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")

# Rename raw data columns to keep final names short
names(cereal) <- c("ID", "Shelf", "Cereal", "size_g",
                   "sugar_raw", "fat_raw", "sodium_raw")
```

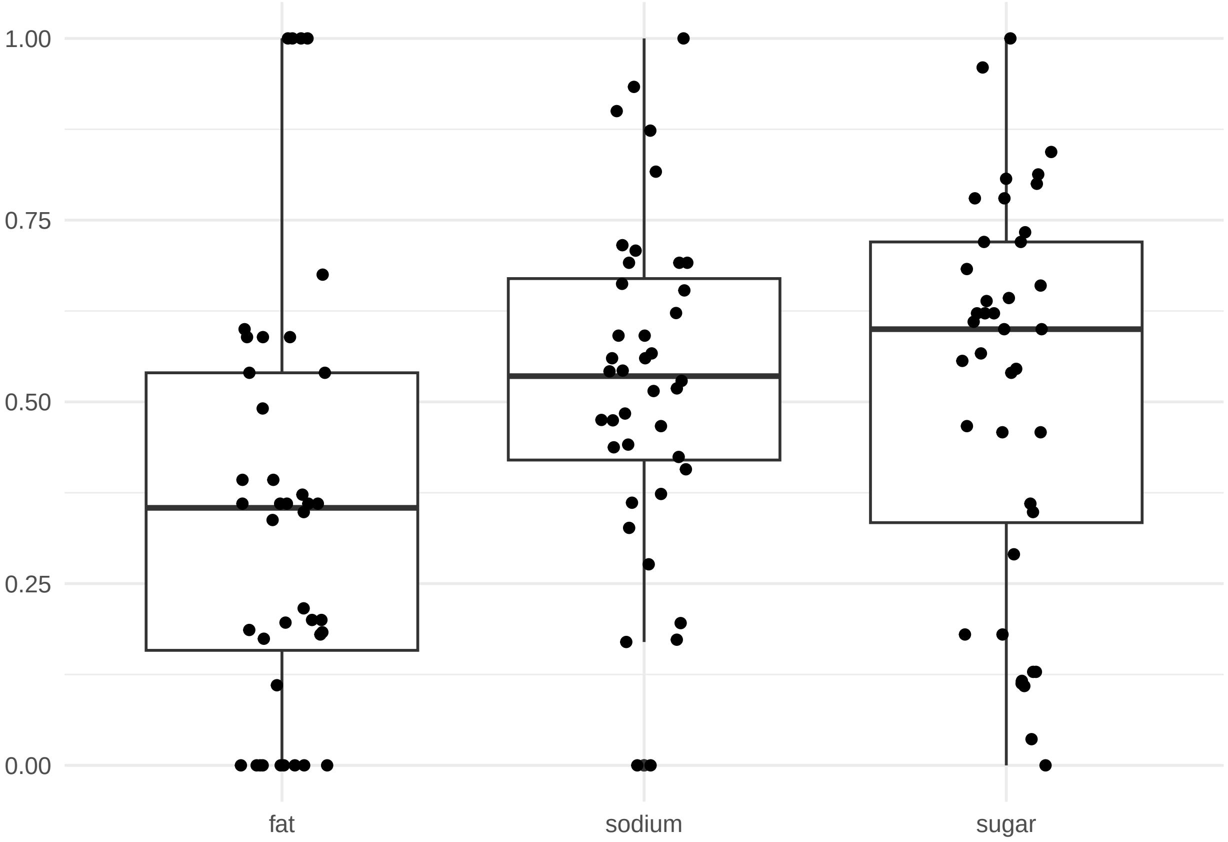
1.1 Recode Data

(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.

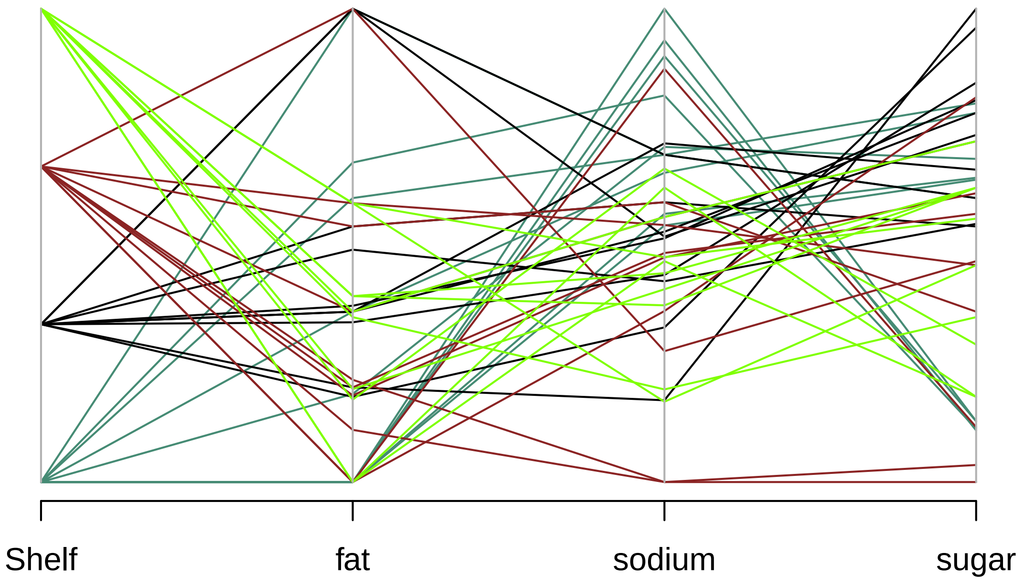
```r
# Re-scale columns in the following list
# Create new columns for re-scaled data
col_trans_list <- c("fat_raw", "sugar_raw", "sodium_raw")

for (col in col_trans_list) {
  col_per <- cereal[col] / cereal["size_g"]
  cereal[str_sub(col, 1, -5)] <- (col_per - min(col_per)) / max(col_per)
}
```

```r
cereal %>%
  pivot_longer(cols = 8:10, names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = indicators, y = values)) +
  geom_boxplot() +
  geom_jitter(width = 0.125) +
  theme_minimal() +
  theme(axis.title.x = element_blank(), axis.title.y = element_blank())
```



Most boxes of cereal contain higher sugar and sodium content than fat content. The median sodium and sugar content for these boxes of cereal is slightly above 0.50, whereas the median fat content is only near 0.37. the sugar content has higher variance than the sodium content.

```r
MASS::parcoord(cereal[c("Shelf", "fat", "sodium", "sugar")],
        col = colors()[12 * cereal$Shelf])
```



The distribution of fat, sugar, and sodium content is not uniform among all shelves. Cereal boxes on certain shelves have different distribution of fat, sugar, and sodium content than those on other shelves.

Content differences do exist between the shelves. Cereal boxes on the top shelf tend to have fat, sodium, and sugar content close to the median. The variance of fat, sodium, and sugar content on the top shelf appears to be lower than that of ceral boxes on other shelves. Cereal boxes on the bottom shelf appear to have the highest sodium content. The location of cereal boxes appears to have an effect on the sodium content of cereal boxes.

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).

It makes most sense to use Shelf as a categorical variable. While there is natural order to shelves, it is likely irrelevant to the price setting. Prime shelves are in the middle of the rack and supermarkets often stock those shelves with the most popular ites. Therefore, desirability scale might differ from the physcial scale. In this situation we belive using thsi variable as categorical makes most sense.

```r
# Set Shelf as a categorical value
cereal$Shelf <- factor(cereal$Shelf, levels = c("1", "2", "3", "4"))

# Estimate linear model
```

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

# Estimate linear model with all interactions
model_cereal_shelves_quadratic <- multinom(
  formula = Shelf ~ fat + sugar + sodium +
    fat:sugar + fat:sodium + sodium:sugar +
    sodium:sugar:fat,
  data = cereal
)
```

```r
# Conduct Anova test on linear model
lrt_cereal_main_effects <- car::Anova(model_cereal_shelves_linear)
lrt_cereal_main_effects
```

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

```r
# Conduct Anova test on the interactions
lrt_cereal_quadratic_effects <- car::Anova(model_cereal_shelves_quadratic)
lrt_cereal_quadratic_effects
```

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

The LRT test for linear response variables shows strong statistical significance for sugar and sodium content (p-values of 4.520699×10^{-5} and 7.0732813×10^{-6} respectievely) but fails to show even weak significance of fat content (p-value of 0.1522)

The LRT test for the interactions further revealed that no two-way or three way interaction achieved statistical significance

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.

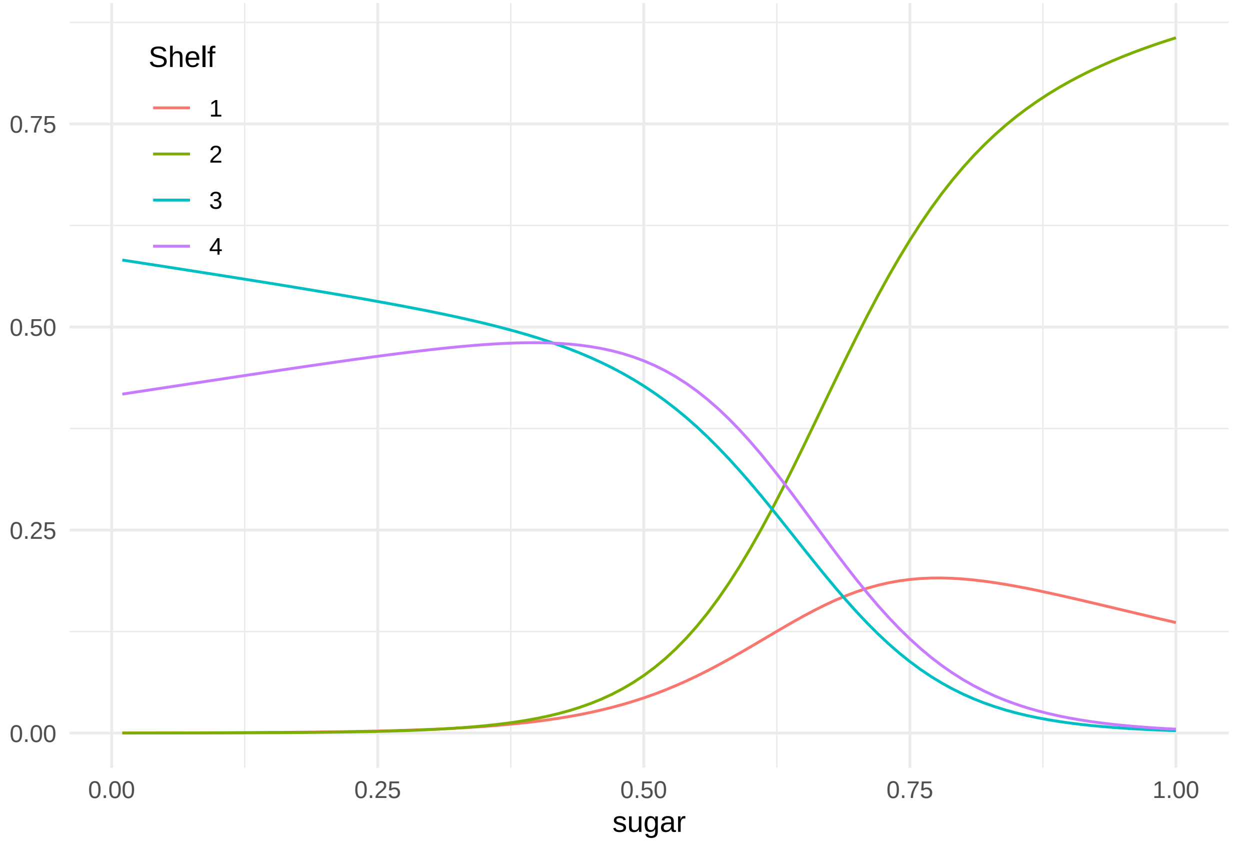
```r
# Estimate new model that removes non-significant fat variable
model_cereal_shelves_trim <- multinom(formula = Shelf ~ sugar + sodium,
                                      data = cereal)

# Create a datframe with Apple Jack data
app_jack <- data.frame(size_g = 28,
                       sugar_raw = 12,
                       fat_raw = 0.5, sodium_raw = 130)

# Use the same normalization procedures as for the main dataframe
for (col in col_trans_list) {
  # Column of variable devided by portion size:
  col_per <- app_jack[col] / app_jack["size_g"]

  # Reference column of variable devided by portion size:
  ref_col <- cereal[col] / cereal["size_g"]
  app_jack[str_sub(col, 1, -5)] <- (col_per - min(ref_col)) / max(ref_col)
}

# Estimate placement of Apple Jack with a trimmed model
aj_shelf_probs_trim <- predict(model_cereal_shelves_trim,
                               newdata = app_jack, type = "probs")
shelf_trim <- aj_shelf_probs_trim[which.max(aj_shelf_probs_trim)]

# Estimate placement of Apple Jack with bloated model
aj_shelf_probs <- predict(model_cereal_shelves_linear,
                          newdata = app_jack, type = "probs")
shelf <- aj_shelf_probs[which.max(aj_shelf_probs)]
```

Using best practices for variable selection we estimated a new model that only contains statistically significant variables: sugar and sodium. Using this model we estimate probabilities of placing Apple Jack on the shelves 1, 2, 3 and 4 respectievely as 4, 59, 16, 21 percentage points. Thus, shelf 2 is clearly the most likely place. If we were to use a full model, that includes insignificant variable Fat, the result would stay the same, but the difference between shelves would be smaller at 5, 47, 20, 27 percentage points.

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
# Get mean values for static values
fat_mean <- mean(cereal$fat)
na_mean <- mean(cereal$sodium)

# Make dataframe with values used fo inference
df_to_plot <- data.frame(fat = rep(fat_mean, times = 100),
                         sodium = rep(na_mean, times = 100),
                         sugar = seq(1, 100) / 100)

# Attach predicted values to this dataframe
df_to_plot <- cbind(df_to_plot,
    predict(model_cereal_shelves_linear, newdata = df_to_plot, type = "probs"))

# Plot the data
shelf_vs_sugar_plot <- df_to_plot %>%
  pivot_longer(cols = c(4:7), names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = sugar, y = values, colour = indicators)) +
  geom_line() +
  theme_minimal() +
  theme(legend.position = c(0.1, 0.8)) + guides(color=guide_legend("Shelf")) +
  theme(axis.title.y = element_blank())

shelf_vs_sugar_plot
```



> For cereals with normalized sugar content up to approximately average, there is roughly eaula chance of finding them on 4’th and 3’d shelfs. Assuming the first shelf is at teh bottom and the 4’th is at the top, an average health-concious adult might see them there. As the normalized sugar content approaches the higher end of the spectrum, the chances of finding this cereal on the second shelf, where a kid might see it, is growing dramatically.

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.

```
odds_ratios <- "fill this in"
```

> ‘Fill this in: What do you learn about each of these variables?’

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.*

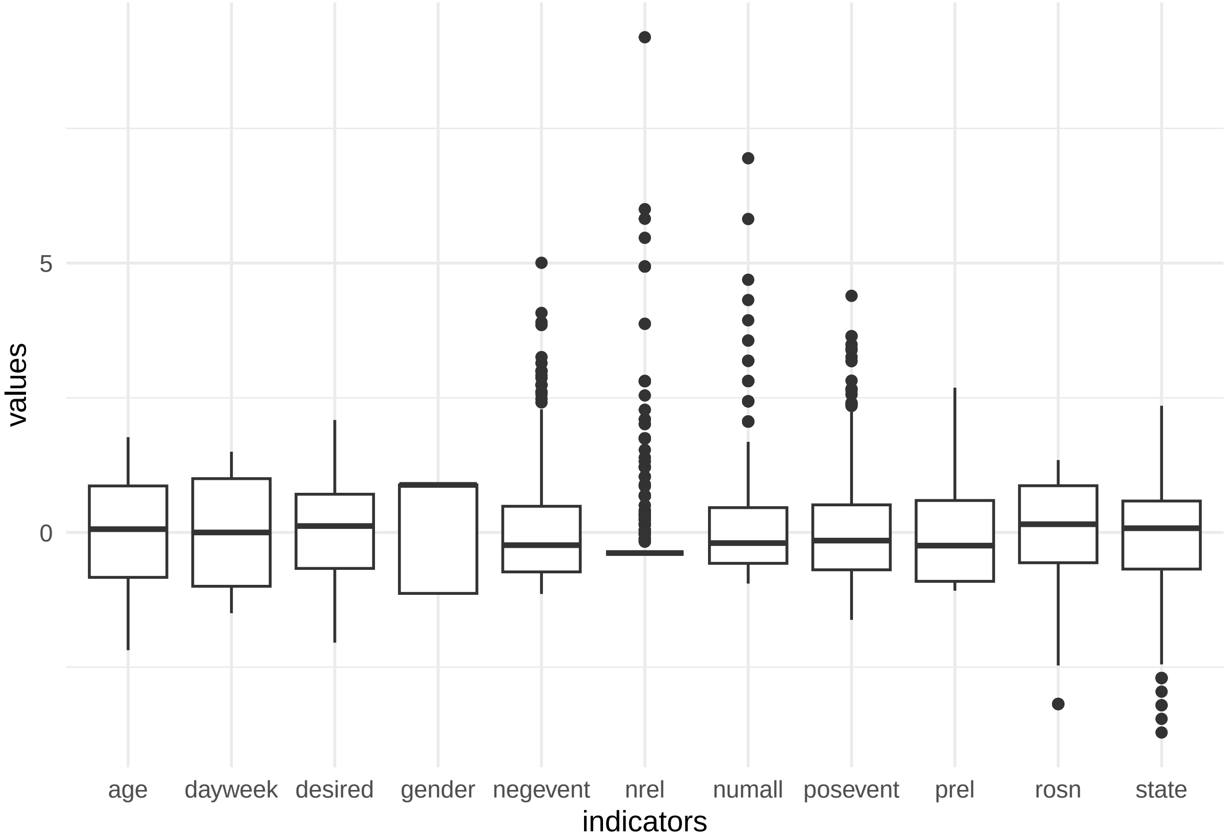
```
drinks <- read_csv("../data/short-questions/DeHartSimplified.csv")
drinks <- read_csv("../data/short-questions/DeHartSimplified.csv")
```

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_scaled <- as.data.frame(scale(drinks))
drinks_scaled %>%
  pivot_longer(cols = 3:13, names_to = "indicators", values_to = "values") %>%
  ggplot(data = ., aes(x = indicators, y = values)) +
  geom_boxplot() +
  theme_minimal()
```

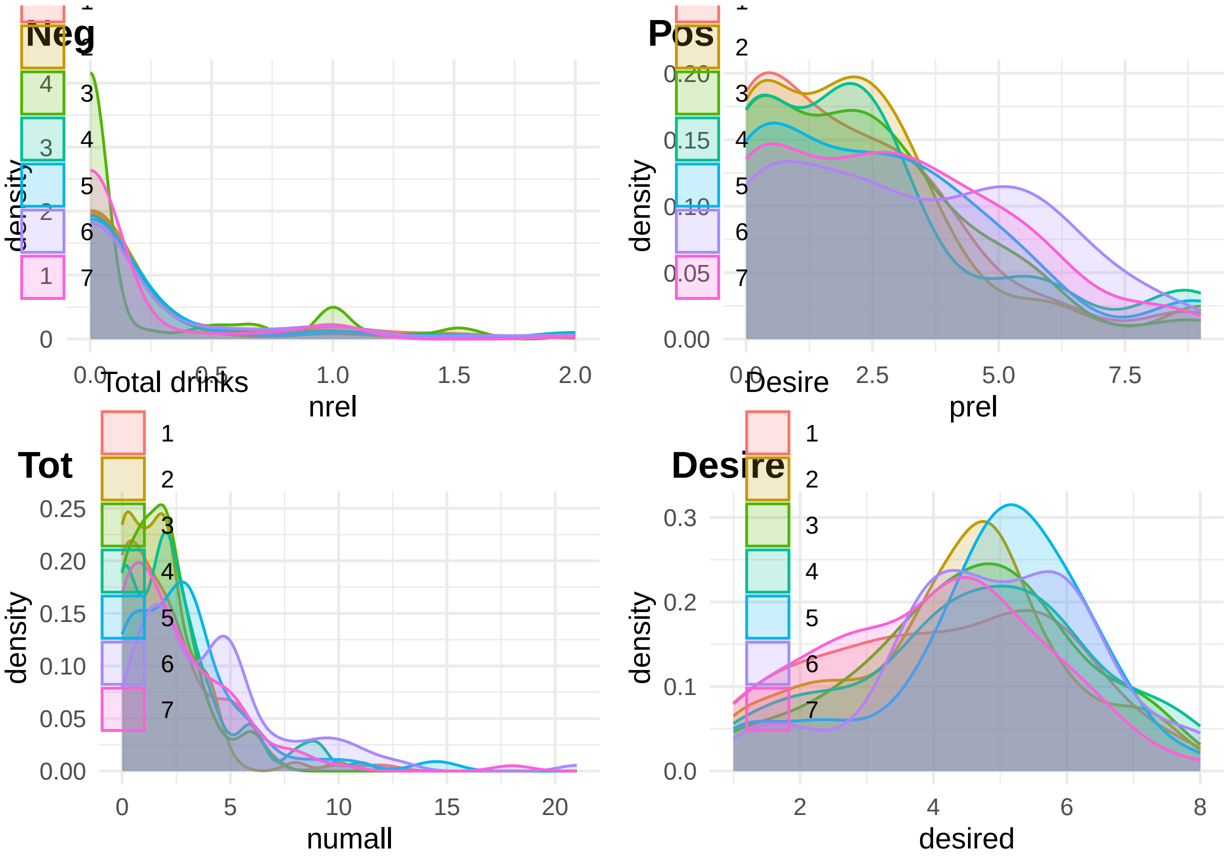
```
## Warning: Removed 7 rows containing non-finite values (`stat_boxplot()`).
```



```
neg_plot <- ggplot(
  data = drinks,
  aes(
    x = nrel,
    fill = as.factor(dayweek),
    colour = as.factor(dayweek)
  )
) +
  geom_density(alpha = 0.2) +
  theme_minimal() +
  theme(legend.position = c(0.15, 0.8)) +
  guides(
    fill = guide_legend(title = "Neg romantic event"),
    colour = guide_legend(title = "Neg romantic event")
  ) +
  xlim(0, 2)

pos_plot <- ggplot(
  data = drinks,
  aes(
    x = prel,
    fill = as.factor(dayweek),
    colour = as.factor(dayweek)
  )
) +
  geom_density(alpha = 0.2) +
  theme_minimal() +
  theme(legend.position = c(0.15, 0.8)) +
  guides(
    fill = guide_legend(title = "Pos romantic event"),
    colour = guide_legend(title = "Pos romantic event")
  )

tot_plot <- ggplot(
  data = drinks,
  aes(
    x = numall,
    fill = as.factor(dayweek),
    colour = as.factor(dayweek)
  )
) +
  geom_density(alpha = 0.2) +
  theme_minimal() +
  theme(legend.position = c(0.15, 0.8)) +
  guides(
    fill = guide_legend(title = "Total drinks"),
    colour = guide_legend(title = "Total drinks")
  )

sest_plot <- ggplot(
  data = drinks,
  aes(
    x = desired,
    fill = as.factor(dayweek),
    colour = as.factor(dayweek)
  )
) +
  geom_density(alpha = 0.2) +
  theme_minimal() +
  theme(legend.position = c(0.15, 0.8)) +
  guides(
    fill = guide_legend(title = "Desire"),
    colour = guide_legend(title = "Desire")
  )
```

```r
ggarrange(NULL, NULL, neg_plot, pos_plot, NULL, NULL, tot_plot, sest_plot,
  labels = c("Neg", "Pos", "", "", "Tot", "Desire", "", ""),
  ncol = 2, nrow = 4, heights = c(0.1, 1, 0.1, 1)
)
```

```
## Warning: Removed 30 rows containing non-finite values (`stat_density()`).

## Warning: Removed 1 rows containing non-finite values (`stat_density()`).

## Warning: Removed 3 rows containing non-finite values (`stat_density()`).
```



```r
stargazer(as.data.frame(drinks), header = FALSE, type = "latex")
```

'Fill this in: What do you learn?'

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.

```r
model_neg_rel_saturday <- glm(formula = as.integer(numall) ~ nrel, family = poisson, data = fil
summary(model_neg_rel_saturday)
```

```
##
```

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
id	623	75.888	49.901	1	160
studyday	623	4.000	2.002	1	7
dayweek	623	4.000	2.002	1	7
numall	622	2.524	2.660	0	21
nrel	623	0.359	0.940	0.000	9.000
prel	623	2.583	2.388	0.000	9.000
negevent	623	0.441	0.387	0.000	2.377
posevent	623	1.048	0.646	0.000	3.883
gender	623	1.562	0.497	1	2
rosn	623	3.436	0.420	2.100	4.000
age	623	34.293	4.512	24.433	42.278
desired	620	4.465	1.694	1.000	8.000
state	620	3.966	0.440	2.333	5.000

```
## Call:
## glm(formula = as.integer(numall) ~ nrel, family = poisson, data = filter(drinks,
##     dayweek == "6"))
##
## 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
```

```
model_neg_rel_tuesday <- glm(formula = as.integer(numall) ~ nrel, family = poisson, data = fil
summary(model_neg_rel_tuesday)
```

```
##
## Call:
## glm(formula = as.integer(numall) ~ nrel, family = poisson, data = filter(drinks,
##     dayweek == "2"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.55994    0.08492   6.594 4.28e-11 ***
## nrel        -0.06843    0.10661  -0.642    0.521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 155.31  on 88  degrees of freedom
## Residual deviance: 154.86  on 87  degrees of freedom
## AIC: 327.48
##
## Number of Fisher Scoring iterations: 5
```

```
model_neg_rel_saturday <- glm(formula = as.integer(numall) ~ nrel, family = poisson, data = fil
summary(model_neg_rel_saturday)
```

```
##
## Call:
## glm(formula = as.integer(numall) ~ nrel, family = poisson, data = filter(drinks,
##     dayweek == "6"))
##
## 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
```

```
model_des_tuesday <- lm(formula = as.numeric(desired) ~ nrel, data = filter(drinks, dayweek ==
summary(model_des_tuesday)
```

```
##
## Call:
## lm(formula = as.numeric(desired) ~ nrel, data = filter(drinks,
##     dayweek == "2"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3220 -0.9887  0.3402  0.6780  3.6780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.32203    0.18626  23.204   <2e-16 ***
## nrel         0.01323    0.20605   0.064    0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.665 on 87 degrees of freedom
## Multiple R-squared:  4.739e-05,  Adjusted R-squared:  -0.01145
## F-statistic: 0.004123 on 1 and 87 DF,  p-value: 0.949
```

```
model_des_saturday <- lm(formula = as.numeric(desired) ~ nrel, data = filter(drinks, dayweek ==
summary(model_des_saturday)
```

```
##
## Call:
## lm(formula = as.numeric(desired) ~ nrel, data = filter(drinks,
##     dayweek == "6"))
##
## 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
```

‘Fill this in: What do you learn?’

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.

```
model_neg_rel_saturday <- glm(formula = as.integer(numall) ~ nrel + rosn + rosn:nrel, family =
summary(model_neg_rel_saturday)
```

```
##
## Call:
## glm(formula = as.integer(numall) ~ nrel + rosn + rosn:nrel, family = poisson,
##     data = filter(drinks, dayweek == "6"))
##
## 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
```

```
model_neg_rel_tuesday <- glm(formula = as.integer(numall) ~ nrel + rosn + rosn:nrel, family =
summary(model_neg_rel_tuesday)
```

```
##
## Call:
## glm(formula = as.integer(numall) ~ nrel + rosn + rosn:nrel, family = poisson,
##     data = filter(drinks, dayweek == "2"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.2166     0.7333  -0.295   0.7677
## nrel          2.4783     1.4350   1.727   0.0842 .
## rosn          0.2230     0.2105   1.059   0.2894
## nrel:rosn    -0.7108     0.4029  -1.764   0.0777 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 155.31  on 88  degrees of freedom
## Residual deviance: 151.57  on 85  degrees of freedom
## AIC: 328.19
##
## Number of Fisher Scoring iterations: 5
```