W271 Lab 2

Tiffany Jaya, Joanna Huang, Shan He, Robert Deng

```
# add packages
library(car)
library(dplyr)
library(Hmisc)
library(knitr)
library(MASS)
library(nnet)
library(ggplot2)
# prevent source code from running off the page
opts_chunk$set(tidy.opts=list(width.cutoff=70), tidy=TRUE)
# remove all objects from current workspace
rm(list = ls())
# set seed number to reproduce results
set.seed(1)
# load data
d <- read.csv("./dataset/cereal_dillons.csv", header=TRUE, sep=",")
```

Introduction

We are under the impression that supermarkets place cereals strategically on shelves in order to increase sales. For this reason, we use a random sample of size 10 from each of the four shelves at a Dillons grocery store in Manhattan, KS to answer our question: Does the cereal's nutritional composition have a significant effect on its shelving placement?

We chose the multinomial regression model to answer this question since the dependent variable, the probability of the cereal being placed on the shelf, has more than two categories. What we found is that cereals with high sugar content tend to be placed on the second shelf (from the bottom) while cereals with low sugar content tend to be placed on the third shelf (from the bottom).

EDA

In order for us to address this question, we first explored the dataset. The dataset contains 40 observations with 7 numerical attributes and no missing values:

- ID: Unique identifier for each cereal
- Shelf: Shelf number, which is numbered from the bottom (1) to the top (4)
- Cereal: Cereal product name with 38 distinct names
- size_g: Serving size with a range of 27-60 grams
- sugar_g: Sugar per serving with a range of 0-20 grams
- fat_g: Fat per serving with a range of 0-5 grams
- sodium_mg: Sodium per serving with a range of 0-330 milligrams

```
# structure of data
str(d)
```

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

```
# summary of data
kable(summary(d[, 4:7]))
```

size_g	sugar_g	fat_g	sodium_mg
Min. :27.00	Min. : 0.0	Min. :0.000	Min. : 0.0
1st Qu.:29.75	1st Qu.: 6.0	1st Qu.:0.500	1st Qu.:157.5
Median :31.00	Median :11.0	Median :1.000	Median :200.0
Mean :37.20	Mean :10.4	Mean :1.200	Mean :195.5
3rd Qu.:51.00	3rd Qu.:14.0	3rd Qu.:1.625	3rd Qu.:262.5
Max. :60.00	Max. :20.0	Max. :5.000	Max. :330.0

```
# list row with missing values
d[!complete.cases(d), ]
```

```
## [1] ID       Shelf    Cereal   size_g   sugar_g  fat_g    sodium_mg
## <0 rows> (or 0-length row.names)
```

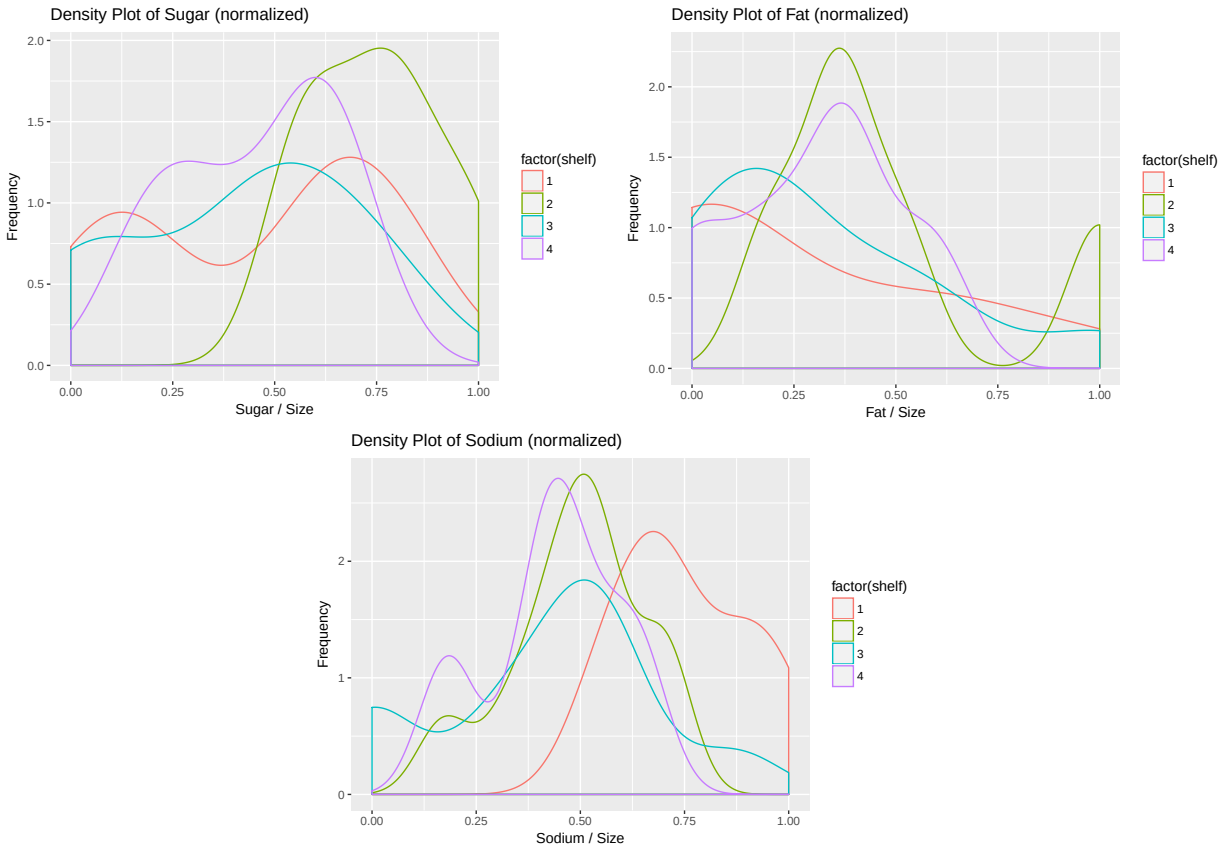
Question A: The explanatory variables need to be re-formatted before proceeding further. First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Second, re-scale each variable to be within 0 and 1.

```
# 1. adjust for different serving sizes
adj.d <- data.frame(shelf = d$Shelf, sugar = d$sugar_g/d$size_g, fat = d$fat_g/d$size_g,
    sodium = (d$sodium_mg/1000)/d$size_g)
# 2. normalize data to be within 0 and 1
normalize <- function(x) {
    (x - min(x))/(max(x) - min(x))
}
norm.d <- data.frame(shelf = adj.d$shelf, sugar = normalize(adj.d$sugar),
    fat = normalize(adj.d$fat), sodium = normalize(adj.d$sodium))
```

Now that we have adjusted the data, we can begin analyzing the explanatory variables of interest, the cereal nutrition's composition: sugar per serving, fat per serving, and sodium per serving.

```
ggplot(norm.d, aes(sugar, colour = factor(shelf))) + geom_density() + ggtitle("Density Plot of
    xlab("Sugar / Size") + ylab("Frequency")

ggplot(norm.d, aes(fat, colour = factor(shelf))) + geom_density() + ggtitle("Density Plot of Fa
    xlab("Fat / Size") + ylab("Frequency")
```

```
ggplot(norm.d, aes(sodium, colour = factor(shelf))) + geom_density() +
    ggtitle("Density Plot of Sodium (normalized)") + xlab("Sodium / Size") +
    ylab("Frequency")
```

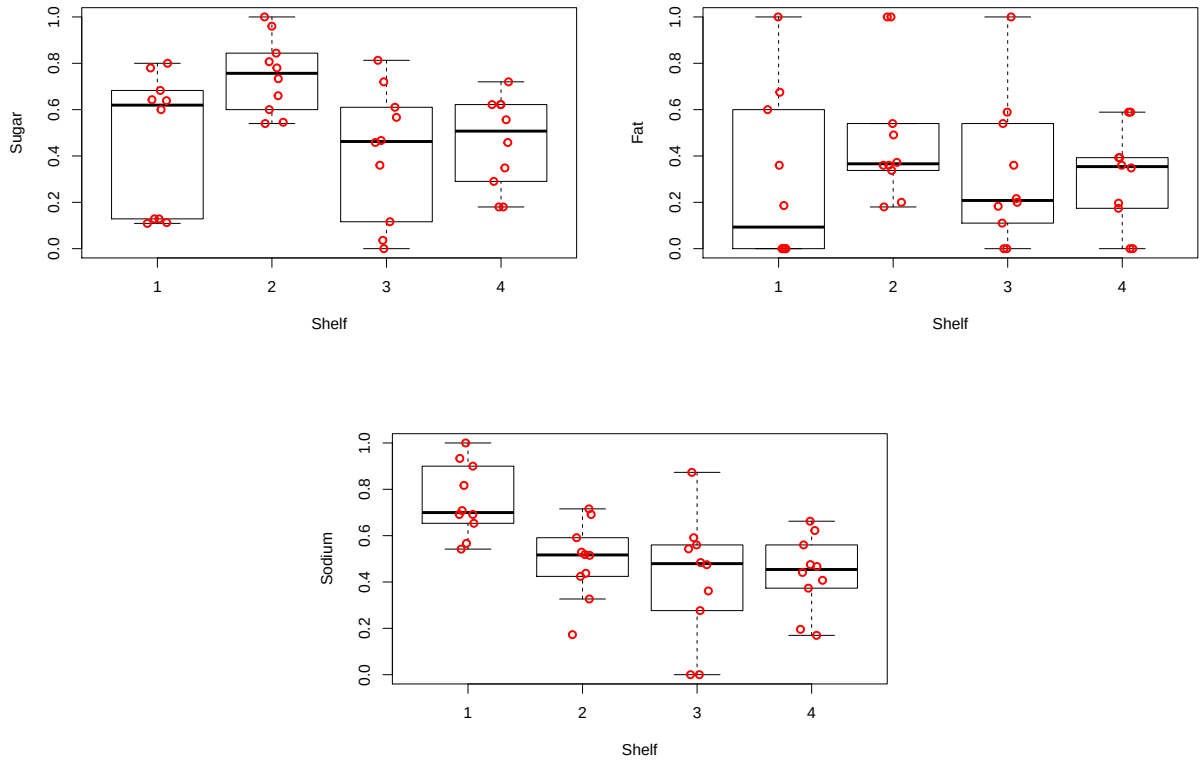


As can be seen from the graph, the distribution of sugar per serving appears to skew left with most products with sugar comprising more than half their size. This is in part due to the second shelf having cereals with higher sugar content compared to those on the other shelves. The density plot of fat per serving and sodium per serving, although appearing bimodal in shape, are not as distinct. The distribution of fat per serving is more right skewed with few cereals having higher fat content overall while the distribution of sodium per serving is more left skewed with few cereals having sodium less than half their content overall.

Question B: Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables.

```
# explanatory variable: sugar
boxplot(formula = sugar ~ shelf, data = norm.d, ylab = "Sugar", xlab = "Shelf",
    pars = list(outpch = NA))
stripchart(x = norm.d$sugar ~ norm.d$shelf, lwd = 2, col = "red", method = "jitter",
    vertical = TRUE, pch = 1, add = TRUE)

# explanatory variable: fat
boxplot(formula = fat ~ shelf, data = norm.d, ylab = "Fat", xlab = "Shelf",
    pars = list(outpch = NA))
```

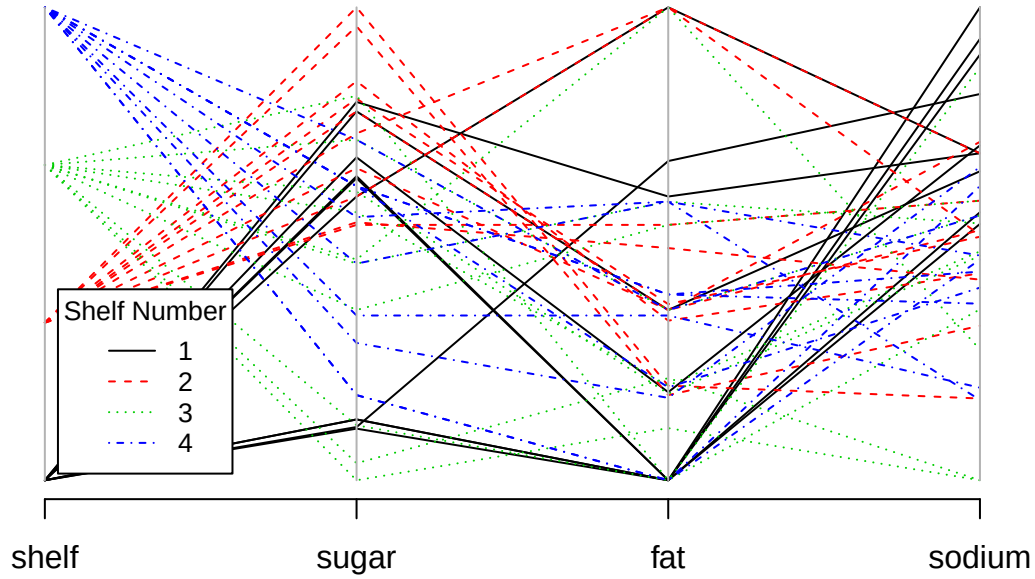
```
stripchart(x = norm.d$fat ~ norm.d$shelf, lwd = 2, col = "red", method = "jitter",
    vertical = TRUE, pch = 1, add = TRUE)

# explanatory variable: sodium
boxplot(formula = sodium ~ shelf, data = norm.d, ylab = "Sodium", xlab = "Shelf",
    pars = list(outpch = NA))
stripchart(x = norm.d$sodium ~ norm.d$shelf, lwd = 2, col = "red", method = "jitter",
    vertical = TRUE, pch = 1, add = TRUE)
```



Although there are a few outliers in the second shelf in terms of fat per serving, it does not appear to be outside of the realm of possibility if compared to the cereals in the first and third shelves.

Also, construct a parallel coordinates plot for the explanatory variables and the shelf number. Discuss if possible content differences exist among the shelves.

```
parcoord(x = norm.d, col = norm.d$shelf, lty = norm.d$shelf)
legend("bottomleft", cex = 0.8, inset = 0.05, title = "Shelf Number", bg = "white",
    legend = seq(1, 4), col = seq(1, 4), lty = seq(1, 4))
```



From this parallel coordinates plot, we see that:

1. Shelf 1 generally has higher sodium levels.
2. Shelf 2 generally has higher sugar levels.
3. Most cereals overlap in sugar and sodium levels but range in fat levels.
4. Shelf 3 has the highest variance overall in sugar, fat and sodium levels.

```
# print out the mean and standard deviation
norm.d %>% group_by(shelf) %>% summarise(mean.sugar = mean(sugar), sd.sugar = sd(sugar),
    mean.fat = mean(fat), sd.fat = sd(fat), mean.sodium = mean(sodium),
    sd.sodium = sd(sodium))
```

```
## # A tibble: 4 x 7
##   shelf mean.sugar sd.sugar mean.fat sd.fat mean.sodium sd.sodium
##   <int>      <dbl>    <dbl>    <dbl>  <dbl>       <dbl>     <dbl>
## 1     1      0.462    0.301    0.282  0.363       0.750     0.156
## 2     2      0.747    0.162    0.484  0.293       0.492     0.163
## 3     3      0.415    0.284    0.320  0.312       0.416     0.269
## 4     4      0.460    0.198    0.304  0.210       0.437     0.162
```

Question 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?

Cereal manufacturers are willing to pay grocery stores more if their product was strategically placed on shelves where it can increase sales. Under this condition, it would be desirable to take ordinality

into account. Based on past studies on store design and visual merchandising, retailers divide vertical shelves into 4 zones: 1) Stoop Level 2) Touch Level 3) Eye Level 4) Stretch Level. Eye Level and Touch Level are the most profitable with adults and children respectively given what is easily seen and obtainable. Furthermore, since it takes a little less physical effort to tip toe for the Stretch Level than crouch down for the Stoop Level, we will assume the former is slightly more profitable as well. Based on this reasoning, we could rank the importance of the shelves as follows: $3 > 2 > 4 > 1$.

However, given very little information about the context of the supermarket and customer demographics, the priority toss up between shelves 2 and 3 make it inconclusive to definitively rank shelves. Marketing science can make arguments for either and statistically we don't have enough information, hence ultimately we will stick to the nominal response model.

Modeling

Since there are 4 shelf categories that a cereal box can be placed on, we will be using a multinomial regression model. Our goal is to model how the nutritional content affects a cereal's corresponding probabilities in each shelf category.

Multinomial logistic regression does not assume normality, linearity, or homoscedasticity. However, it does assume independence among the dependent variable choices meaning that the membership in one category is not related to membership in another category. As of right now, we can assume that independence is met since the nutritional content of cereal A, for example, does not affect the nutritional content of cereal B and therefore its product placement on the shelf. Additionally, multinomial logistic regression assumes no complete separation.

Question 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.

```
# set shelf 2 as the base level
norm.d$shelf <- factor(norm.d$shelf)
norm.d$shelf <- relevel(norm.d$shelf, ref = "2")
# estimate multinomial regression model
m <- multinom(formula = shelf ~ sugar + fat + sodium, data = norm.d)
```

```
## # 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
```

```
summary(m)
```

```
## Call:
## multinom(formula = shelf ~ sugar + fat + sodium, data = norm.d)
##
## Coefficients:
```

```
##   (Intercept)      sugar       fat    sodium
## 1    -6.89779  -2.692069 -4.063019 17.486515
## 3    14.78681 -14.912714 -4.621863 -7.495409
## 4    14.39434 -14.090538 -4.934381 -7.189731
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 1    6.486685 5.051034 2.306998 7.095999
## 3    5.513107 5.059891 2.752076 5.558649
## 4    5.496190 4.989776 2.745046 5.522355
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

```
# perform LRT to examine importance of each variable
Anova(m)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: 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
```

```
# select standard deviations as proper c values for coefficient
# interpretation
c.1 <- c(1, sd(adj.d[(adj.d$shelf == 1), ]$sugar), sd(adj.d[(adj.d$shelf ==
    1), ]$fat), sd(adj.d[(adj.d$shelf == 1), ]$sodium))
c.2 <- c(1, sd(adj.d[(adj.d$shelf == 2), ]$sugar), sd(adj.d[(adj.d$shelf ==
    2), ]$fat), sd(adj.d[(adj.d$shelf == 2), ]$sodium))
c.3 <- c(1, sd(adj.d[(adj.d$shelf == 3), ]$sugar), sd(adj.d[(adj.d$shelf ==
    3), ]$fat), sd(adj.d[(adj.d$shelf == 3), ]$sodium))
```

From the coefficients table in the output, we obtain the estimated model as:

$$log(\frac{\hat{\pi}_{\text{shelf}=1}}{\hat{\pi}_{\text{shelf}=2}}) = -6.90 - 2.69(\text{sugar}) - 4.06(\text{fat}) + 17.49(\text{sodium})$$

```
c.1
```

```
## [1] 1.000000000 0.167295660 0.033583554 0.001667284
```

```
exp(c.1 * coefficients(m)[1, ])
```

```
## (Intercept)       sugar         fat      sodium
## 0.001010015 0.637391357 0.872449404 1.029584153
```

- The estimated odds of being in shelf 1 vs shelf 2 change by 0.64 times for a 0.17 grams per

grams of serving increase in sugar content holding the other variable constant.
- The estimated odds of being in shelf 1 vs shelf 2 change by 0.87 times for a 0.04 grams per grams of serving increase in fat content holding the other variable constant.
- The estimated odds of being in shelf 1 vs shelf 2 change by 1.03 times for a 0.002 grams per grams of serving increase in sodium content holding the other variable constant.

$$log(\frac{\hat{\pi}_{\text{shelf}=3}}{\hat{\pi}_{\text{shelf}=2}}) = 14.78 - 14.91(\text{sugar}) - 4.62(\text{fat}) - 7.49(\text{sodium})$$

```
c.2
```

```
## [1] 1.000000000 0.090010188 0.027143538 0.001745598
```

```
exp(c.2 * coefficients(m)[2, ])
```

```
##  (Intercept)        sugar          fat       sodium
## 2.641371e+06 2.612451e-01 8.820966e-01 9.870012e-01
```

- The estimated odds of being in shelf 3 vs shelf 2 change by 0.26 times for a 0.09 grams per grams of serving increase in sugar content holding the other variable constant.
- The estimated odds of being in shelf 3 vs shelf 2 change by 0.88 times for a 0.03 grams per grams of serving increase in fat content holding the other variable constant.
- The estimated odds of being in shelf 3 vs shelf 2 change by 0.99 times for a 0.002 grams per grams of serving increase in sodium content holding the other variable constant.

$$log(\frac{\hat{\pi}_{\text{shelf}=4}}{\hat{\pi}_{\text{shelf}=2}}) = 14.39 - 14.09(\text{sugar}) - 4.93(\text{fat}) - 7.19(\text{sodium})$$

```
c.3
```

```
## [1] 1.000000000 0.157700574 0.028913593 0.002886223
```

```
exp(c.3 * coefficients(m)[3, ])
```

```
##  (Intercept)        sugar          fat       sodium
## 1.783946e+06 1.083828e-01 8.670396e-01 9.794627e-01
```

- The estimated odds of being in shelf 4 vs shelf 2 change by 0.11 times for a 0.16 grams per grams of serving increase in sugar content holding the other variable constant.
- The estimated odds of being in shelf 4 vs shelf 2 change by 0.87 times for a 0.03 grams per grams of serving increase in fat content holding the other variable constant.
- The estimated odds of being in shelf 4 vs shelf 2 change by 0.98 times for a 0.003 grams per grams of serving increase in sodium content holding the other variable constant.

Based on the likelihood ratio test, sugar and sodium are important explanatory variables in determining where the cereal will be placed on the shelf. On the other hand, there is not sufficient evidence that the fat variable has an effect on the shelf placement.

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

To look at whether we need interactions within our model, we can create a saturated model and run an Anova test to see if any interactions have significance.

```
saturated.m <- multinom(formula = shelf ~ sugar * fat * sodium, data = norm.d)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 32.627653
## iter  20 value 30.478214
## iter  30 value 29.636600
## iter  40 value 28.728753
## iter  50 value 28.106187
## iter  60 value 27.929829
## iter  70 value 27.712664
## iter  80 value 27.550300
## iter  90 value 27.220767
## iter 100 value 26.956905
## final  value 26.956905
## stopped after 100 iterations
```

```
Anova(saturated.m)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: shelf
##                  LR Chisq Df Pr(>Chisq)
## sugar             19.3548  3  0.0002309 ***
## fat                6.0331  3  0.1100102
## sodium            31.0155  3  8.437e-07 ***
## sugar:fat          3.0105  3  0.3900057
## sugar:sodium       2.5803  3  0.4609509
## fat:sodium         3.1675  3  0.3665160
## sugar:fat:sodium   2.0432  3  0.5634965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Anova table indicates that the main effects are significant, but that the interaction effect is not.

Question 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.

```
# Apple Jack's content
serving <- 28
sugar <- 12
fat <- 0.5
sodium <- 130
# predict Apple Jack's shelf placement
predict(object = m, data.frame(sugar = (sugar/serving - min(adj.d$sugar))/(max(adj.d$sugar) -
    min(adj.d$sugar)), fat = (fat/serving - min(adj.d$fat))/(max(adj.d$fat) -
    min(adj.d$fat)), sodium = (sodium/serving - min(adj.d$sodium))/(max(adj.d$sodium) -
    min(adj.d$sodium))), type = "probs")
```

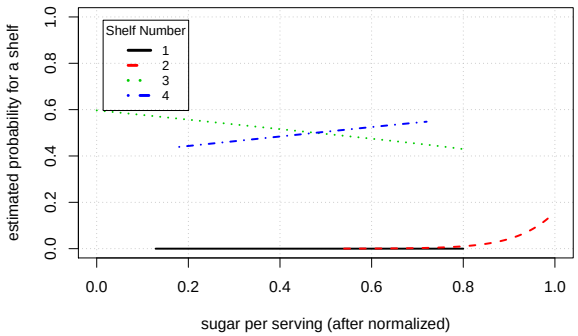
```
## 2 1 3 4
## 0 1 0 0
```

Apple Jack will most likely be placed on the first shelf.

Question 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(m)
mean.fat <- mean(norm.d$fat)
mean.sodium <- mean(norm.d$sodium)

curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat +
    mean.sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat +
    mean.sodium) + exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat +
    mean.sodium)), ylab = "estimated probability for a shelf", xlab = "sugar per serving (after
    ylim = c(0, 1), xlim = c(0, 1), type = "n", panel.first = grid(col = "gray",
        lty = "dotted"))

# pi.hat.shelf.1
curve(expr = exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat + mean.sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat + mean.sodium)),
    col = 1, lty = 1, lwd = 2, n = 1000, add = TRUE, xlim = c(norm.d$sugar[norm.d$shelf ==
        1]), max(norm.d$sugar[norm.d$shelf == 1]))

# pi.hat.shelf.2
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat +
    mean.sodium) + exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat +
    mean.sodium) + exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat +
    mean.sodium)), col = 2, lty = 2, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sugar[n
    2]), max(norm.d$sugar[norm.d$shelf == 2])))

# pi.hat.shelf.3
curve(expr = exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat + mean.sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat + mean.sodium)),
    col = 3, lty = 3, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sugar[norm.d$shelf ==
        3]), max(norm.d$sugar[norm.d$shelf == 3])))

# pi.hat.shelf.4
curve(expr = exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 2] * x + mean.fat + mean.sodium) +
```

```
    exp(beta.hat[2, 1] + beta.hat[2, 2] * x + mean.fat + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 2] * x + mean.fat + mean.sodium)),
    col = 4, lty = 4, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sugar[norm.d$shelf ==
        4]), max(norm.d$sugar[norm.d$shelf == 4])))

# legend
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4), lwd = rep(3, 4))
```



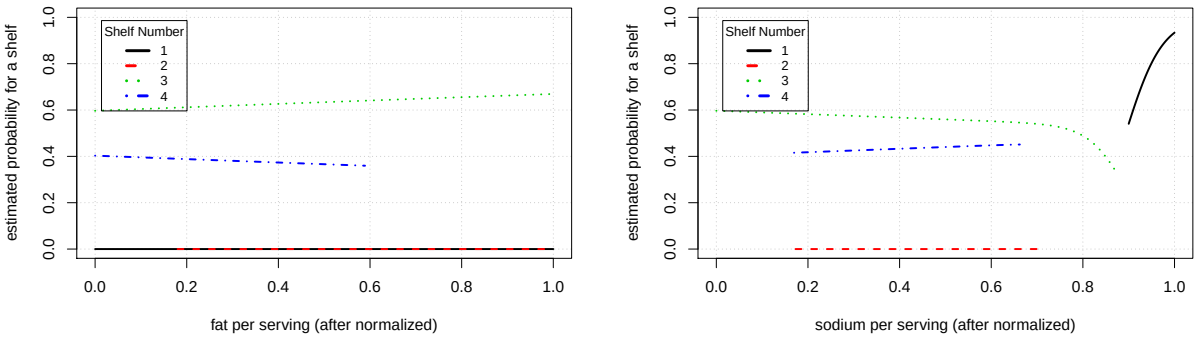
The estimated probability that the cereal is placed on the third shelf is highest for smaller serving of sugar. Then, the estimated probability that the cereal is placed on the fourth shelf is highest when the sugar level is slightly above the middle. When the sugar serving is large, the estimated probability that the cereal is placed on the second shelf is highest.

If we looked at the other explanatory variables individually, they are not as good of an indicator for the estimated probability for a shelf as much as sugar.

```
mean.sugar <- mean(norm.d$sugar)
# estimated multinomial regression model for the data where ...  fat is
# the only explanatory variable included in the model
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar +
    mean.sodium) + exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar +
    mean.sodium) + exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar +
    mean.sodium)), ylab = "estimated probability for a shelf", xlab = "fat per serving (after
    ylim = c(0, 1), xlim = c(0, 1), type = "n", panel.first = grid(col = "gray",
        lty = "dotted"))
# pi.hat.shelf.1
curve(expr = exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar + mean.sodium)),
    col = 1, lty = 1, lwd = 2, n = 1000, add = TRUE, xlim = c(norm.d$fat[norm.d$shelf ==
        1]), max(norm.d$fat[norm.d$shelf == 1]))
# pi.hat.shelf.2
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar +
    mean.sodium) + exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar +
```

```
    mean.sodium) + exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar +
    mean.sodium)), col = 2, lty = 2, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$fat[no
    2]), max(norm.d$fat[norm.d$shelf == 2])))
# pi.hat.shelf.3
curve(expr = exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar + mean.sodium)),
    col = 3, lty = 3, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$fat[norm.d$shelf ==
        3]), max(norm.d$fat[norm.d$shelf == 3])))
# pi.hat.shelf.4
curve(expr = exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar + mean.sodium)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[2, 1] + beta.hat[2, 3] * x + mean.sugar + mean.sodium) +
    exp(beta.hat[3, 1] + beta.hat[3, 3] * x + mean.sugar + mean.sodium)),
    col = 4, lty = 4, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$fat[norm.d$shelf ==
        4]), max(norm.d$fat[norm.d$shelf == 4])))
# legend
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4), lwd = rep(3, 4))

# sodium is the only explanatory variable included in the model
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar +
    mean.fat) + exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar +
    mean.fat) + exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar +
    mean.fat)), ylab = "estimated probability for a shelf", xlab = "sodium per serving (after r
    ylim = c(0, 1), xlim = c(0, 1), type = "n", panel.first = grid(col = "gray",
        lty = "dotted"))
# pi.hat.shelf.1
curve(expr = exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar + mean.fat)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar + mean.fat)),
    col = 1, lty = 1, lwd = 2, n = 1000, add = TRUE, xlim = c(norm.d$sodium[norm.d$shelf ==
        1]), max(norm.d$sodium[norm.d$shelf == 1]))
# pi.hat.shelf.2
curve(expr = 1/(1 + exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar +
    mean.fat) + exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar +
    mean.fat) + exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar +
    mean.fat)), col = 2, lty = 2, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sodium[no
    2]), max(norm.d$sodium[norm.d$shelf == 2])))
# pi.hat.shelf.3
curve(expr = exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar + mean.fat)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar + mean.fat)),
    col = 3, lty = 3, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sodium[norm.d$shelf ==
```

```
        3]), max(norm.d$sodium[norm.d$shelf == 3])))
# pi.hat.shelf.4
curve(expr = exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar + mean.fat)/(1 +
    exp(beta.hat[1, 1] + beta.hat[1, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[2, 1] + beta.hat[2, 4] * x + mean.sugar + mean.fat) +
    exp(beta.hat[3, 1] + beta.hat[3, 4] * x + mean.sugar + mean.fat)),
    col = 4, lty = 4, lwd = 2, n = 1000, add = TRUE, xlim = c(min(norm.d$sodium[norm.d$shelf ==
        4]), max(norm.d$sodium[norm.d$shelf == 4])))
# legend
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4), lwd = rep(3, 4))
```



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

```
# standard deviation
sd.d <- apply(norm.d[2:4], 2, sd)
round(sd.d, 2)
```

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

```
# wald ci
conf.beta <- confint(m, level = 0.95)
ci.or.shelf.1 <- exp(sd.d * conf.beta[2:4, 1:2, 1])
ci.or.shelf.2 <- exp(sd.d * conf.beta[2:4, 1:2, 2])
ci.or.shelf.3 <- exp(sd.d * conf.beta[2:4, 1:2, 3])

# odd ratio for shelf 1 vs shelf 2:
round(1/exp(sd.d * beta.hat[1, 2:4]), 2)
```

```
##  sugar    fat sodium
##   2.06   3.37   0.02
```

```
round(data.frame(low = 1/ci.or.shelf.1[, 2], up = 1/ci.or.shelf.1[, 1]),
    2)
```

```
##         low    up
## sugar  0.14 29.66
## fat    0.87 13.03
## sodium 0.00  0.44
```

```
# odd ratio for shelf 3 vs shelf 2:
round(1/exp(sd.d * beta.hat[2, 2:4]), 2)
```

```
##  sugar    fat sodium
##  55.40   3.98   5.60
```

```
round(data.frame(low = 1/ci.or.shelf.2[, 2], up = 1/ci.or.shelf.2[, 1]),
    2)
```

```
##         low     up
## sugar  3.84 799.84
## fat    0.79  19.99
## sodium 0.46  68.49
```

```
# odd ratio for shelf 4 vs shelf 2:
round(1/exp(sd.d * beta.hat[3, 2:4]), 2)
```

```
##  sugar    fat sodium
##  44.40   4.37   5.22
```

```
round(data.frame(low = 1/ci.or.shelf.3[, 2], up = 1/ci.or.shelf.3[, 1]),
    2)
```

```
##         low     up
## sugar  3.19 617.74
## fat    0.88  21.85
## sodium 0.43  62.81
```

- The estimated odds of shelf 1 vs shelf 2 change by 2.06 times (with 95% CI of 0.14 and 29.66) for a 0.27 decrease in sugar per serving holding the other variables constant. The estimated odds of shelf 3 vs shelf 2 change by 55.40 times (with 95% CI of 3.84 and 799.84) for a 0.27 decrease in sugar per serving holding the other variables constant. The estimated odds of shelf 4 vs shelf 2 change by 44.40 times (with 95% CI of 3.19 and 617.74) for 0.27 decrease in sugar per serving holding the other variables constant.
- The estimated odds of shelf 1 vs shelf 2 change by 3.37 times (with 95% CI of 0.87 and 13.03) for a 0.30 decrease in fat per serving holding the other variables constant. The estimated odds of shelf 3 vs shelf 2 change by 3.98 times (with 95% CI of 0.79 and 19.99) for a 0.30 decrease in fat per serving holding the other variables constant. The estimated odds of shelf 4 vs shelf 2 change by 4.37 times (with 95% CI of 0.88 and 21.85) for a 0.30 decrease in fat per serving holding the other variables constant.
- The estimated odds of shelf 1 vs shelf 2 change by 0.02 times (with 95% CI of 0 and 0.44) for a 0.23 decrease in sodium per serving holding the other variables constant. The estimated odds of shelf 3 vs shelf 2 change by 5.60 times (with 95% CI of 0.46 and 68.49) for a 0.23 decrease in sodium per serving holding the other variables constant. The estimated odds of shelf shelf 4 vs shelf 2 change by 5.22 times (with 95% CI of 0.43 and 62.81) for a 0.23 decrease in sodium per serving holding the other variables constant.

We see that cereals with high content of sugar is more likely to be placed at shelf 2 while cereals with high content of sodium is more likely placed in either shelf 1 or 2. We can relate these results back to the parallel coordinates plot. The large variance of sugar and sodium content in the third and fourth shelves can also be seen in the boxplots done earlier during our EDA. Lastly, sugar as an explanatory variable seems to be a stronger indicative of shelf placement, and the graphs in the previous sections seem to validate this intuition.

Conclusion

Our findings seem to confirm that cereals are strategically placed depending on the target market with sugar per serving being the best indicator of the shelf placement. Cereals with high sugar content are placed on the second shelf, also known as the "touch level", reachable by children. Healthier alternatives with lower sugar are placed on the third shelf where it hits at approximately adult eye level. If the cereal was slightly more sweet than the ones on the third shelf, it was placed at fourth shelf within stretching distance of an average adult height. The last shelf closest to the floor appears to store a variety of different types of cereals with different nutritional content. Supplemental sales data can confirm the importance of the shelf order that we have found in this initial research.