W271 - Lab 2

Beau Kramer, Kathryn Papandrew

10/12/2018

```
library(car)
library(nnet)
library(ggplot2)
library(tidyr)
library(Hmisc)
library(MASS)

raw_data <- read.csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQ2QpkcFZmMulpj-Kt71WDso1mdZIYJMO
```

```
head(raw_data)
```

```
##   ID Shelf                              Cereal size_g sugar_g fat_g
## 1  1     1 Kellog's Razzle Dazzle Rice Crispies     28      10     0
## 2  2     1            Post Toasties Corn Flakes     28       2     0
## 3  3     1               Kellogg's Corn Flakes     28       2     0
## 4  4     1              Food Club Toasted Oats     32       2     2
## 5  5     1                    Frosted Cheerios     30      13     1
## 6  6     1             Food Club Frosted Flakes     31      11     0
##   sodium_mg
## 1       170
## 2       270
## 3       300
## 4       280
## 5       210
## 6       180
```

Lab 2 - *Strategic Placement of Products in Grocery Stores*

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

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular boxes of cereals on specific shelves, grocery stores may better attract customers to them. To investigate this further, a random sample of size 10 was taken from each of four shelves at a Dillons grocery store in Manhattan, KS. These data are given in the **cereal_dillons.csv** file. The response variable is the shelf number, which is numbered from bottom (1) to top (4), and the explanatory variables are the sugar, fat, and sodium content of the cereals.

Part A

The explanatory variables need to be reformatted before proceeding further.

Divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Rescale each variable to be within 0 and 1

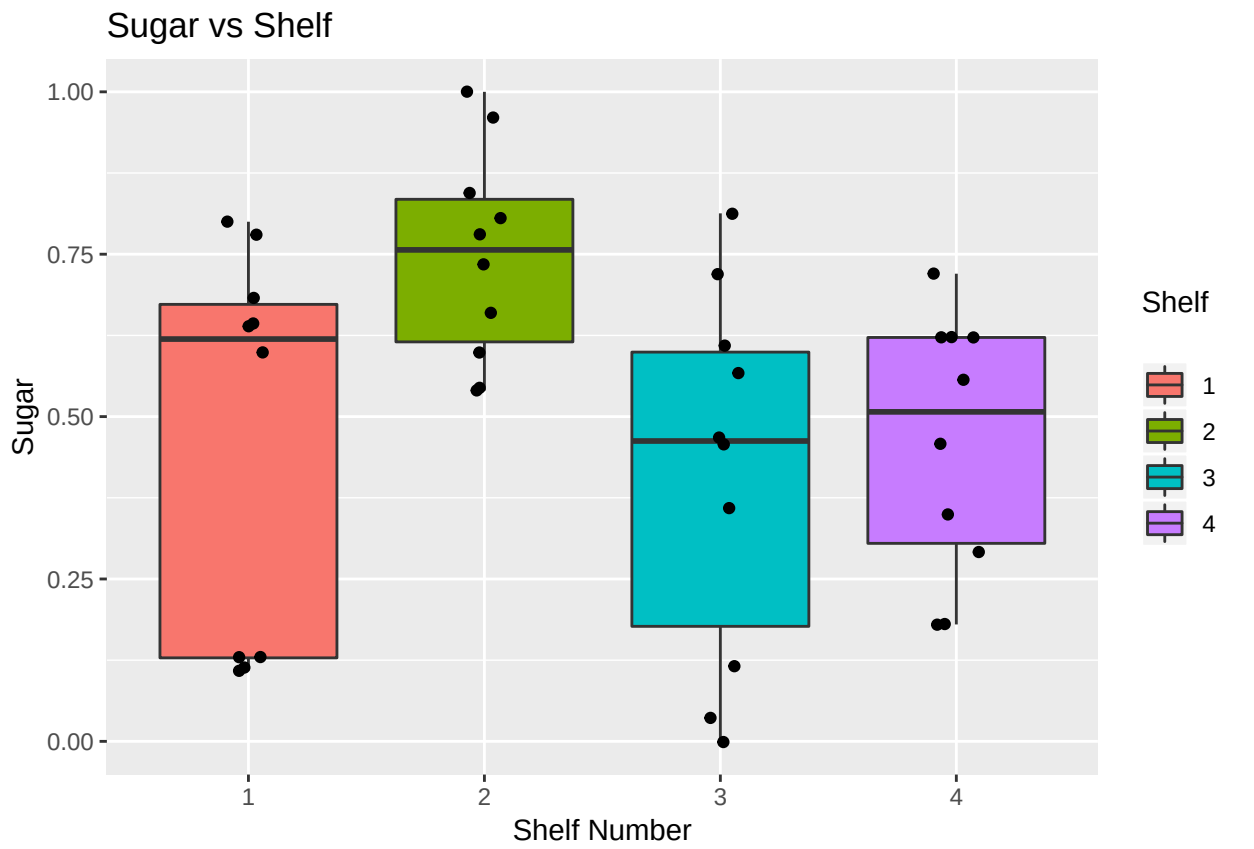
```
rescale <- function(x){(x - min(x))/(max(x)-min(x))}

df <- data.frame(Shelf=raw_data$Shelf,
                     sugar=rescale(raw_data$sugar_g / raw_data$size_g),
                     fat=rescale(raw_data$fat_g / raw_data$size_g),
                     sodium=rescale(raw_data$sodium_mg / raw_data$size_g))
summary(df)
```

```
##      Shelf          sugar             fat             sodium
##  Min.   :1.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.75   1st Qu.:0.3339   1st Qu.:0.1582   1st Qu.:0.4200
##  Median :2.50   Median :0.6000   Median :0.3542   Median :0.5354
##  Mean   :2.50   Mean   :0.5209   Mean   :0.3476   Mean   :0.5240
##  3rd Qu.:3.25   3rd Qu.:0.7200   3rd Qu.:0.5400   3rd Qu.:0.6696
##  Max.   :4.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

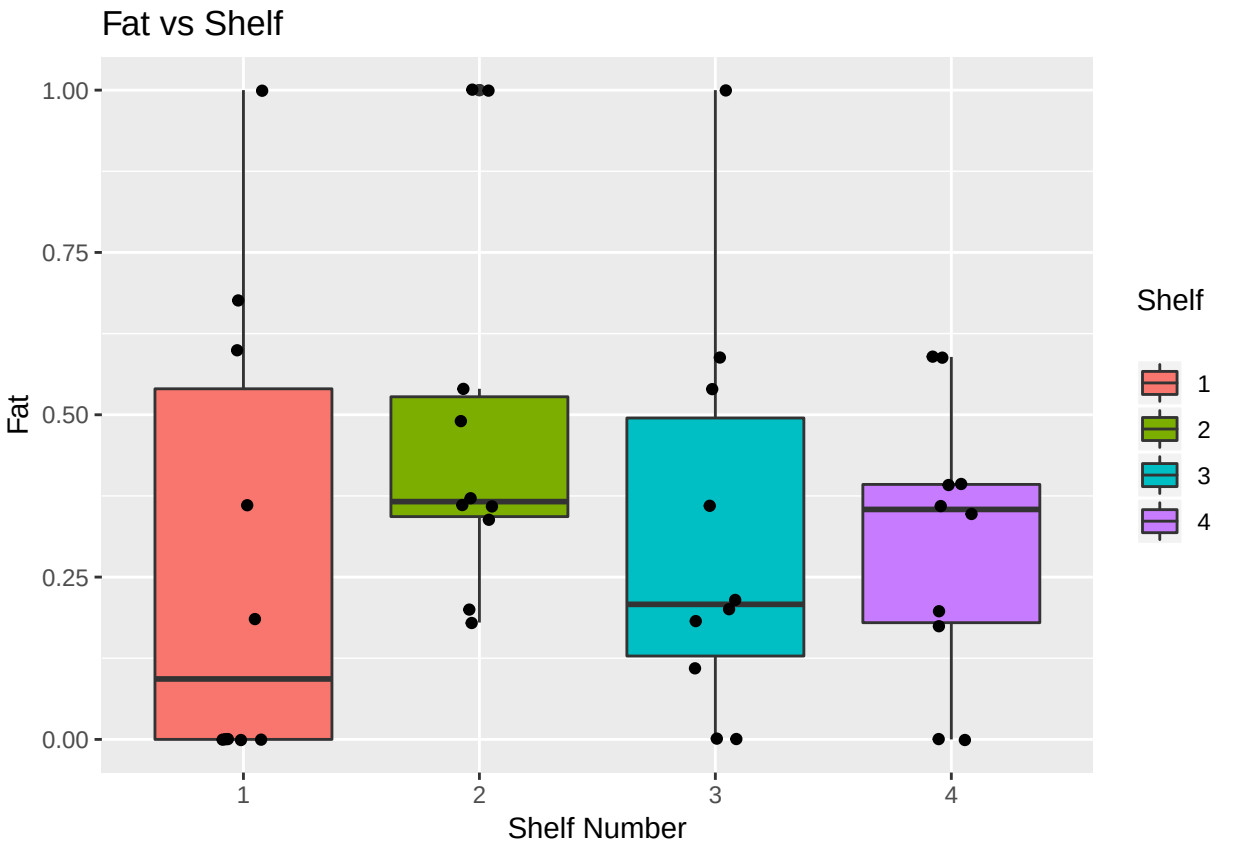
Part B

B.1 - Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables

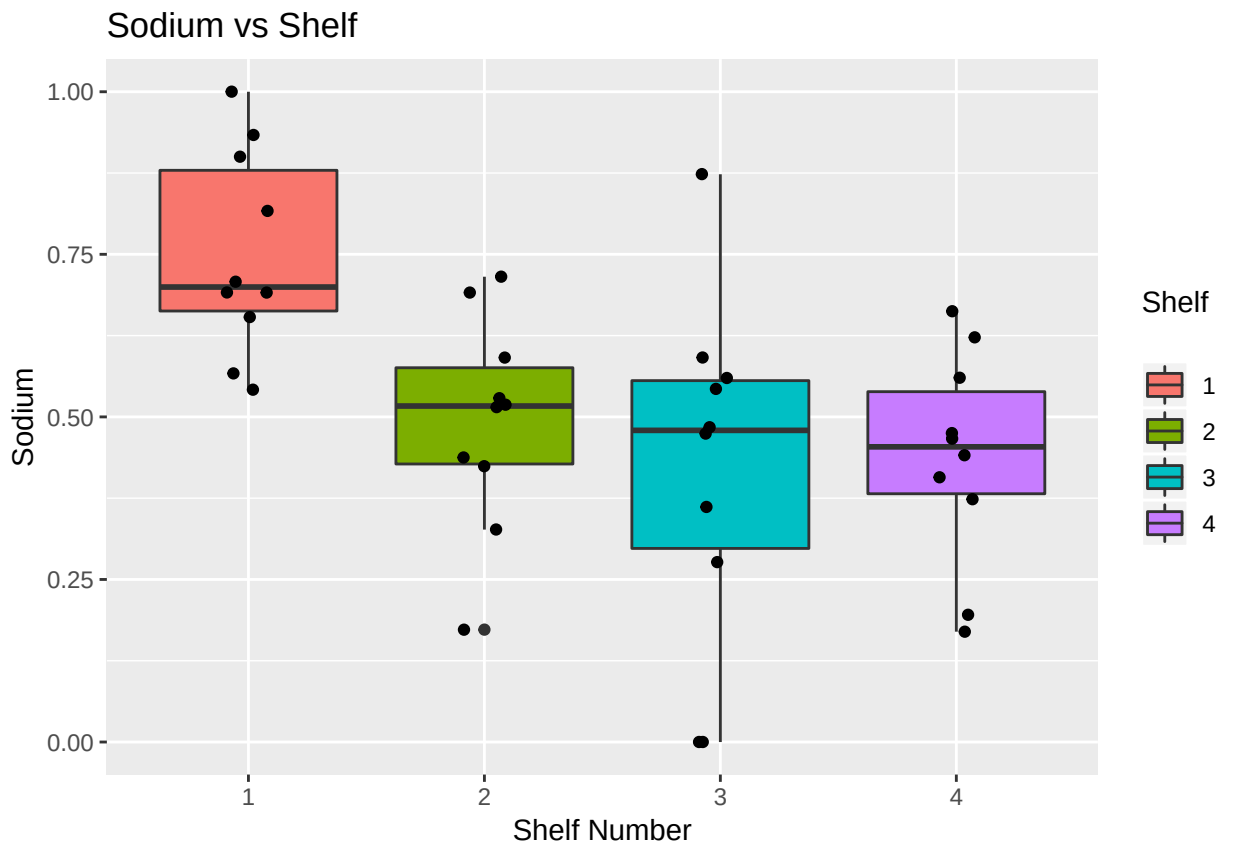
```
p1 <- ggplot(df, aes(factor(Shelf),sugar)) + geom_boxplot(aes(fill=factor(Shelf))) + geom_jitter(width =
p2 <- ggplot(df, aes(factor(Shelf),fat)) + geom_boxplot(aes(fill=factor(Shelf))) + geom_jitter(width = (
p3 <- ggplot(df, aes(factor(Shelf),sodium)) + geom_boxplot(aes(fill=factor(Shelf))) + geom_jitter(width
p1
```


Sugar vs Shelf
1.00
0.75
0.50
0.25
0.00
Sugar
1
2
3
4
Shelf Number
Shelf
1
2
3
4

```
p2
```


Fat vs Shelf
1.00
0.75
0.50
0.25
0.00
Fat
1
2
3
4
Shelf Number
Shelf
1
2
3
4

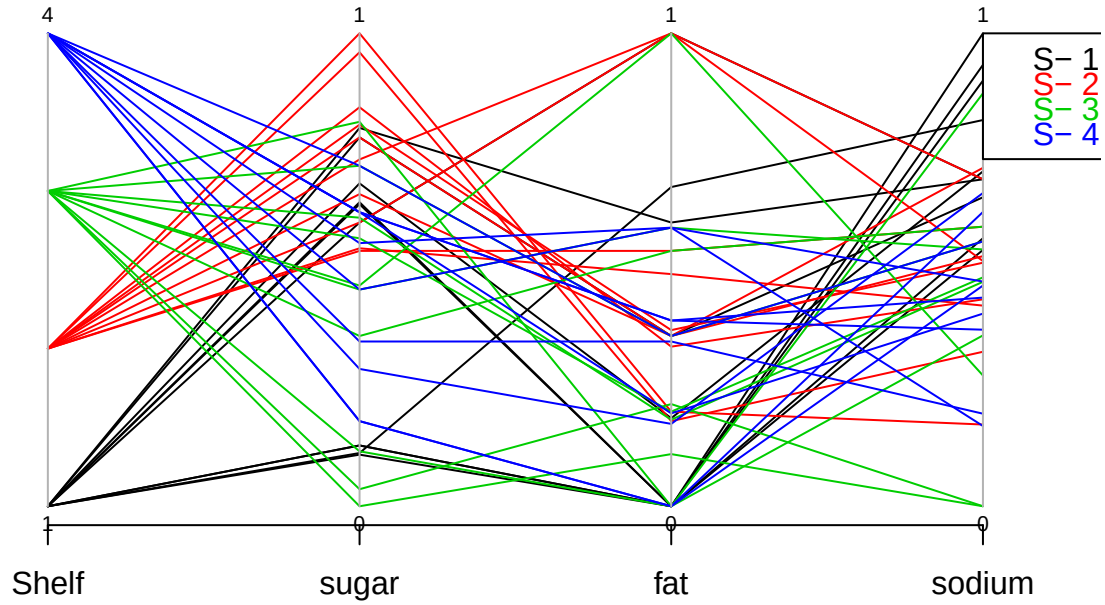
```
p3
```



B.2 - Construct a *parallel coordinates plot* for the explanatory variables and the shelf number

Discuss if possible content differences exist among the shelves.

```
parcoord(df,col=c(as.numeric(df$Shelf)),var.label=TRUE)
par(mar=c(1.1, 1.1, 1.1, 1.1), xpd=TRUE)
#par(xpd = T, mar = par()$mar + c(0,0,0,5))
legend(4, 1, legend = c('S- 1', 'S- 2', 'S- 3','S- 4'),
       text.col = c(1, 2,3,4))
```



The only clear content difference among shelves is for sugar. It appears that the lower shelves, and in particular shelf 2, have high sugar content. The relationship between shelves and fat and sodium seems more mixed.

Part C

The response has values of $1, 2, 3,$ *and* 4.

C.1 Under what setting would it be desirable to take into account ordinality?

Ordinality is used when there is a natural ordering to a categorical variable's levels. Here, ordinality would be helpful when all observations (i.e. cereals) are abiding by the same ordinality in terms of the value associated with each category. For example, with the shelves, if all cereal has the highest placement value associated with the bottom (1) shelf, and that value decrements with each subsequent shelf (i.e. shelves two (2) through four (4)), then ordinality is useful because of the natural ordering.

C.2 Do you think that this setting occurs here?

Though there is a natural ordering to shelf position/location, the logic above implies that the use of ordinal response is not desirable here. The cereals are not all of the same type (e.g. all children's cereals) so ordinality does not apply consistently. For example, from the "perspective" of an adult cereal shelf 4 is the highest ranked and shelf 1 is the lowest ranked. For a children's cereal, this order would be reveresed with shelf 1 having the highest rank and shelf 4 having the lowest.

Part D

D.1 Estimate a multinomial regression model with linear forms of the sugar, fat, and sodium variables

```
mod.fit <- multinom(formula= Shelf~ sugar + fat + sodium, data=df)
```

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

```
summary(mod.fit)
```

```
## Call:
## multinom(formula = Shelf ~ sugar + fat + sodium, data = df)
##
## 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
```

D.2 Perform LRTs to examine the importance of each explanatory variable

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

```
## 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
```

According to the Likelihood Ratio Test (LRT), the only two variables that are important given the other variables are in the model (for $\alpha = 0.05$) are `sugar` with a p-value of 0.0000452 and `sodium` with a p-value of 0.00000707. There is not sufficient evidence that, given the other variables are in the model, `fat` is important to include.

Part E - Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables)

```
mod.fit.interactions <- multinom(formula= Shelf~ sugar + fat + sodium + sugar:fat
                                 + sugar:sodium + sodium:fat + sugar:sodium:fat, data=df)
```

```
## # 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
```

```
Anova(mod.fit.interactions)
```

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

As can be seen above in the LRT, no interactions are signficant to threshold of $\alpha = 0.05$. The `sugar:fat` interaction has p-value of 0.357, `sugar:sodium` interaction has p-value of 0.388, `fat:sodium` has p-value of 0.368, and the three-way interaction of `sugar:fat:sodium` has p-value of 0.46. Therefore, we can say that, given the other variables are in the model, there is sufficient evidence that the interactions are not important.

Part F - Kellogg's Apple Jacks

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.

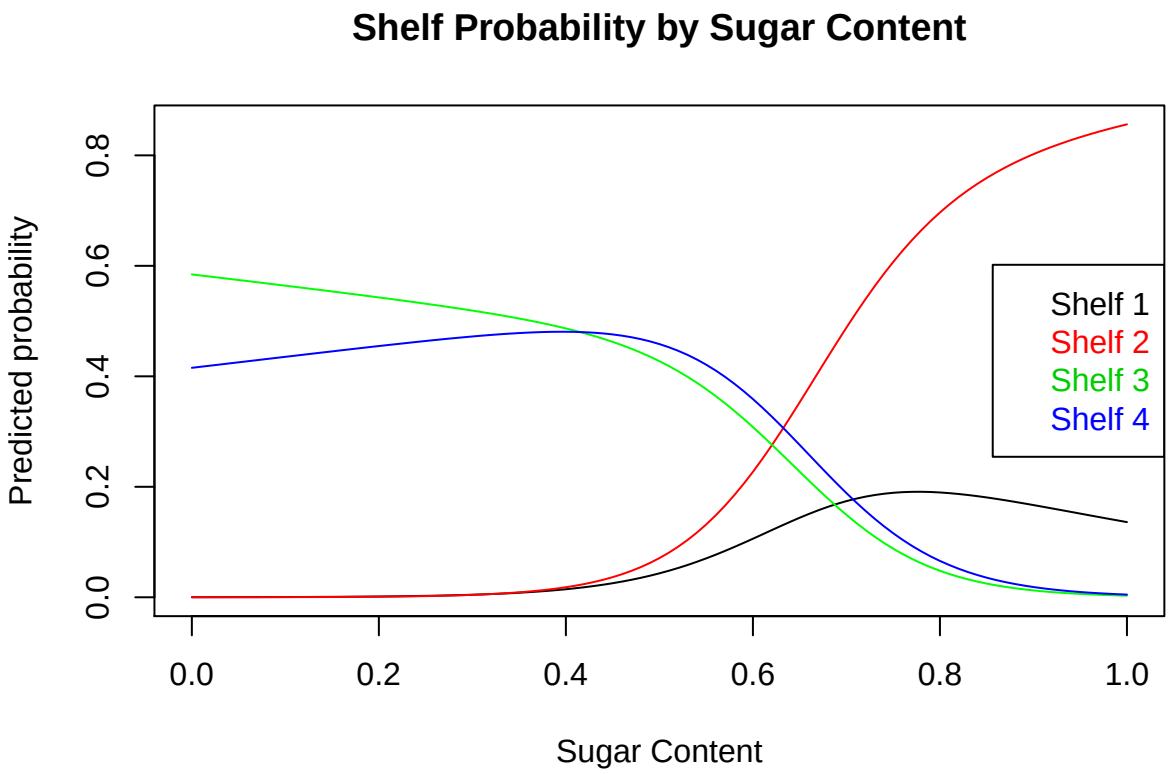
```
newdata <- data.frame(sugar=(12/28 - min(raw_data$sugar))/(max(raw_data$sugar)-min(raw_data$sugar))
                      , fat=(0.5/28 - min(raw_data$fat))/(max(raw_data$fat)-min(raw_data$fat)),
                      sodium=(130/28 - min(raw_data$sodium))/(max(raw_data$sodium)-min(raw_data$sodium))
pi.hat <- predict(object=mod.fit, newdata=newdata, type="probs")
data.frame("Prob of Shelf1"=pi.hat[1],"Prob of Shelf2"=pi.hat[2],"Prob of Shelf3"=pi.hat[3],"Prob of Sh
```

```
##   Prob.of.Shelf1 Prob.of.Shelf2 Prob.of.Shelf3 Prob.of.Shelf4
## 1   4.203155e-10   3.507451e-07      0.5918304      0.4081693
```

Part G

G.1 - 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.

```
x <- seq(0,1,0.01)
simulated.data <- data.frame(sugar=x,
                             fat=mean(df$fat),
                             sodium=mean(df$sodium))
pi.hat.sim <- predict(mod.fit, newdata = simulated.data, type = "probs")

plot.new()
plot(x, pi.hat.sim[, 1], type = "l", col = "black", ylim = range(min(pi.hat.sim),
    max(pi.hat.sim)), xlab = "Sugar Content", ylab = "Predicted probability",
    main = "Shelf Probability by Sugar Content")
lines(x, pi.hat.sim[, 2], col = "red")
lines(x, pi.hat.sim[, 3], col = "green")
lines(x, pi.hat.sim[, 4], col = "blue")
legend("right",legend = c('Shelf 1', 'Shelf 2', 'Shelf 3','Shelf 4'), text.col = c(1,2,3,4))
```

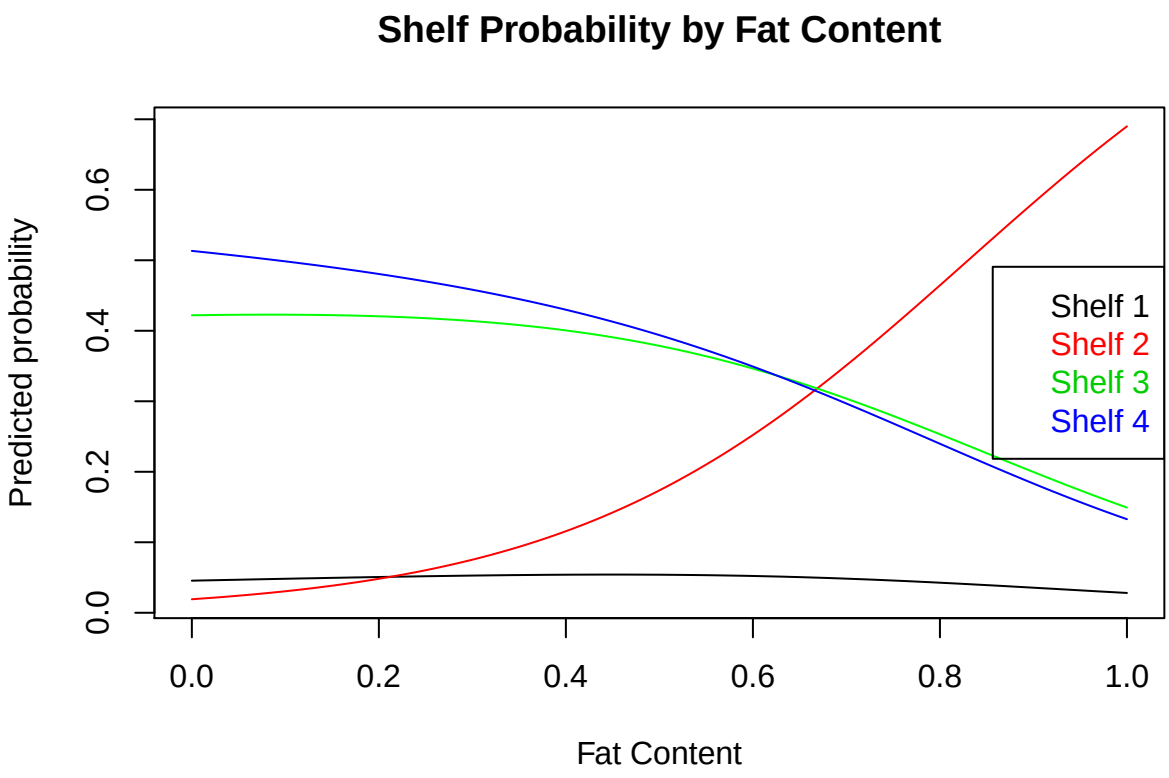


G.2 - Interpret the plot with respect to sugar content.

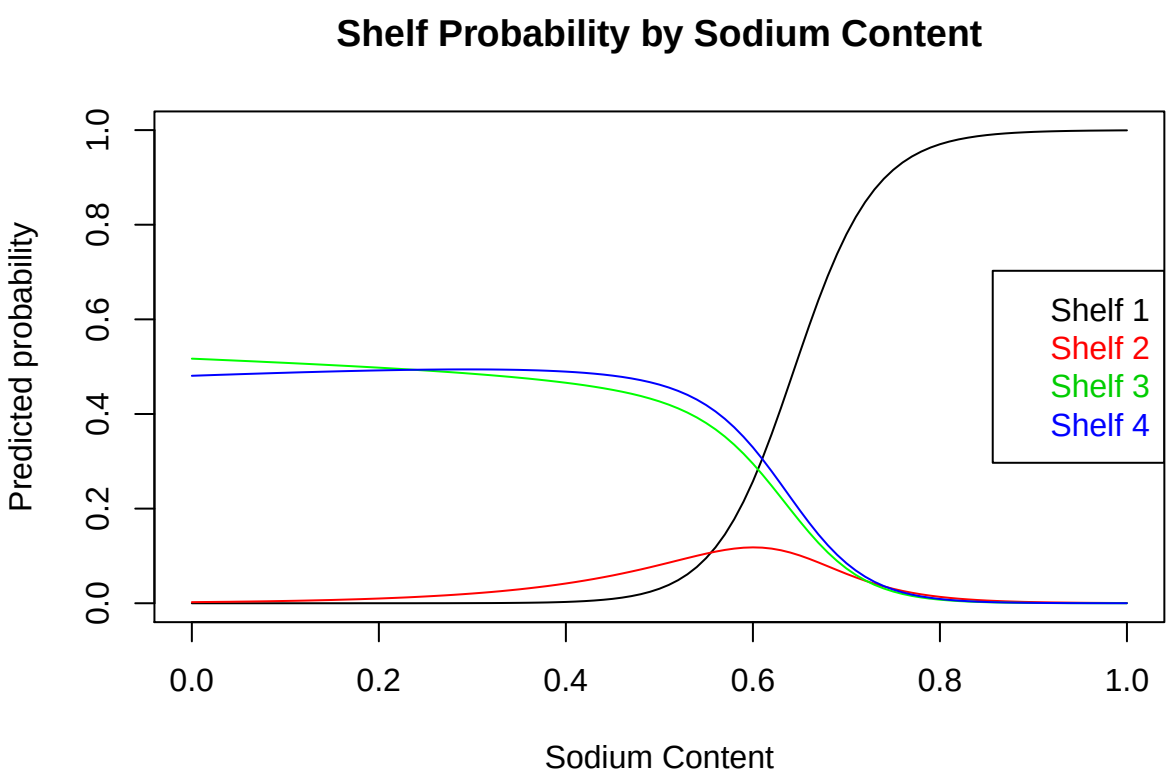
The plot shows that low sugar content implies a greater probability of assignment to a higher shelf (3,4) and high sugar content implies a greater probability of assignment to a lower shelf (1,2). This latter observation is especially true for shelf 2.

Below we also created comparable graphs for sodium and fat in order to help with visualizing Part F.

```
x <- seq(0,1,0.01)
simulated.data <- data.frame(sugar=mean(df$sugar),
                             fat=x,
                             sodium=mean(df$sodium))
pi.hat.sim <- predict(mod.fit, newdata = simulated.data, type = "probs")

plot.new()
plot(x, pi.hat.sim[, 1], type = "l", col = "black", ylim = range(min(pi.hat.sim),
    max(pi.hat.sim)), xlab = "Fat Content", ylab = "Predicted probability",
    main = "Shelf Probability by Fat Content")
lines(x, pi.hat.sim[, 2], col = "red")
lines(x, pi.hat.sim[, 3], col = "green")
lines(x, pi.hat.sim[, 4], col = "blue")
legend("right",legend = c('Shelf 1', 'Shelf 2', 'Shelf 3','Shelf 4'), text.col = c(1,2,3,4))
```




```
x <- seq(0,1,0.01)
simulated.data <- data.frame(sugar=mean(df$sugar),
                             fat=mean(df$fat),
                             sodium=x)
pi.hat.sim <- predict(mod.fit, newdata = simulated.data, type = "probs")

plot.new()
plot(x, pi.hat.sim[, 1], type = "l", col = "black", ylim = range(min(pi.hat.sim),
    max(pi.hat.sim)), xlab = "Sodium Content", ylab = "Predicted probability",
    main = "Shelf Probability by Sodium Content")
lines(x, pi.hat.sim[, 2], col = "red")
lines(x, pi.hat.sim[, 3], col = "green")
lines(x, pi.hat.sim[, 4], col = "blue")
legend("right",legend = c('Shelf 1', 'Shelf 2', 'Shelf 3','Shelf 4'), text.col = c(1,2,3,4))
```



Part 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.df <- apply(X=df[,c(2,3,4)], MARGIN=2, FUN=sd)
c.value <- sd.df
# coefficients(mod.fit)
beta.hat2 <- coefficients(mod.fit)[1 , 2:4]
beta.hat3 <- coefficients(mod.fit)[2 , 2:4]
beta.hat4 <- coefficients(mod.fit)[3 , 2:4]

conf.beta <- confint(object=mod.fit, level=0.95)
# data.frame("Odds 2 to 1"=round(exp(c.value*beta.hat2), 2))
ci.or2 <- round(exp(c.value * conf.beta[2:4,1:2,1]), 2)
# ci.or2
data.frame(estimateOR2=round(exp(c.value*beta.hat2), 2), standard_deviation=sd.df,lower=round(exp(c.valu
```

```
##         estimateOR2 standard_deviation lower upper
## sugar          2.06           0.2692078  0.14 29.68
## fat            3.37           0.2990292  0.87 13.04
## sodium         0.02           0.2298359  0.00  0.44
```

```
# data.frame("Odds 3 to 1" = round(exp(c.value*beta.hat3), 2)
ci.or3 <- round(exp(c.value * conf.beta[2:4,1:2,2]), 2)
# ci.or3
```

```
data.frame(estimateOR3=round(exp(c.value*beta.hat3), 2), standard_deviation=sd.df,lower=round(exp(c.valu
```

```
##         estimateOR3 standard_deviation lower upper
## sugar          0.04           0.2692078  0.00  0.49
## fat            0.85           0.2990292  0.21  3.49
## sodium         0.00           0.2298359  0.00  0.12
```

```
# data.frame("Odds 4 to 1" = round(exp(c.value*beta.hat4), 2)
ci.or4 <- round(exp(c.value * conf.beta[2:4,1:2,3]), 2)
# ci.or4
data.frame(estimateOR4=round(exp(c.value*beta.hat4), 2), standard_deviation=sd.df,lower=round(exp(c.valu
```

```
##         estimateOR4 standard_deviation lower upper
## sugar          0.05           0.2692078  0.00  0.61
## fat            0.77           0.2990292  0.19  3.16
## sodium         0.00           0.2298359  0.00  0.13
```

Sugar

The most significant odds of a shelf vs. shelf 1 for sugar is the estimated odds of a cereal being on shelf 2 vs. shelf 1. The estimated odds change by 2.06 times for a 0.26 change in the rescaled proportion (i.e. a shift in proportion of sugar scaled for the range of sugar by serving in this dataset) holding all other variables constant. By doing the simple linear transformation back to the nominal proportion of sugar in a serving size, we can also interpret this odds ratio as the estimated odds of a cereal being on shelf 2 vs. shelf 1 change by 2.06 times for a 0.15 change in the proportion of sugar in a serving size. For example, if a cereal with 2 grams of sugar in a serving size of 10 grams and another cereal has 3.5 grams of sugar in a serving size of 10 grams, it has just over 2x estimated odds of being on shelf 2 than shelf 1 over the first cereal holding all other variables constant.

This odds ratio is substantiated with the graph created in part G, as we can see that for every 0.2 increase in the rescaled proportion of sugar, there is approximately double the odds that the cereal will appear on shelf 2 vs. shelf 1.

Fat

Again, we see that the estimated odds ratio for shelf 2 vs. shelf 1 is the most noticeable. For a 0.299 change in the rescaled proportion (same idea as with sugar), the estimated odds change by 3.37 times for shelf 2 vs. shelf 1 holding all other variables constant. Again, we can do a linear transformation back to the nominal proportion of fat in a serving to interpret this as the estimated odds of a cereal appearing on shelf 2 vs. shelf 1 change by 3.37 times for a 0.027 change in the proportion of fat in a serving while holding all other variables constant.

This odds ratio is also substantiated with the graph created in part G. For example, for shelf 2, starting with an estimated probability of 0.05 with a fat content (proportion) of 0.2, we see that the estimated probability is just under 0.2 when fat content is 0.5 (delta of 0.3), effectively just under quadruple the original propability.

Sodium

Shelf 1 dominates for cereals with increased sodium content, therefore with increases in sodium, we do not see much of an change in estimated odds of a cereal appearing on anything but shelf 1. Consider the estimated odds of a cereal appearing on shelf 2 vs. shelf 1. For an increase of 0.22 of the rescaled milligrams/gram serving proportion, the estimated odds change by 0.02 times holding all other variables constant. Applying the same unscaling linear transformation, we can also interpret this odds ratio as the estimated odds of a cereal appearing on shelf 2 vs. shelf 1 changes by 0.02 times for an increase in 2.46 mg/g. Or in other words, a proportion change of 0.00246 sodium in a serving, holding all other variables constant.

This low variability in estimated odds is observed in the graph created in part G. For cereals between 0.0 and 0.6 mg/g serving of sodium, there is almost no change in the probability of on which shelf it appears.

Then, at approximately 0.6 mg/g serving, there is a large shift where the estimated probability changes for all shelves and remains fairly constant through to the limit of 1.

Note: Below is the function used to "unscale" the proportions to put the estimated odds ratios back in terms of the de-normalized, original proportions

```
unscale <- function(x, y) {x*(max(y)-min(y))+min(y)}
unscale(0.2298359, (raw_data$sodium_mg/raw_data$size_g))
```

```
## [1] 2.462527
```