# W271 Lab 2

*Tiffany Jaya, Joanna Huang, Shan He, Robert Deng*

```
# add packages
library(dplyr)
library(knitr)
library(MASS)
library(nnet)
library(Hmisc)
library(car)
# 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, a random sample of size 10 was taken from each of the four shelves at a Dillons grocery store in Manhattan, KS in order to answer our question: Does the cereal's nutritional composition have a significant effect on its shelving placement?

TODO: methodology being employed, highlight the results

## 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: Reformat the data

**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 Symptoms Seizures Cardiac deficiency 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))
```

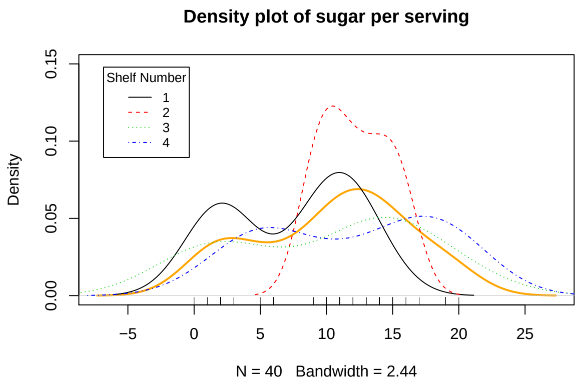
### Univariate analysis

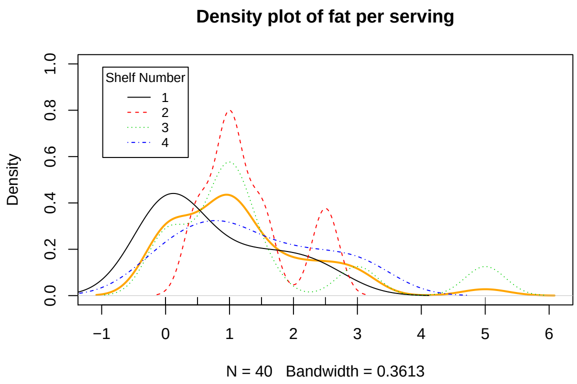
Now that we have adjusted the data to account for the different serving size and normalized to be within 0 and 1, we can begin analyzing the explanatory variables of interest, the cereal nutrition's composition: sugar per serving, fat per serving, and sodium per serving.

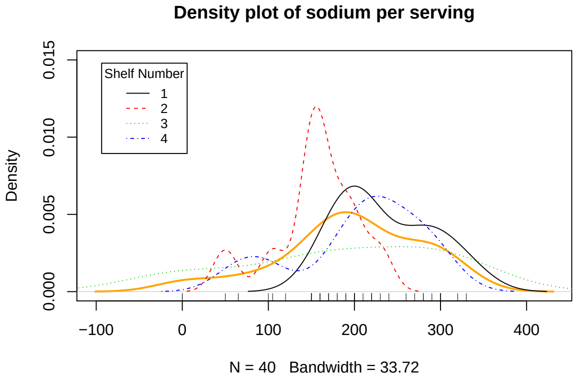
```
# explanatory variable: sugar
plot(density(d$sugar_g), col = "orange", lwd = 2, ylim = c(0, 0.15), main = "Density plot of su
lapply(seq(1, 4), function(shelf) lines(density(d[which(d$Shelf == shelf),
    ]$sugar_g), col = shelf, lty = shelf))
rug(d$sugar_g)
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4))

# explanatory variable: fat
plot(density(d$fat_g), col = "orange", lwd = 2, ylim = c(0, 1), main = "Density plot of fat pe
lapply(seq(1, 4), function(shelf) lines(density(d[which(d$Shelf == shelf),
    ]$fat_g), col = shelf, lty = shelf))
rug(d$fat_g)
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4))

# explanatory variable: sodium
plot(density(d$sodium_mg), col = "orange", lwd = 2, ylim = c(0, 0.015),
    main = "Density plot of sodium per serving")
lapply(seq(1, 4), function(shelf) lines(density(d[which(d$Shelf == shelf),
    ]$sodium_mg), col = shelf, lty = shelf))
rug(d$sodium_mg)
legend("topleft", cex = 0.8, inset = 0.05, title = "Shelf Number", legend = seq(1,
    4), col = seq(1, 4), lty = seq(1, 4))
```







The orange density plot in each graph signifies the density plot for all shelves. As can be seen from the graph, the distribution of sugar per serving appears to be bimodal with the main peak having a wider spread than the smaller one. This is in part due to the second shelf having cereals with higher sugar content compared to those on the other shelves. Also, it is possible that within one shelf, cereal types change from one end of the shelf to the other, attributing to the bimodal shape. For example, cereals with higher sugar, fat, and sodium contents are more popular amongst children while adults might prefer a healthier alternative. It comes to no surprise that the second shelf also has a bimodal shape with cereals having higher fat and sodium content than those placed on the other shelves. In contrast, the third shelf seems to carry healthier alternative cereals with higher fat and lower sugar and sodium level. 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 positively skewed with few cereals having higher fat content overall while the distribution of sodium per serving is more negatively skewed with few cereals having less sodium content overall.

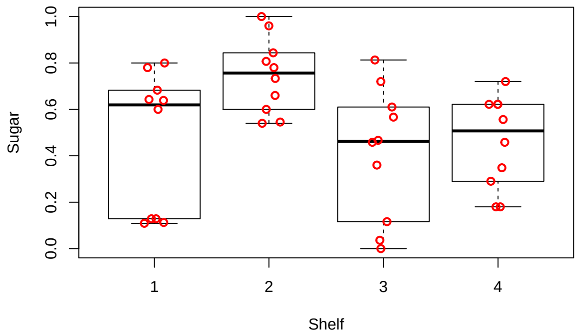
## Question B: Bivariate analysis

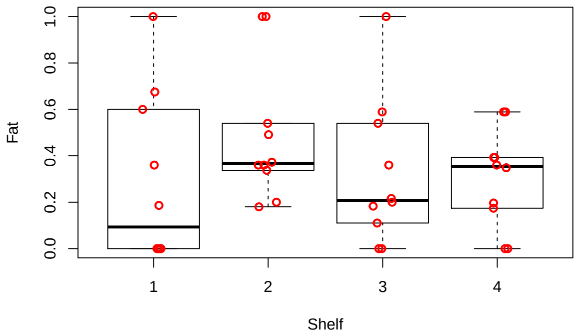
### Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables.

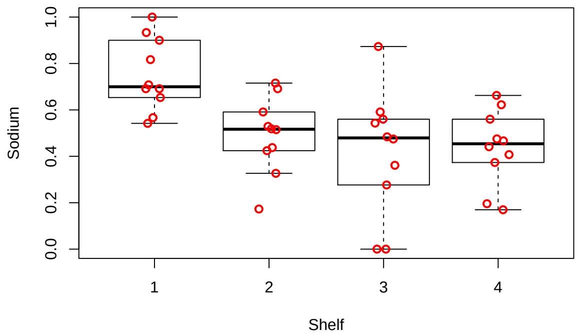
```r
# explanatory variable: sugar
boxplot(formula = sugar ~ shelf, data = norm.d, ylab = "Sugar", xlab = "Shelf",
    pars = list(outpch = NA))
stripchart(x = norm.d$sugar ~ norm.d$shelf, lwd = 2, col = "red", method = "jitter",
    vertical = TRUE, pch = 1, add = TRUE)

# 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)
```



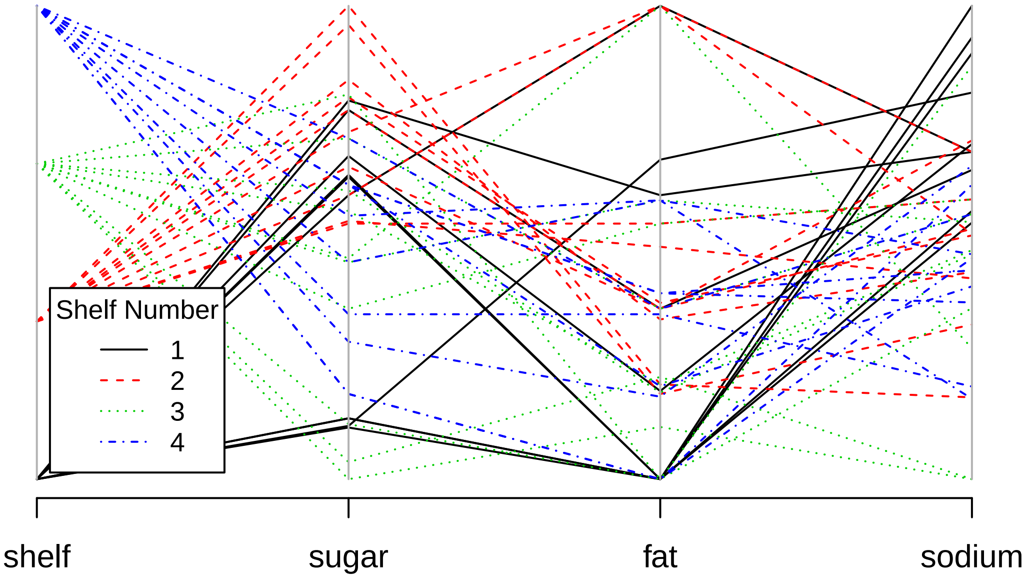




Although there are few outliers in the second shelf in terms of fat per serving, it does not appear to be outside of the realm of possibility if compared to the cereals in the first and third shelves.

**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.

```
# print out the mean and standard deviation of each explanatory
# variable
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
```

What we noticed is similar to what we have noticed in our univariate analysis. The second shelf has higher sugar, fat, and sodium per serving in general. In addition, the first shelf has higher sodium levels as well as broader range in terms of sugar and fat content similar to the third shelf.

## 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 supermarkets 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. In the case of the 40 samples randomly chosen at Dillons grocery store, it seems like this is the case. Cereals with higher sugar, fat, and sodium levels are placed on the second shelf, and the bimodal shape in the density plot diagrams mentioned earlier appears to indicate that there are cereal options for children and adults alike. The third shelf has healthier alternatives with higher fat and lower sugar and sodium servings, and the first and fourth shelves are similar except that the first shelf has higher sodium content with broader deviation. Based on this reasoning, we could rank the importance of the shelves as follows: $2 > 3 > 1 > 4$. Without sales data to backup our reasonings, the ordinality of the shelves that we derive from guess work can be considered a moot point.

```r
# apply ordinality
norm.d$shelfFO <- factor(norm.d$shelf, levels = c(2, 3, 1, 4), ordered = TRUE)
levels(norm.d$shelfFO)
```

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

## Question D: 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.

**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
# 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 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(m)
```

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

```
# 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
```

```
# estimate multinomial regression model if shelves have ordinality
m.fo <- multinom(formula = shelfFO ~ 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
```

```r
summary(m.fo)
```

```
## Call:
## multinom(formula = shelfFO ~ sugar + fat + sodium, data = norm.d)
##
## Coefficients:
##   (Intercept)      sugar       fat    sodium
## 3    14.78681 -14.912714 -4.621863 -7.495409
## 1    -6.89779  -2.692069 -4.063019 17.486515
## 4    14.39434 -14.090538 -4.934381 -7.189731
##
## Std. Errors:
##   (Intercept)    sugar      fat   sodium
## 3    5.513107 5.059891 2.752076 5.558649
## 1    6.486685 5.051034 2.306998 7.095999
## 4    5.496190 4.989776 2.745046 5.522355
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

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

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

TODO: explain Fat not significant

## Question E:

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

## 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, fat = fat/serving,
    sodium = sodium/serving), type = "probs")
```

```
##            1            2            3            4
## 1.000000e+00 1.802745e-32 5.913707e-44 2.325186e-43
```

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.

## 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.

TODO: need to make correction because of normalization made above ˆ

```
# odd ratios
round(exp(coefficients(m)[, -1]), 2)
```

```
##   sugar   fat sodium
## 2 14.78 58.25      0
## 3  0.00  0.57      0
## 4  0.00  0.42      0
```

```
# confidence intervals
conf.beta <- confint(object = m, level = 0.95)
ci.OR.shelf2 <- exp(conf.beta[2:4, 1:2, 1])
ci.OR.shelf3 <- exp(conf.beta[2:4, 1:2, 2])
ci.OR.shelf4 <- exp(conf.beta[2:4, 1:2, 3])
round(data.frame(low = ci.OR.shelf2[, 1], up = ci.OR.shelf2[, 2]), 2)
```

```
##         low        up
## sugar  0.00 294843.41
## fat    0.63   5360.63
## sodium 0.00      0.03
```

```
round(data.frame(low = ci.OR.shelf3[, 1], up = ci.OR.shelf3[, 2]), 2)
```

```
##         low    up
## sugar  0.00  0.07
## fat    0.01 65.11
## sodium 0.00  0.00
```

```
round(data.frame(low = ci.OR.shelf4[, 1], up = ci.OR.shelf4[, 2]), 2)
```

```
##        low    up
## sugar    0  0.16
## fat      0 46.76
## sodium   0  0.00
```

## Conclusion