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

Harry Xu and Rich Ung

Introduction

Strategic Placement of Products in Grocery Stores

We receive the following prompt from **Question 12 of chapter 3 (on page 189 and 190)** within Bilder and Loughin's *"Analysis of Categorical Data with R"*:

`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.`

By using multicategory response models, we can maximize sales by placing the cereal in the shelf that best fits its attributes (such as sugar, fat, and sodium content).

Exploratory Data Analysis

In order to perform our analyses, we first load the required R libraries:

```
knitr::opts_chunk$set(echo = TRUE, tidy.opts=list(width.cutoff=60), tidy=TRUE)
library(package = MASS)
library(package = car)
library(package = nnet)
```

Next, we load the data and perform a quick EDA:

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

```
##   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
```

```
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 ...
```

```
summary(cereal)
```

```
##        ID            Shelf                                      Cereal
##  Min.   : 1.00   Min.   :1.00   Capn Crunch's Peanut Butter Crunch: 2
##  1st Qu.:10.75   1st Qu.:1.75   Food Club Toasted Oats            : 2
##  Median :20.50   Median :2.50   Basic 4                           : 1
##  Mean   :20.50   Mean   :2.50   Capn Crunch                       : 1
##  3rd Qu.:30.25   3rd Qu.:3.25   Cinnamon Grahams                  : 1
##  Max.   :40.00   Max.   :4.00   Cocoa Pebbles                     : 1
##                                 (Other)                           :32
##      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
##
```

We can see that our dataset has 40 observations of 7 variables, with no missing values for any of our variables. It looks like each row corresponds to a particular cereal, where "Cereal" contains the name of the cereal and "Shelf" contains the shelf that the cereal is located. We can also see that we have the serving size, sugar, fat, and sodium contents for each cereal.

Modeling & Questions

We go through our modeling as we go through parts a through h within this question:

Part A

Question

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.[12] Below is code we use to re-format the data after the data file is read into an object named `cereal`:

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

Answer

By executing the code chunk above, we have already re-formatted the dataset by dividing each explanatory variable by its serving size to account for the different serving sizes among the cereals, and re-scaling each variable to be between 0 and 1 in order to help with the convergence of parameter estimates.

Part B

Question

Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables. Below is code that can be used for plots involving sugar:

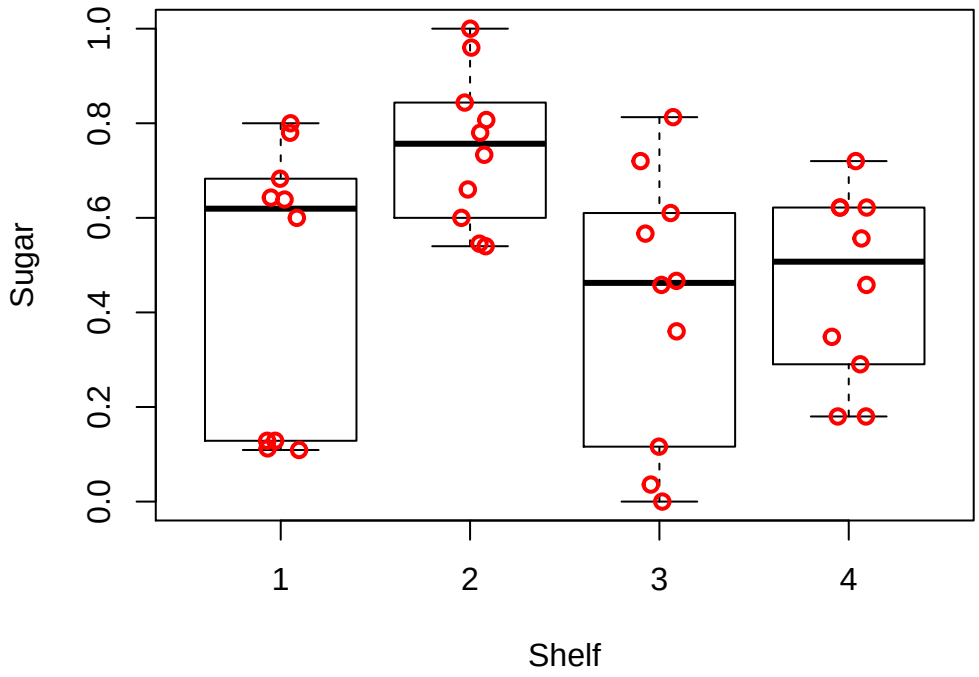
```
boxplot(formula = sugar ~ Shelf, data = cereal2, ylab = "Sugar",
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sugar ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```

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

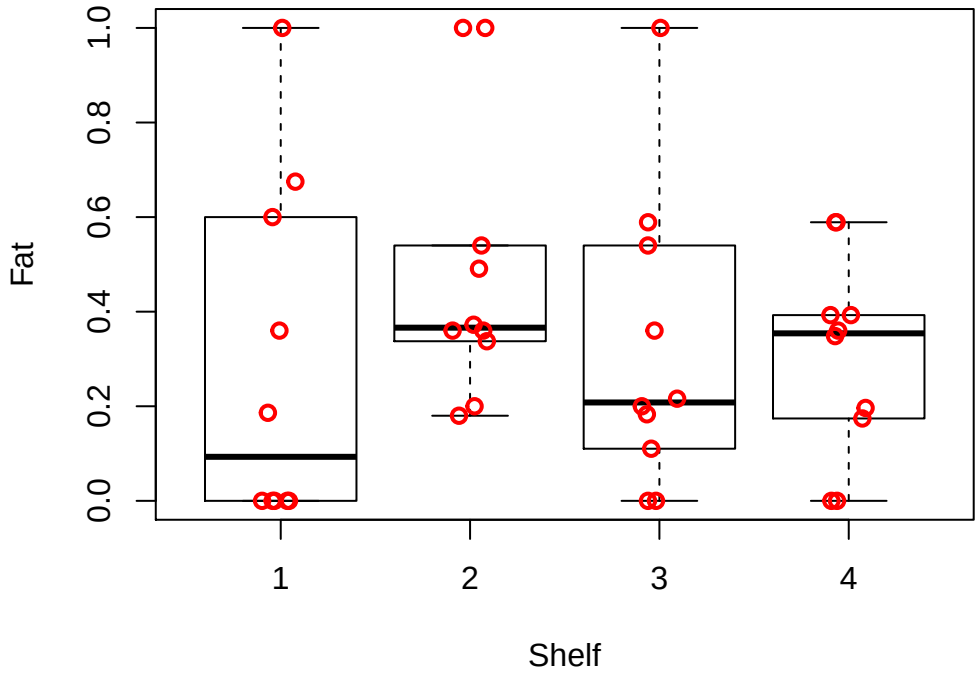
Answer

Below are the side-by-side box plots with dot plots overlaid for each of the explanatory variables:

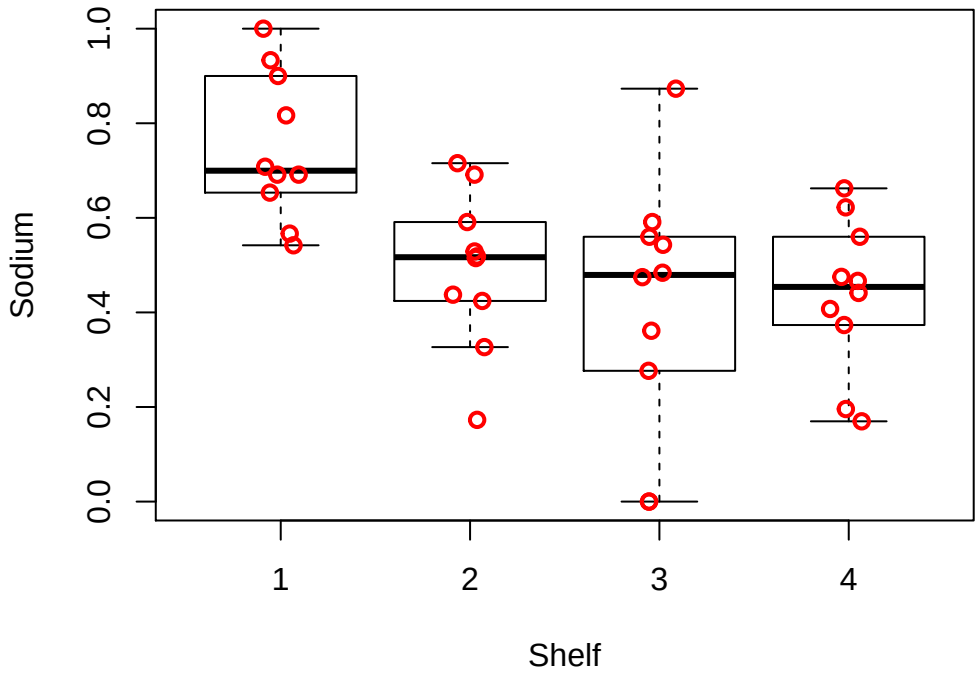
```
boxplot(formula = sugar ~ Shelf, data = cereal2, ylab = "Sugar",
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sugar ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```




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

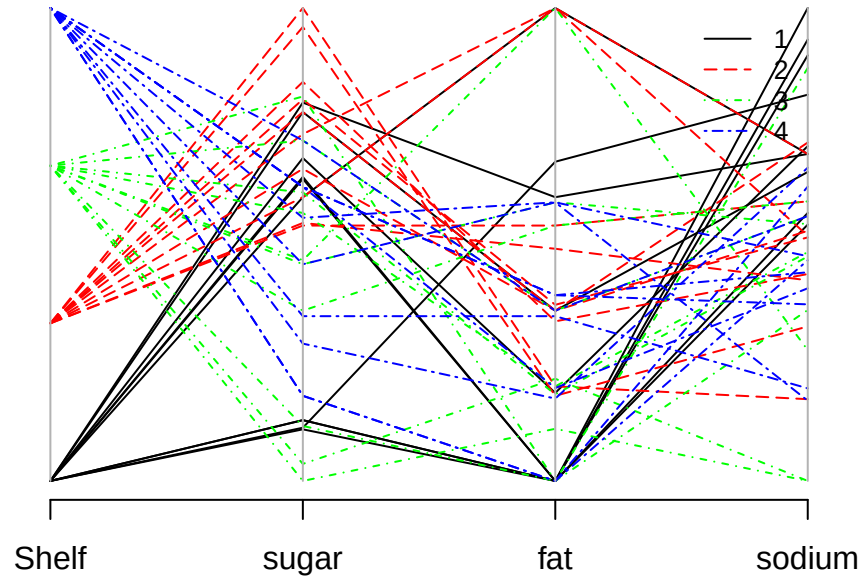



```
boxplot(formula = sodium ~ Shelf, data = cereal2, ylab = "Sodium",
    xlab = "Shelf", pars = list(outpch = NA))
stripchart(x = cereal2$sodium ~ cereal2$Shelf, lwd = 2, col = "red",
    method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```



Below is a parallel coordinates plot for the explanatory variables and the shelf number:

```
cereal2.colors <- ifelse(test = cereal2$Shelf == 1, yes = "black",
    no = ifelse(test = cereal2$Shelf == 2, yes = "red", ifelse(test = cereal2$Shelf ==
        3, yes = "green", no = "blue")))

cereal2.lty <- ifelse(test = cereal2$Shelf == 1, yes = "solid",
    no = ifelse(test = cereal2$Shelf == 2, yes = "longdash",
        ifelse(test = cereal2$Shelf == 3, yes = "dotdash", no = "twodash")))

parcoord(x = cereal2, col = cereal2.colors, lty = cereal2.lty)  # Plot
legend(x = 3.5, y = 1, legend = c("1", "2", "3", "4"), lty = c("solid",
    "longdash", "dotdash", "twodash"), col = c("black", "red",
    "green", "blue"), cex = 0.8, bty = "n")
```



There appears to be some content difference clustered by shelves. As the above parallel coordinates plot shows, Shelf 4 Cereals appear to have the lowest Sugar, Fat and Sodium content – which suggest this shelf contains "healthier" types of cereals. Shelf 1 appears to have the highest Sugar content and perhaps the most "unhealthiest" cereals. This appears to be the case by examining the brands of the Cereals placed on Shelf 4 vs. brands of Cereals placed on Shelf 1

```
cereal[cereal$Shelf == 4, ]
```

```
##    ID Shelf                                     Cereal size_g sugar_g
## 31 31     4                         Total Raisin Bran     55      19
## 32 32     4                     Food Club Wheat Crunch     60       6
## 33 33     4                       Oatmeal Crisp Raisin     55      19
## 34 34     4                      Food Club Bran Flakes     31       5
## 35 35     4                                Cookie Crisp     30      12
## 36 36     4                Kellogg's All Bran Original      31       6
## 37 37     4                   Food Club Low Fat Granola     55      14
## 38 38     4                  Oatmeal Crisp Apple Cinnamon     55      19
## 39 39     4 Post Fruit and Fiber - Dates, Raisons, Walnuts     55      17
## 40 40     4                         Total Corn Flakes     30       3
##    fat_g sodium_mg
## 31   1.0       240
## 32   0.0       300
## 33   2.0       220
## 34   0.5       220
## 35   1.0       180
## 36   1.0        65
## 37   3.0       100
## 38   2.0       260
## 39   3.0       280
## 40   0.0       200
```

```
cereal[cereal$Shelf == 1, ]
```

```
##    ID Shelf                              Cereal size_g sugar_g fat_g
## 1   1     1 Kellog's Razzle Dazzle Rice Crispies     28      10   0.0
## 2   2     1            Post Toasties Corn Flakes     28       2   0.0
## 3   3     1                Kellogg's Corn Flakes     28       2   0.0
## 4   4     1               Food Club Toasted Oats     32       2   2.0
## 5   5     1                     Frosted Cheerios     30      13   1.0
## 6   6     1             Food Club Frosted Flakes     31      11   0.0
## 7   7     1                           Capn Crunch     27      12   1.5
## 8   8     1   Capn Crunch's Peanut Butter Crunch     27       9   2.5
## 9   9     1                      Post Honeycomb      29      11   0.5
## 10 10     1               Food Club Crispy Rice     33       2   0.0
##    sodium_mg
## 1        170
## 2        270
## 3        300
## 4        280
## 5        210
## 6        180
## 7        200
## 8        200
## 9        220
## 10       330
```

Part C

Question

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

It would be desirable to take into account ordinality when the variable has a natural ordering to their levels. In other words, it would be desirable to take into account ordinality if response levels can be arranged so that category 1 $<$ category 2 $<$ $\cdots$ $<$ category J in some conceptual scale of measurement (e.g., amount of agreement). Since the shelf has a natural ordering to their levels, bottom (1) to top (4), it would make sense to take into account ordinality.

Part D

Question

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.

Answer

```
mod.fit.ord <- polr(formula = as.factor(Shelf) ~ sugar + fat +
    sodium, data = cereal2, method = "logistic")
summary(mod.fit.ord)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = as.factor(Shelf) ~ sugar + fat + sodium, data = cereal2,
##     method = "logistic")
##
## Coefficients:
##           Value Std. Error  t value
## sugar  -1.61101     1.2830 -1.25565
## fat    -0.05123     0.9657 -0.05305
## sodium -4.85950     1.6302 -2.98094
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -4.7534  1.4837    -3.2037
## 2|3 -3.3435  1.3810    -2.4210
## 3|4 -1.9823  1.2867    -1.5407
##
## Residual Deviance: 98.52912
## AIC: 110.5291
```

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

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##        LR Chisq Df Pr(>Chisq)
## sugar    1.6794  1  0.1950069
## fat      0.0028  1  0.9577007
## sodium  11.5685  1  0.0006708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the LRT for the multionmial regression model show that the only significant variable appears to be Sodium in predicting which shelf the Cereal will be placed on. Because of the large test statistic value for sodium, there is sufficient evidence that sodium is an important explanatory variable. Even though the other variables are statistically insignificant, the test for sodium is conditional on the other variables being in the model.

Part E

Question

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

Answer

```
mod.fit.ord2 <- polr(formula = as.factor(Shelf) ~ sugar + fat +
    sodium + sugar:fat + sugar:sodium + fat:sodium + sugar:fat:sodium,
    data = cereal2, method = "logistic")
Anova(mod.fit.ord2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.factor(Shelf)
##                  LR Chisq Df Pr(>Chisq)
## sugar              1.1760  1  0.2781685
## fat                0.0419  1  0.8377311
## sodium            11.1699  1  0.0008314 ***
## sugar:fat          0.1014  1  0.7501457
## sugar:sodium       0.3945  1  0.5299556
## fat:sodium         0.2607  1  0.6096643
## sugar:fat:sodium   0.1077  1  0.7427907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above Anova test after adding in all the different interactions between the explanatory variables, we can see that none of the interactions have a statistically significant result (in fact, all the interactions have a p-value of above 0.50). This shows that there are no significant interactions among the explanatory variables (including the interaction among all three variables).

Part F

Question

Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed to- ward 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

```
predict(object = mod.fit.ord, newdata = data.frame(sugar = (((12/28) -
    min(cereal$sugar_g))/(max(cereal$sugar_g) - min(cereal$sugar_g))),
    fat = (((0.5/28) - min(cereal$fat_g))/(max(cereal$fat_g) -
        min(cereal$fat_g))), sodium = (((130/28) - min(cereal$sodium_mg))/(max(cereal$sodium_mg
        min(cereal$sodium_mg)))), type = "probs")
```

```
##           1           2           3           4
## 0.009468087 0.028204574 0.094803728 0.867523610
```

By predicting the shelf probabilities based on the serving size, sugar content, fat content, and sodium content of Apple Jacks cereal, we can see that there is about a 0.9% chance of it being on shelf 1, a 2.8% chance of it being on shelf 2, a 9.5% chance of it being on shelf 3, and a 86.8% chance of it being on shelf 4. We can conclude that Apple Jacks has a higher probability of appearing on the higher shelves, with the highest probability of it being on the top (4) shelf.

Part G

Question

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.

Answer

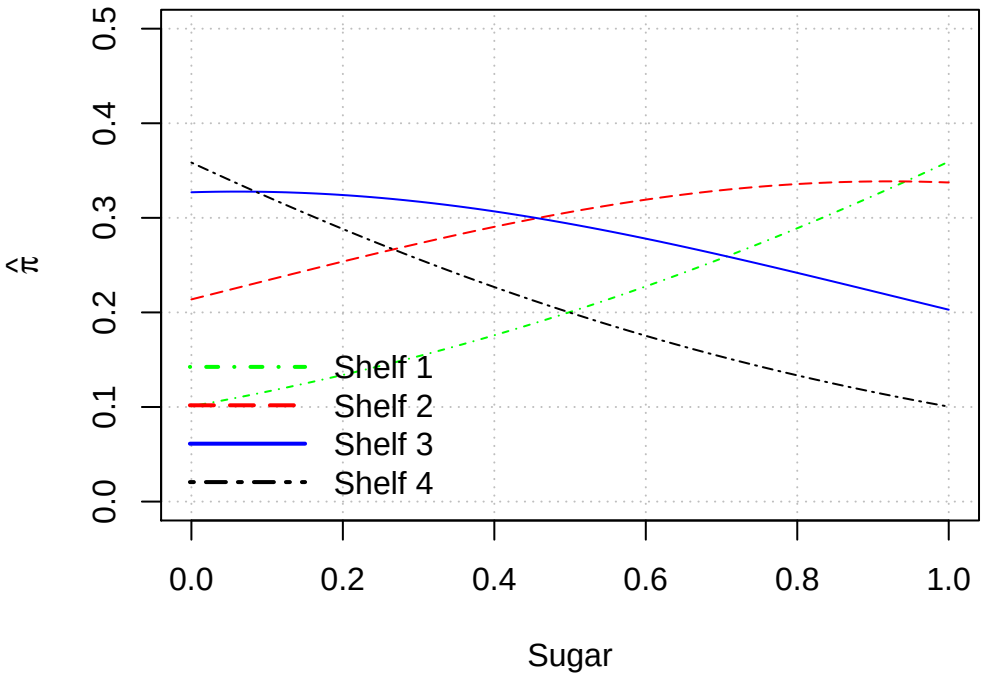
```
mod.fit.sugar <- polr(formula = as.factor(Shelf) ~ sugar + sodium +
    fat, data = cereal2, method = "logistic")
summary(mod.fit.sugar)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = as.factor(Shelf) ~ sugar + sodium + fat, data = cereal2,
##     method = "logistic")
##
## Coefficients:
##           Value Std. Error  t value
## sugar  -1.61101     1.2830 -1.25565
## sodium -4.85950     1.6302 -2.98094
## fat    -0.05123     0.9657 -0.05305
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -4.7534  1.4837    -3.2037
## 2|3 -3.3435  1.3810    -2.4210
## 3|4 -1.9823  1.2867    -1.5407
##
## Residual Deviance: 98.52912
## AIC: 110.5291
```

```
beta.hat <- c(-mod.fit.ord$coefficients, mod.fit.ord$zeta)

curve(1/(1 + exp(beta.hat[3] + beta.hat[1] * x) + exp(beta.hat[4] +
    beta.hat[1] * x)), ylab = expression(hat(pi)), xlab = "Sugar",
    ylim = c(0, 0.5), xlim = c(min(cereal2$sugar), max(cereal2$sugar)),
    col = "black", lty = "solid", lwd = 2, n = 1000, type = "n",
```

```
    panel.first = grid(col = "gray", lty = "dotted"))

# Shelf 1
curve(expr = plogis(q = mod.fit.sugar$zeta[1] - mod.fit.sugar$coefficients[1] *
    x - mod.fit.sugar$coefficients[2] * mean(cereal2$sodium) -
    mod.fit.sugar$coefficients[3] * mean(cereal2$fat)), col = "green",
    type = "l", add = TRUE, lty = "dotdash", n = 1000)

# Shelf 2
curve(expr = plogis(q = mod.fit.sugar$zeta[2] - mod.fit.sugar$coefficients[1] *
    x - mod.fit.sugar$coefficients[2] * mean(cereal2$sodium) -
    mod.fit.sugar$coefficients[3] * mean(cereal2$fat)) - plogis(q = mod.fit.sugar$zeta[1] -
    mod.fit.sugar$coefficients[1] * x - mod.fit.sugar$coefficients[2] *
    mean(cereal2$sodium) - mod.fit.sugar$coefficients[3] * mean(cereal2$fat)),
    col = "red", type = "l", add = TRUE, lty = "longdash", n = 1000)

# Shelf 3
curve(expr = plogis(q = mod.fit.sugar$zeta[3] - mod.fit.sugar$coefficients[1] *
    x - mod.fit.sugar$coefficients[2] * mean(cereal2$sodium) -
    mod.fit.sugar$coefficients[3] * mean(cereal2$fat)) - plogis(q = mod.fit.sugar$zeta[2] -
    mod.fit.sugar$coefficients[1] * x - mod.fit.sugar$coefficients[2] *
    mean(cereal2$sodium) - mod.fit.sugar$coefficients[3] * mean(cereal2$fat)),
    col = "blue", type = "l", add = TRUE, lty = "solid", n = 1000)

# Shelf 4
curve(expr = 1 - plogis(q = mod.fit.sugar$zeta[3] - mod.fit.sugar$coefficients[1] *
    x - mod.fit.sugar$coefficients[2] * mean(cereal2$sodium) -
    mod.fit.sugar$coefficients[3] * mean(cereal2$fat)), col = "black",
    type = "l", add = TRUE, lty = "twodash", n = 1000)

legend(x = "bottomleft", legend = c("Shelf 1", "Shelf 2", "Shelf 3",
    "Shelf 4"), lty = c("dotdash", "longdash", "solid", "twodash"),
    col = c("green", "red", "blue", "black"), bty = "n", lwd = c(2,
        2, 2, 2), seg.len = 4)
```



By holding the Sodium and Fat variables constant at their mean values, we can plot the probability of being on each Shelf (1-4) relative to the explanatory variable Sugar. As we noted before, it appears that Shelf 4 contains the "healthy" cereals whereas Self 1 contains the "unhealthy" cereals. We can see that in the above probability curves since as the Sugar variable increases, the probability of being on Shelf 4 decreases while the probability of being on Shelf 1 increases.

Part H

Question

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

Answer

First, we take a look at the stanard deviations of the explanatory variables in order to find an appropriate c-value for our odds ratios interpretation and calculation:

```
# Obtain standard deviation of the explanatory variables and
# normalize the values to obtain c-values used for the odd
# ratios and confidence intervals
(sd.cereal <- round(apply(X = cereal[, -c(1, 2, 3, 4)], MARGIN = 2,
    FUN = sd)))
```

```
##   sugar_g     fat_g sodium_mg
##         6         1        82
```

Looking at the above standard deviations, we decided to calculate odds ratios and corresponding confidence intervals according to a 5g increase in sugar per serving, a 1g increase in fat per serving,

and 100mg increease in sodium per serving. This would provide a clearer interpretation with an increase that's close to an explanatory variable's standard deviation.

```
sugar_1g_conv <- ((5 - min(cereal$sugar_g))/(max(cereal$sugar_g) -
    min(cereal$sugar_g)))
fat_1g_conv <- ((1 - min(cereal$fat_g))/(max(cereal$fat_g) -
    min(cereal$fat_g)))
sodium_1mg_conv <- ((100 - min(cereal$sodium_mg))/(max(cereal$sodium_mg) -
    min(cereal$sodium_mg)))
c.value <- c(sugar_1g_conv, fat_1g_conv, sodium_1mg_conv)

# Gather Odds Ratios
round(exp(c.value * (-mod.fit.ord$coefficients)), 2)
```

```
##  sugar    fat sodium
##   1.50   1.01   4.36
```

```
# Calculate Confidence Intervals for Odds Ratios
conf.beta <- confint(object = mod.fit.ord, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##
## Re-fitting to get Hessian
```

```
ci <- exp(c.value * (-conf.beta))
round(data.frame(low = ci[, 2], up = ci[, 1]), 2)
```

```
##         low    up
## sugar  0.82  2.93
## fat    0.69  1.48
## sodium 1.80 12.88
```

Based on the above results, we interpret the following ratios:

- The estimated odds of shelf location being below a particular level change by about 1.50 times for a 5g increase in sugar per serving size, holding the other variables constant.
- The estimated odds of shelf location being below a particular level change by about 1.01 times for a 1g increase in fat per serving size, holding the other variables constant.
- The estimated odds of shelf location being below a particular level change by about 4.36 times for a 100mg increase in sodium per serving size, holding the other variables constant.
- With 95% confidence, the odds of shelf location being below a particular level change by 0.82 to 2.93 times when sugar is increased by 5g per serving size, holding the other variables constant.
- With 95% confidence, the odds of shelf location being below a particular level change by 0.69 to 1.48 times when fat is increased by 1g per serving size, holding the other variables constant.
- With 95% confidence, the odds of shelf location being below a particular level change by 1.80 to 12.88 times when sodium is increased by 100mg per serving size, holding the other variables constant.

This relates to the plots constructed because...

Conclusion

In conclusion, we estimated a ordinal response regression model in order to conduct strategic placement of cereal boxes in grocery stores. We chose an ordinal response regression model because our response variable, `Shelf`, is a categorical variable with 4 different categories, which also has a natural ordering to their levels. Using this model, we are able to recommend a shelf based on the characteristics of the cereal, such as the serving size, sugar content, fat content, and sodium content. This would allow us to better place a cereal box on a shelf where we can maximize revenue and increase profit.