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

Tiffany Jaya, Joanna Huang, Shan He, Robert Deng

```r
setwd('/Users/arthur/Downloads/w271_Lab2_2018Summer/')
# add packages
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(knitr)
library(nnet)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```r
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

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

```
library(car)
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# 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
cereal <- read.csv("cereal_dillons.csv", header=TRUE, sep=",")
```

Introduction

In this project, we look into grocery stores and how the product contents can give us insight into how stores strategically place items to draw customer attention. We look specifically at one type of item – breakfast cereal. Our dataset includes a random sample of 40 shelves at a Dillons grocery store in Manhattan, KS. We aim to answer the research question: Does the cereal's nutritional composition have a significant effect on its shelving placement?

EDA

In order for us to answer this question, we explored the dataset that contains 40 observations. This sample was obtained by taking a random sample of 10 cereal products taken from each of four shelves at a Dillons grocery store in Manhattan, KS. Looking at the dataset, there are six numerical attributes with no missing values: - Shelf: Shelf number, which is numbered from bottom (1) to top (4) - Cereal: Cereal product name with 38 distinct names - Size.g: Serving size with a range of 27-60 grams - Sugar.g: Sugar content with a range of 0-16 grams - Fat.g: Fat content with a range of 0-5 grams - Sodium.mg: Sodium content with a range of 0-330 milligrams

(Answer 12a) Before proceeding further, we rescale each variable to standardize the measurements so we can make accurate comparisons across the content attributes.

```
stand01 <- function(x) {
    (x - min(x))/(max(x) - min(x))
}
cereal2 <- data.frame(Shelf = cereal$Shelf, sugar = round(cereal$sugar_g/cereal$size_g,
    2), fat = round(cereal$fat_g/cereal$size_g, 2), sodium = round((cereal$sodium_mg/1000)/cere
    3))
```

```r
table(cereal2$sugar)
```

```
## 
##    0 0.02 0.06 0.07  0.1 0.16 0.19  0.2 0.25 0.26  0.3 0.31 0.33 0.34 0.35 
##    1    1    3    2    2    1    1    1    2    1    2    2    2    1    4 
## 0.36 0.37 0.38  0.4 0.41 0.43 0.44 0.45 0.47 0.53 0.56 
##    1    1    1    2    1    2    1    2    1    1    1
```

Checking our values, we may consider binning to have more values under each sugar content value.

Bivariate analysis

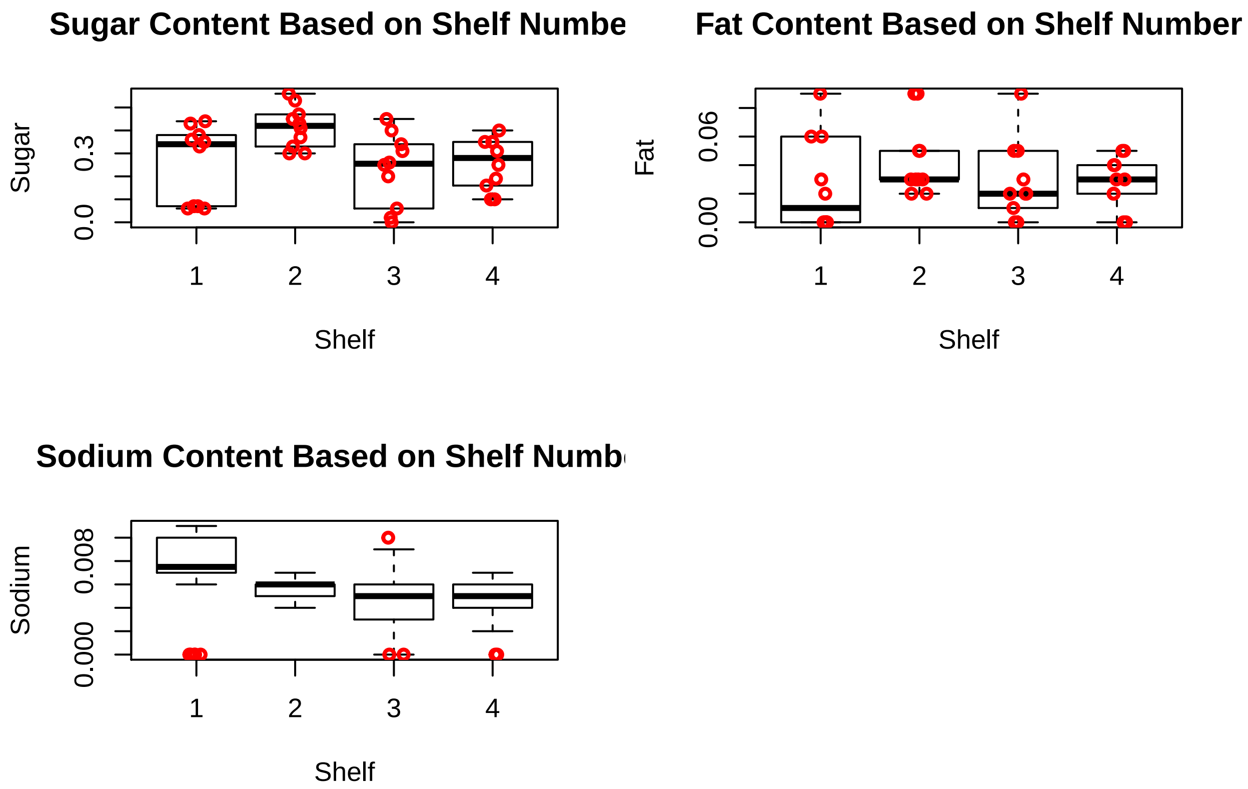
(Answer 12b)

```r
par(mfrow = c(2, 2))

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

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



Include brief description of plots

```r
# Parallel coordinate plot
library(package = MASS)
```

```
## Warning: package 'MASS' was built under R version 3.4.4
```

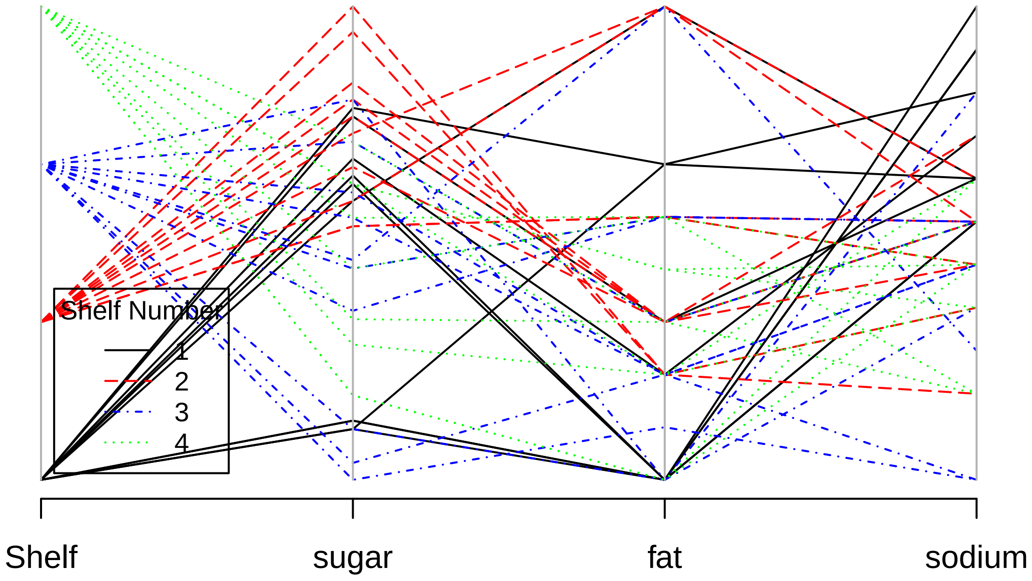
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Colors by condition
shelf.colors <- ifelse(cereal2$Shelf == 1, yes = "black", no = ifelse(cereal2$Shelf ==
    2, yes = "red", no = ifelse(cereal2$Shelf == 3, yes = "blue", no = "green")))

# Line type by condition:
shelf.lty <- ifelse(cereal2$Shelf == 1, yes = "solid", no = ifelse(cereal2$Shelf ==
    2, yes = "longdash", no = ifelse(cereal2$Shelf == 3, yes = "dotdash",
    no = "dotted")))
# Create plot for book
parcoord(x = cereal2, col = shelf.colors, lty = shelf.lty)  # Plot

legend("bottomleft", inset = 0.05, legend = c("1", "2", "3", "4"), lty = c("solid",
    "longdash", "dotdash", "dotted"), col = c("black", "red", "blue", "green"),
    cex = 0.8, bty = "L", title = "Shelf Number")
```



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.

Choosing a Model

(Answer 12c) 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 each shelf category. Assuming that shelving has the potential of increasing or decreasing the likelihood that a product is purchased, an ordinal response model may be fitting. Since there is no given information on the ordering of the shelves, we will assume that customers have a tendency to look from the top to bottom, thus leading to the order of: Shelf 1 < Shelf 2 < Shelf 3 < Shelf 4.

Validating Assumptions

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. Additionally, multinomial logistic regression assumes no complete separation.

Applying Model

(Answer 12d) Estimate a multinomial regression model with linear forms of the sugar, fat, and sodium variables. Perform LRTs to examine the important of each explanatory variable.

```
m <- multinom(formula = Shelf ~ sugar_g + fat_g + sodium_mg, data = cereal)
```

```
## # weights:  20 (12 variable)
## initial  value 55.451774 
## iter  10 value 49.667625
## final  value 49.238414 
## converged
```

```
summary(m)
```

```
## Call:
## multinom(formula = Shelf ~ sugar_g + fat_g + sodium_mg, data = cereal)
## 
## Coefficients:
##   (Intercept)    sugar_g     fat_g    sodium_mg
## 2   0.7876557 0.19205944 0.3228634 -0.016417137
## 3   0.8037972 0.05032418 0.4889536 -0.008474307
## 4  -0.3581925 0.12413008 0.3195256 -0.005503545
## 
## Std. Errors:
##   (Intercept)    sugar_g     fat_g   sodium_mg
## 2    1.662418 0.11299740 0.5637985 0.007639325
## 3    1.635870 0.09551477 0.5349099 0.006439051
## 4    1.773441 0.09597892 0.5408380 0.006822662
## 
## Residual Deviance: 98.47683 
## AIC: 122.4768
```

```
Anova(m)
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: Shelf
##           LR Chisq Df Pr(>Chisq)  
## sugar_g     4.0485  3    0.25627  
## fat_g       0.8870  3    0.82857  
## sodium_mg   6.3034  3    0.09775 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Answer 12e) Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables)

```
# cor(cereal2[,c('sugar', 'fat', 'sodium')])
library(psych)
```

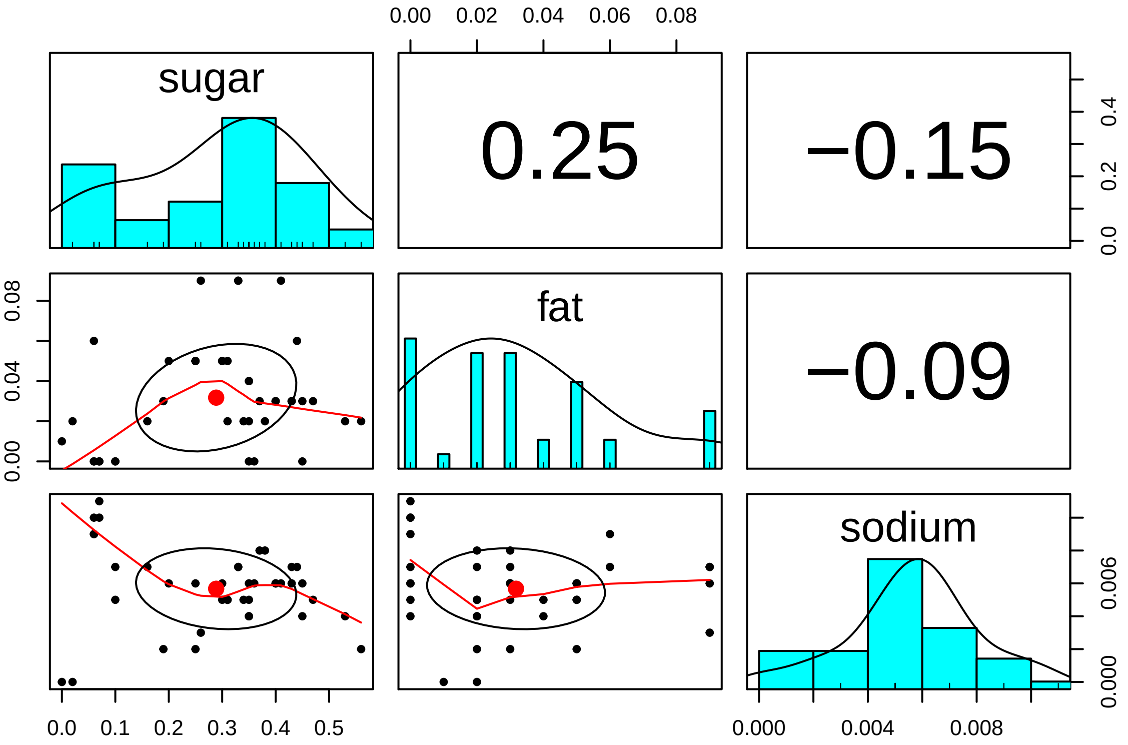
```
## Warning: package 'psych' was built under R version 3.4.4
```

```
## 
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
## 
##     logit
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(cereal2[, c("sugar", "fat", "sodium")])
```



Looking at the correlation between the three explanatory variables, there does not appear to be any strong relationships as the correlation coefficients are all less than 0.5 in magnitude.

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.

```
m0 <- multinom(formula = Shelf ~ 1, data = cereal2)
```

```
## # weights:  8 (3 variable)
## initial  value 55.451774
## final  value 55.451774
## converged
```

```
m2 <- multinom(formula = Shelf ~ sugar * fat * sodium, data = cereal2)
```

```
## # weights:  36 (24 variable)
## initial  value 55.451774
## iter  10 value 48.609837
## iter  20 value 47.303673
## iter  30 value 46.770662
## iter  40 value 46.403603
## iter  50 value 45.244626
## iter  60 value 44.881981
## iter  70 value 44.502518
## iter  80 value 44.072839
## iter  90 value 43.909860
## iter 100 value 43.558892
## final  value 43.558892
## stopped after 100 iterations
```

```
# summary(cereal_m1)
Anova(m2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                   LR Chisq Df Pr(>Chisq)
## sugar              10.3831  3    0.01557 *
## fat                -2.5568  3    1.00000
## sodium              3.3694  3    0.33809
## sugar:fat          -6.4662  3    1.00000
## sugar:sodium       -2.4613  3    1.00000
## fat:sodium         -2.4319  3    1.00000
## sugar:fat:sodium    1.1285  3    0.77019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# anova(cereal_base, cereal_saturated, test='Chi')
```

Answer 12f Kellog's Apple Jacks 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 12g 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 repsect to sugar content.

Answer 12h Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretation back to the plots constructed for this exercise.

Conclusion