W271 Assignment 2

Due Sunday 18 October 2020 11:59pm

Amber Chen

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

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

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

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

```
# read in data set
cereal <- read.csv("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
```

```
par(mfrow = c(1, 3))

boxplot(formula = sugar ~ Shelf, data = cereal2, ylab = "Sugar",
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sugar ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)

boxplot(formula = fat ~ Shelf, data = cereal2, ylab = "Fat",
```

```
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$fat ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)

boxplot(formula = sodium ~ Shelf, data = cereal2, ylab = "Sodium",
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sodium ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```

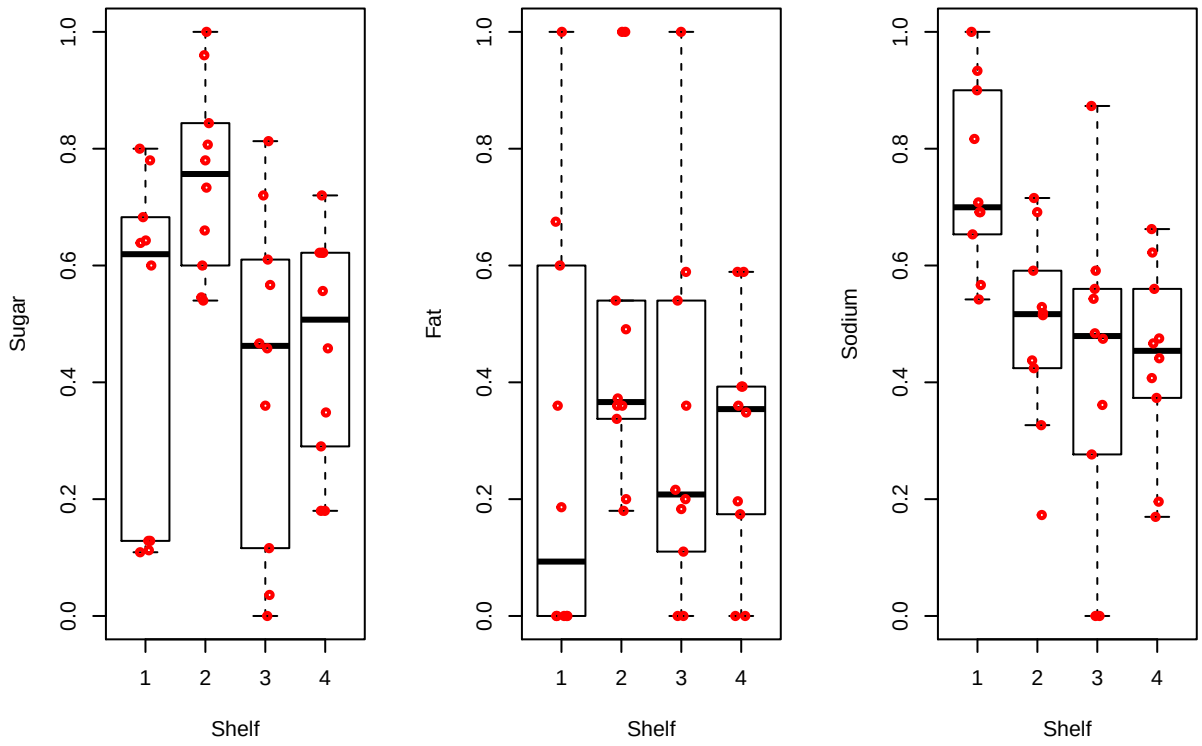


Figure 1: Boxplots of all explanatory variables by shelf numbers

```
par(mar = c(5.1, 4.1, 4.1, 8.1), xpd = TRUE)
clusterR = hclust(dist(cereal2$Shelf))
colDir = cutree(clusterR, 4) + 1
parcoord(cereal2[c(-1)], col = colDir, lty = 1)
legend("right", title = "Shelf Number", c("1", "2", "3", "4"),
    col = c("red", "green", "blue", "cyan"), lty = 1, inset = c(-0.2,
        0))
```

From the boxplots and the parallel coordinates plot (Fig.1 and 2), we can see some differences in the three contents among the four shelves:

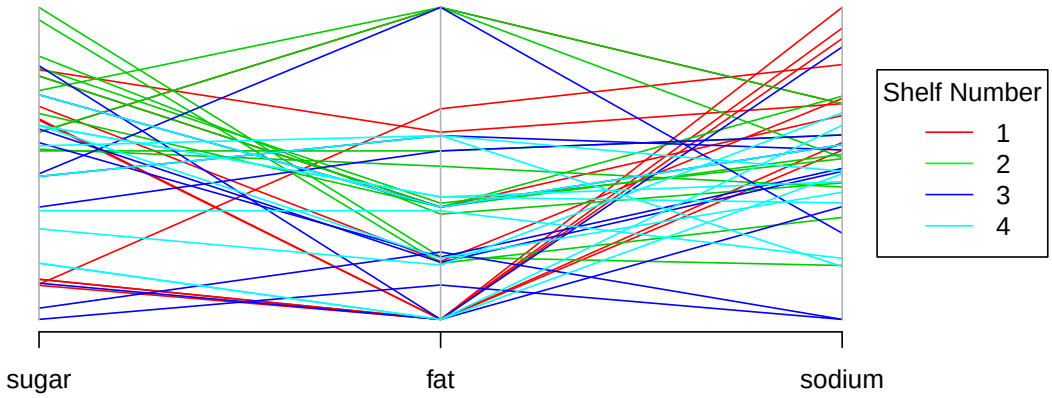


Figure 2: Parallel Coordinates Plot For Shelf Placement of Cereal Products

1. Cereals on the first shelf have higher sodium content per serving than cereals on other shelves, but lower level of fat.
2. Cereals on the second shelf have higher sugar content per serving than cereals on other shelves.
3. The forth shelf has cereals that are low to medium level of all three contents compared to other shelves.

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

It would be desirable to take ordinality into account when there is some ordering of the data across the levels. In our study, the response value can take ordinality into account by ordering average sugar content on the shelves. For example, Shelf #1 has the lowest average sugar content, Shelf #2 and #3 has higher sugar content by order, and Shelf #4 has the highest sugar content among all shelves. Charts in section 1.1 showed this setting does not occur in our data.

Estimate multinomial regression model The model for Shelf #2 vs #1 is definied as follow. The model is similar for Shelf #3 vs #1 and Shelf #4 vs #1

$$log\left(\frac{\widehat{\pi}_{shelf2}}{\widehat{\pi}_{shelf1}}\right) = \beta_0 + \beta_1 Sugar + \beta_2 Fat + \beta_3 Sodium$$

```
cereal.mod <- 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(cereal.mod)
```

```
## 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
```

For Shelf #2 vs #1, the estimated multinomial regression model is:

$$log\left(\frac{\widehat{\pi}_{shelf2}}{\widehat{\pi}_{shelf1}}\right) = 6.9 + 2.693071Sugar + 4.0647092Fat - 17.49373Sodium$$

For Shelf #3 vs #1, the estimated multinomial regression model is:

$$log\left(\frac{\widehat{\pi}_{shelf3}}{\widehat{\pi}_{shelf1}}\right) = 21.68068 - 12.216442Sugar - 0.5571273Fat - 24.9785Sodium$$

For Shelf #4 vs #1, the estimated multinomial regression model is:

$$log\left(\frac{\widehat{\pi}_{shelf4}}{\widehat{\pi}_{shelf1}}\right) = 21.288343 - 11.39371Sugar - 0.870118Fat - 24.67385Sodium$$

To examine the importance of each explanatory variable, we set the hypothesis as follow and perform LRT:

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

$$H_a : \beta_{jr} \neq 0, \quad \text{for some } j$$

```
Anova(cereal.mod, test = "LR")
```

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

The p-values of the variables, sugar and sodium, are smaller than the critical value 0.05, so we can conclude there is strong evidence that the sugar and sodium content of a cereal are related to the shelf placement given that the other explanatory variables are in the model.

To examine the significance of interactions among the explanatory variables, we define a multinomial regression model with interaction term as follow and perform LRT:

```
cereal.mod2 <- multinom(formula = Shelf ~ sugar + fat + sodium +
    sugar:fat + fat:sodium + sugar: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
```

```
summary(cereal.mod2)
```

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

```
Anova(cereal.mod2)
```

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

Sugar and sodium are still the only two variables with p-value less than 0.05 in the LRT test given that other explanatory variable and all interaction terms are in the model. The p-values for fat,

all two-way intereactions and the three-way interaction variables are large, so there is no sufficient evidence that shows interactions among the explanatory variables are significant.

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

```
AppJk.pi.hat <- predict(cereal.mod, newdata = data.frame(sugar = 12/28,
    fat = 0.5/28, sodium = 0.13/28), type = "probs")
AppJk.pi.hat
```

```
##             1             2             3             4
## 4.178599e-08 1.304578e-04 5.109095e-01 4.889600e-01
```

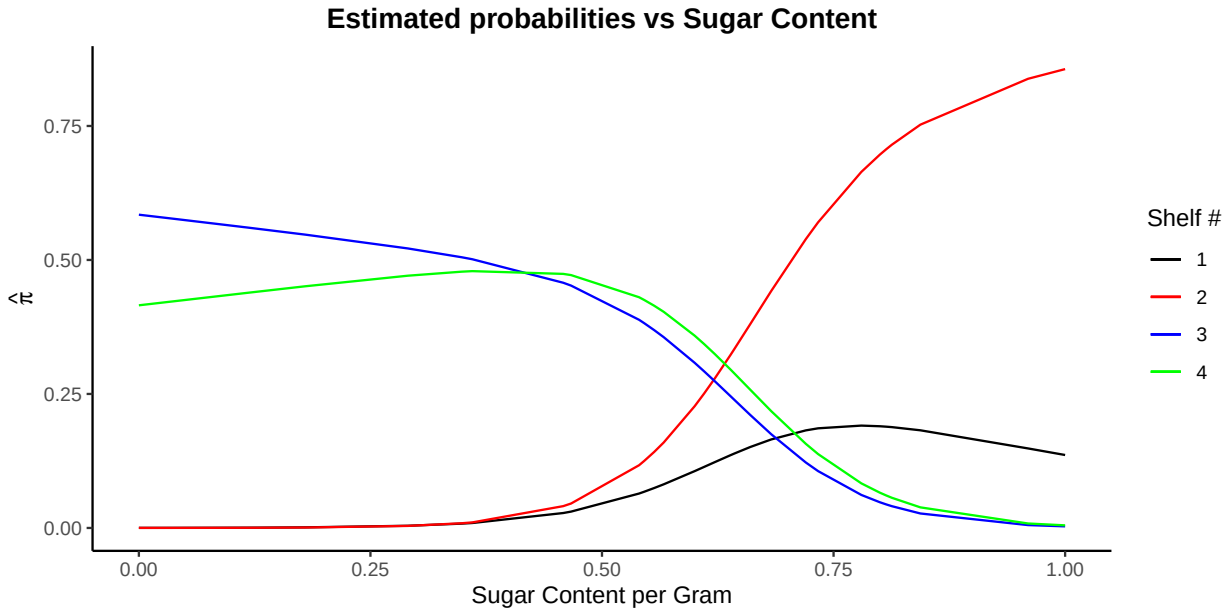
The probabilities for placing Apple Jacks:

On shelf #1: 0.00000004178599 On shelf #2: 0.0001304578 On shelf #3: 0.5109095 On shelf #4: 0.48896

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

```
# predict estimated probabilities when fat content and sodium
# content are the mean overalls
pi.hat <- predict(cereal.mod, newdata = data.frame(sugar = cereal2$sugar,
    fat = mean(cereal2$fat), sodium = mean(cereal2$sodium)),
    type = "probs")
```

```
# plot
ggplot(data = data.frame(cereal2$sugar, pi.hat), aes(x = cereal2$sugar)) +
    # add fit line
geom_line(mapping = aes(y = pi.hat[, 1], color = "1")) + geom_line(mapping = aes(y = pi.hat[,
    2], color = "2")) + geom_line(mapping = aes(y = pi.hat[,
    3], color = "3")) + geom_line(mapping = aes(y = pi.hat[,
    4], color = "4")) + scale_color_manual(name = "Shelf #",
    values = c(`1` = "black", `2` = "red", `3` = "blue", `4` = "green")) +
    # pretty formatting
labs(title = "Estimated probabilities vs Sugar Content", x = "Sugar Content per Gram",
    y = expression(hat(pi))) + theme_classic() + theme(plot.title = element_text(hjust = 0.5,
    face = "bold"), plot.subtitle = element_text(hjust = 0.5)) +
    theme(legend.position = "right")
```



We have below observations from the plot:

1. Shelf #2 has the highest probability of high sugar cereal products
2. Shelf #3 has the highest probability of low sugar cereal products
3. Shelf #4 has a similar probability distribution as Shelf #3 but has a lower probability of low sugar cereals.
4. Shelf #1 has a similar probability distribution as Shelf #2 when sugar content is low, but has a much lower probability of high sugar cereals than Shelf #2.
5. Shelf #1 has a slightly higher probability of high sugar cereals than Shelf #3 and #4
6. For medium sugar cereal product placement (sugar = 0.5 per gram), Shelf #1 has the lowest probability, Shelf #2 has a bit higher probability, and then Shelf #3 and #4 have the highest.

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

```
# set c = 0.1 for sugar, fat, and sodium
c = 0.1

# calculate odds ratios for shelf #2 vs #1, shelf #3 vs #1,
# and shelf #4 vs #1
OR.beta1 = exp(c * summary(cereal.mod)$coefficients[, 2:4])

# calculate odds ratio CI for shelf #3 vs #2 and shelf #4 vs
# #2
OR.beta2 = exp(c * (summary(cereal.mod)$coefficients[2:3, 2:4] -
    summary(cereal.mod)$coefficients[1, 2:4]))
```

```r
# calculate odds ratio CI for shelf #4 vs #3
OR.beta3 = exp(c * (summary(cereal.mod)$coefficients[3, 2:4] -
    summary(cereal.mod)$coefficients[2, 2:4]))

# combine OR matrices into one
OR.beta.mtx <- rbind(OR.beta1, OR.beta2, OR.beta3)
rownames(OR.beta.mtx) <- c("2 vs 1", "3 vs 1", "4 vs 1", "3 vs 2",
    "4 vs 2", "4 vs 3")

# build a function to integrate the CI matrices for
# combinations into one matrix
transform.df <- function(CImatrix) {
    df <- as.data.frame(CImatrix)
    comb.df <- unite(df, CI, c("2.5 %", "97.5 %"), sep = ", ")
    comb.df <- t(comb.df[-1, ])
    colnames(comb.df) <- c("sugar CI", "fat CI", "sodium CI")
    comb.df
}

# calculate odds ratio CI for shelf #2 vs #1, shelf #3 vs #1,
# and shelf #4 vs #1
CI.beta <- confint(object = cereal.mod, level = 0.95)
CI.beta1 <- exp(c * CI.beta)
CI.beta.mtx <- rbind(transform.df(CI.beta1[, , 1]), transform.df(CI.beta1[,
    , 2]), transform.df(CI.beta1[, , 3]))

# calculate odds ratio CI for shelf #3 vs #2, shelf #4 vs #2,
# and shelf #4 vs #3
CI.beta3v2 <- exp(c * (CI.beta[, , 2] - CI.beta[, , 1]))
CI.beta.mtx <- rbind(CI.beta.mtx, transform.df(CI.beta3v2))
CI.beta4v2 <- exp(c * (CI.beta[, , 3] - CI.beta[, , 1]))
CI.beta.mtx <- rbind(CI.beta.mtx, transform.df(CI.beta4v2))
CI.beta4v3 <- exp(c * (CI.beta[, , 3] - CI.beta[, , 2]))
CI.beta.mtx <- rbind(CI.beta.mtx, transform.df(CI.beta4v3))
rownames(CI.beta.mtx) <- c("2 vs 1", "3 vs 1", "4 vs 1", "3 vs 2",
    "4 vs 2", "4 vs 3")

OR.beta.mtx
```

```
##            sugar       fat     sodium
## 2 vs 1 1.3090571 1.5015095 0.17388298
## 3 vs 1 0.2947452 0.9458108 0.08226171
## 4 vs 1 0.3200203 0.9166663 0.08480636
## 3 vs 2 0.2251584 5.4393524 0.05478601
## 4 vs 2 0.2131324 0.7002493 0.48772091
## 4 vs 3 1.0857524 0.9691857 1.03093363
```

```r
# inverse of OR matrix
1/OR.beta.mtx
```

```
##             sugar       fat     sodium
## 2 vs 1 0.7639086 0.6659965  5.7509942
## 3 vs 1 3.3927614 1.0572939 12.1563238
## 4 vs 1 3.1248022 1.0909096 11.7915679
## 3 vs 2 4.4413183 0.1838454 18.2528354
## 4 vs 2 4.6919200 1.4280629  2.0503529
## 4 vs 3 0.9210203 1.0317940  0.9699946
```

```r
CI.beta.mtx
```

```
##        sugar CI
## 2 vs 1 "0.486360190609854, 3.5233772226957"
## 3 vs 1 "0.113079343828199, 0.768263284439357"
## 4 vs 1 "0.123176665102488, 0.831431562470281"
## 3 vs 2 "0.23250123264901, 0.218047411866835"
## 4 vs 2 "0.253262227214845, 0.235975744270198"
## 4 vs 3 "1.08929412687104, 1.08222217475487"
##        fat CI
## 2 vs 1 "0.955288584159479, 2.36005196485228"
## 3 vs 1 "0.589172171537896, 1.51833047655708"
## 4 vs 1 "0.572053715110685, 1.468877918525"
## 3 vs 2 "0.616747840712747, 0.643346205579042"
## 4 vs 2 "0.598828170457007, 0.622392193223141"
## 4 vs 3 "0.970944899888046, 0.96742964802747"
##        sodium CI
## 2 vs 1 "0.0432664259690363, 0.698816449626022"
## 3 vs 1 "0.0168812537763203, 0.400858207089563"
## 4 vs 1 "0.0174648399889782, 0.411805629248596"
## 3 vs 2 "0.390169823326784, 0.573624457901765"
## 4 vs 2 "0.403658023463204, 0.589290119700212"
## 4 vs 3 "1.03457007520831, 1.02730996139138"
```

Intepretations of the estimated odds ratios:

- The estimated odds of placement on Shelf #2 vs. on Shelf #1 change by 1.309 times for every 0.1 increase in sugar content holding the other variables constant. On the other hand, the estimated odds of placement on Shelf #3 and Shelf #4 vs Shelf #1 change by 3.393 and 3.125 times, respectively, for every 0.1 decrease in sugar content holding the other varaibles constant. These are in-line with the first 5 observations from the plot in exercise 1.4.
- With 95% confidence, the odds of placement on Shelf #2 vs on Shelf #1 change by 0.4864 to 3.5234 times for every 0.1 increase in sugar content holding the other variables constant; the odds of placement on Shelf #3 vs on Shelf #1 change by ($1/0.768 = 1.301$ to $1/0.1131 = 8.843$)

times for every 0.1 decrease in sugar content; and the odds of placement on Shelf #4 vs on Shelf #1 change by ($1/0.831 = 1.203$ to $1/0.123 = 8.118$) times for every 0.1 decrease in sugar content.

- The estimated odds of placement on Shelf #2 vs. on Shelf #1 change by 5.751 times for every 0.1 decrease in sodium content, and by ($1/0.6988 = 1.431$ to $1/0.0433 = 23.11$) times with 95% confidence, holding the other variables constant. Similarly, 12.156 times for shelf #3 vs. #1 and 11.782 times for shelf #4 vs. #1. These are in-line with our observations from the parallel coordinates plot (Fig.2) and the sodium boxplot in exercise 1.1. Shelf #1 has cereal products that have highest sodium content, and Shelf #3 has cereal products that are the lowest in sodium content.

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

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

The researchers stated the following hypothesis:

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

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

```
# read in data set
dehart <- read.table("DeHartSimplified.csv", header = TRUE, sep = ",",
    na.strings = " ")
glimpse(dehart)
```

```
## Rows: 623
## Columns: 13
## $ id       <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4...
## $ studyday <int> 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6...
## $ dayweek  <int> 6, 7, 1, 2, 3, 4, 5, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1...
## $ numall   <int> 9, 1, 1, 2, 2, 1, 4, 3, 4, 0, 4, 7, 4, 1, 0, 1, 3, 1, 0, 1...
## $ nrel     <dbl> 1.0000000, 0.0000000, 1.0000000, 0.0000000, 1.3333333, 1.0...
## $ prel     <dbl> 0.0000000, 0.0000000, 0.0000000, 1.0000000, 0.3333333, 0.0...
## $ negevent <dbl> 0.4000000, 0.2500000, 0.2666667, 0.5333333, 0.6633333, 0.5...
## $ posevent <dbl> 0.5250000, 0.7000000, 1.0000000, 0.6083333, 0.6933333, 0.6...
## $ gender   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ rosn     <dbl> 3.3, 3.3, 3.3, 3.3, 3.3, 3.3, 3.3, 3.9, 3.9, 3.9, 3.9, 3.9...
## $ age      <dbl> 39.48528, 39.48528, 39.48528, 39.48528, 39.48528, 39.48528...
## $ desired  <dbl> 5.666667, 2.000000, 3.000000, 3.666667, 3.000000, 4.000000...
## $ state    <dbl> 4.000000, 2.777778, 4.222222, 4.111111, 4.222222, 4.333333...
```

```
summary(dehart)
```

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

- The dataset has 623 values and 13 variables (one is id, not meaningful for our analysis). There is one missing value in variable numall, 3 missing values in desired and 3 missing variables in state. The number of missing values are small relative to the size of the dataset, and the missing values are in the explanatory and response variables which we want to study for our question. Therefore, we can remove the lines that have missing values and work with complete cases

```
# remove missing values
dehart = dehart[complete.cases(dehart), ]
describe(dehart)
```

```
## dehart
##
##  13  Variables      618  Observations
## ---------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      618        0       89        1    76.06    56.78        7       17
##      .25      .50      .75      .90      .95
```

```
##       33       60      123      148      153
##
## lowest :   1   2   4   5   7, highest: 153 154 155 156 160
## ---------------------------------------------------------------------------
## studyday
##        n  missing distinct     Info     Mean      Gmd
##      618        0        7     0.98    3.997    2.293
##
## lowest : 1 2 3 4 5, highest: 3 4 5 6 7
##
## Value          1     2     3     4     5     6     7
## Frequency     89    89    87    88    88    89    88
## Proportion 0.144 0.144 0.141 0.142 0.142 0.144 0.142
## ---------------------------------------------------------------------------
## dayweek
##        n  missing distinct     Info     Mean      Gmd
##      618        0        7     0.98    3.995    2.285
##
## lowest : 1 2 3 4 5, highest: 3 4 5 6 7
##
## Value          1     2     3     4     5     6     7
## Frequency     88    89    88    89    88    89    87
## Proportion 0.142 0.144 0.142 0.144 0.142 0.144 0.141
## ---------------------------------------------------------------------------
## numall
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      618        0       18     0.97    2.519    2.635     0.00     0.00
##      .25      .50      .75      .90      .95
##     1.00     2.00     3.75     6.00     8.00
##
## lowest :  0  1  2  3  4, highest: 13 14 15 18 21
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency    141   111   131    80    49    43    24     5     9     7     7
## Proportion 0.228 0.180 0.212 0.129 0.079 0.070 0.039 0.008 0.015 0.011 0.011
##
## Value         11    12    13    14    15    18    21
## Frequency      4     2     1     1     1     1     1
## Proportion 0.006 0.003 0.002 0.002 0.002 0.002 0.002
## ---------------------------------------------------------------------------
## nrel
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      618        0       33    0.551   0.3615    0.629        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
##      618        0       68    0.982    2.586    2.614    0.000    0.000
##      .25      .50      .75      .90      .95
##    0.500    2.000    4.000    6.000    7.915
##
## 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
##      618        0      130    0.996   0.4435   0.4128   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.1688   0.3500   0.6333   1.0000   1.1575
##
## 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
##      618        0      215        1    1.048   0.7073    0.200    0.300
##      .25      .50      .75      .90      .95
##    0.600    0.950    1.367    1.942    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
##      618        0        2     0.74    1.558    0.494
##
## Value          1     2
## Frequency    273   345
## Proportion 0.442 0.558
## ---------------------------------------------------------------------------
## rosn
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      618        0       17    0.993    3.434   0.4671      2.7      2.9
##      .25      .50      .75      .90      .95
##      3.2      3.5      3.8      3.9      4.0
##
## lowest : 2.1 2.4 2.5 2.7 2.8, highest: 3.6 3.7 3.8 3.9 4.0
##
## Value        2.1   2.4   2.5   2.7   2.8   2.9   3.0   3.1   3.2   3.3   3.4
## Frequency      7     7    14     7    21    35    42    21    28    42    34
## Proportion 0.011 0.011 0.023 0.011 0.034 0.057 0.068 0.034 0.045 0.068 0.055
##
```

```
## Value        3.5   3.6   3.7   3.8   3.9   4.0
## Frequency     84    62    48    63    48    55
## Proportion 0.136 0.100 0.078 0.102 0.078 0.089
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      618        0       89        1    34.28    5.184    26.24    27.80
##      .25      .50      .75      .90      .95
##    30.53    34.57    38.19    39.95    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
##      618        0       22    0.996    4.464    1.923    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
##      618        0       25    0.993    3.965   0.4895    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
## ---------------------------------------------------------------------------
```

```
summary(dehart)
```

```
##        id            studyday        dayweek          numall
##  Min.   :  1.00   Min.   :1.000   Min.   :1.000   Min.   : 0.000
##  1st Qu.: 33.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 1.000
##  Median : 60.00   Median :4.000   Median :4.000   Median : 2.000
##  Mean   : 76.06   Mean   :3.997   Mean   :3.995   Mean   : 2.519
##  3rd Qu.:123.00   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.: 3.750
##  Max.   :160.00   Max.   :7.000   Max.   :7.000   Max.   :21.000
##       nrel             prel          negevent         posevent
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.500   1st Qu.:0.1688   1st Qu.:0.600
##  Median :0.0000   Median :2.000   Median :0.3500   Median :0.950
##  Mean   :0.3615   Mean   :2.586   Mean   :0.4435   Mean   :1.048
##  3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:0.6333   3rd Qu.:1.367
```

```
##  Max.   :9.0000   Max.   :9.000   Max.   :2.3767   Max.   :3.883
##      gender           rosn            age           desired
##  Min.   :1.000   Min.   :2.100   Min.   :24.43   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:3.200   1st Qu.:30.53   1st Qu.:3.333
##  Median :2.000   Median :3.500   Median :34.57   Median :4.667
##  Mean   :1.558   Mean   :3.434   Mean   :34.28   Mean   :4.464
##  3rd Qu.:2.000   3rd Qu.:3.800   3rd Qu.:38.19   3rd Qu.:5.667
##  Max.   :2.000   Max.   :4.000   Max.   :42.28   Max.   :8.000
##      state
##  Min.   :2.333
##  1st Qu.:3.667
##  Median :4.000
##  Mean   :3.965
##  3rd Qu.:4.222
##  Max.   :5.000
```

By examining the results from summary(), we can see the removal of missing values does not change the characteristics of the three variables, numall, desired and state. Therefore, the trimmed dataset is a good representation of the full dataset and we can continue our analysis with the trimmed.

```
dehart <- dehart[dehart$dayweek == 6, ]
describe(dehart$id)
```

```
## dehart$id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       89        0       89        1    75.89    57.38      7.8     16.8
##      .25      .50      .75      .90      .95
##     33.0     60.0    123.0    144.8    152.6
##
## lowest :   1   2   4   5   7, highest: 153 154 155 156 160
```

We limit the our study to obeservation from only Saturday to eliminate seasonality. Generally speaking, most of people would drink more alcohols on weekends than during weekdays. We can avoid the seasonality factor in our model by examining the relationship between negative romantic relationship events and drinking behaviour on same day of a week. In this case, we picked Saturday for our study since 89/89 participants have data on Saturday.

```
# Histogram of Number of Drinks Consumed
p1 <- ggplot(dehart, aes(x = numall)) + geom_histogram(aes(y = ..density..),
    fill = "#0072B2", colour = "black") + ggtitle("Number of Drinks Consumed") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))

# Distribution of desire to drink
p2 <- ggplot(dehart, aes(x = desired)) + geom_histogram(aes(y = ..density..),
    fill = "#0072B2", colour = "black") + ggtitle("Desire to Drink") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```

```r
# Histogram of negative romantic-relationship events
p3 <- ggplot(dehart, aes(x = nrel)) + geom_histogram(aes(y = ..density..),
    fill = "#0072B2", colour = "black") + ggtitle("Negative romantic-relationship events") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))

# Distribution of Trait (Long Term) Self-esteem
p4 <- ggplot(dehart, aes(x = rosn)) + geom_histogram(aes(y = ..density..),
    fill = "#0072B2", colour = "black") + ggtitle("Trait (Long Term) Self-esteem") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))

grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```

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

Figure 3: Histograms of all relevant variables

Above shows histograms of all relevant variables for our study.

- The variable, numall, number of drinks consumed has a concentration between 0 and 5. There are few data points over 5 drinks to less than 15, but one extreme data at 21. The variable does not exhibit a normal distribution.
- The variable, desired, desire to drink sort of follow a normal distribution with concentration around 4-6 and heavy tails.
- The variable, nrel, negative romantic-relationship events are mostly concentrated at 0, and have only few data points above 0.
- The variable, rosn, trait self-esteem has an increasing number of data points as self-esteem score increases. It does not exhibit a normal distribution pattern.

Next we run below scatter plots to examine the relationships among the relevant variables: 1. Number of Drinks Consumed vs. Negative Romantic-relationship. Number of drinks consumed varies from 0 to 21 when individuals have 0 negative romantic event (nrel = 0). Excluding nrel = 0, the plot does not show an obvious relationship between the two variables. There might be other factors affecting individuals' drinking behaviour, but they are outside of this research question.

2. Desire to drink vs. negative romantic-relationship. Similar to the plot on left-hand side, desire to drink varies from the lowest to the highest when individuals have 0 negative romantic event (nrel = 0). Excluding nrel = 0, the plot does not show an obvious relationship between the two variables.

3. Plot the two response variables together. As expected, desire to drink has a significant positive relationship with number of drinks consumed

4. Trait (Long-term) Self-esteem vs. Number of Drinks Consumed does not show a obvious relationship

```
# create plot for nrel by numall
ggplot(dehart, aes(x = nrel, y = numall)) + # add points
geom_point(size = 2, shape = 23, fill = "blue") + # pretty formatting
labs(title = "Number of drinks consumed vs. negative romantic-relationship",
    x = "Negative romantic-relationship events", y = "Number of drinks consumed") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5,
    face = "bold"), plot.subtitle = element_text(hjust = 0.5))
```

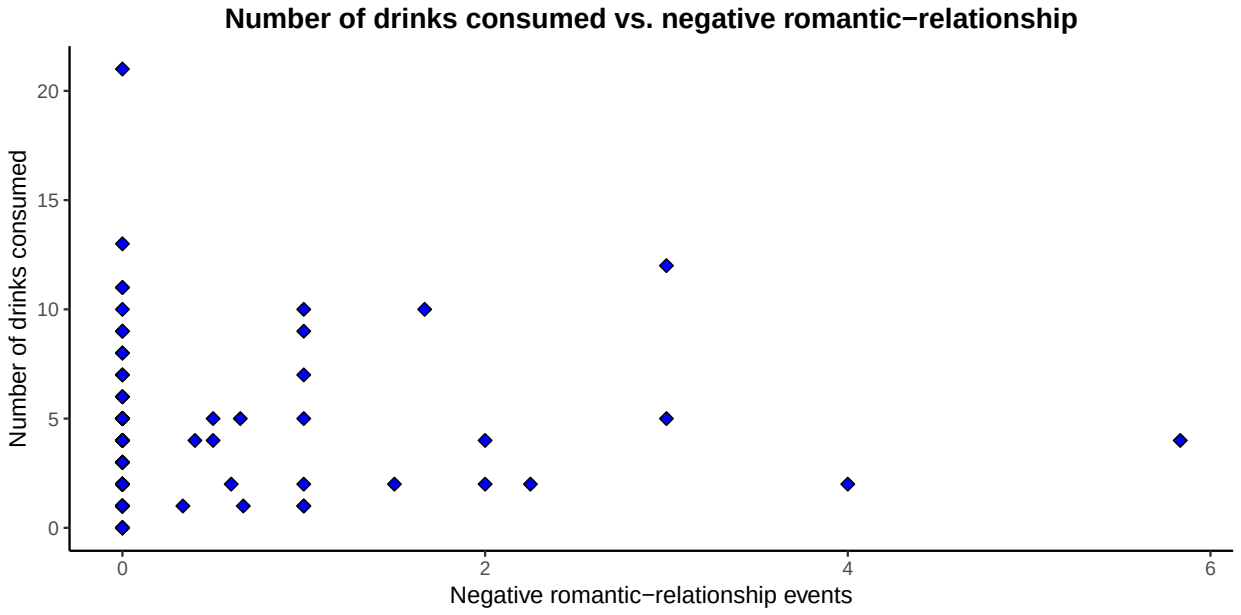


Figure 4: Number of drinks consumed vs negative romantic-relationship

```
# create plot for nrel by numall
ggplot(dehart, aes(x = nrel, y = desired)) + # add points
geom_point(size = 2, shape = 23, fill = "blue") + # pretty formatting
labs(title = "Desire to drink vs. negative romantic-relationship",
    x = "Negative romantic-relationship events", y = "Desire to drink") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5,
    face = "bold"), plot.subtitle = element_text(hjust = 0.5))
```

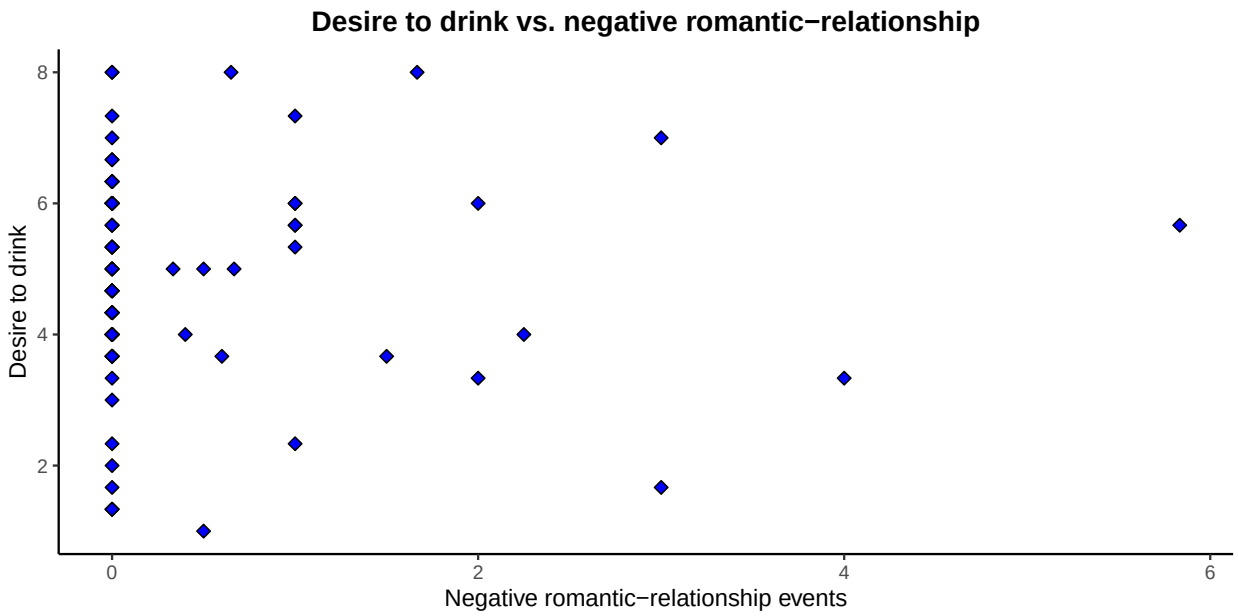


Figure 5: Desire to drink vs negative romantic-relationship

```
# create plot for nrel by numall
ggplot(dehart, aes(x = desired, y = numall)) + # add points
geom_point(size = 2, shape = 23, fill = "blue") + # pretty formatting
labs(title = "Desire to Drink vs Number of Drinks Consumed",
    x = "Desire to drink", y = "Number of drinks consumed") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5,
    face = "bold"), plot.subtitle = element_text(hjust = 0.5))
```

```
# create plot for rosn by nrel
ggplot(dehart, aes(x = rosn, y = numall)) + # add points
geom_point(size = 2, shape = 23, fill = "blue") + # pretty formatting
labs(title = "Trait (Long-term) Self-esteem vs. Number of Drinks Consumed",
    x = "Trait (long term) self-esteem", y = "Number of drinks consumed") +
    theme_classic() + theme(plot.title = element_text(hjust = 0.5,
    face = "bold"), plot.subtitle = element_text(hjust = 0.5))
```

2.2 (2 points): The researchers hypothesize that negative interactions with romantic partners would be associated with alcohol consumption and an increased desire to drink. Using appropriate

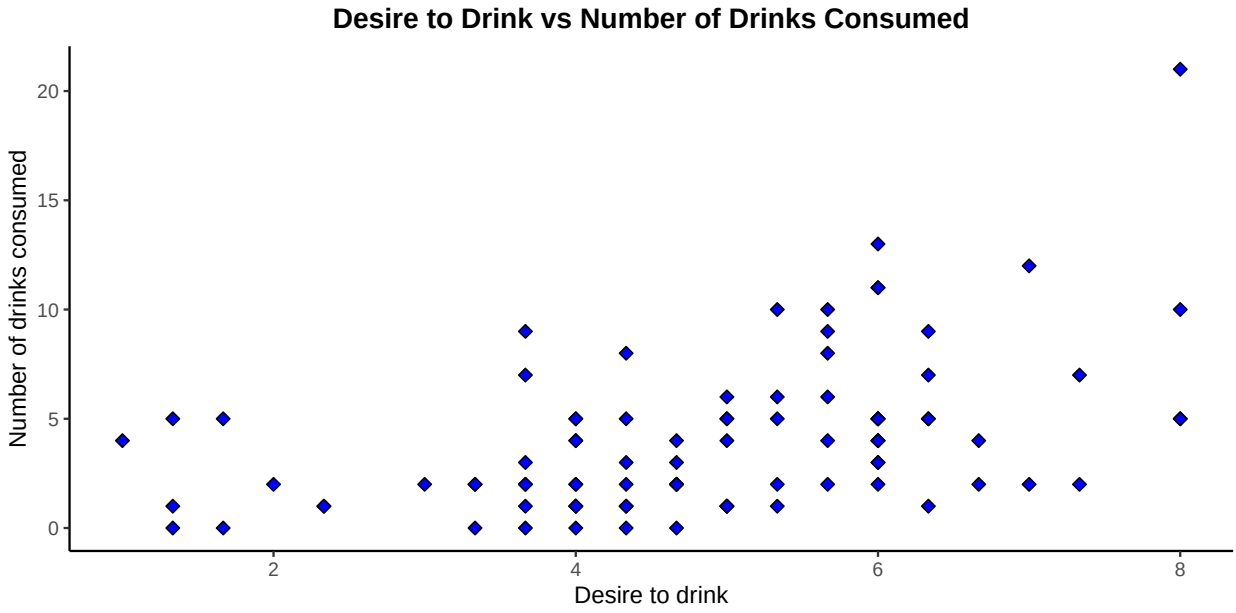


Figure 6: Desire to drink vs number of drinks consumed

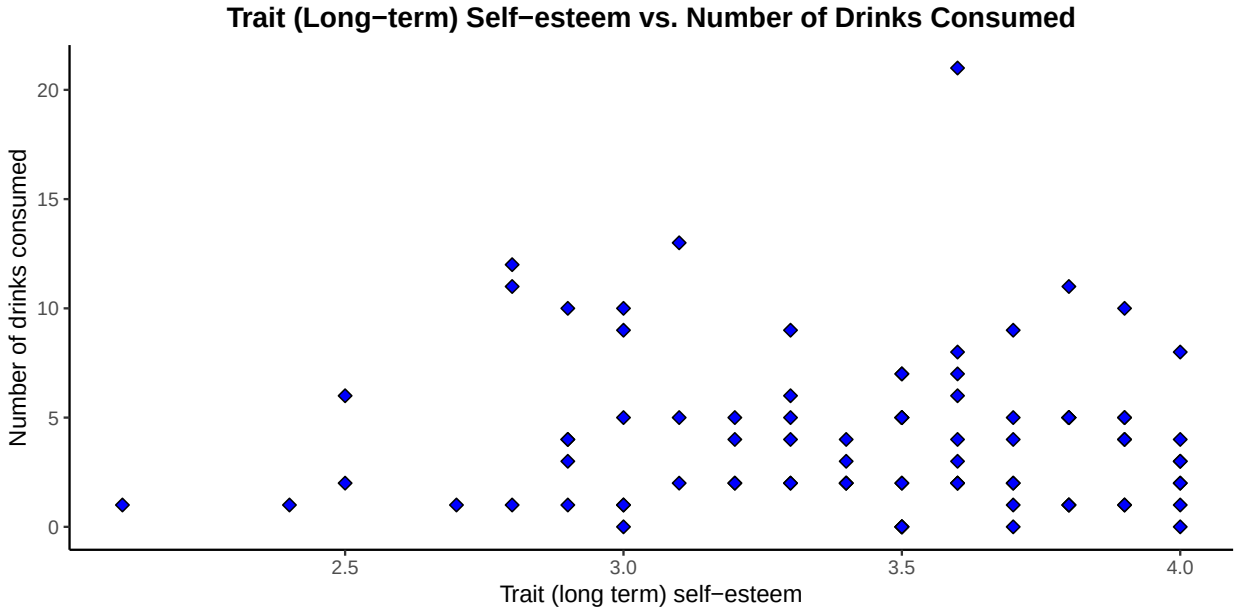


Figure 7: Desire to drink / Trait self-esteem vs number of drinks consumed

models, evaluate the evidence that negative relationship interactions are associated with higher alcohol consumption and an increased desire to drink.

Model 1: Response variable - alcohol consumption

Since the reponse variable is a counting variable, the poisson regression would be the appropriate model. We define the model as below:

$$log(\mu) = \beta_0 + \beta_1 nrel$$

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

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

The estimated poisson regression model is

$$log(\mu) = 1.39 + 0.4971 nrel$$

```
exp(dehart.mod$coefficients[2])
```

```
##     nrel
## 1.050961
```

The positive coefficient of nrel shows that negative interactions with romantic partners positively relate to number of drinks consumed. Therefore, one event increase in negative interactions with romantic partners leads to an estimated 1.05961 increase in alcoholic beverage consumption.

```
anova(dehart.mod, test = "LR")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: numall
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    88     250.34
## nrel  1  0.90934        87     249.43   0.3403
```

We further conduct LRT to examine the significance of nrel in the model. From the result, we can conclude that there is no strong evidence that nrel is significant since the p-value is greater than 0.05.This result is in-line with our observation in the EDA stage.

Model 2: Response variable - desire to drink

Since the response variable in this case is a score, we can use a linear regression model to examine the relationship between the desire to drink and negative romantic interations.

$$y = \beta_0 + \beta_1 nrel$$

```
dehart.mod.desire <- lm(formula = desired ~ nrel, data = dehart)
summary(dehart.mod.desire)
```

```
##
## Call:
## lm(formula = desired ~ nrel, data = dehart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8467 -0.8453  0.1533  1.1547  3.1547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.845267   0.184642  26.241   <2e-16 ***
## nrel        0.002914   0.178607   0.016    0.987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.604 on 87 degrees of freedom
## Multiple R-squared:  3.059e-06,  Adjusted R-squared:  -0.01149
## F-statistic: 0.0002662 on 1 and 87 DF,  p-value: 0.987
```

The estimated model is as follow

$$y = 4.845267 + 0.002914nrel$$

Similar to the result from the poisson regression model, the result of the linear regression shows a positive relationship between negative romantic interactions and desire to drink. However the p-value of nrel is very large, indicating that we cannot reject the null hypothesis that $\beta_1 = 0$. This suggests that this positive relationship is statistically insignificant.

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

We continue our study with Model 1.

To examine whether this relationship is evident for individuals with high trait self-esteem, we include the trait self-esteem variable and an interaction term between trait self-esteem and negative romantic-relationship events.

$$log(\mu) = \beta_0 + \beta_1 nrel + \beta_2 rosn + \beta_3 nrel : rosn$$

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

```
##
## Call:
## glm(formula = numall ~ nrel + rosn + nrel:rosn, family = poisson(link = "log"),
##     data = dehart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8324  -1.6025  -0.1471   0.5059   5.9811
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.32343    0.46367   2.854  0.00431 **
## nrel         1.07253    0.45716   2.346  0.01897 *
## rosn         0.01642    0.13403   0.123  0.90248
## nrel:rosn   -0.28731    0.13036  -2.204  0.02752 *
## ---
## 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: 244.30  on 85  degrees of freedom
## AIC: 507.7
##
## Number of Fisher Scoring iterations: 5
```

```
anova(dehart.mod, test = "LR")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: numall
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         88     250.34
## nrel       1   0.9093        87     249.43  0.34029
## rosn       1   0.4122        86     249.02  0.52086
## nrel:rosn  1   4.7191        85     244.30  0.02983 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above results, we yield below estimated regression model:

$$log(\mu) = 1.32343 + 1.07253nrel + 0.01642rosn - 0.28731nrel : rosn$$

The LRT result shows that the interaction is the only significant term in the model as its p-value is less than 0.05. This interaction term has a negative coefficient, while the coefficients of nreal and rosn are positive. This means that if a person has a higher trait self-esteem, the interaction effect will increase and offset the effect of the other two variables. Hence, the positive relationship between negative romantic-relationship event and alcohol consumption would not be evident for individuals with high trait self-esteem.