Lab 1, Short Questions

Contents


```
library(tidyverse)
library(patchwork)
library(ggplot2)
library(MASS)
library(car)
```

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')
```

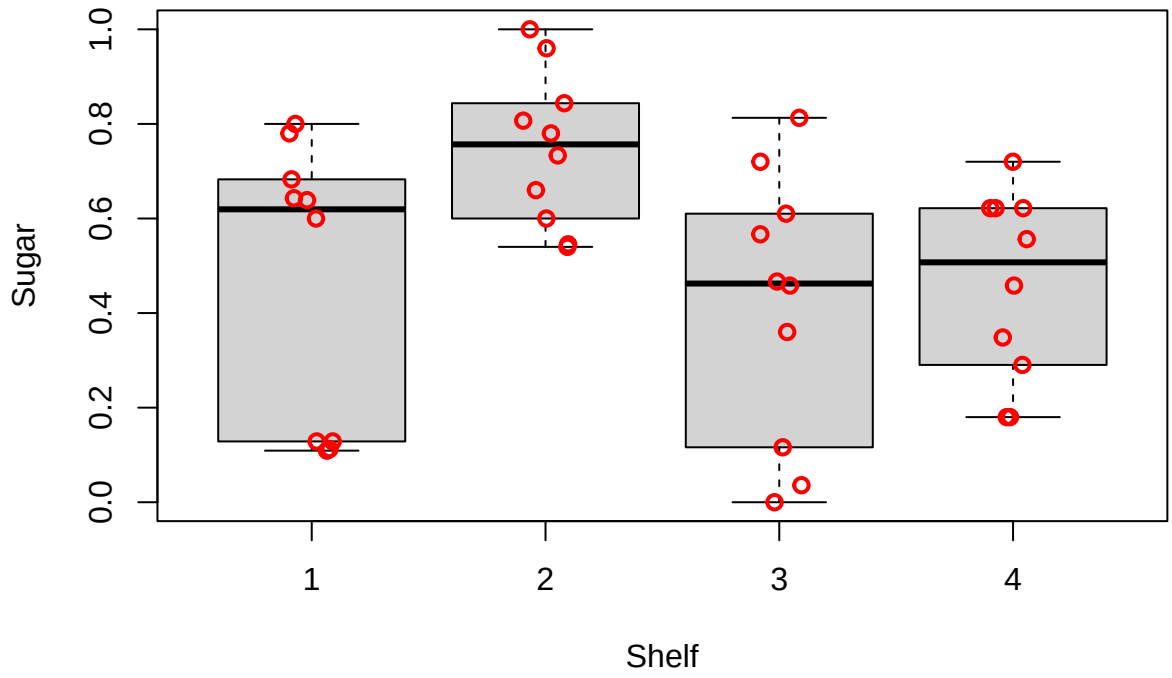
1.1 Recode Data

(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

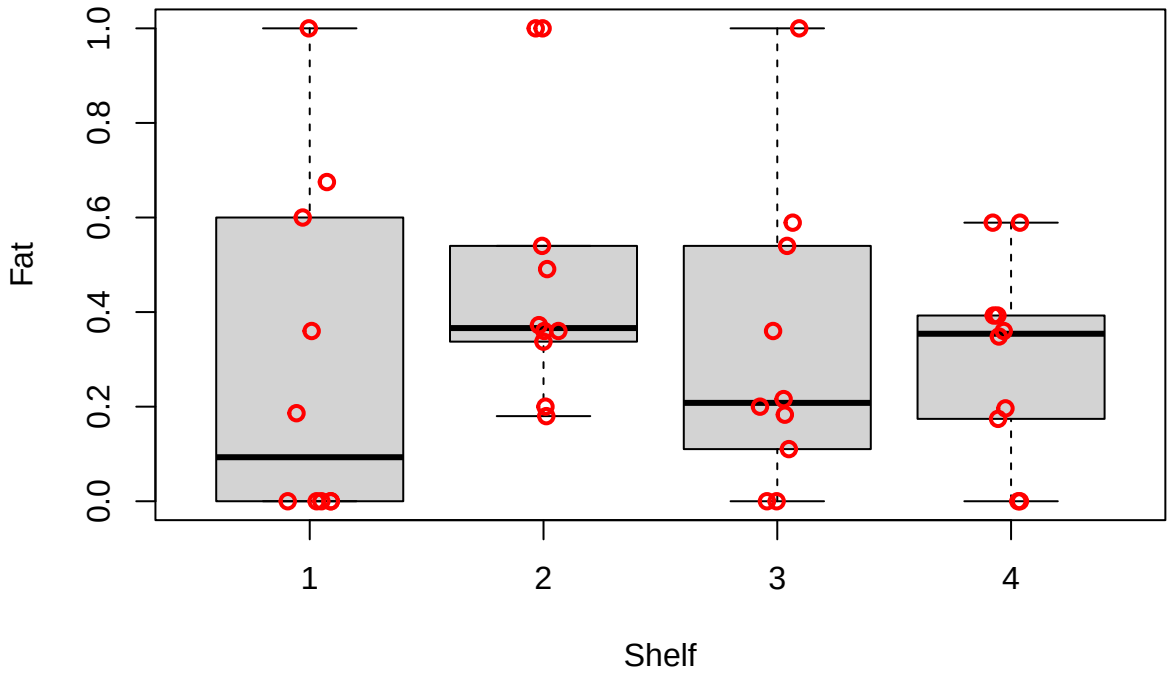
for the different serving sizes among the cereals. Second, rescale each variable to be within 0 and 1. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables. Also, construct a parallel coordinates plot for the explanatory variables and the shelf number. Discuss whether possible content differences exist among the shelves.

```
standardize <- function (x) { (x - min(x))/( max(x) - min(x)) }

rescaled_cereal <- cereal %>%
  dplyr::mutate(sugar = standardize(sugar_g/size_g),
                fat = standardize(fat_g/size_g),
                sodium = standardize(sodium_mg/size_g))
```

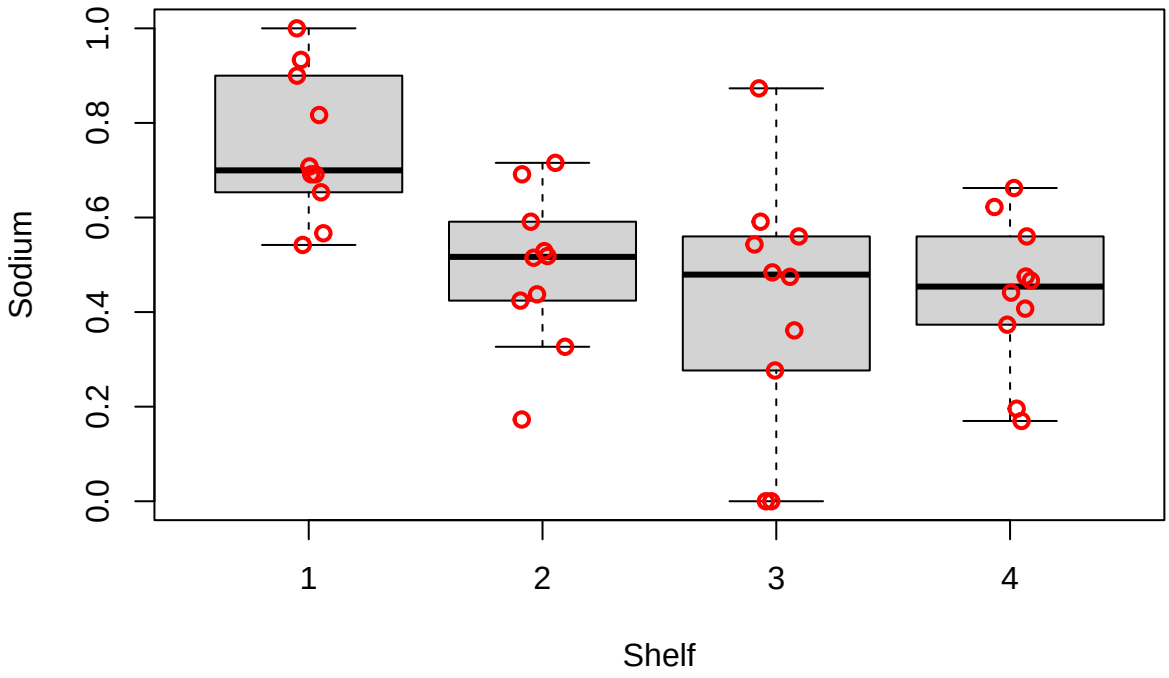
```
boxplot(formula=sugar ~ Shelf, data=rescaled_cereal, ylab="Sugar",
        xlab="Shelf", pars=list(outpch =NA))
stripchart(x=rescaled_cereal$sugar ~ rescaled_cereal$Shelf, lwd=2, col="red",
           method="jitter", vertical=TRUE ,pch=1, add=TRUE)
```




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

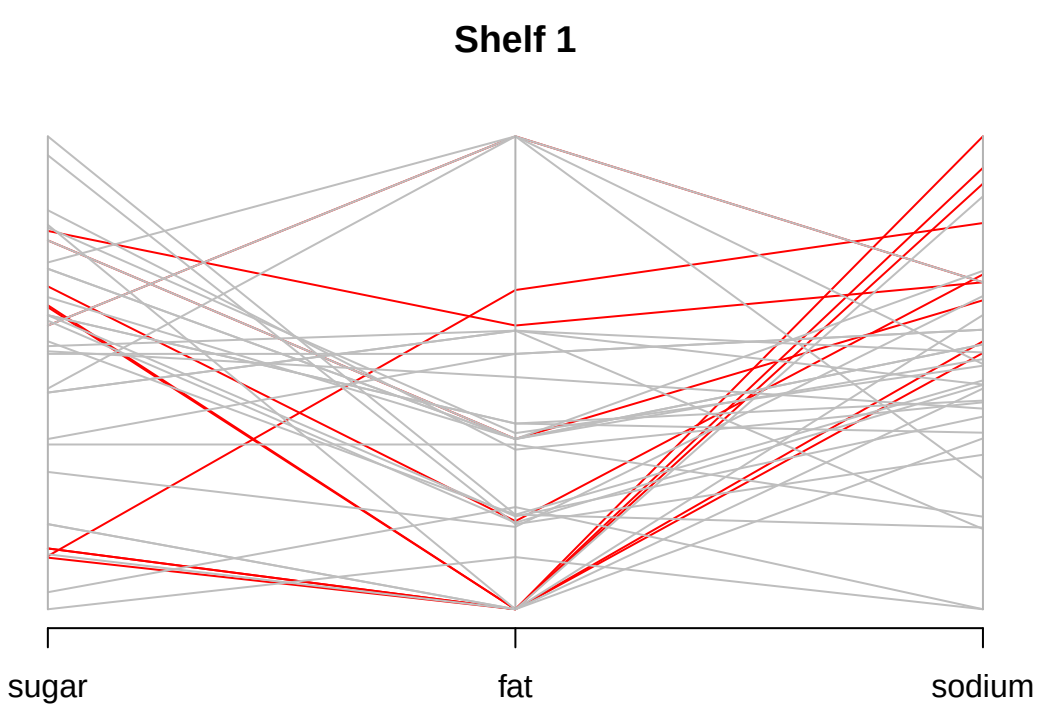



```
boxplot(formula=sodium ~ Shelf, data=rescaled_cereal, ylab="Sodium",
        xlab="Shelf", pars=list(outpch =NA))
stripchart(x=rescaled_cereal$sodium ~ rescaled_cereal$Shelf, lwd=2, col="red",
           method="jitter", vertical=TRUE, pch=1, add=TRUE)
```

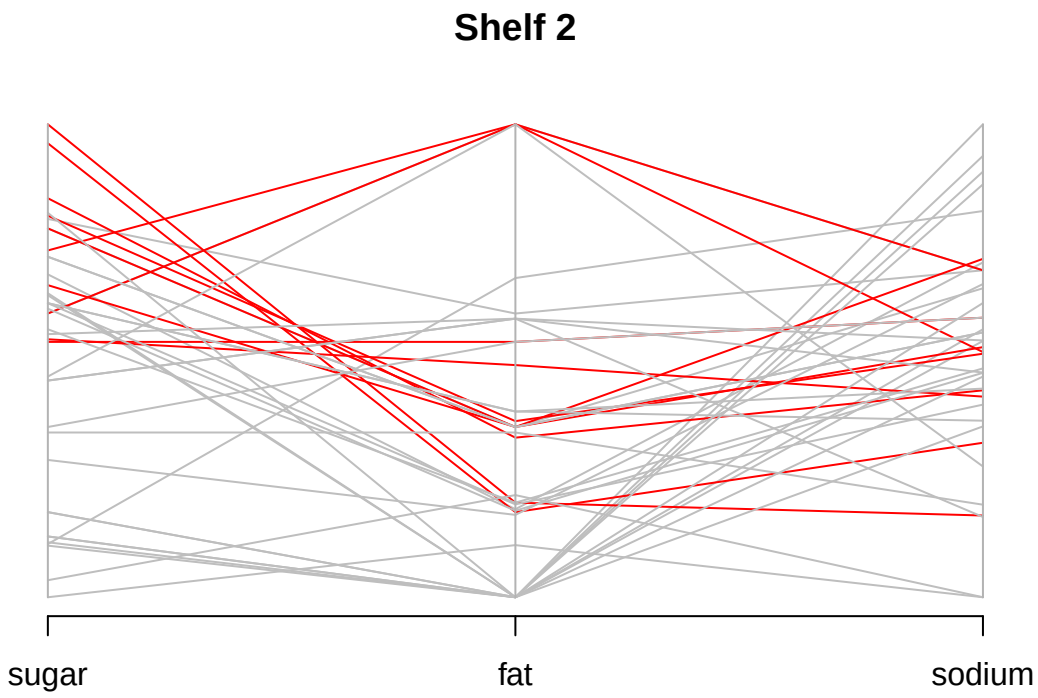


Observations:: 1. Cereal boxes placed on the second shelf appear to have the highest amount of sugar, ranging from ~0.6 to 1.0 in standardized terms. The distribution of sugar content is much wider on the other shelves, going as low as 0.1. 2. Shelf 1 has the widest distribution of fat content. Shelves 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 3. 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.

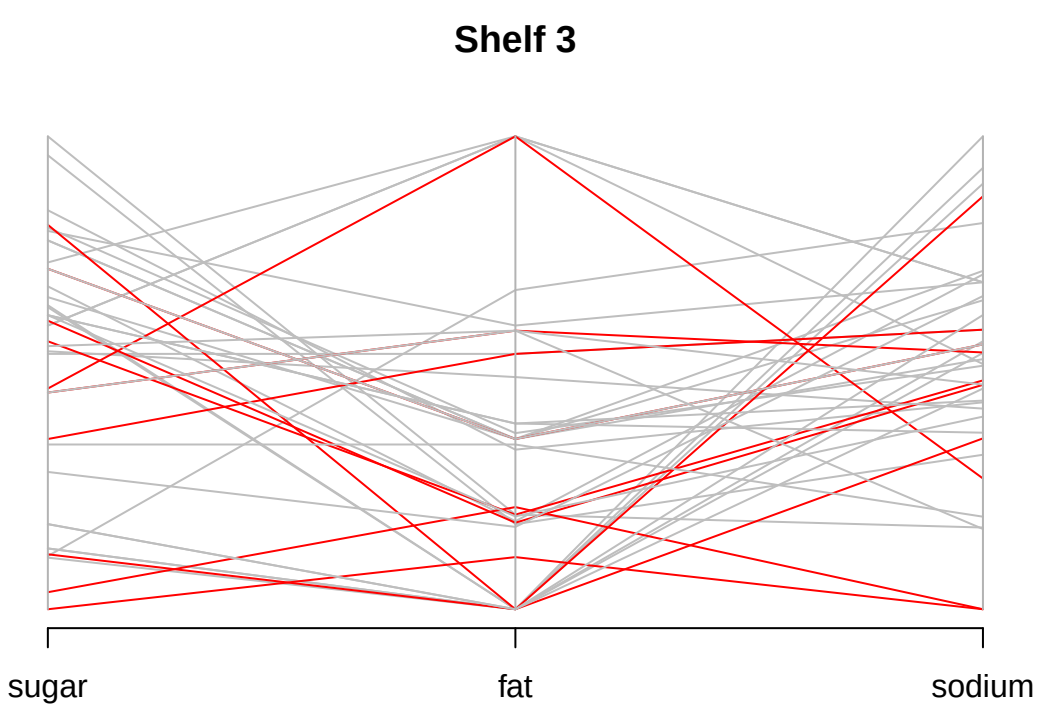
```
is1 <- ifelse(rescaled_cereal$Shelf==1,"red","grey")
is2 <- ifelse(rescaled_cereal$Shelf==2,"red","grey")
is3 <- ifelse(rescaled_cereal$Shelf==3,"red","grey")
is4 <- ifelse(rescaled_cereal$Shelf==4,"red","grey")

MASS::parcoord(x = rescaled_cereal[,8:10], col = is1, main = "Shelf 1")
```



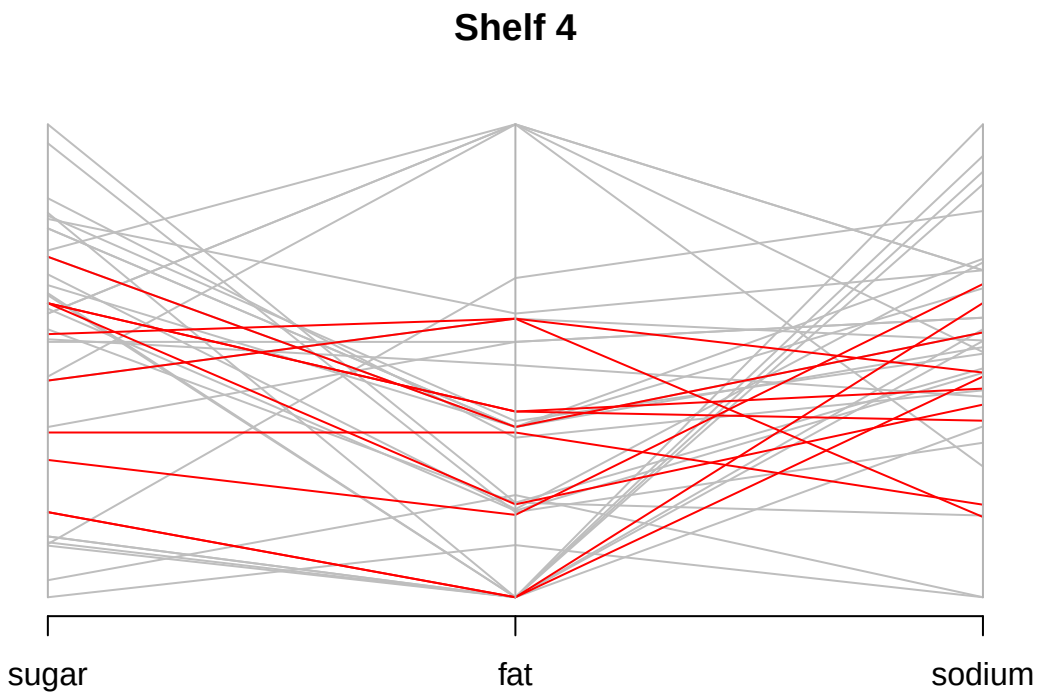

```
MASS::parcoord(x = rescaled_cereal[,8:10], col = is2, main = "Shelf 2")
```




```
MASS::parcoord(x = rescaled_cereal[,8:10], col = is3, main = "Shelf 3")
```




```
MASS::parcoord(x = rescaled_cereal[,8:10], col = is4, main = "Shelf 4")
```



Observations: 1. Shelf 1 has the highest amount of sodium content. 2. Shelf 2 has the highest sugar and also a few boxes with the highest fat content. 3. Shelf 3's sugar, fat, amd sodium contents cover the entire range from low to high 4. Shelf 4 has low to medium amount of sugar, far, 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).

We can imagine the shelves to be physically ordered such that shelf 1 is below eye level and shelf 4 is above eye level. Items placed on lower shelves are often cheaper and may be of lower nutritional quality than items placed on shelves at the eye level or above. Moreover, one of the major consumer segments for cereals is children and cereal boxes' packaging often shows cartoon characters looking downward. These cereal brands are often the tastier ones, with medium to high sugar and fat content, and are placed on shelves at the adults' eye level. The idea here is that children will look up to find these cartoon characters looking at them and will ask the adults for a box, who will find the box at their eye level and can easily pick them up. The boxes above the adults' eye level are often the "healthier" ones and will need the person to take an extra step to reach up and grab the box.

Grocery stores use the concept of friction from behavioral psychology to make the decisions around product placement on shelves. In the setup described above, cereals with high sugar and fat content slowly get consumers hooked onto the taste. Manufacturers can then use this feature of human beings craving sugar and fat to price these brands of cereal higher. Cheaper cereal brands (ones with high sodium) and the healthier options are places on shelves 1 and 4, introducing a point of friction in the shopping experience when the consumer either has to crouch down or crane up to grab a box.

Given this context, the ordinality of the shelves should be taken into account with $1 < 2 < 3 < 4$.

```
rescaled_cereal$Shelf.order <- factor(rescaled_cereal$Shelf, levels = c(1,2,3,4))
model_cereal_shelves_linear    <- polr(formula = Shelf.order ~ sugar + fat + sodium,
                                       data = rescaled_cereal, method = "logistic")

model_cereal_shelves_quadratic <- polr(formula = Shelf.order ~ sugar + fat + sodium +
                                                               sugar:fat +
                                                               fat:sodium +
                                                               sodium:sugar +
                                                               sugar:fat:sodium,
                                       data = rescaled_cereal, method = "logistic")
```

```
lrt_cereal_main_effects <- Anova(model_cereal_shelves_linear)
lrt_cereal_main_effects
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf.order
##        LR Chisq Df Pr(>Chisq)
## sugar    1.6794  1  0.1950069
## fat      0.0028  1  0.9577007
## sodium  11.5685  1  0.0006708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test shows sufficient evidence for sodium to be an important explanatory variable in shelf placement. This follows from our initial EDA where high sodium content was clearly a feature of cereal boxes placed on Shelf 1.

```
lrt_cereal_quadratic_effects <- Anova(model_cereal_shelves_quadratic)
lrt_cereal_quadratic_effects
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf.order
##                  LR Chisq Df Pr(>Chisq)
## sugar              1.1760  1  0.2781685
## fat                0.0419  1  0.8377311
## sodium            11.1699  1  0.0008314 ***
## sugar:fat          0.1014  1  0.7501457
## fat:sodium         0.2607  1  0.6096643
```

```
## sugar:sodium       0.3945  1  0.5299556
## sugar:fat:sodium   0.1077  1  0.7427907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio tests for the model with interacted variables only show sufficient evidence of importance for sodium.

```
## we can also use nominal_test() to check whether
## the ordinal logistic regression as constructed is valid or not.

## taken from live session 4 solutions

#refit prop odds model using ordinal package to allow use of nomial_test function
lrt_cereal_main_effects2 <- ordinal::clm(Shelf.order ~  fat + sodium + sugar,
                                         data = rescaled_cereal)
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
```

1.2.1 TODO: why are we not getting any output for sugar?

'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.

```
size   <- 28
# rescaled

sugar_scaled_data <- rescaled_cereal$sugar_g/rescaled_cereal$size_g
fat_scaled_data <-  rescaled_cereal$fat_g/rescaled_cereal$size_g
sodium_scaled_data <-  rescaled_cereal$sodium_mg/rescaled_cereal$size_g

sugar  <- ((12/size) - min(sugar_scaled_data)) /
  (max(sugar_scaled_data) - min(sugar_scaled_data))

fat    <- ((0.5/size) - min(fat_scaled_data)) /
```

```
  (max(fat_scaled_data) - min(fat_scaled_data))

sodium <- ((130/size) - min(sodium_scaled_data)) /
  (max(sodium_scaled_data) - min(sodium_scaled_data))

predict.data <- data.frame(sugar  = sugar,
                           fat    = fat,
                           sodium = sodium)

aj_shelf_probs <- predict(object = model_cereal_shelves_linear,
                          newdata = predict.data, type = "probs")
aj_shelf_probs
```

```
##         1         2         3         4
## 0.1986192 0.3051178 0.2946317 0.2016313
```

```
predict(object = model_cereal_shelves_linear,
                          newdata = predict.data, type = "class")
```

```
## [1] 2
## Levels: 1 2 3 4
```

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

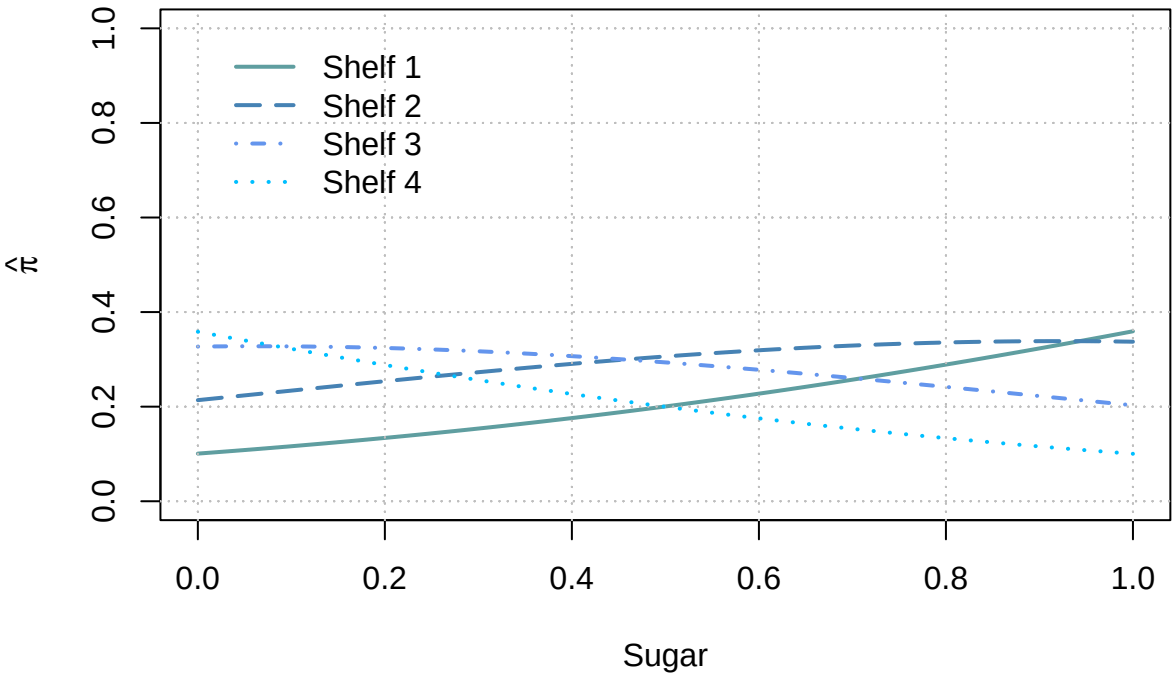
Based on the predicted probabilites, the model predicts apple jacks to be placed on Shelf 2.

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.

```
# 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(rescaled_cereal$sodium),
                                          sugar = x,
                                          fat = mean(rescaled_cereal$fat)),
                     type = "probs")[,1],
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(rescaled_cereal$sugar), max(rescaled_cereal$sugar)),
      ylim = c(0,1), col = "#5F9EA0", lty = "solid", lwd = 2, n = 1000,
      panel.first = grid(col = "gray", lty = "dotted"))

curve(expr = predict(object = model_cereal_shelves_linear,
                     newdata = data.frame(sodium = mean(rescaled_cereal$sodium),
                                          sugar = x,
                                          fat = mean(rescaled_cereal$fat)),
                     type = "probs")[,2],
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(rescaled_cereal$sugar), max(rescaled_cereal$sugar)),
```

```
      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(rescaled_cereal$sodium),
                                          sugar = x,
                                          fat = mean(rescaled_cereal$fat)),
                     type = "probs")[,3],
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(rescaled_cereal$sugar), max(rescaled_cereal$sugar)),
      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(rescaled_cereal$sodium),
                                          sugar = x,
                                          fat = mean(rescaled_cereal$fat)),
                     type = "probs")[,4],
      ylab = expression(hat(pi)), xlab = "Sugar",
      xlim = c(min(rescaled_cereal$sugar), max(rescaled_cereal$sugar)),
      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","longdash","dotdash", "dotted"),
       col=c("#5F9EA0","#4682B4","#6495ED", "#00BFFF"), bty="n", lwd = c(2,2,2,2))
```



For mean values of fat and sodium, the plot shows that at low levels of sugar content, the cereal is most likely to be placed on shelf 4 followed by shelf 3. At higher levels of sugar content, the cereal is more likely to be placed on shelves 1 and 2. Apple Jacks, with the normalized sugar content of 0.77 is most likely to be placed on shelf 2.

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.

```
c.value   <- apply(X = rescaled_cereal[,8:10], MARGIN = 2, FUN = sd)
print("Using 1 standard deviation as c unit of change:")
```

```
## [1] "Using 1 standard deviation as c unit of change:"
```

```
c.value
```

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

```
print("Odds ratios")
```

```
## [1] "Odds ratios"
```

```
odds_ratios <- round(exp(c.value * (-model_cereal_shelves_linear$coefficients)), 2)
odds_ratios
```

```
##  sugar    fat sodium
##   1.54   1.02   3.06
```

```
print("LR Confidence Intervals")
```

```
## [1] "LR Confidence Intervals"
```

```
conf.beta   <- confint(object = model_cereal_shelves_linear, level = 0.95)
```

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

```
##
## Re-fitting to get Hessian
```

```
ci <- exp(c.value * (-conf.beta))
round (data.frame (low = 1/ci[,1] , up = 1/ci[ ,2]) , 2)
```

```
##         low   up
## sugar  0.31 1.24
## fat    0.56 1.75
## sodium 0.14 0.64
```

> With 95% confidence, the odds of a cereal box being placed below a particular level change by 0.31 to 1.23 times when sugar content is decreased by 0.27, holding the other variables constant. With 95% confidence, the odds of a cereal box being placed below a particular level change by 0.56 to 1.75 times when fat content is decreased by 0.299, holding the other variables constant. With 95% confidence, the odds of a cereal box being placed below a particular level change by 0.14 to 0.64 times when sugar content is decreased by 0.23, holding the other variables constant.

The confidence intervals for sodium are the only intervals that do not contain 1. Thus, there is sufficient evidence that it is an important explanatory variable in the linear model.

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?'