf 1:** The estimated odds of a cereal being on shelf 3 vs. shelf 1 change by 0.04 times for a 0.27 increase in sugar holding the other variables constant. With 95% confidence, the odds of level 3 instead of level 1 change by 0.0028 to 0.49 times when sugar changes by 0.27, while holding the other variables constant. The estimated odds of a cereal being on shelf 3 vs. shelf 1 change by 0.85 times for a 0.30 (rounded) increase in fat holding the other variables constant. With 95% confidence, the odds of level 3 instead of level 1 change by 0.21 to 3.49 times when fat changes by 0.30, while holding the other variables constant. The estimated odds of a cereal being on shelf 3 vs. shelf 1 change by 0.003 times for a 0.23 increase in sodium holding the other variables constant. With 95% confidence, the odds of shelf 3 instead of shelf 1 change by 0.0001 to 0.1223 times when sodium changes by 0.23, while holding the other variables constant.

Relationship to previous plots: Figure 3 shows that probability of being on shelf 3 decreases with higher levels of sugar, which supports the odds ratio less than 1. In fact, sugar and sodium have odds ratios that are only 0.04 and 0.003, respectively. The difference between the variables in shelf 3 versus shelf 1 are best depicted in the boxplots for these variables as shelf 1 has higher sugar and sodium values relative to shelf 3. This is also apparent in the parallel coordinate plot that shows the difference between sodium values for shelf 1 versus sodium values for shelf 3. For fat, which has an odds ratio of 0.85, it is more difficult to tell the relative difference for all of the plots.

**Shelf 4 Vs. Shelf 1:** The estimated odds of a cereal being on shelf 4 vs. shelf 1 change by 0.05 times for a 0.27 increase in sugar holding the other variables constant. With 95% confidence, the odds of level 4 instead of level 1 change by 0.0036 to 0.6084 times when sugar changes by 0.27, while holding the other variables constant. The estimated odds of a cereal being on shelf 4 vs. shelf 1 change by 0.77 times for a 0.30 (rounded) increase in fat holding the other variables constant. With 95% confidence, the odds of level 3 instead of level 1 change by 0.19 to 3.16 times when fat changes by 0.30, while holding the other variables constant. The estimated odds of a cereal being on shelf 4 vs. shelf 1 change by 0.003 times for a 0.23 increase in sodium holding the other variables constant. With 95% confidence, the odds of shelf 4 instead of shelf 1 change by 0.0001 to 0.1301 times when sodium changes by 0.23, while holding the other variables constant.

Relationship to previous plots: Figure 3 shows that probability of being on shelf 4 decreases with higher levels of sugar, which supports the odds ratio less than 1. When comparing shelf 4 to shelf 1, sugar and sodium have odds ratios that are only 0.05 and 0.003, respectively. The difference between sodium on the two shelves is apparent in both the sodium boxplot and the parallel coordinate plots for shelf 1 and shelf 4. The difference between sugar on both the parallel coordinate plot and the boxplot is less pronounced but there are difference between the median values in the boxplot. Fat, which has an odds ratio of 0.77, is presented best in the parallel coordinate plots, as the fat values seem to be higher on shelf 1 relative to shelf 4.

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

Read the example **'Alcohol Consumption'** in chapter 4.2.2 of the textbook(Bilder and Loughin's "Analysis of Categorical Data with R). This is based on a study in which moderate-to-heavy drinkers (defined as at least 12 alcoholic drinks/week for women, 15 for men) were recruited to keep a daily record of each drink that they consumed over a 30-day study period. Participants also completed a variety of rating scales covering daily events in their lives and items related to self-esteem. The data are given in the *DeHartSimplified.csv* data set. Questions 24-26 of chapter 3 of the textbook also relate to this data set and give definitions of its variables: the number of drinks consumed (`numall`), positive romantic-relationship events (`prel`), negative romantic-relationship events (`nrel`), age (`age`), trait (long-term) self-esteem (`rosn`), state (short-term) self-esteem (`state`).

The researchers stated the following hypothesis:

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

```
#drinks <- read_csv('../data/short-questions/DeHartSimplified.csv')
drinks <- read_csv('DeHartSimplified.csv')
```

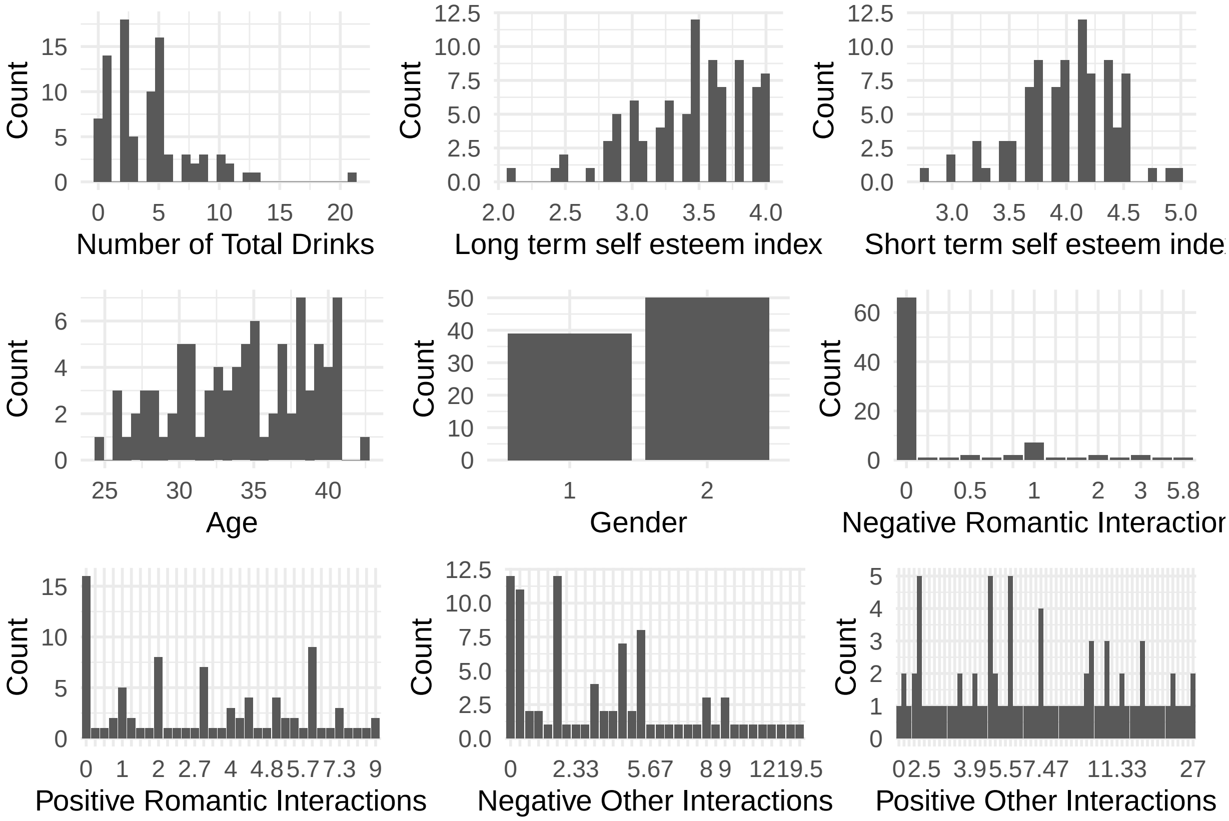
## 2.1 EDA

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

---

```
drinks_subset <- drinks %>%
  dplyr::filter(dayweek == 6) %>%
  dplyr::select(numall, nrel, prel, rosn,
                negevent, posevent, age, gender, state) %>% # additional controls
  dplyr::mutate(nrel = round(nrel, 1),
                prel = round(prel, 1),
                negother = round(10*negevent - nrel,2),
                posother = round(10*posevent - prel,2)) %>%
  dplyr::select(-c("negevent", "posevent"))

drinks_subset$gender <- as_factor(drinks_subset$gender)
## As per the paper, 1 = Male and 2 = Female
```

Starting with some basic histograms:



A few key things are revealed and by the histograms and bar charts:

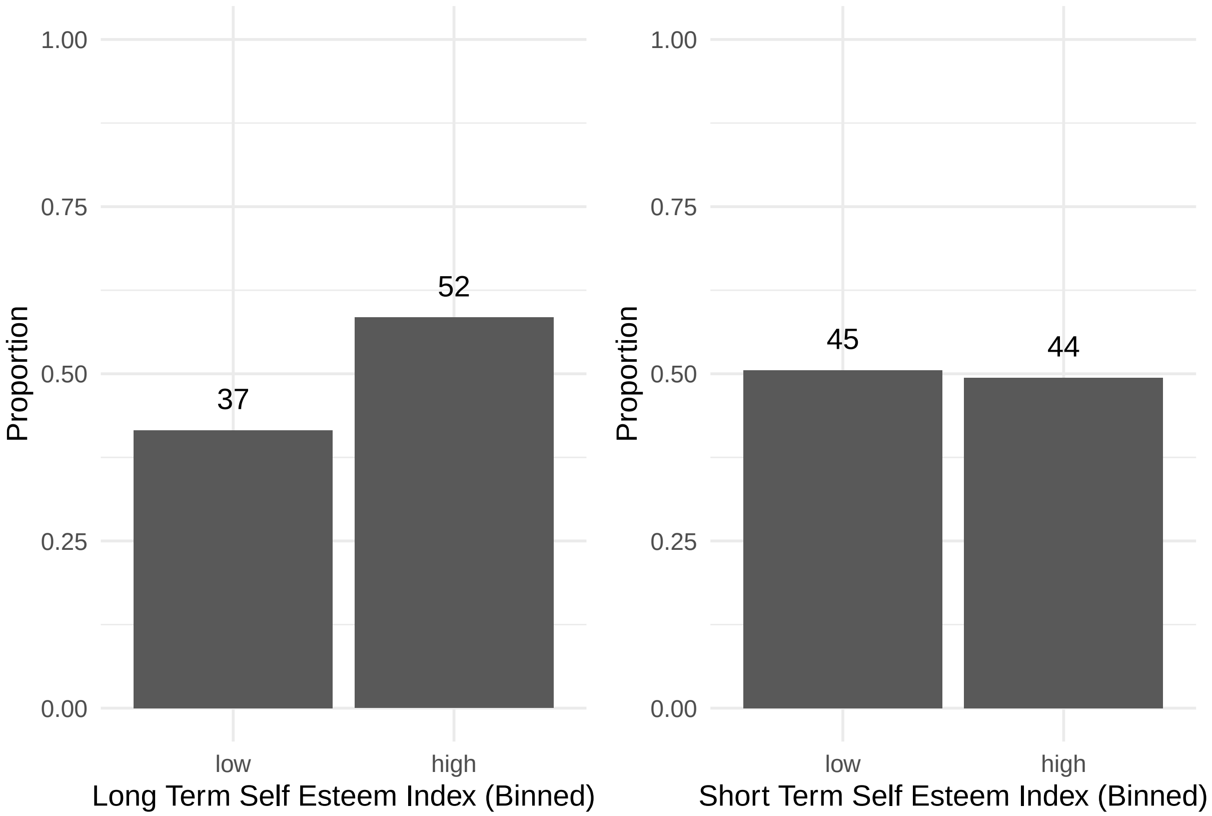
- Most people drink less than 5 drinks a day and the number of people drinking more than six or more drinks a day is drastically lower (3 and lower) than the number drinking five or less drinks a day.

```
table(drinks_subset$numall)
```

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

- Since the short and long term self esteem indices are discrete, we can bin them into low and high using a middle point. We used the mean value of the variables to determine the cut-off point. (We compared mean and median values for both variables and found them to be very close for each variable). A higher proportion of people in the study have high long-term self esteem and people with lower long-term self esteem. The short-term self esteem is approximately equally distributed in the dataset.

```
drinks_subset$rosn_binned <- ifelse(drinks_subset$rosn < mean(drinks_subset$rosn),
                                    "low", "high") %>% as_factor()
drinks_subset$state_binned <- ifelse(drinks_subset$state < mean(drinks_subset$state),
                                     "low", "high") %>% as_factor()
```



- Most study participants are between the ages of 25 and 41.
- There are more females in the study than males.
- Of the 89 people in the sample, 66 have had 0 negative relationship interactions (`nrel`). This skewed distribution will have implications for the models constructed with this variable.
- A lot more people have had positive romantic interactions, with only 15 with no positive romantic interactions.
- The two variables `negother` and `posother` are created by subtracting the `nrel` and `prel` from `negevent` and `posevent` respectively. This is done to isolate the effects of romantic and all other interactions. A higher count of people have had minimal negative other interactions.

**Why was the study limited to observations from only one day?**

By limiting observations to one day, we reduce the scope of the problem from one involving the time component and can therefore assume constant conditions across the subjects. If we were to include data from all days, we would need to model the impact of time series component.

## 2.2 Hypothesis One

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

---

As shown above, only 23 people in the sample had non-zero negative romantic relationships on the Saturday of the study. Given the large skew in the data, it will be difficult to detect a statistically significant relationship between negative interactions, `nrel`, and alcohol consumption, `numall`. Below we use multiple models to try and detect the relationship:

```r
simple_model <- glm(formula = numall ~ nrel,
              family = poisson(link = "log"), data = drinks_subset)

summary(simple_model)
```

```
##
## Call:
## glm(formula = numall ~ nrel, family = poisson(link = "log"),
##     data = drinks_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8330  -1.3203  -0.5295   0.4744   5.9612
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.38951    0.05719  24.298   <2e-16 ***
## nrel         0.05087    0.05083   1.001    0.317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 249.39  on 87  degrees of freedom
## AIC: 508.79
##
## Number of Fisher Scoring iterations: 5
```

The simplest model, regressing the number of negative romantic interactions on the number of overall drinks consumed shows a positive relationship between the variables. However, this estimate is not statistically significant. It is important to note that the `nrel` variable does not have much variability in its values since a majority of the respondents answered 0 negative romantic interactions on the day of the study.

We can add other variables to the model to evaluate whether the relationship between these two variables changes. We are including interaction terms below because the impact of a negative romantic interaction on the number of total drinks consumed may be affected by one or more of these factors.

```
model2 <- glm(formula = numall ~ nrel + negother + posother + prel
              + state_binned + rosn_binned + nrel:state_binned
              + nrel:rosn_binned, family = poisson(link = "log"),
              data = drinks_subset)

summary(model2)
```

```
##
## Call:
## glm(formula = numall ~ nrel + negother + posother + prel + state_binned +
##     rosn_binned + nrel:state_binned + nrel:rosn_binned, family = poisson(link = "log"),
##     data = drinks_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4578  -1.3647  -0.3666   0.8181   5.1407
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.681367   0.144348  11.648  < 2e-16 ***
## nrel                   -0.080902   0.107581  -0.752 0.452045
## negother               -0.055894   0.017118  -3.265 0.001094 **
## posother                0.008572   0.009237   0.928 0.353428
## prel                    0.014080   0.022207   0.634 0.526055
## state_binnedhigh       -0.565932   0.121513  -4.657  3.2e-06 ***
## rosn_binnedhigh         0.079482   0.117776   0.675 0.499769
## nrel:state_binnedhigh  0.409081   0.122613   3.336 0.000849 ***
## nrel:rosn_binnedhigh  -0.105759   0.120389  -0.878 0.379687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.34  on 88  degrees of freedom
## Residual deviance: 214.75  on 80  degrees of freedom
## AIC: 488.15
##
## Number of Fisher Scoring iterations: 5
```

In the more involved model, `nrel` is negatively correlated with the number of total drinks consumed and the estimate remains statistically insignificant. The changing sign of the estimate is of concern and points to its unreliability in explaining the number of total drinks consumed.

Other variables are show statistical significance in the model: `negother` (other negative interactions), `state_binnedhigh` (the short term self-esteem index), and `nrel:statebinnedhigh` (the interaction of negative romantic relationships and high long-term self-esteem).

```
Anova(model2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: numall
##                    LR Chisq Df Pr(>Chisq)
## nrel                 2.6206  1  0.1054875
## negother            11.5175  1  0.0006894 ***
## posother             0.8496  1  0.3566656
## prel                 0.4005  1  0.5268174
## state_binned        13.2051  1  0.0002792 ***
## rosn_binned          0.1436  1  0.7047543
## nrel:state_binned   11.9792  1  0.0005380 ***
## nrel:rosn_binned     0.7681  1  0.3808172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio tests show the same variables to be statistically significant as those shown by the `summary()` function. Interesting points to note are:

- Negative Romantic Interactions are not correlated with the total number of drinks
- Negative Other Interactions are negatively correlated with the total number of drinks, counter to the researchers' intuition.
- The short term self-esteem index, individually as well as interacted with the number of negative romantic interactions, is correlated with the total number of drinks and is statistically significant.

### 2.3 Hypothesis Two

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

---

The model shown above includes the `rosn_binned` term to account for long term high self-esteem. We see that the interaction between `nrel:rosn_binned` is statistically insignificant via both the `summary()` function and the LR tests' `Anova()` function. The researchers' hypothesis that, "the relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem" can mean one of two things:

- Individuals with high self-esteem do not show high volumes of drinking, or
- The relationship between `nrel:rosn_binned` and `numall` is statistically insignificant

The output of the `Anova()` function validates point #2 above. To test point #1, we would need to collect more data to reduce the amount of skew towards 0 within `nrel` in the dataset.