w271 Lab 2: Cereal Shelf Placement

Jessica Hays Fisher, Alice Lam, Marshall Ratliff, Paul Varjan

6/3/2018

Contents


```r
#Load libraries and insert a function to tidy up the code when they are printed out
library(vcd, quietly=T)
library(nnet, quietly=T)
library(car, quietly=T)
library(Hmisc, quietly=T)
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(skimr, quietly=T)
library(MASS, quietly=T)

rm(list = ls())
library(knitr, quietly=T)
```

```
##
## Attaching package: 'knitr'

## The following object is masked from 'package:skimr':
##
##     kable
```

```
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

```
cereal <- read.csv("cereal_dillons.csv")
str(cereal)
```

```
## '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 ...
```

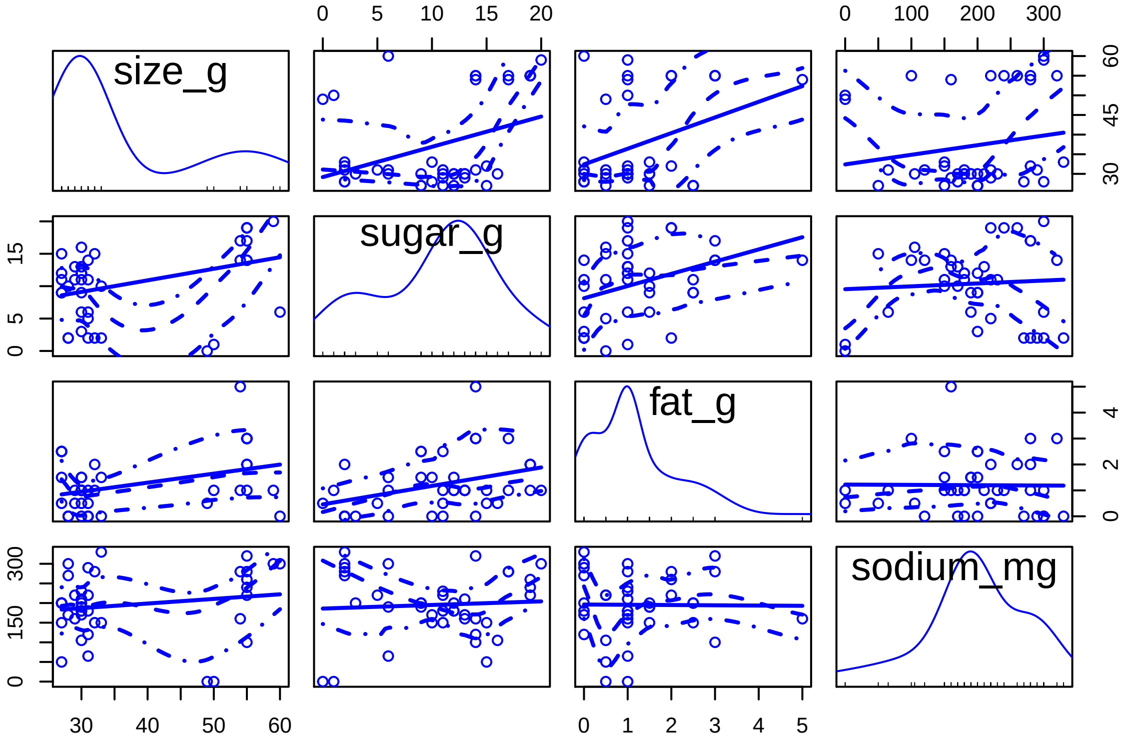
```
# Examine the data to check data validity before proceeding
# with the questions.
skim(cereal)
```

```
## Skim summary statistics
##  n obs: 40
##  n variables: 7
##
## -- Variable type:factor ------
##  variable missing complete  n n_unique                     top_counts
##    Cereal       0       40 40       38 Cap: 2, Foo: 2, Bas: 1, Cap: 1
##  ordered
##    FALSE
##
## -- Variable type:integer -----
##   variable missing complete  n  mean    sd p0    p25   p50    p75 p100
##         ID       0       40 40  20.5 11.69  1  10.75  20.5  30.25   40
##      Shelf       0       40 40   2.5  1.13  1   1.75   2.5   3.25    4
##     size_g       0       40 40  37.2 11.79 27  29.75  31     51      60
##  sodium_mg       0       40 40 195.5 81.67  0 157.5  200    262.5   330
##    sugar_g       0       40 40  10.4  5.67  0   6     11     14      20
##      hist
##
##
##
##
##
##
## -- Variable type:numeric -----
```

```
##   variable missing complete  n mean  sd p0 p25 p50  p75 p100     hist
##      fat_g       0       40 40  1.2 1.1  0 0.5   1 1.62    5
```

There are 7 variables with 40 observations evenly distributed across 4 shelves. There's no missing data. There are 38 types of cereal, with sugar content ranging 0 to 20 gram, fat content ranging from 0 to 5 gram, sodium content from 0 to 330 gram, serving size ranging 27 to 60 gram.

```r
# suppress warnings
oldw <- getOption("warn")
options(warn = -1)
scatterplotMatrix(~size_g + sugar_g + fat_g + sodium_mg, data = cereal)
```




```r
# restore old warning level
options(warn = oldw)
```

There is not much clear relationship between any of the variables in the scatterplot matrix, with fairly horizontal lines between more pairs indicating a general lack of correlation. The matrix together with a lack of strong supporting intuition suggest that there is no clear value of interaction terms between the explanatory variables.

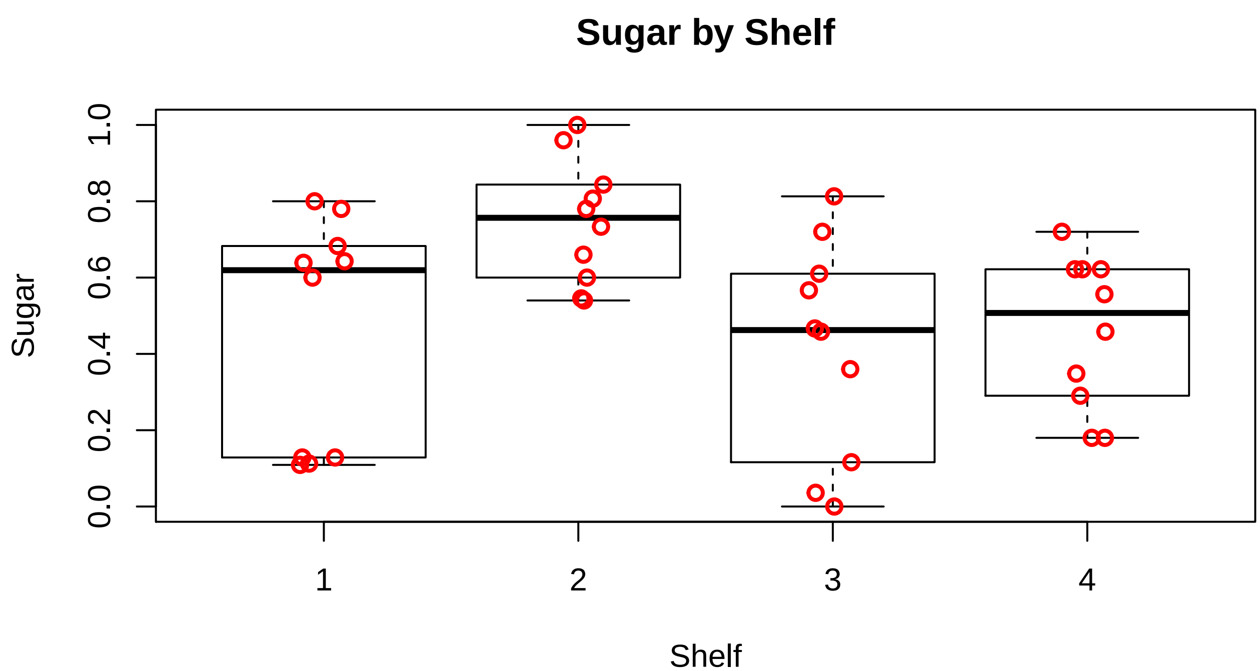
Part (b) involves additional EDA, so we save the rest of our EDA for this section.

(a) The explanatory variables need to be re-formatted before proceeding further. First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Second, re-scale each variable to be within 0 and 1.

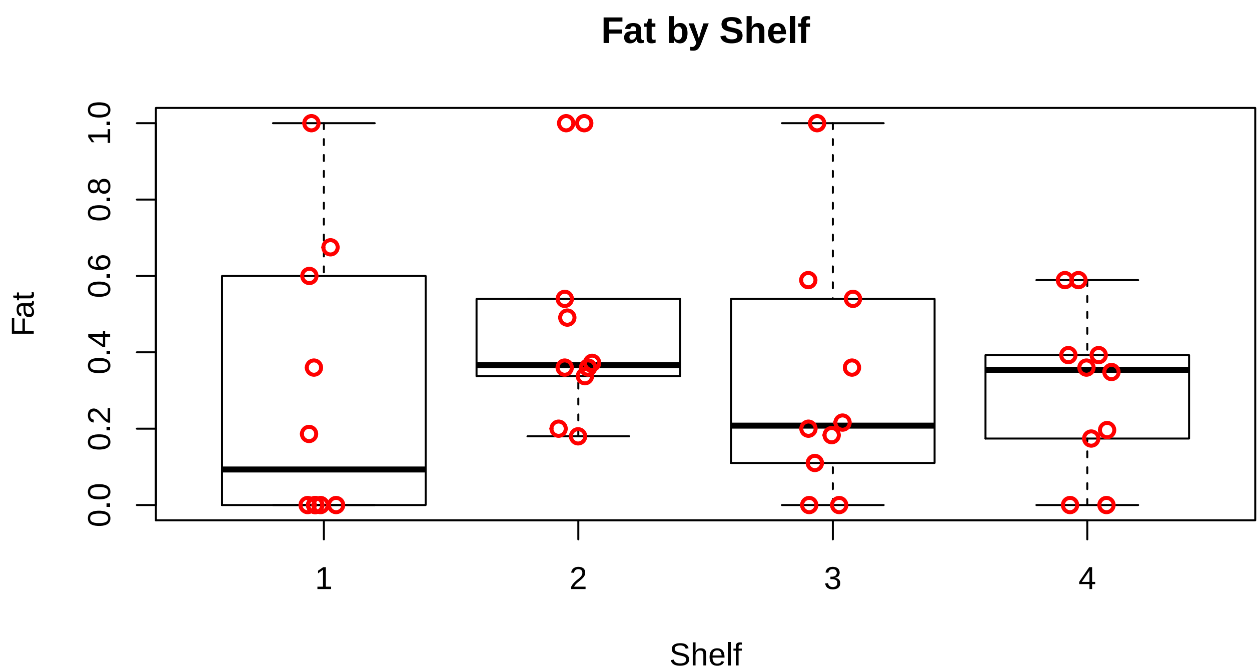
```
standardize <- function(x) {
    (x - min(x))/(max(x) - min(x))
}
cereal2 <- data.frame(Shelf = cereal$Shelf, Cereal = cereal$Cereal,
    sugar = standardize(cereal$sugar_g/cereal$size_g), fat = standardize(cereal$fat_g/cereal$si
    sodium = standardize(cereal$sodium/cereal$size_g))
```

(b) 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 if possible content differences exist among the shelves.

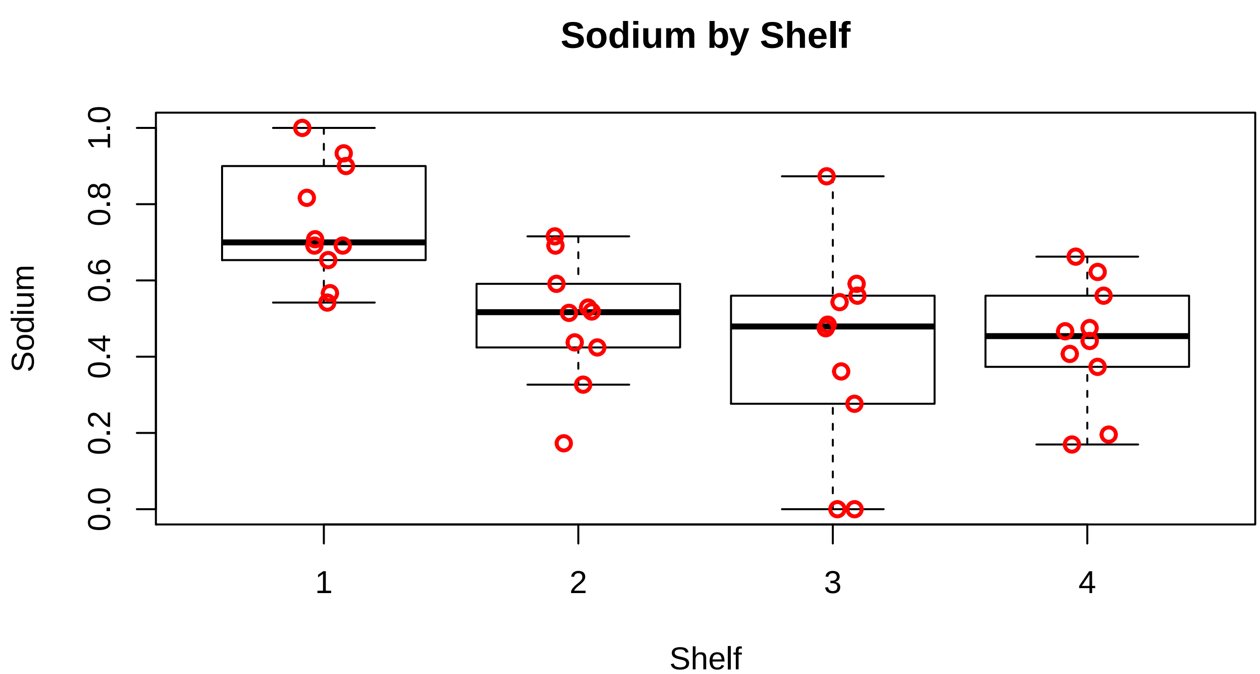
```
boxplot(formula = sugar ~ Shelf, data = cereal2, ylab = "Sugar",
    xlab = "Shelf", main = "Sugar by Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sugar ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```




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

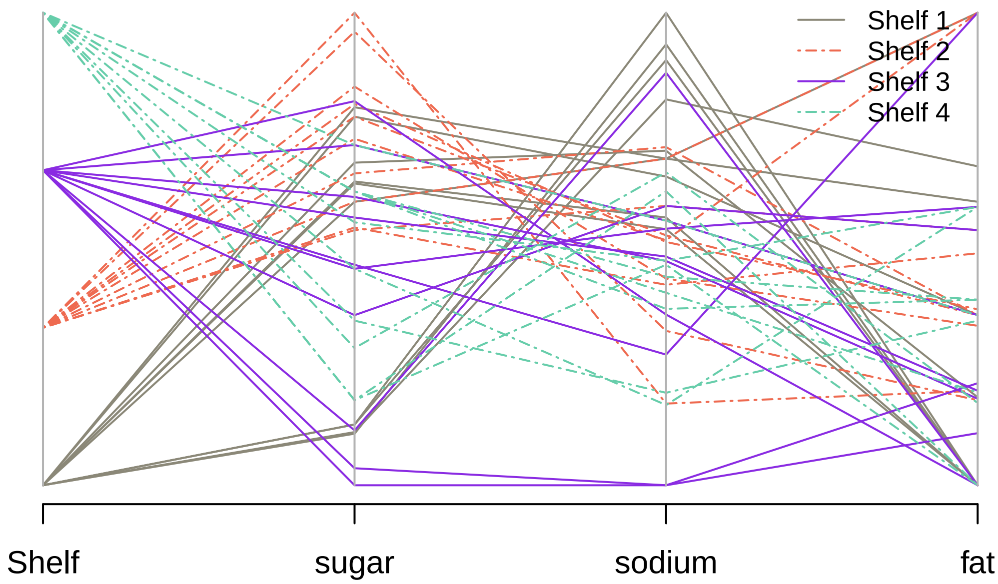



```
boxplot(formula = sodium ~ Shelf, data = cereal2, ylab = "Sodium",
    xlab = "Shelf", main = "Sodium by Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sodium ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```




```
cereal3 <- data.frame(cereal2[1], cereal2[3], cereal2[5], cereal2[4])

# Colors by condition:
cereal.colors <- ifelse(test = cereal2$Shelf == 1, yes = "cornsilk4",
    no = ifelse(test = cereal2$Shelf == 2, yes = "coral2", no = ifelse(test = cereal2$Shelf ==
```

```
        3, yes = "blueviolet", no = "aquamarine3")))
# Line type by condition:
cereal.lty <- ifelse(test = cereal2$Shelf == 1, yes = "solid",
    no = ifelse(test = cereal2$Shelf == 2, yes = "twodash", no = ifelse(test = cereal2$Shelf ==
        3, yes = "solid", no = "twodash")))

parcoord(x = cereal3, col = cereal.colors, lty = cereal.lty)  # Plot
legend(x = 3.35, y = 1.05, legend = c("Shelf 1", "Shelf 2", "Shelf 3",
    "Shelf 4"), lty = c("solid", "twodash", "solid", "twodash"),
    col = c("cornsilk4", "coral2", "blueviolet", "aquamarine3"),
    cex = 0.8, bty = "n")
```



High sugar content seems to be most prevalent among Shelf 2. In addition, the cereals with lowest sugar content on shelf 2 had elevated fat and sodium content comparatively, indicating there is still an inflated flavor profile corresponding to likely less healthy but popular cereals.

The other shelves have a pretty wide spread of sugar content, with means roughly in the same places. Of note is Shelf 1's bimodal distribution of sugar, with one cluster of cereals with nearly no sugar and the other cluster having above average sugar content. Without that low sugar cluster, the rest of the shelf would have a mean sugar content much closer to Shelf 2.

Fat content seems to be pretty evenly distributed across shelves. In cereal this most likely corresponds to contents like nuts and oilseeds. There is a heavy occurrence of fat content at both extremes (1 and 0). Shelf 1 has so many 0 fat score cereals that its mean is lower than the others. Perhaps also notable is that shelf 2 is the only shelf with no cereals with a 0 score for fat, and that shelf 4 is the only shelf with no cereals with a 1 score, but visually, that information does not add muchin light of the rest of the fat content plots.

Sodium content is notably highest on Shelf 1, but otherwise the other shelves have a more or less similar mean, with Shelf 3 showing the most breadth of sodium levels within that shelf.

(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 this occurs here?

Answer: If we believed that there was a natural ordering to the shelves, or that they could be arranged in an order such that shelf 1 $<$ shelf 2 $<$ shelf 3, etc. - then it would be desirable to take into account ordinality (especially if we believed that the "distance" between each level was constant). However, we do not believe that is the case with this data, as it is not clear whether being on a low shelf is objectively better than on a high shelf, or vice versa. There are attractors/detractors from each shelf height and for different customers - for example, children are at the height of lower shelves than adults are - but that ordering is not universal and therefore not desireable to take into account in our modeling. If other data could be brought in that demonstrated the desirability or marketability of each shelf had some order (which probably does exist), that could also be used as a factor for ordinality.

For example, the most significant factors for shelf ordering are probably target audience of the product and some metric of difficulty/ease of reaching a given shelf. If we had stats on shelf heights and arm lengths of target customer groups it may be possible to rank the shelves in a meaningful way. It seems likely that shelves 2 or 3 is highest priority for most products, however given the sensitivity to children for this product segment clearly, shelf 2 is highest priority, shelf 4 is lowest, and it is probably difficult to distinguish shelves 1 and 3 but perhaps 1 continues to cater more toward children and 3 more toward adults. This is reflected by the clustering of sugary cereals on shelves 1 & 2. In any case, there is still no immediately apparent ordinality aside from a clear distinction between shelf 2 and shelf 4, thus it seems inappropriate to take into accound ordinality in this example.

(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.

```r
levels(as.factor(cereal2$Shelf))
```

```
## [1] "1" "2" "3" "4"
```

```r
mod1 <- multinom(as.factor(Shelf) ~ sugar + fat + sodium, data = cereal2)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 37.329384
## iter  20 value 33.775257
## iter  30 value 33.608495
## iter  40 value 33.596631
## iter  50 value 33.595909
## iter  60 value 33.595564
## iter  70 value 33.595277
## iter  80 value 33.595147
## final  value 33.595139
## converged
```

```r
summary(mod1)
```

```
## Call:
## multinom(formula = as.factor(Shelf) ~ sugar + fat + sodium, data = cereal2)
##
## Coefficients:
##   (Intercept)      sugar        fat    sodium
## 2    6.900708   2.693071  4.0647092 -17.49373
## 3   21.680680 -12.216442 -0.5571273 -24.97850
## 4   21.288343 -11.393710 -0.8701180 -24.67385
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 2    6.487408 5.051689 2.307250 7.097098
## 3    7.450885 4.887954 2.414963 8.080261
## 4    7.435125 4.871338 2.405710 8.062295
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

```r
Anova(mod1)
```

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

- We cannot use mcprofile package for likelihood ratio, as nnet package author does not believe that one at a time intervals should be calculated. We use value of c equal to 1 standard deviation instead. *

In line with the EDA above, we see that sugar and sodium are the key discriminating factors both in terms of likelihood ratios and statistical significance. Specifically, we see that increases in sodium levels correspond to decreased likelihood of all other shelves compared to the base-case of shelf 1, and increases in sugar levels correspond to decreased likelihood of shelves 3 & 4 compared to the base-case of shelf 1, but conversely a somewhat increased likelihood of shelf 2 relative to the base-case of shelf 1.

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

```r
## Create expanded models including interaction terms between
## all pairs of explanatory variables and triple interaction
```

```r
## between all 3.

modA <- multinom(as.factor(Shelf) ~ sugar + fat + sodium + sugar:fat,
    data = cereal2)
```

```
## # weights:  24 (15 variable)
## initial  value 55.451774
## iter  10 value 37.518284
## iter  20 value 33.008954
## iter  30 value 31.923790
## iter  40 value 31.590382
## iter  50 value 31.417819
## iter  60 value 31.212591
## iter  70 value 31.158365
## iter  80 value 31.068476
## iter  90 value 31.042861
## iter 100 value 30.998934
## final  value 30.998934
## stopped after 100 iterations
```

```r
Anova(modA)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##           LR Chisq Df Pr(>Chisq)
## sugar      22.7648  3  4.521e-05 ***
## fat         5.2836  3     0.1522
## sodium     30.8407  3  9.183e-07 ***
## sugar:fat   5.1924  3     0.1582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
modB <- multinom(as.factor(Shelf) ~ sugar + fat + sodium + sugar:sodium,
    data = cereal2)
```

```
## # weights:  24 (15 variable)
## initial  value 55.451774
## iter  10 value 36.577939
## iter  20 value 33.026993
## iter  30 value 32.740384
## iter  40 value 32.604061
## iter  50 value 32.452790
## iter  60 value 32.427677
## iter  70 value 32.423013
## iter  80 value 32.420834
## iter  90 value 32.420382
## iter 100 value 32.420219
## final  value 32.420219
```

```
## stopped after 100 iterations
```

```r
Anova(modB)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##              LR Chisq Df Pr(>Chisq)
## sugar         22.7648  3  4.521e-05 ***
## fat            6.1167  3     0.1061
## sodium        26.6197  3  7.073e-06 ***
## sugar:sodium   2.3498  3     0.5030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
modC <- multinom(as.factor(Shelf) ~ sugar + fat + sodium + fat:sodium,
    data = cereal2)
```

```
## # weights:  24 (15 variable)
## initial  value 55.451774
## iter  10 value 36.936754
## iter  20 value 32.546388
## iter  30 value 32.211409
## iter  40 value 32.085217
## iter  50 value 31.709427
## iter  60 value 31.095835
## iter  70 value 30.872944
## iter  80 value 30.734369
## iter  90 value 30.653914
## iter 100 value 30.639405
## final  value 30.639405
## stopped after 100 iterations
```

```r
Anova(modC)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##            LR Chisq Df Pr(>Chisq)
## sugar       19.2525  3  0.0002424 ***
## fat          5.2836  3  0.1521727
## sodium      26.6197  3  7.073e-06 ***
## fat:sodium   5.9115  3  0.1159978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
modD <- multinom(as.factor(Shelf) ~ sugar + fat + sodium + fat:sodium:sugar,
    data = cereal2)
```

```
## # weights:  24 (15 variable)
## initial  value 55.451774
```

```
## iter  10 value 37.410894
## iter  20 value 33.855213
## iter  30 value 33.056977
## iter  40 value 32.715781
## iter  50 value 32.677949
## iter  60 value 32.592524
## iter  70 value 32.578128
## iter  80 value 32.550159
## iter  90 value 32.541053
## iter 100 value 32.535394
## final  value 32.535394
## stopped after 100 iterations
```

```r
Anova(modD)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##                  LR Chisq Df Pr(>Chisq)
## sugar             22.7648  3  4.521e-05 ***
## fat                5.2836  3     0.1522
## sodium            26.6197  3  7.073e-06 ***
## sugar:fat:sodium   2.1195  3     0.5480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
modE <- multinom(as.factor(Shelf) ~ sugar + fat + sodium + sugar:fat +
    sugar:sodium + fat:sodium + fat:sodium:sugar, 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
```

```r
Anova(modE)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##                  LR Chisq Df Pr(>Chisq)
```

```
## sugar              19.2525  3  0.0002424 ***
## fat                 6.1167  3  0.1060686
## sodium             30.8407  3  9.183e-07 ***
## sugar:fat           3.2309  3  0.3573733
## sugar:sodium        3.0185  3  0.3887844
## fat:sodium          3.1586  3  0.3678151
## sugar:fat:sodium    2.5884  3  0.4595299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the significance values of the likelihood ratio tests for each additional interaction term none of them are even marginally significant, and the individual term significances remain fairly stable in significance levels. As such we can reject incorporating any of these interaction terms in our final model.

(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.

```
stand.new <- function(meas, serv_size, comparison) {
    (meas/serv_size - min(comparison))/(max(comparison) - min(comparison))
}

newdata <- data.frame(sugar = stand.new(12, 28, cereal$sugar_g/cereal$size_g),
    fat = stand.new(0.5, 28, cereal$fat_g/cereal$size_g), sodium = stand.new(130,
        28, cereal$sodium_mg/cereal$size_g))

round(predict(object = mod1, newdata = newdata, type = "probs",
    se.fit = TRUE), 7)
```

```
##         1         2         3         4
## 0.0532685 0.4719426 0.2004274 0.2743615
```

From the above prediction, we see that Kellogg's Apple Jacks are most likely to be placed on shelf 2, given a relatively elevated level of sugar and a sodium level that falls in the first quantile making shelf 1 fairly unlikely. While it fits squarely in the range of sugary cereals, it does not have an extreme sugar level, so shelves 3 & 4 are still somewhat possible. As mentioned previously in part (c), sugary cereals are likely to be placed on shelf 2 for close proximity to children's eye-level, which would presumably aid in sales for Kellogg's Apple Jacks.

(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.

```
beta.hat <- coefficients(mod1)
beta.hat
```

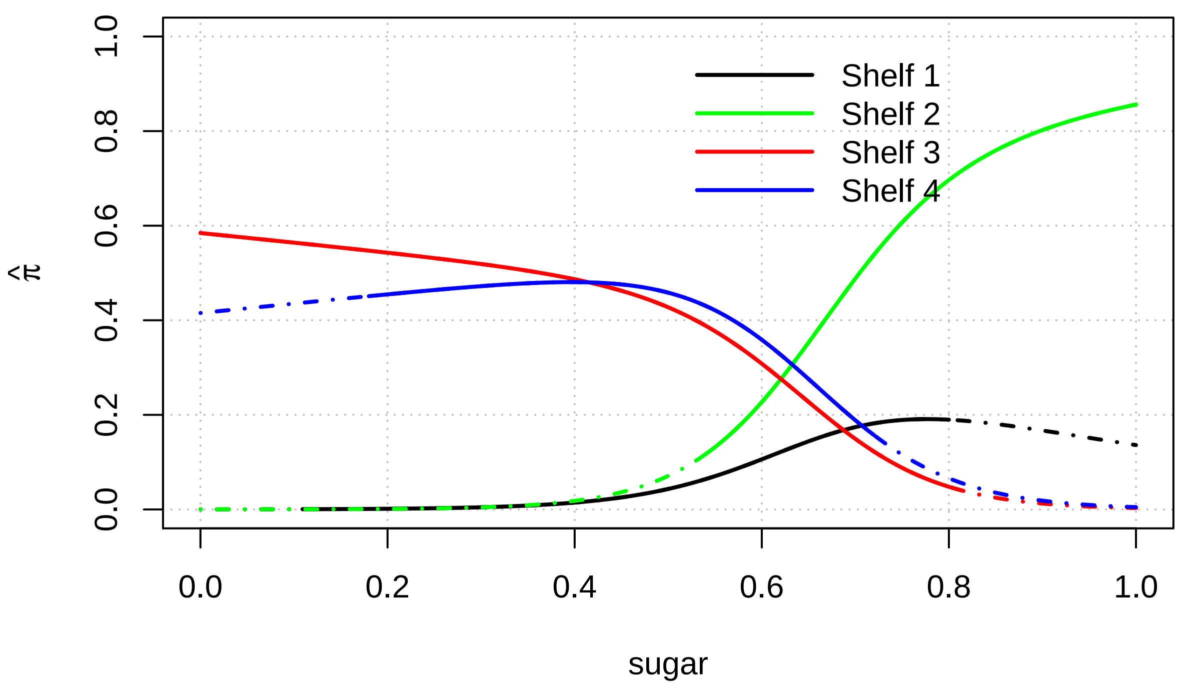
```
##   (Intercept)      sugar        fat    sodium
## 2    6.900708   2.693071  4.0647092 -17.49373
## 3   21.680680 -12.216442 -0.5571273 -24.97850
## 4   21.288343 -11.393710 -0.8701180 -24.67385
```

```
mean_fat <- mean(cereal2$fat)
mean_sodium <- mean(cereal2$sodium)


# Create plotting area first to make sure get the whole
# region with respect to x-axis
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 2] * x) +
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x)), ylab = expression(hat(pi)),
    xlab = "sugar", xlim = c(min(cereal2$sugar), max(cereal2$sugar)),
    ylim = c(0, 1), col = "black", lty = "solid", lwd = 2, n = 1000,
    type = "n", panel.first = grid(col = "gray", lty = "dotted"))

## Plot each pi_j Shelf1
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 2] * x +
    beta.hat[1, 3] * mean_fat + beta.hat[1, 4] * mean_sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x + beta.hat[2, 3] *
        mean_fat + beta.hat[2, 4] * mean_sodium) + exp(beta.hat[3,
    1] + beta.hat[3, 2] * x + beta.hat[3, 3] * mean_fat + beta.hat[3,
    4] * mean_sodium)), col = "black", lty = "solid", lwd = 2,
    n = 1000, add = TRUE, xlim = c(min(cereal2$sugar[cereal2$Shelf ==
        1]), max(cereal2$sugar[cereal2$Shelf == 1])))

curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 2] * x +
    beta.hat[1, 3] * mean_fat + beta.hat[1, 4] * mean_sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x + beta.hat[2, 3] *
        mean_fat + beta.hat[2, 4] * mean_sodium) + exp(beta.hat[3,
    1] + beta.hat[3, 2] * x + beta.hat[3, 3] * mean_fat + beta.hat[3,
    4] * mean_sodium)), col = "black", lty = "dotdash", lwd = 2,
    n = 1000, add = TRUE, xlim = c(0, 1))

# Shelf2
curve(expr = exp(beta.hat[1, 1] + beta.hat[1, 2] * x + beta.hat[1,
    3] * mean_fat + beta.hat[1, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
```

```
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "green",
    lty = "solid", lwd = 2, n = 1000, add = TRUE, xlim = c(min(cereal2$sugar[cereal2$Shelf ==
        2]), max(cereal2$sugar[cereal2$Shelf == 2])))

curve(expr = exp(beta.hat[1, 1] + beta.hat[1, 2] * x + beta.hat[1,
    3] * mean_fat + beta.hat[1, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "green",
    lty = "dotdash", lwd = 2, n = 1000, add = TRUE, xlim = c(0,
        1))

# Shelf3
curve(expr = exp(beta.hat[2, 1] + beta.hat[2, 2] * x + beta.hat[2,
    3] * mean_fat + beta.hat[2, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "red",
    lty = "solid", lwd = 2, n = 1000, add = TRUE, xlim = c(min(cereal2$sugar[cereal2$Shelf ==
        3]), max(cereal2$sugar[cereal2$Shelf == 3])))

curve(expr = exp(beta.hat[2, 1] + beta.hat[2, 2] * x + beta.hat[2,
    3] * mean_fat + beta.hat[2, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "red",
    lty = "dotdash", lwd = 2, n = 1000, add = TRUE, xlim = c(0,
        1))

# Shelf4
curve(expr = exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3,
    3] * mean_fat + beta.hat[3, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "blue",
    lty = "solid", lwd = 2, n = 1000, add = TRUE, xlim = c(min(cereal2$sugar[cereal2$Shelf ==
        4]), max(cereal2$sugar[cereal2$Shelf == 4])))
```

```r
curve(expr = exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3,
    3] * mean_fat + beta.hat[3, 4] * mean_sodium)/(1 + exp(beta.hat[1,
    1] + beta.hat[1, 2] * x + beta.hat[1, 3] * mean_fat + beta.hat[1,
    4] * mean_sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] *
    x + beta.hat[2, 3] * mean_fat + beta.hat[2, 4] * mean_sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + beta.hat[3, 3] *
        mean_fat + beta.hat[3, 4] * mean_sodium)), col = "blue",
    lty = "dotdash", lwd = 2, n = 1000, add = TRUE, xlim = c(0,
        1))
```

```r
legend(x = 0.5, y = 1, legend = c("Shelf 1", "Shelf 2", "Shelf 3",
    "Shelf 4"), lty = c("solid", "solid", "solid", "solid"),
    col = c("black", "green", "red", "blue"), bty = "n", lwd = c(2,
        2, 2), seg.len = 4)
```



This chart shows the predicted probabilities of which shelf a box of cereal would be found on when sugar content is the only explanatory variable included in the model, for average levels of fat and sodium. In the chart, solid lines are drawn for sugar levels between the min and max of cereals on each shelf, and dashed lines extend the curves to sugar levels outside this range.

In particular, the plot shows that for relatively low sugar levels, assuming average levels of fat and sodium, that shelf 3 or shelf 4 are vastly more likely than the other two shelves, but roughly equivalent to one another, while for higher sugar content, shelf 2 becomes dominant in likelihood while shelf 1 becomes more likely than the remaining two shelves. This corresponds with the hypothesis that sugary cereals target children, who are closest to shelves 1 & 2, but for whom shelves

3 & 4 are too difficult to either see or reach. It is also notable that the increase in likelihood for shelf 1 is clear but substantially subdued which can almost certainly be attributed to the assumption of average sodium levels, given that this is the key explanatory variable for shelf 1 against all other shelves including shelf 2.

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

```
sd.cereal <- apply(X = cereal2[, -c(2)], MARGIN = 2, FUN = sd)
c.value <- c(sd.cereal)[2:4]
# Estimated standard deviations for each explanatory variable
round(c.value, 2)
```

```
##  sugar    fat sodium
##   0.27   0.30   0.23
```

```
conf.beta <- confint(object = mod1, level = 0.95)
ci.OR <- exp(c.value * conf.beta[2:4, 1:2, ])

# coefficients(mod1)
beta.hat2 <- coefficients(mod1)[1, 2:4]
beta.hat3 <- coefficients(mod1)[2, 2:4]
beta.hat4 <- coefficients(mod1)[3, 2:4]

# OR for j = 2 (Shelf 2 vs Shelf 1)
print("OR for j = 2 vs j = 1")
```

```
## [1] "OR for j = 2 vs j = 1"
```

```
mid = exp(c.value * beta.hat2)
round(mid, 4)
```

```
##  sugar    fat sodium
## 2.0647 3.3719 0.0179
```

```
round(1/mid, 4)
```

```
##   sugar     fat  sodium
##  0.4843  0.2966 55.7393
```

```
round(data.frame(low = ci.OR[, 1, 1], mid = mid, up = ci.OR[,
    2, 1]), 4)
```

```
##           low    mid      up
## sugar  0.1436 2.0647 29.6795
## fat    0.8722 3.3719 13.0360
## sodium 0.0007 0.0179  0.4388
```

```r
# Significant variables for j = 2 vs j = 1
round(data.frame(low = 1/ci.OR[3, 2, 1], mid = 1/mid[3], up = 1/ci.OR[3,
    1, 1]), 4)
```

```
##             low     mid       up
## sodium 2.2788 55.7393 1363.371
```

```r
# OR for j = 3 (Shelf 3 vs Shelf 1)
print("OR for j = 3 vs j = 1")
```

```
## [1] "OR for j = 3 vs j = 1"
```

```r
mid = exp(c.value * beta.hat3)
round(mid, 4)
```

```
##  sugar    fat sodium
## 0.0373 0.8465 0.0032
```

```r
round(1/mid, 4)
```

```
##    sugar      fat   sodium
##  26.8096   1.1813 311.3613
```

```r
round(data.frame(low = ci.OR[, 1, 2], mid = mid, up = ci.OR[,
    2, 2]), 4)
```

```
##           low    mid     up
## sugar  0.0028 0.0373 0.4918
## fat    0.2056 0.8465 3.4861
## sodium 0.0001 0.0032 0.1223
```

```r
# Significant variables for j = 3 vs j = 1
round(data.frame(low = 1/ci.OR[c(1, 3), 2, 2], mid = 1/mid[c(1,
    3)], up = 1/ci.OR[c(1, 3), 1, 2]), 4)
```

```
##           low      mid         up
## sugar  2.0334  26.8096   353.4806
## sodium 8.1747 311.3613 11859.3180
```

```r
# OR for j = 4 (Shelf 4 vs Shelf 1)
print("OR for j = 3 vs j = 1")
```

```
## [1] "OR for j = 3 vs j = 1"
```

```r
mid = exp(c.value * beta.hat4)
round(mid, 4)
```

```
##  sugar    fat sodium
## 0.0465 0.7709 0.0034
```

```r
round(1/mid, 4)
```

```
##    sugar      fat   sodium
##  21.4833   1.2972 290.3058
```

```r
round(data.frame(low = ci.OR[, 1, 3], mid = mid, up = ci.OR[,
    2, 3]), 4)
```

```
##           low    mid     up
## sugar  0.0036 0.0465 0.6084
## fat    0.1882 0.7709 3.1574
## sodium 0.0001 0.0034 0.1301
```

```r
# Significant variables for j = 4 vs j = 1
round(data.frame(low = 1/ci.OR[c(1, 3), 2, 3], mid = 1/mid[c(1,
    3)], up = 1/ci.OR[c(1, 3), 1, 3]), 4)
```

```
##           low      mid         up
## sugar  1.6437  21.4833   280.7812
## sodium 7.6838 290.3058 10968.2197
```

```r
cereal2$new_shelf <- relevel(as.factor(cereal2$Shelf), "2")
mod.fit <- multinom(new_shelf ~ sugar + fat + sodium, data = cereal2)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 33.794856
## iter  20 value 33.616990
## iter  30 value 33.595713
## iter  40 value 33.595185
## iter  50 value 33.595142
## final  value 33.595141
## converged
```

```r
conf.beta.new <- confint(object = mod.fit, level = 0.95)
ci.OR.new <- exp(c.value * conf.beta.new[2:4, 1:2, ])
beta.hat3.new <- coefficients(mod.fit)[2, 2:4]
beta.hat4.new <- coefficients(mod.fit)[3, 2:4]

# OR for j = 3 (Shelf 3 vs Shelf 2)
print("OR for j = 3 vs j = 2")
```

```
## [1] "OR for j = 3 vs j = 2"
```

```r
mid = exp(c.value * beta.hat3.new)
round(mid, 4)
```

```
##  sugar    fat sodium
## 0.0180 0.2511 0.1786
```

```r
round(1/mid, 4)
```

```
##   sugar     fat  sodium
## 55.4021  3.9831  5.5997
```

```r
round(data.frame(low = ci.OR.new[, 1, 2], mid = mid, up = ci.OR.new[,
    2, 2]), 4)
```

```
##           low    mid     up
## sugar  0.0013 0.0180 0.2606
## fat    0.0500 0.2511 1.2597
## sodium 0.0146 0.1786 2.1843
```

```
# Significant variables for j = 3 vs j = 2
round(data.frame(low = 1/ci.OR.new[1, 2, 2], mid = 1/mid[1],
    up = 1/ci.OR.new[1, 1, 2]), 4)
```

```
##          low     mid       up
## sugar 3.8375 55.4021 799.8359
```

```
# OR for j = 4 (Shelf 4 vs Shelf 2)
print("OR for j = 4 vs j = 2")
```

```
## [1] "OR for j = 4 vs j = 2"
```

```
mid = exp(c.value * beta.hat4.new)
round(mid, 4)
```

```
##  sugar    fat sodium
## 0.0225 0.2287 0.1916
```

```
round(1/mid, 4)
```

```
##   sugar     fat  sodium
## 44.4019  4.3733  5.2198
```

```
round(data.frame(low = ci.OR.new[, 1, 3], mid = mid, up = ci.OR.new[,
    2, 3]), 2)
```

```
##         low  mid   up
## sugar  0.00 0.02 0.31
## fat    0.05 0.23 1.14
## sodium 0.02 0.19 2.31
```

```
# Significant variables for j = 4 vs j = 2
round(data.frame(low = 1/ci.OR.new[1, 2, 3], mid = 1/mid[1],
    up = 1/ci.OR.new[1, 1, 3]), 4)
```

```
##          low     mid       up
## sugar 3.1915 44.4019 617.7447
```

```
cereal2$new_shelf <- relevel(as.factor(cereal2$Shelf), "3")
mod.fit <- multinom(new_shelf ~ sugar + fat + sodium, data = cereal2)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774
## iter  10 value 35.514143
## iter  20 value 33.667925
## iter  30 value 33.598476
## iter  40 value 33.595194
## iter  50 value 33.595146
```

```
## final  value 33.595139
## converged
```

```r
conf.beta.new <- confint(object = mod.fit, level = 0.95)
ci.OR.new <- exp(c.value * conf.beta.new[2:4, 1:2, ])
beta.hat4.new <- coefficients(mod.fit)[3, 2:4]

# OR for j = 4 (Shelf 3 vs Shelf 3)
print("OR for j = 4 vs j = 3")
```

```
## [1] "OR for j = 4 vs j = 3"
```

```r
mid = exp(c.value * beta.hat4.new)
round(mid, 4)
```

```
##  sugar    fat sodium
## 1.2481 0.9107 1.0723
```

```r
round(1/mid, 4)
```

```
##  sugar    fat sodium
## 0.8012 1.0981 0.9325
```

```r
round(data.frame(low = ci.OR.new[, 1, 3], mid = mid, up = ci.OR.new[,
    2, 3]), 4)
```

```
##           low    mid     up
## sugar  0.4451 1.2481 3.5001
## fat    0.3259 0.9107 2.5445
## sodium 0.4062 1.0723 2.8310
```

Using the confidence intervals computed for the odds ratio for each explanatory variable for each pair of shelves, we can conduct hypothesis tests to determine whether a given explanatory variable has a statistically significant effect on the odds ratio between the two shelves, in particular by determining if the ratio of 1 is outside the bounds of the 95% confidence interval, in which case we can reject the null hypothesis that the variable in question has no discriminating power between the two shelves. For shelf 2 vs shelf 1, we see that only sodium yields a significant result, telling us with 95% confidence that the odds of a cereal being on shelf 1 instead of shelf 2 change by between 2.28 and 1363.37 times for a 0.23 of scaled sodium. That sugar is not significant is marginally surprising, however looking at the boxplots, we note that a large portion of the cereals on shelf 1 have comparable sugar levels to those on shelf 2, and it is only due to a cluster with very low sugar levels on shelf 1 that any large difference is apparent, however clearly sodium is signficantly different between the shelves in the boxplots, thus this result is expected. For shelf 3 vs shelf 1, both sugar and sodium are significant, saying with 95% confidence that the odds of cereal being on shelf 1 instead of shelf 3 change by between 2.03 and 353.48 times for a 0.27 increase in scaled sugar as well as 95% confidence that the odds of cereal being on shelf 1 instead of shelf 3 change by between 8.17 and 11859.32 times for a 0.23 increase in scaled sodium. Unsurprisingly from looking at the boxplots, both sugar and sodium are significant discriminating factors between cereals on shelf 3 vs shelf 1, given that they have fairly low sugar content and even lower sodium than those on shelf 2. For shelf 4 vs shelf 1, again both sugar and sodium are significant, saying with 95% that the odds of cereal being on shelf 1 instead of shelf 4 change by between 1.64 and 280.78 times for a 0.27 increase in scaled sugar as well as 95% confidence that the odds of cereal being on shelf 1 instead of shelf 4 change by between 7.68 and 10968.22 times for a 0.27 increase in scaled sodium, which is close to that of shelf 3 for the same reasons. For shelf 3 vs shelf 2, only sugar is significant, telling us with 95% confidence that the odds of a cereal being on shelf 2 instead of shelf 3 change by between 3.84 and 799.84 times for a 0.27 increase in scaled sugar. For shelf 4 vs shelf 2, again only sugar is significant, telling us with 95% confidence that the odds of a cereal being on shelf 2 instead of shelf 4 change by between 3.19 and 617.74 times for a 0.27 increase in scaled sugar. Both of these correspond to the fairly similar levels of sugar in cereals on shelves 3 and 4 and relatively elevated sugar levels of cereals on shelf 2. For shelf 4 vs shelf 3, there are no significant variables, corresponding to the very similar values across all three explanatory variables for cereals on these two shelves.