Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Michael Berger, John Gao and Thomas Hamnett

Strategic Placement of Products in Grocery Stores

Answer **Question 12 of chapter 3 (on page 189 and 190)** of Bilder and Loughin's *"Analysis of Categorical Data with R"*. Here is the background of this analysis, taken as an excerpt from this question:

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of items – 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.

Introduction

In this paper, we will examine shelf placement of breakfast cereals in grocery stores, using a random sample of 10 cereals from each of four shelves at a Dillons grocery store in Manhattan, KS. Specifically, we will model the probability of a cereal being placed on a specific shelf given the explanatory variables of its sugar, fat, and sodium content. And we will use that model to estimate shelf placement of an out of sample cereal given its sugar, fat, and sodium content.

The key question being asked is, if probability of shelf placement of a cereal can be realiably modeled using its sugar, fat, and sodium content. The question is motivated by grocery stores' desire to best attract customers to cereal products, by placing particular boxes of cereal on specific shelves.

In order to conduct this analysis, we first standarized variables to enable interpretation of results, with each variable in the sample data bounded by 0 and 1 (inclusive). We then modeled log odds ratios of shelf placement using nominal multinomial logistic regression modeling the log odds ratio of shelf levels 2, 3, and 4 versus the base shelf level of 1, as a function of cereal sugar, fat, and sodium content. We chose a nominal model vs an ordinal model since a priori we do not know the natural order of 'worst' to 'best' shelf level. We did not use interaction terms since no interaction terms were identified as significant, and we failed to reject the hypothesis that the model with interaction terms performed better than the model without interaction terms.

Using our model, we found the estimated probability of Apple Jacks appearing on shelf 2 is 98%, shelf 1 is 2%, and is neglible for shelves 3 and 4. We note that the sugar level of Apple Jacks is outside the sample data bounds, potentially calling into question the validity of the estimate (if additional sample data that was similar to Apple Jacks changed our model parameters). We will note this as a key caveat for interpreting our probability estimate of Apple Jacks shelf placement.

We also note that a one standard deviation increase in sodium has a large impact on odds ratios of shelf 1 vs shelves 2, 3, and 4, and one sd increase in sugar has a large impact on odds ratios of shelf 1 vs shelves 3 and 4. One sd increase in fat content has a much lower impact on odds ratios than the other two explanatory variables. This aligns with our observation that the model parameters for sugar and sodium are found to be highly significant, while parameter for fat is not found to be significant (statistically different from 0 at an alpha = 0.05 level). The analysis can provide insight into how grocery stores tend to place cereals on shelf level given its sugar, fat, and sodium content.

##a. The explanatory variables need to be reformatted before proceeding further.

```
- 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.
- Some sample code is provided
```

Answer:

Based on the lab discription, no introduction, exploratory data analysis and conclusion are needed. We load the libraries we use, the data and inspect the data only quickly before attempting the first question.

```r
# Libraries
library(knitr)
library(stargazer)
library(car)
library(dplyr)
library(Hmisc)
library(MASS)
library(nnet)
# We use the gridExtra library to call the grid.arrange() function
# Details: https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf
#library(gridExtra)
# We use the kableExtra library to format tables
# Details: https://cran.r-project.org/web/packages/kableExtra/kableExtra.pdf
library(kableExtra)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# Loading data
cereals <- read.csv('cereal_dillons.csv', sep = ',', header = TRUE)

# Inspecting data
str(cereals)
```

```
## 'data.frame':    40 obs. of  7 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Shelf    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Cereal   : Factor w/ 38 levels "Basic 4","Capn Crunch",..: 17 34 19 13 16 9 2 3 30 8 ...
##  $ size_g   : int  28 28 28 32 30 31 27 27 29 33 ...
##  $ sugar_g  : int  10 2 2 2 13 11 12 9 11 2 ...
##  $ fat_g    : num  0 0 0 2 1 0 1.5 2.5 0.5 0 ...
##  $ sodium_mg: int  170 270 300 280 210 180 200 200 220 330 ...
```

```r
kable(summary(cereals), "latex", booktabs = T) %>%
  kable_styling(font_size = 7)
```

ID	Shelf	Cereal	size_g	sugar_g	fat_g	sodium_mg
Min. : 1.00	Min. :1.00	Capn Crunch's Peanut Butter Crunch: 2	Min. :27.00	Min. : 0.0	Min. :0.000	Min. : 0.0
1st Qu.:10.75	1st Qu.:1.75	Food Club Toasted Oats : 2	1st Qu.:29.75	1st Qu.: 6.0	1st Qu.:0.500	1st Qu.:157.5
Median :20.50	Median :2.50	Basic 4 : 1	Median :31.00	Median :11.0	Median :1.000	Median :200.0
Mean :20.50	Mean :2.50	Capn Crunch : 1	Mean :37.20	Mean :10.4	Mean :1.200	Mean :195.5
3rd Qu.:30.25	3rd Qu.:3.25	Cinnamon Grahams : 1	3rd Qu.:51.00	3rd Qu.:14.0	3rd Qu.:1.625	3rd Qu.:262.5
Max. :40.00	Max. :4.00	Cocoa Pebbles : 1	Max. :60.00	Max. :20.0	Max. :5.000	Max. :330.0
NA	NA	(Other) :32	NA	NA	NA	NA

```r
any(is.na(cereals))
```

```
## [1] FALSE
```

```r
head(cereals, 4)
```

```
##   ID Shelf                               Cereal size_g sugar_g fat_g
## 1  1     1 Kellog's Razzle Dazzle Rice Crispies     28      10     0
## 2  2     1               Post Toasties Corn Flakes     28       2     0
## 3  3     1                   Kellogg's Corn Flakes     28       2     0
## 4  4     1                  Food Club Toasted Oats     32       2     2
##   sodium_mg
## 1       170
## 2       270
## 3       300
## 4       280
```

We note: - No missing values in the data. - ID variable acts as index, which we can remove. - `Shelf` is coded as an integer, although it should be categorical with four levels (1: bottom, to 4: top). - There are 38 different cereals. Hence, `Cereal` is a categorical variable with 38 levels. Nearly each observation accounts for one cereal type. - There are different sizes for the cereals, which are store din `size_g` and influence the three explanatory variables `sigar_g`, `fat_g` and `sodium_mg`.

We now use the provided code to rescale the variables.

```r
# Defining the standardization function as in the book, p. 190
# This function takes the difference of each value to the minimum and divides
# it then by the range between minimum and maximum. This will then let the variable
# be between 0 and 1.
stand01 <- function (x) { (x - min(x)) / (max(x) - min(x))}

# Applying the function to the data.
cereal2 <- data.frame(Shelf = cereals$Shelf,

                      sug_std = stand01(x = cereals$sugar_g / cereals$size_g),

                      fat_std = stand01(x = cereals$fat_g/cereals$size_g),

                      sod_std = stand01(x = cereals$sodium_mg/cereals$size_g))
```

```r
describe(cereal2)
```

```
## cereal2
##
##  4  Variables      40  Observations
## ---------------------------------------------------------------------------
## Shelf
##        n  missing distinct     Info     Mean      Gmd
##       40        0        4    0.938      2.5    1.282
##
## Value         1    2    3    4
## Frequency    10   10   10   10
## Proportion 0.25 0.25 0.25 0.25
## ---------------------------------------------------------------------------
## sug_std
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       32    0.999   0.5209   0.3062   0.1054   0.1158
##      .25      .50      .75      .90      .95
##   0.3339   0.6000   0.7200   0.8075   0.8496
##
## lowest : 0.0000000 0.0360000 0.1090909 0.1125000 0.1161290
## highest: 0.8068966 0.8129032 0.8437500 0.9600000 1.0000000
## ---------------------------------------------------------------------------
## fat_std
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       20    0.985   0.3476   0.3319   0.0000   0.0000
##      .25      .50      .75      .90      .95
##   0.1582   0.3542   0.5400   0.7075   1.0000
##
## lowest : 0.0000000 0.1102041 0.1741935 0.1800000 0.1830508
## highest: 0.5400000 0.5890909 0.6000000 0.6750000 1.0000000
## ---------------------------------------------------------------------------
## sod_std
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       35    0.999    0.524   0.2583   0.1612   0.1934
##      .25      .50      .75      .90      .95
##   0.4200   0.5354   0.6696   0.8223   0.9017
##
## lowest : 0.0000000 0.1696970 0.1728395 0.1956989 0.2765432
## highest: 0.8166667 0.8731183 0.9000000 0.9333333 1.0000000
## ---------------------------------------------------------------------------
```

We rescaled the three variables, as desired, in order to transform each of the variables to be bounded by 0 and 1, inclusive.

##b. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables.

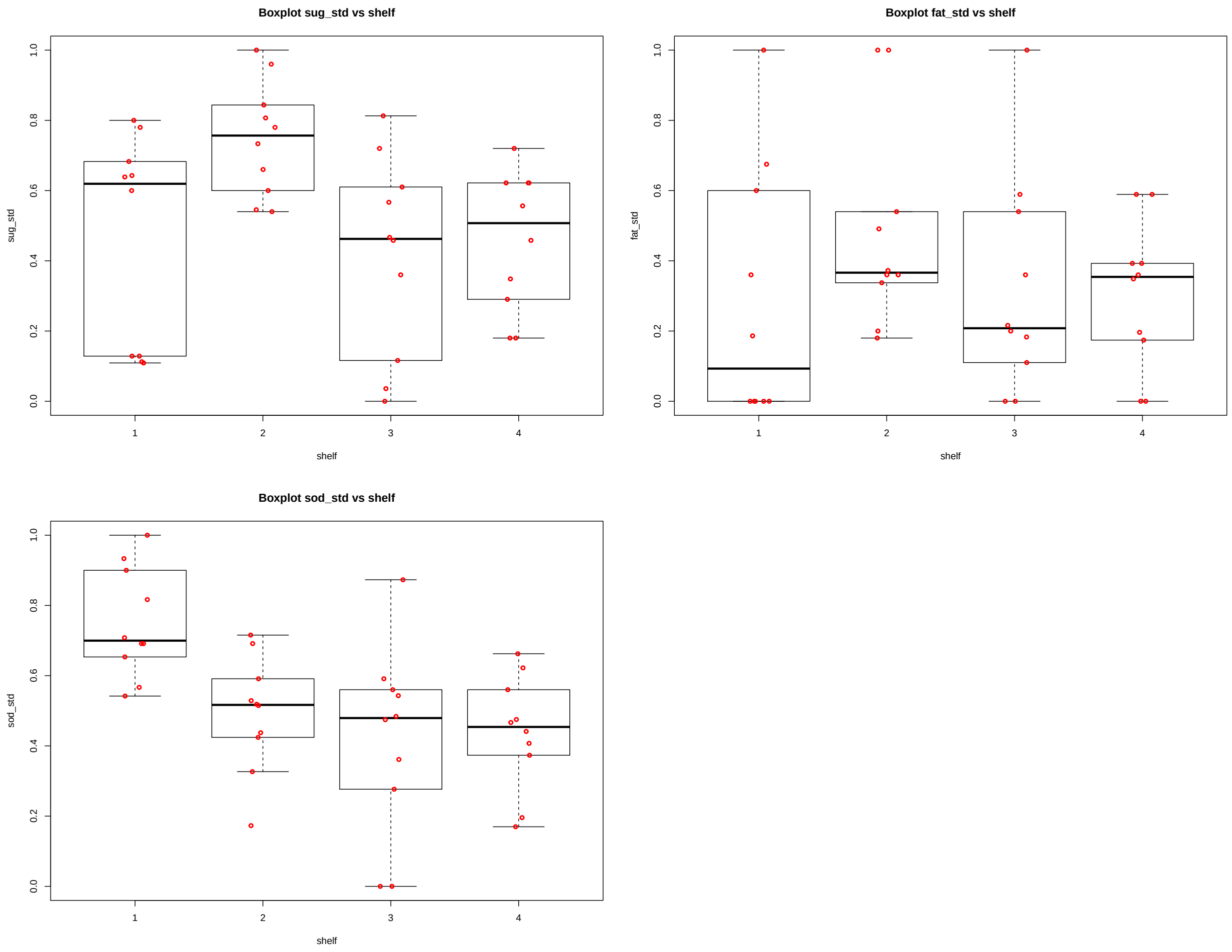
```
- Some sample code is provided
```

- Also, construct a **parallel coordinates plot** for the explanatory variables and the shelf r

Answer:

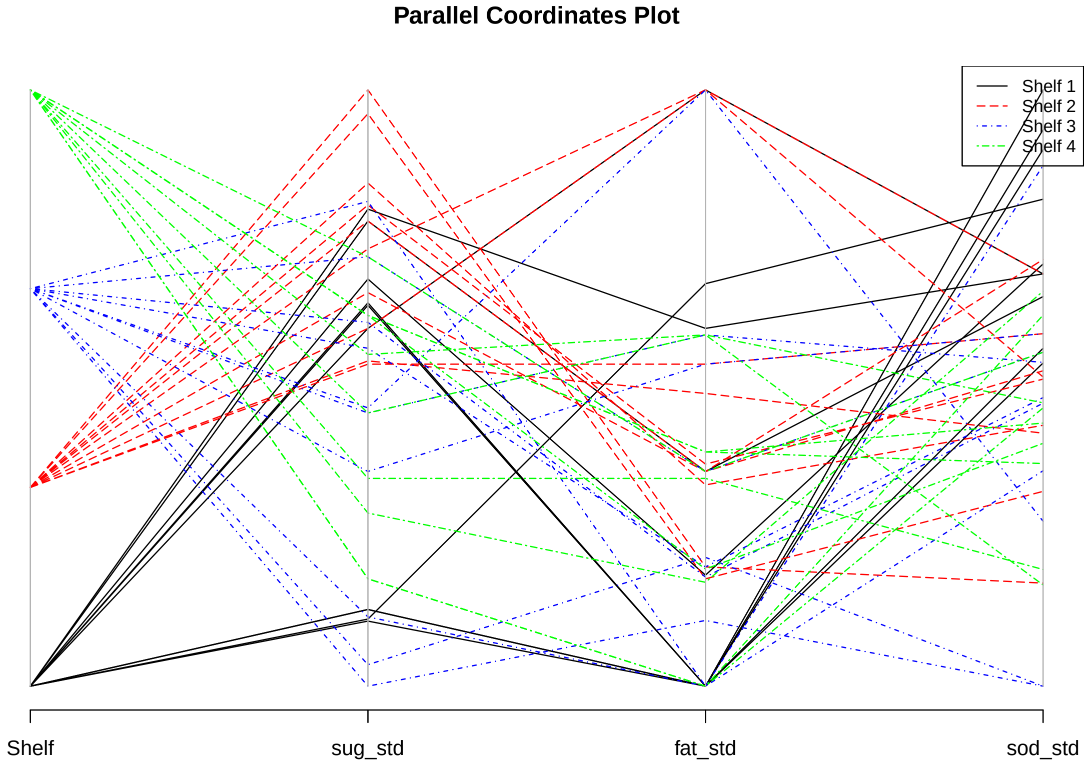
We construct three boxplots for the three explanatory variables `sugar`, `fat` and `sodium` using the rescaled data and the sample code provided.

```r
# Producing boxplots
par(mfrow = c(2, 2))

for (col in 1:3) {
    column = names(cereal2)[col + 1]
    boxplot(cereal2[, column] ~ cereal2$Shelf, ylab = column,
        xlab = "shelf", pars = list(outpch = NA), main = paste0("Boxplot ",
            column, " vs shelf"))
    stripchart(cereal2[, column] ~ cereal2$Shelf, lwd = 2, col = "red",
        method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
}
```



We now construct a parallel coordinates plot for the explanatory variables and the shelf number, using the code from the book. We interpret both plots together.

```r
# Colors by condition
cereal2.colors <- ifelse(test = cereals$Shelf == 1, yes = "black",
                         no = ifelse(test = cereals$Shelf == 2, yes = "red",

                                     no = ifelse(test = cereals$Shelf == 3,
                                                 yes = "blue",
                                                 no = "green")))

# Line type by condition
cereal2.lty <- ifelse(test = cereals$Shelf == 1, yes = "solid",
                          no = ifelse(test = cereals$Shelf == 2, yes = "longdash",
                          no = ifelse(test = cereals$Shelf == 3, yes = "dotdash",
                                      no = "twodash")))
# Create plot
parcoord(x = cereal2, col = cereal2.colors, lty = cereal2.lty,
         main = 'Parallel Coordinates Plot')
legend("topright", legend = c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
       lty = c("solid", "longdash", "dotdash", "twodash"),
       col = c("black", "red", "blue", "green"), cex = 0.8)
```



We note the following content differences between shelves: - Sugar: Shelf 2 has the highest median in sugar and also the smallest range between first and third quartile. The other shelves are quite similar. - Fat: Shelf 2 and shelf 4 have higher median fat values compared to 1 and 3. Shelf 2 also shows also outliers on the upper tail, while for shelf 4 the observations are more “normally” distributed. - Sodium: Shelf 1 has the highest median in sodium, while 2 to 4 seem to be quite similar.

There appear to be observable differences in values of the three variables based on the shelf level.

##c. The response has values of $1, 2, 3$, and 4. Under what setting would it be desirable to take into account ordinality. Do you think that this setting occurs here?

Answer:

Ordinality should be taken into account when there exists a natural ordering between the levels of the categorical variable.

This seems not be the case here, as a priori it is unknown to us which shelf position would be considered “better” compared to any of the others. Hence we think that the shelf position should be treated as a nominal categorical variable.

##d. Estimate a **multinomial regression model with linear forms of the sugar, fat, and sodium variables**. Perform **LRTs** to examine the importance of each explanatory variable.

Answer:

Estimating multinomial regression model:

```
# Transforming Shelf to factor
cereal2$Shelf <- factor(cereal2$Shelf, levels = c("1", "2", "3",
    "4"))


mod.fit <- multinom(Shelf ~ sug_std + fat_std + sod_std, 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(mod.fit)
```

```
## Call:
## multinom(formula = Shelf ~ sug_std + fat_std + sod_std, data = cereal2)
##
## Coefficients:
##   (Intercept)    sug_std    fat_std   sod_std
## 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)  sug_std  fat_std  sod_std
## 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
```

We use shelf 1 as baseline category.

Now we perform LRTs on each explanatory variable.

```r
Anova(mod.fit)
```

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

We find that `sugar` and `sodium` are statistically significant at the <0.1% significance level, while `fat` does not show significance even to the 10% level. Hence, the sugar and sodium content seem to impact the probability to be in a certain shelf.

##e. Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables).

Answer:

```r
# Model with up to three interactions

mod.fit2 <- multinom(Shelf ~ .^3, 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(mod.fit2)
```

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

```
Anova(mod.fit2)
```

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

We see from the LRT that no interaction between the three explanatory variables up to three interactions is significant. And the only varialbes that appear significant in the model with interactions are the same as those that appear significant in the original model without interactions (`sugar` and `sodium`).

We will next compare both models using the anova function. When comparing both models, we see that the model with the interaction terms is not significant.

```
anova(mod.fit, mod.fit2, test = "Chisq")
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: Shelf
##                            Model Resid. df Resid. Dev   Test    Df
## 1     sug_std + fat_std + sod_std       108   67.19028
## 2 (sug_std + fat_std + sod_std)^3        96   53.54496 1 vs 2    12
##   LR stat.   Pr(Chi)
## 1
## 2 13.64531 0.3239288
```

Hence, the model with no interactions would be preferred by us.

##f. 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.

Answer:

```
apj_size <- 28
apj_sug <- 22
apj_fat <- 0.5
apj_sod <- 130

# new standardization function, to take vector of previous
# standardized values
stand02 <- function(x, v, w) {
    (x - min(v/w))/(max(v/w) - min(v/w))
}

newdata <- data.frame(sug_std = stand02(apj_sug/apj_size, cereals$sugar_g,
    cereals$size_g), fat_std = stand02(apj_fat/apj_size, cereals$fat_g,
    cereals$size_g), sod_std = stand02(apj_sod/apj_size, cereals$sodium_mg,
    cereals$size_g))
newdata
```

```
##    sug_std   fat_std   sod_std
## 1 1.414286 0.1928571 0.4333333
```

We note that the new observation has a sugar value which, when standardized using the original data, exceeds the maximum of the original data by 41%. Hence, we get a standardized sugar value of 1.41. Since this will be an outlier in the "sugar dimension", it might have a strong impact on the predicted shelf via the originally fit regression model. We will note this as a key caveat in interpreting our probability estimates of Apple Jacks shelf placement.

```
pi.hat <- predict(object = mod.fit, newdata, type = "probs")
round(pi.hat, 5)
```

```
##       1       2       3       4
## 0.01959 0.98031 0.00003 0.00007
```

We note that shelf 2 is by far the most likely, with an estimated probability of shelf 2 at 98% . As we see from the parallel coordinates plot, this shelf is the one which is also strongly associated with the sugar level.

##g. 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.

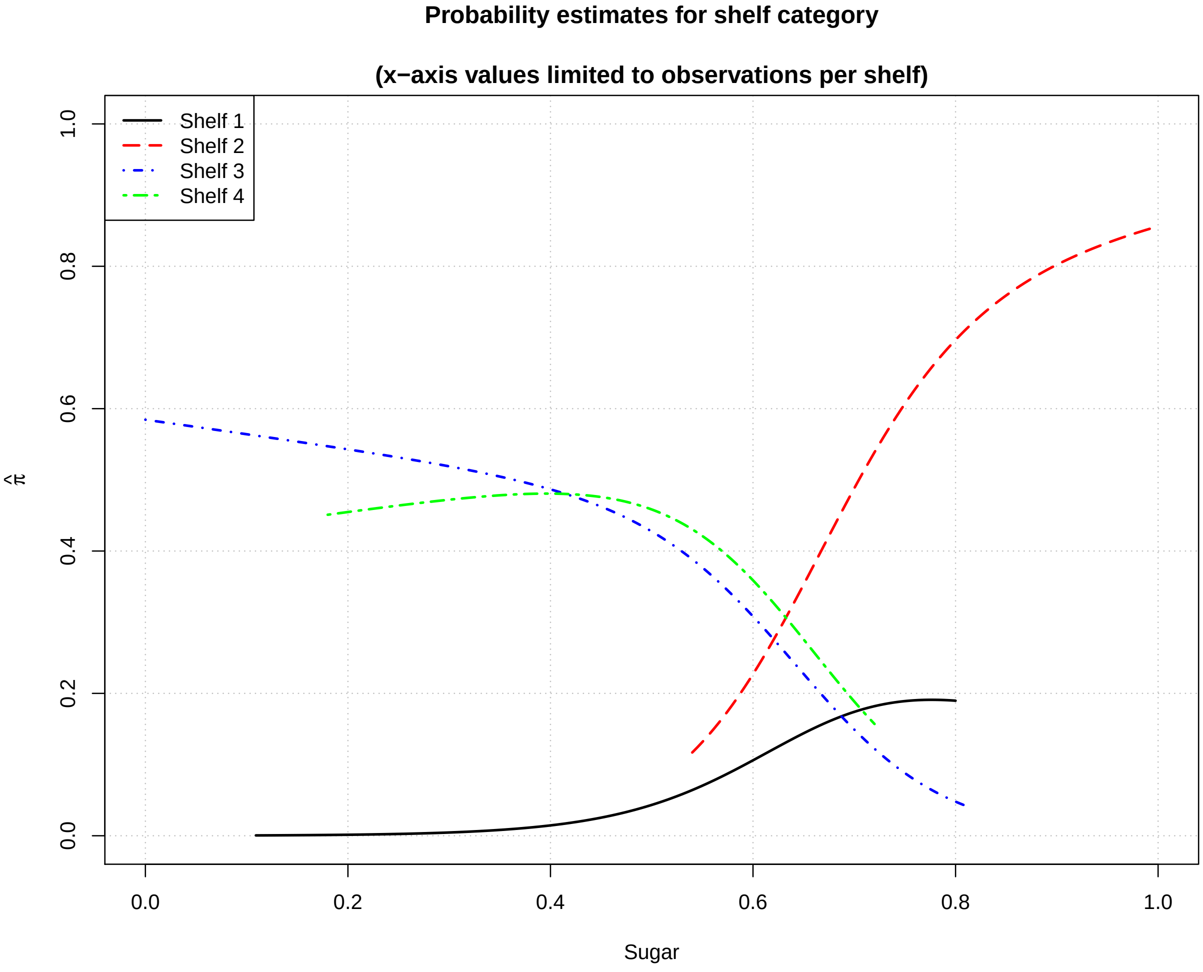
Answer:

We will first create the plot of estimated probability of shelf level vs sugar content (standardized), bounding our probability curves by the range of standardized sugar values by shelf in our sample data (later, we will look at an expanded range).

```
intercept <- coef(mod.fit)[, "(Intercept)"]
beta.sugar <- coef(mod.fit)[, "sug_std"]
beta.fat <- coef(mod.fit)[, "fat_std"]
beta.sodium <- coef(mod.fit)[, "sod_std"]

fat_mean <- mean(cereal2$fat_std)
sodium_mean <- mean(cereal2$sod_std)

intercept
```

```
##         2         3         4
##  6.900708 21.680680 21.288343
```

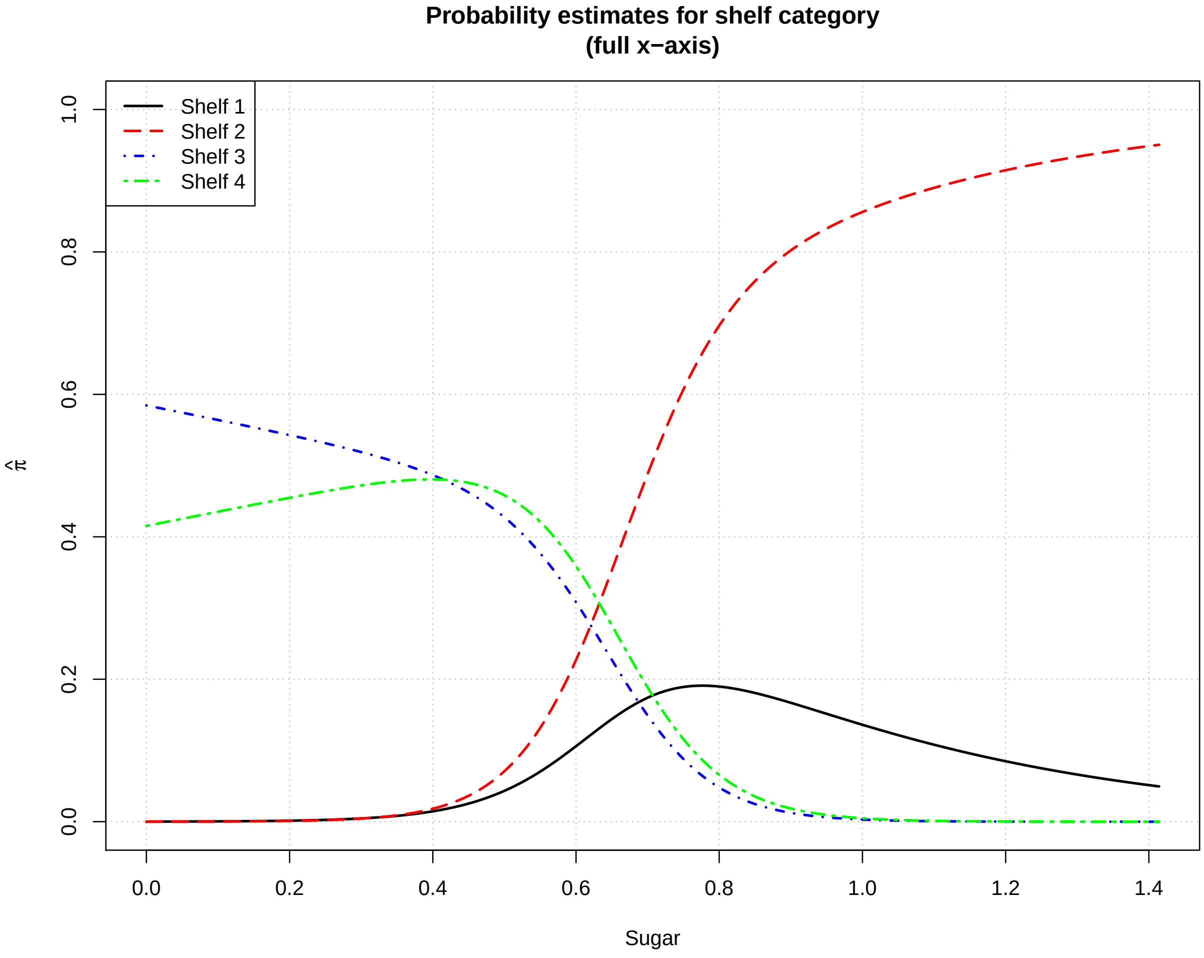
```
# Create plotting area first to make sure get the whole region with respect to x-axis

curve(expr = 1/(1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                          fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
      ylab = expression(hat(pi)), xlab = "Sugar", ylim = c(0, 1),
      xlim = c(min(cereal2$sug_std), max(cereal2$sug_std)), col = "black",

      lty = "solid", lwd = 2, n = 1000, type = "n",
      panel.first = grid(col = "gray", lty = "dotted"),
      main = 'Probability estimates for shelf category
      \n(x-axis values limited to observations per shelf)')

# Plot each pi_j
curve(expr = 1/(1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                          fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
      col = "black", lty = "solid", lwd = 2, n = 1000, add = TRUE,
      xlim = c(min(cereal2$sug_std[cereal2$Shelf == "1"]),
               max(cereal2$sug_std[cereal2$Shelf == "1"])))  # Shelf 1

curve(expr = exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean)/
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "red", lty = "longdash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(min(cereal2$sug_std[cereal2$Shelf == "2"]),
           max(cereal2$sug_std[cereal2$Shelf == "2"])))  # Shelf 2

curve(expr = exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                   fat_mean + beta.sodium[2] * sodium_mean) /
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "blue", lty = "dotdash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(min(cereal2$sug_std[cereal2$Shelf == "3"]),
           max(cereal2$sug_std[cereal2$Shelf == "3"])))  # Shelf 3

curve(expr = exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                   fat_mean + beta.sodium[3] * sodium_mean)/
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "green", lty = "twodash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(min(cereal2$sug_std[cereal2$Shelf == "4"]),
           max(cereal2$sug_std[cereal2$Shelf == "4"])))  # Shelf 4

legend('topleft', legend = c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
       lty=c("solid","longdash","dotdash", "twodash"),
       col=c("black","red", "blue", "green"), lwd = c(2,2,2,2))
```



We see that for a high sugar level it is likely that the cereal will be om shelf 2 or, much less likely, in shelf 1. For a low sugar level it is more likely to be in shelf 3 or 4. The probability estimates for shelf 3 and 4 are very similar to each other.

Now, we will plot the same probability curves with a lower bound of 0 and an upper bound of the Apple Jacks sugar level (transformed into the standardized variable).

```
# Create plotting area first to make sure get the whole region with respect to x-axis

curve(expr = 1/(1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                          fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
      ylab = expression(hat(pi)), xlab = "Sugar", ylim = c(0, 1),
      xlim = c(0,max(newdata$sug_std)), col = "black", lty = "solid", lwd = 2,
      n = 1000, type = "n", panel.first = grid(col = "gray", lty = "dotted"),
      main = 'Probability estimates for shelf category\n(full x-axis)')

# Plot each pi_j
curve(expr = 1/(1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                          fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "black", lty = "solid", lwd = 2, n = 1000, add = TRUE,
  xlim = c(0,max(newdata$sug_std)))

curve(expr = exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean)/
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "red", lty = "longdash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(0,max(newdata$sug_std)))

curve(expr = exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                   fat_mean + beta.sodium[2] * sodium_mean) /
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "blue", lty = "dotdash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(0,max(newdata$sug_std)))

curve(expr = exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                   fat_mean + beta.sodium[3] * sodium_mean)/
        (1 + exp(intercept[1] + beta.sugar[1]*x + beta.fat[1] *
                   fat_mean + beta.sodium[1] * sodium_mean) +
                  exp(intercept[2] + beta.sugar[2]*x + beta.fat[2] *
                        fat_mean + beta.sodium[2] * sodium_mean) +
                  exp(intercept[3] + beta.sugar[3]*x + beta.fat[3] *
                        fat_mean + beta.sodium[3] * sodium_mean)),
  col = "green", lty = "twodash", lwd = 2, n = 1000, add = TRUE,
  xlim = c(0,max(newdata$sug_std)))

legend('topleft', legend = c("Shelf 1", "Shelf 2", "Shelf 3", "Shelf 4"),
       lty=c("solid","longdash","dotdash", "twodash"),
       col=c("black","red", "blue", "green"), lwd = c(2,2,2,2))
```



Here we see the probability distribution over the whole range from 0 to the standardized sugar content of the new observation for Kellogs Apple Jacks (which is at the far right of the graph). The graph supports the results of the model that based on standardized sugar levels shelf 2 would be the shelf with the by far highest probability associated with it for the Kellogs Apple Jacks cereals.

```
# mean fat and sodium, maximum standarized sugar for shelf 2
# shelf 2
exp(intercept[1] + beta.sugar[1] * 1 + beta.fat[1] * fat_mean +
    beta.sodium[1] * sodium_mean)/(1 + exp(intercept[1] + beta.sugar[1] *
    1 + beta.fat[1] * fat_mean + beta.sodium[1] * sodium_mean) +
    exp(intercept[2] + beta.sugar[2] * 1 + beta.fat[2] * fat_mean +
        beta.sodium[2] * sodium_mean) + exp(intercept[3] + beta.sugar[3] *
    1 + beta.fat[3] * fat_mean + beta.sodium[3] * sodium_mean))
```

```
##         2
## 0.8561038
```

```
# mean fat and sodium, standardized Apple Jacks sugar for
# shelf 2 shelf 2
exp(intercept[1] + beta.sugar[1] * newdata$sug_std + beta.fat[1] *
    fat_mean + beta.sodium[1] * sodium_mean)/(1 + exp(intercept[1] +
    beta.sugar[1] * newdata$sug_std + beta.fat[1] * fat_mean +
    beta.sodium[1] * sodium_mean) + exp(intercept[2] + beta.sugar[2] *
    newdata$sug_std + beta.fat[2] * fat_mean + beta.sodium[2] *
    sodium_mean) + exp(intercept[3] + beta.sugar[3] * newdata$sug_std +
    beta.fat[3] * fat_mean + beta.sodium[3] * sodium_mean))
```

```
##         2
## 0.9504717
```

We note that at the maximum sample standardized sugar content and mean fat and sodium content, the maximum probability of a shelf level based on sugar content is ~85% (for shelf 2, at standarized sugar = 1). Extending the model to include Apple Jacks sugar content at the mean fat and sodium content pushes this probability to 95%; however, since the model was not trained on cereals with sugar content as high as Apple Jacks, we are assuming that the modeled relationships hold for values outside of sample data values. Unless we train the model on additional data that inclues sugar levels as high as Apple Jacks, we must caveat that we are extending the model to a range outside of observed sample.

##h. Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretations back to the plots constructed for this exercise.

Answer:

We are using the Wald CI as in the book.

```
conf.beta <- confint(mod.fit, level = 0.95)
conf.beta
```

```
## , , 2
##
##                  2.5 %    97.5 %
## (Intercept)  -5.814378 19.615793
## sug_std      -7.208058 12.594200
## fat_std      -0.457418  8.586836
## sod_std     -31.403783 -3.583672
##
## , , 3
##
##                  2.5 %    97.5 %
## (Intercept)   7.077214 36.284146
## sug_std     -21.796655 -2.636228
## fat_std      -5.290368  4.176114
## sod_std     -40.815515 -9.141475
##
## , , 4
##
```

```
##                   2.5 %    97.5 %
## (Intercept)   6.715766 35.860921
## sug_std     -20.941357 -1.846063
## fat_std      -5.585224  3.844988
## sod_std     -40.475656 -8.872038
```

```
sd.cereals2 <- apply(cereal2[, -1], 2, sd)  # 1 sd increase of normalized values
c.value <- c(sd.cereals2)
c.value
```

```
##   sug_std   fat_std   sod_std
## 0.2692078 0.2990292 0.2298359
```

```
beta.hat2 <- coefficients(mod.fit)[1, 2:4]
beta.hat3 <- coefficients(mod.fit)[2, 2:4]
beta.hat4 <- coefficients(mod.fit)[3, 2:4]

OR2 <- exp(c.value * beta.hat2)
OR3 <- exp(c.value * beta.hat3)
OR4 <- exp(c.value * beta.hat4)
ci.OR2 <- exp(c.value * conf.beta[2:4, 1:2, 1])
ci.OR3 <- exp(c.value * conf.beta[2:4, 1:2, 2])
ci.OR4 <- exp(c.value * conf.beta[2:4, 1:2, 3])

# Odds of 2 instead of 1
round(data.frame(low = 1/ci.OR2[, 2], mean = 1/OR2, up = 1/ci.OR2[,
    1]), 2)
```

```
##          low  mean      up
## sug_std 0.03  0.48    6.96
## fat_std 0.08  0.30    1.15
## sod_std 2.28 55.74 1363.37
```

```
# Odds of 3 instead of 1
round(data.frame(low = 1/ci.OR3[, 2], mean = 1/OR3, up = 1/ci.OR3[,
    1]), 2)
```

```
##          low   mean       up
## sug_std 2.03  26.81   353.48
## fat_std 0.29   1.18     4.86
## sod_std 8.17 311.36 11859.32
```

```
# Odds of 4 instead of 1
round(data.frame(low = 1/ci.OR4[, 2], mean = 1/OR4, up = 1/ci.OR4[,
    1]), 2)
```

```
##          low   mean       up
## sug_std 1.64  21.48   280.78
## fat_std 0.32   1.30     5.31
## sod_std 7.68 290.31 10968.22
```

Calculating the odds ratios, we can see that (holding all other variables constant):

- Increasing sugar by 1 standard deviation, the odds of being in shelf 1 relative to shelf 2 is decreased by roughly half, while the odds of being in shelf 1 relative to 3 and 4 increases by roughly 25 and 20 times, respectively.
- Increasing fat by 1 standard deviation, the odds of being in shelf 1 relative to shelf 2 is decreased by roughly two-thirds, while the odds of being in shelf 1 relative to 3 and 4 is increased by about 18 and 30%, respectively.
- Increasing sodium by 1 standard deviation, the odds of being in shelf 1 relative to shelf 2, 3 and 4 increases by 54, 310, and 289 times, respectively.

We note that increases in one standard deviation for sodium appears to have a much greater impact on the odds of being in shelves 2, 3, and 4 (vs shelf 1) than one standard deviation increases for sugar or fat. A one sd increase in sugar also has a large impact on odds of being in shelves 3 and 4 versus 1. Increasing fat by one sd has less of a relative impact on these odds ratios.

For sodium, the results are expected, as indicated by the boxplots with group 1 having a higher median value than all other shelves, with the inner-quartile range nearly not overlapping with any shelves. Based on medians for sugar, we can see from the boxplots that our calculated odds ratios are consistent with our expectations. For fat, we are surprised when comparing the mean odds ratio to the boxplots for shelves 3 and 4, but the confidence intervals are consistent with our expectations given how wide the inner-quartile ranges are.

Conclusion

The question we investigated in this report concerns the probaility of shelf placement for cereals given its sugar, fat, and sodium content; namely, can observable factors such as sugar, fat, and sodium help predict the probability shelf placement of cereals in grocery stores. Our question and analysis is based on random samples of 10 cereals from each of four shelf levels from Dillons grocery store in Manhattan, KS to inform this question. This analysis is important to show how grocery stores attempt to attract customers to certain cereals through shelf placement decisions. Grocery store managers might use this analysis in order to inform their own shelf placement decisions, especially if the Dillons that generated the source data was found to be successful in attracting customers using its shelf placement of cereals.

After examining the data and standardizing to create an interpretable analysis, we evaluated candidate model specifications to predict shelf level from sugar, fat, and sodium content. We noted that shelf level should be modeled as nominal instead of ordinal since a priori we do not know what the natural ordering from ‘worst’ to ‘best’ is for shelf levels. We used a nominal multinomial logistic regression model in order to model log odds of base shelf level (chosen as shelf 1) versus other shelves (2, 3, and 4). We also considered a model with interacton effects but ultimately discarded this model. The main reasons we discarded this model were that no interaction terms were determined to be signficant (and the same terms - sugar and sodium - appeared as significant in models with and witout interaction terms), and we failed to reject the null hypothesis that the model without interactions performed better than the model with interactions. These were the main factors that we chose a nominal multinomial logistic regression model with standardized sugar, fat, and sodium explanatory varialbes as our reference model in this analysis.

The main result of our analysis is that probability of cereal shelf placement is related to its sugar, fat, and sodium content. High sugar levels are associated with shelf 2 placement, while lower sugar levels are associated with shelves 3 and 4. High fat content is associated with shelves 3 and 4, while low fat content is associated with shelf 2. High sodium content is associated with shelves 3 and 4, while low sodium content is associated with shelf 1.

The sensitivity of odd ratios on a one standard deviation increase in explanatory variables is very different based on the variable. Sodium shows high sensitivity, where the odds of being in shelf 1 vs shelf 3 increase by over 300x by a one sd increase in sodium. Sugar is also sensitive, where the odds of being in shelf 1 vs shelf 4 increase by 25x by a one sd increase in sugar. The one sd increase in sugar is also associated with a 50% reduction in odds of being in shelf 1 vs shelf 2. Fat is least sensitive, where the odds of being in shelf 1 vs shelf 4 increase by 1.3x with a one sd increase in fat. The one sd increase in fat is also associated with a ~2/3 reduction in odds of being in shelf 1 vs shelf 2. The higher sensitivity of odds ratios to sodium and sugar content vs fat is in line with our observation that the parameter estimates of these variables are highly significant, whereas the fat parameter estimate is not significant at alpha = 0.05 level.

In summary, this analysis helps provide a statistical, data-driven framework for understanding attributes of cereals in relation to its shelf placement in grocery stores. It quantifies the probabilities of shelf placement based on sugar, fat and sodum content of a particular cereal, and adds value for grocery store managers looking to make shelf placement decisions that mirror decisions of other grocery stores (using Dillons in Manhattan, KS as a proxy). Based on our findings, we find that a cereal with very high sugar content (and fat and sodium content within observed range), such as Apple Jacks, has a very high probability of being placed on shelf 2. This analysis illuminates that shelf placement decisions by grocery store managers can be informed by sugar, fat, and sodium content of cereals. And if the sample data represents best practices, the model can help grocery stores execute on their goals of better attracting customers to cereals through use of shelf placement.