Lab 1, Short Questions

Contents


```
library(tidyverse)
library(patchwork)
```

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')
head(cereal) #TODO
```

```
## # A tibble: 6 x 7
##      ID Shelf Cereal                                size_g sugar_g fat_g sodium~1
##   <dbl> <dbl> <chr>                                  <dbl>   <dbl> <dbl>    <dbl>
## 1     1     1 Kellog's Razzle Dazzle Rice Crispies     28      10     0      170
## 2     2     1 Post Toasties Corn Flakes                28       2     0      270
## 3     3     1 Kellogg's Corn Flakes                    28       2     0      300
## 4     4     1 Food Club Toasted Oats                   32       2     2      280
```

```
## 5     5     1 Frosted Cheerios                          30      13      1      210
## 6     6     1 Food Club Frosted Flakes                  31      11      0      180
## # ... with abbreviated variable name 1: sodium_mg
```

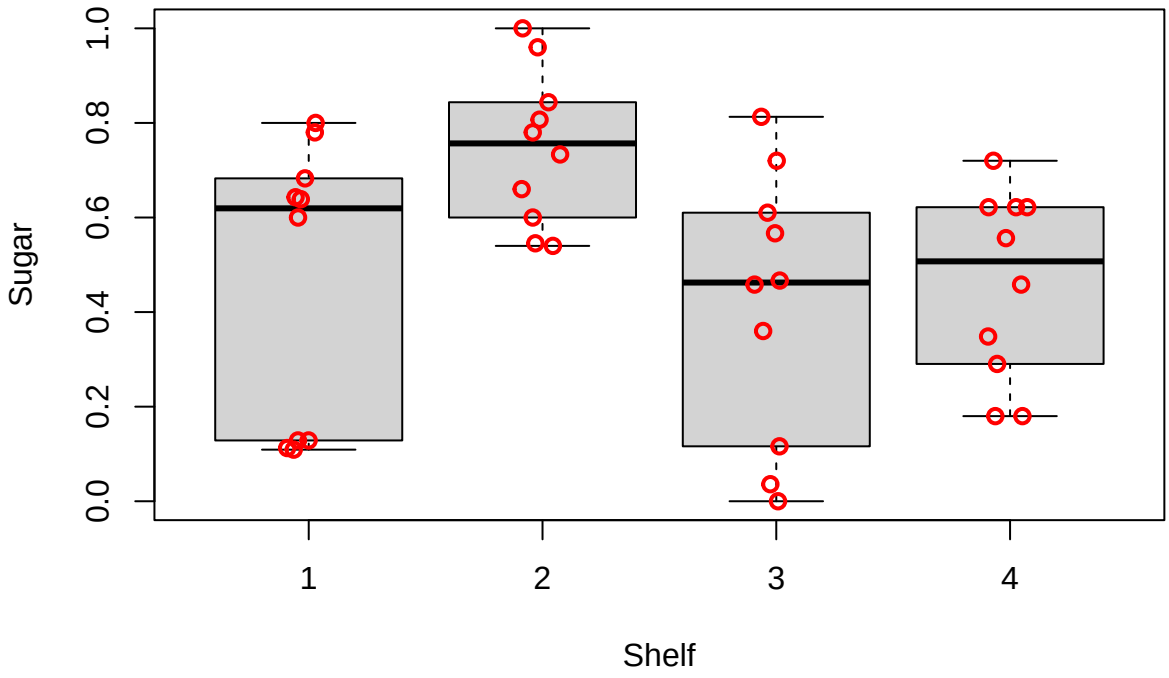
1.1 Recode Data

(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.

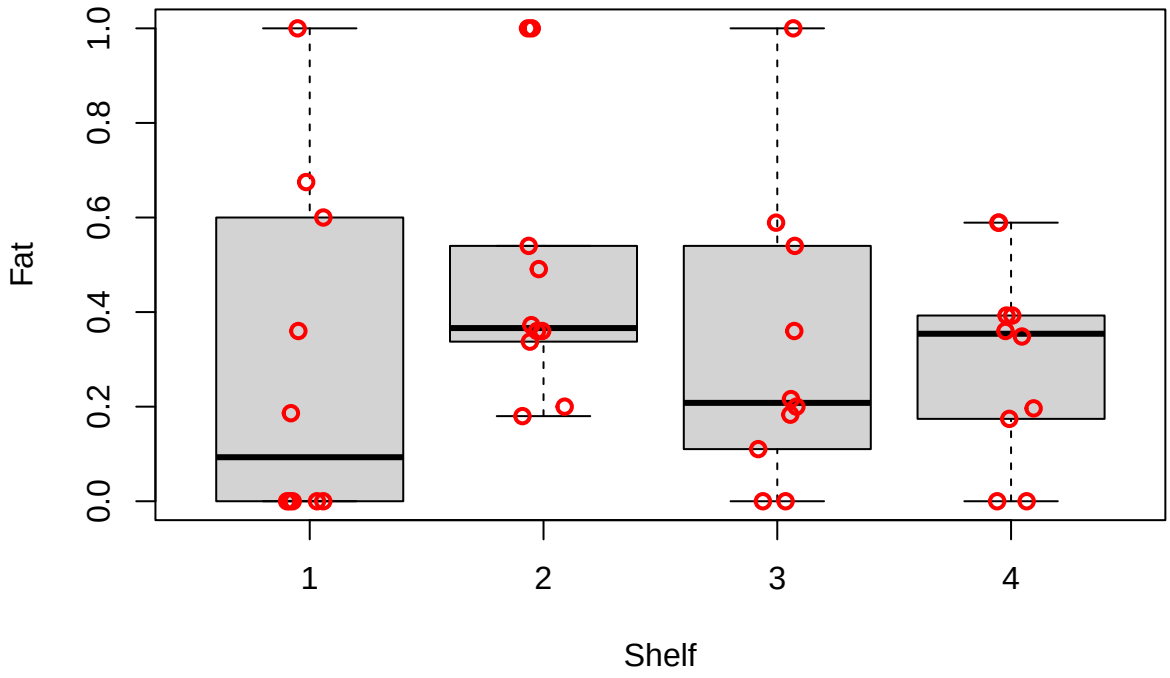
```
# Rescaling each variable to be within 0 and 1
stand01 <- function (x) { (x - min(x))/( max(x) - min(x)) }
cereal2<-data.frame(Shelf=cereal$Shelf, sugar=stand01(x=cereal$sugar_g/cereal$size_g),
                      fat = stand01 (x = cereal$fat_g / cereal$size_g),
                      sodium = stand01 (x = cereal$sodium_mg / cereal$size_g))
head(cereal2)
```

```
##   Shelf     sugar   fat    sodium
## 1     1 0.6428571 0.000 0.5666667
## 2     1 0.1285714 0.000 0.9000000
## 3     1 0.1285714 0.000 1.0000000
## 4     1 0.1125000 0.675 0.8166667
## 5     1 0.7800000 0.360 0.6533333
## 6     1 0.6387097 0.000 0.5419355
```

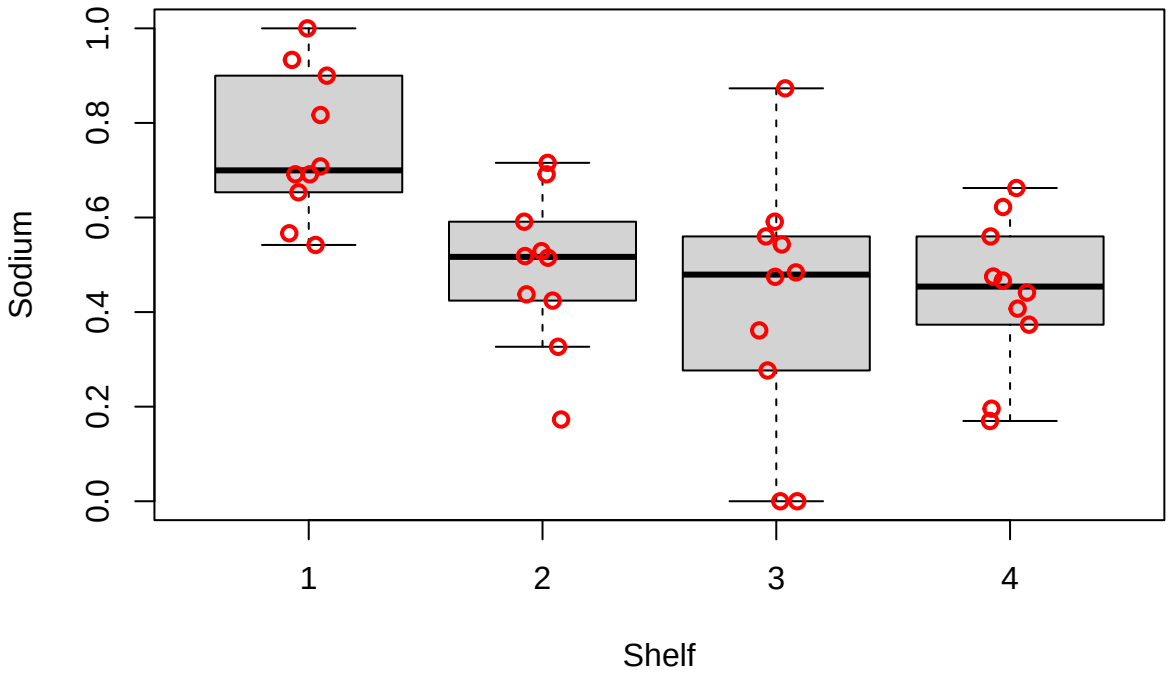
```
# Constructing side-by-side box plots with dot plots overlaid
boxplot(formula=sugar ~ Shelf, data=cereal2, ylab="Sugar",
        xlab="Shelf", pars=list(outpch =NA))
stripchart(x=cereal2$sugar ~ cereal2$Shelf, lwd=2, col="red",
           method="jitter", vertical=TRUE ,pch=1, add=TRUE)
```




```
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)
```



'Fill in: What do you observe in these boxplots?'

```
# Constructing a parallel coordinates plot
library(package = MASS)
```

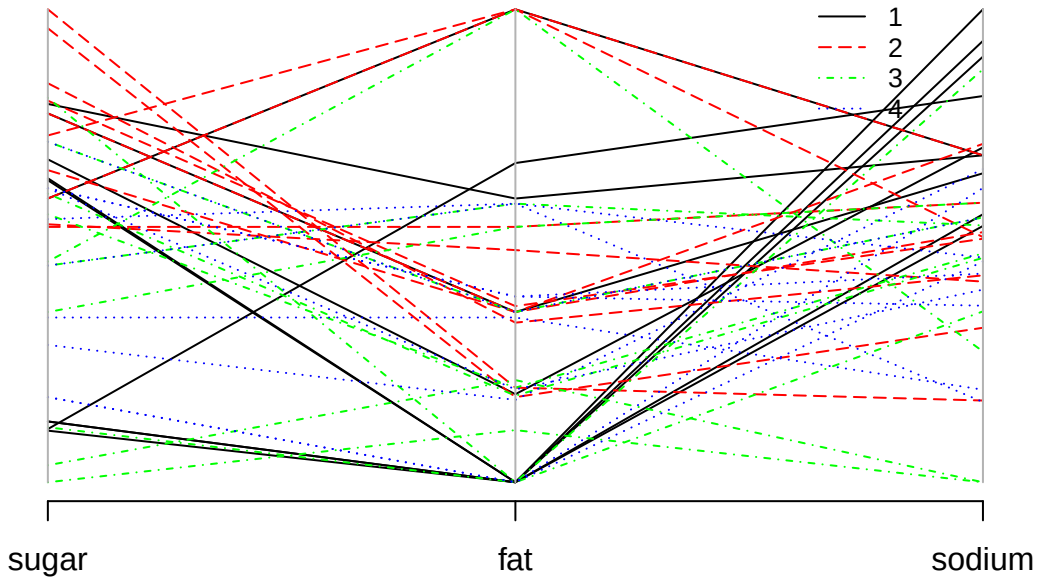
```
##
## Attaching package: 'MASS'
```

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

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

```
# Colors by condition:
cereal2.colors <- ifelse(test = cereal2$Shelf==1, yes = "black",
  no = ifelse(test = cereal2$Shelf==2, yes = "red",
  no = ifelse(test = cereal2$Shelf==3, yes = "green", no = "blue")))
# Line type by condition:
cereal2.lty<-ifelse(test = cereal2$Shelf==1, yes = "solid",
  no = ifelse(test = cereal2$Shelf==2, yes = "longdash",
  no = ifelse(test = cereal2$Shelf==3, yes = "dotdash", no = "dotted")))
# Plot
```

```
parcoord(x = cereal2[,2:4], col = cereal2.colors, lty = cereal2.lty)
legend(x = 2.6, y = 1.05, legend = c("1", "2", "3", "4"),
       lty = c("solid", "longdash", "dotdash", "dotted"),
  col=c("black", "red", "green", "blue"), cex=0.8, bty="n")
```



'Fill in: What do you observe in these parallel coordinates plots?'

Fill in: Do content differences exist between the shelves?'

'Majority of the food items tend to have a "V" shape with high suger, low-to-medium fat and high sodium.'
'Low-to-medium sugar cereals generally available on shelf 1, 3 & 4 and high sugar cereals on shelf 2. High fat cereals generally available on self 1, 3 & 4 and low fat on shelf 2 & 3. Low sodium cereals are generally available on slef 2, 3 & 4 and high sodium on shelf 1'

'Shelf 1 tend to have low fat and high sodium. Shelf 2 tend to have high sugar, high-to-medium fat and medium sodium. Shelf 3 tend to have all types of cereals. Sheld 4 tend to have medium sugar, low-to-medium fat and medium sodium cereals.'

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).

Fill in: What do you think about ordinal data?'

,,

```
model_cereal_shelves_linear    <- 'fill this in'
model_cereal_shelves_quadratic <- 'fill this in'
```

```
lrt_cereal_main_effects <- 'fill this in'
```

```
lrt_cereal_quadratic_effects <- 'fill this in'
```

'Fill in: Write about what you learn as a result of these tests, using inline code evaluation.'

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.

```
aj_shelf_probs <- 'fill this in'
```

'Fill this in: Where does your model predict apple jacks will be placed?'

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.

```
shelf_vs_sugar_plot <- 'fill this in'
```

'Fill this in: What message does your plot give?'

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')
```

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.

'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.

'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.

'Fill this in: What do you learn?'