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

Aditi Khullar
Daniel Vanlunen
XT Nguyen

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.

setup

```
d <- read.csv(file = 'cereal_dillons.csv')
d$Shelf_factor <- as.factor(d$Shelf)
str(d)
```

```
## 'data.frame':    40 obs. of  8 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",..: 22 34 18 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 ...
##  $ Shelf_factor: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

```
describe(d)
```

```
## d
##
##  8  Variables      40  Observations
## ---------------------------------------------------------------------------
## ID
```

```
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       40        1     20.5    13.67     2.95     4.90
##      .25      .50      .75      .90      .95
##    10.75    20.50    30.25    36.10    38.05
##
## lowest :  1  2  3  4  5, highest: 36 37 38 39 40
## ------------------------------------------------------------------------
## Shelf
##        n  missing distinct     Info     Mean      Gmd
##       40        0        4    0.938      2.5    1.282
##
## Value          1    2    3    4
## Frequency     10   10   10   10
## Proportion 0.25 0.25 0.25 0.25
## ------------------------------------------------------------------------
## Cereal
##        n  missing distinct
##       40        0       38
##
## lowest : Basic 4                            Capn Crunch                     Capn Crunch's
## highest: Post Toasties Corn Flakes          Rice Chex                       Rice Crispies
## ------------------------------------------------------------------------
## size_g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       13    0.981     37.2    12.02    27.00    27.00
##      .25      .50      .75      .90      .95
##    29.75    31.00    51.00    55.00    55.20
##
## Value         27    28    29    30    31    32    33    49    50    54
## Frequency      5     3     2     9     5     2     2     1     1     2
## Proportion 0.125 0.075 0.050 0.225 0.125 0.050 0.050 0.025 0.025 0.050
##
## Value         55    59    60
## Frequency      6     1     1
## Proportion 0.150 0.025 0.025
## ------------------------------------------------------------------------
## sugar_g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       17    0.994     10.4    6.521     1.95     2.00
##      .25      .50      .75      .90      .95
##     6.00    11.00    14.00    17.20    19.00
##
## Value          0     1     2     3     5     6     9    10    11    12
## Frequency      1     1     5     1     1     3     3     2     4     3
## Proportion 0.025 0.025 0.125 0.025 0.025 0.075 0.075 0.050 0.100 0.075
##
## Value         13    14    15    16    17    19    20
## Frequency      3     4     2     1     2     3     1
```

```
## Proportion 0.075 0.100 0.050 0.025 0.050 0.075 0.025
## ------------------------------------------------------------------------
## fat_g
##        n  missing distinct     Info     Mean      Gmd
##       40        0        8    0.958      1.2    1.178
##
## Value        0.0   0.5   1.0   1.5   2.0   2.5   3.0   5.0
## Frequency      9     5    12     4     3     3     3     1
## Proportion 0.225 0.125 0.300 0.100 0.075 0.075 0.075 0.025
## ------------------------------------------------------------------------
## sodium_mg
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       40        0       23    0.997    195.5       92     47.5     96.5
##      .25      .50      .75      .90      .95
##    157.5    200.0    262.5    300.0    301.0
##
## lowest :   0  50  65 100 105, highest: 280 290 300 320 330
## ------------------------------------------------------------------------
## Shelf_factor
##        n  missing distinct
##       40        0        4
##
## Value          1    2    3    4
## Frequency     10   10   10   10
## Proportion 0.25 0.25 0.25 0.25
## ------------------------------------------------------------------------
```

a

a. The explanatory variables need to be reformatted before proceeding further.

First

Divide each explanatory variable by its serving size to account for the different serving sizes among the cereals.

```
d <- d %>%
  mutate(
    sugar_g_per_serving = sugar_g/size_g,
    fat_g_per_serving = fat_g/size_g,
    sodium_mg_per_serving = sodium_mg/size_g
  )
```

Second

Rescale each variable to be within 0 and 1.

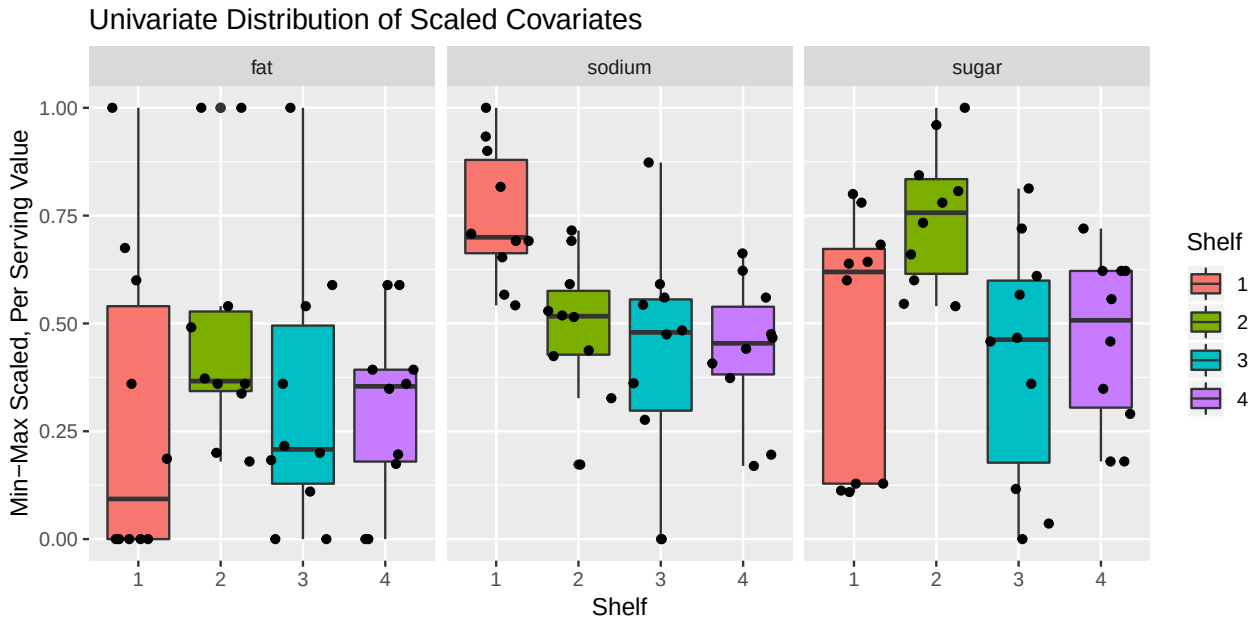
```
# min max scale a data point
#    if newrow, scale according to values in d
min_max_scale <- function(x, newrowcolname) {
  if (missing(newrowcolname)) {
    return((x - min(x)) / (max(x) - min(x)))
  } else {
    return((x - min(d[, newrowcolname])) /
             (max(d[, newrowcolname]) -
                min(d[, newrowcolname])
              )
           )
  }
}

d <- d %>%
  mutate(
    sugar = min_max_scale(sugar_g_per_serving),
    fat = min_max_scale(fat_g_per_serving),
    sodium = min_max_scale(sodium_mg_per_serving)
  )
```

b

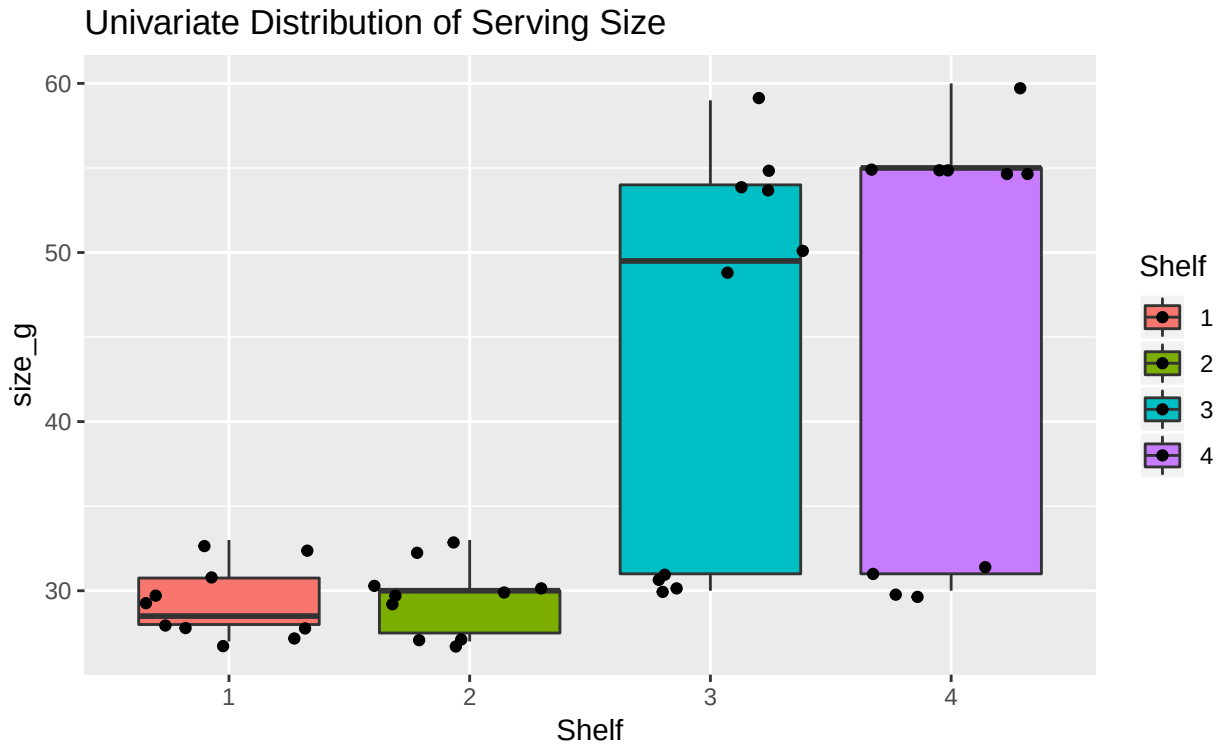
box plots

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

```
# scaled covariates
d %>%
  dplyr::select(ID, Shelf, Cereal, sugar, fat, sodium) %>%
  gather(var, value, -ID, -Shelf, -Cereal) %>%
  ggplot(aes(x=factor(Shelf), y = value)) +
  geom_boxplot(aes(fill=factor(Shelf))) +
  geom_jitter(height = 0) +
  facet_grid( . ~ var) +
  labs(title= "Univariate Distribution of Scaled Covariates",
       x="Shelf",
       y="Min-Max Scaled, Per Serving Value") +
  scale_fill_discrete(name="Shelf")
```



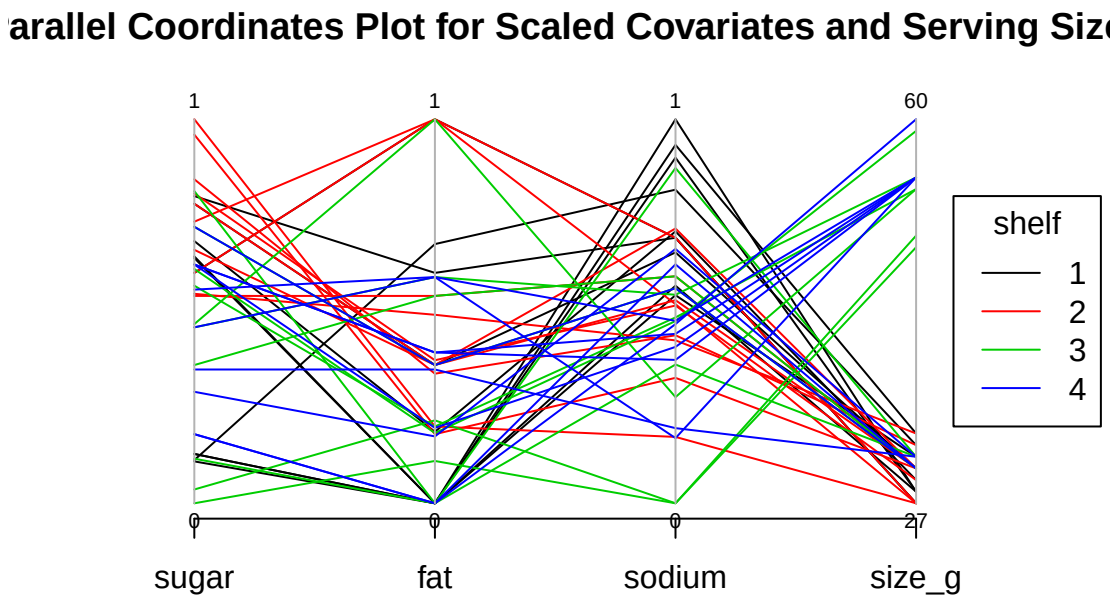

```
# serving size
d %>%
  ggplot(aes(x=factor(Shelf), y = size_g, fill=factor(Shelf))) +
  geom_boxplot() +
  geom_jitter() +
  labs(title= "Univariate Distribution of Serving Size",
       x="Shelf") + scale_fill_discrete(name="Shelf")
```



parallel coordinates plot

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

```
#
par(mar=c(5.1,4.1,4.1,8.1), xpd=T)
d %>%
  dplyr::select(sugar, fat, sodium, size_g) %>%
  parcoord(col = d$Shelf_factor, var.label = T,
           main="Parallel Coordinates Plot for Scaled Covariates and Serving Size")
legend(x= "right", inset = c(-.2,0),
       legend = c("1","2", "3", "4"),
       col=c(1,2,3,4),
       title="shelf",
       lty=1
       )
```



Shelf 1 appears to have high sodium content and small serving sizes. Shelf 2 has high sugar and small serving sizes. Shelf 3 and 4 have large serving sizes.

c

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 that this setting occurs here?

It would make sense to account for ordinality if the variable were ordinal (e.g. Likert) and if we believe the proportional odds assumption of the ordinal logisitc regression. Here, it does not make sense to account for ordinality. We have no reason to believe that there is a natural ordering of shelves: we wouldn't expect a function of the covariates to make us believe the item deserves a

higher shelf or a lower shelf. Rather we would expect a function of the covariates might point us to a particular shelf. It makes more sense to analyze each shelf separately.

d

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.

```
m <- nnet::multinom(
  formula = Shelf ~  fat + sodium + sugar,
  data = d, trace = F)
summary(m)
```

```
## Call:
## nnet::multinom(formula = Shelf ~ fat + sodium + sugar, data = d,
##     trace = F)
##
## Coefficients:
##   (Intercept)        fat    sodium      sugar
## 2    6.900708  4.0647092 -17.49373   2.693071
## 3   21.680680 -0.5571273 -24.97850 -12.216442
## 4   21.288343 -0.8701180 -24.67385 -11.393710
##
## Std. Errors:
##   (Intercept)      fat   sodium    sugar
## 2    6.487408 2.307250 7.097098 5.051689
## 3    7.450885 2.414963 8.080261 4.887954
## 4    7.435125 2.405710 8.062295 4.871338
##
## Residual Deviance: 67.19028
## AIC: 91.19028
```

```
car::Anova(m)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##        LR Chisq Df Pr(>Chisq)
## fat      5.2836  3     0.1522
## sodium  26.6197  3  7.073e-06 ***
## sugar   22.7648  3  4.521e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(nnet::multinom(
  formula = Shelf ~  fat,
  data = d, trace =F))
```

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

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##     LR Chisq Df Pr(>Chisq)
## fat   2.8448  3     0.4162
```

Sodium and sugar both have a significant impact on the log odds of being on a particular shelf holding fat and the other variable (sodium for sugar and sugar for sodium) constant. The impact of fat is not statistically significant whether or not we account for other variables.

e

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

```
# Check each two-way intercation for significance
car::Anova(nnet::multinom(
  formula = Shelf ~  fat + sodium + sugar + fat:sugar,
  data = d, trace = F))
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##           LR Chisq Df Pr(>Chisq)
## fat         5.2836  3     0.1522
## sodium     30.8407  3  9.183e-07 ***
## sugar      22.7648  3  4.521e-05 ***
## fat:sugar   5.1924  3     0.1582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(nnet::multinom(
  formula = Shelf ~  fat + sodium + sugar + fat:sodium,
  data = d, trace = F))
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##            LR Chisq Df Pr(>Chisq)
## fat          5.2836  3  0.1521727
## sodium      26.6197  3  7.073e-06 ***
## sugar       19.2525  3  0.0002424 ***
## fat:sodium   5.9115  3  0.1159978
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(nnet::multinom(
  formula = Shelf ~  fat + sodium + sugar + sodium:sugar,
  data = d, trace = F))
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##              LR Chisq Df Pr(>Chisq)
## fat            6.1167  3     0.1061
## sodium        26.6197  3  7.073e-06 ***
## sugar         22.7648  3  4.521e-05 ***
## sodium:sugar   2.3498  3     0.5030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# check three way interaction for significance
car::Anova(nnet::multinom(
  formula = Shelf ~  fat + sodium + sugar + fat:sodium + fat:sugar + sodium:sugar + fat:sodium
  data = d, trace = F))
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Shelf
##                  LR Chisq Df Pr(>Chisq)
## fat                6.1167  3  0.1060686
## sodium            30.8407  3  9.183e-07 ***
## sugar             19.2525  3  0.0002424 ***
## fat:sodium         3.1586  3  0.3678151
## fat:sugar          3.2309  3  0.3573733
## sodium:sugar       3.0185  3  0.3887844
## fat:sodium:sugar   2.5884  3  0.4595299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After accounting for individual covariate effects, each two way interaction is not significant as can be seen from the first three `Anova` tests above showing p values of .16 for `fat:sugar`, .12 for `fat:sodium`, and .50 for `sodium:sugar`. Also, the three way interaction is not significant after accounting for the individual covariates and the two-way interactions as can be seen in the final `Anova` test above showing a .46 p-value.

f

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.

```
cor(d %>% dplyr::select(fat,sodium,sugar))
```

```
##                fat     sodium      sugar
## fat     1.0000000 -0.0661432  0.2397225
## sodium -0.0661432  1.0000000 -0.1635699
## sugar   0.2397225 -0.1635699  1.0000000
```

Fat does not have a statistically significant impact on the log odds of being on a given shelf as seen above. However, we are interested in how much the different variables affect the probability of which shelf a cereal is on. These estimates could be biased if we do not include fat because fat is correlated with the other covariates and the outcomes. Therefore, to be able to interpret the effects of sodium and sugar as holding fat constant, we choose to use the formula `Shelf ~ fat + sugar + sodium`. This model was stored in `m` above.

To get predictions using this model, we transform the new cereal's covariates the same way we scaled the covariates in the training data (using the min and max from the training data to scale to 0-1).

```
aj <- data.frame(Cereal="Apple Jacks", size_g=28, sugar_g=12, fat_g=.5, sodium_mg=130) %>%
  mutate(
    sugar = min_max_scale(sugar_g/size_g,"sugar_g_per_serving"),
    fat = min_max_scale(fat_g/size_g,"fat_g_per_serving"),
    sodium = min_max_scale(sodium_mg/size_g,"sodium_mg_per_serving")
  )
```

```
predict(m, type="probs", newdata=aj)
```

```
##          1          2          3          4
## 0.05326849 0.47194264 0.20042742 0.27436145
```

According to our model, Apple Jacks as a 5.3% chance of being on shelf 1, 47.2% chance of shelf 2, 20.0% chance of shelf 3, and 27.4% change of shelf 4.

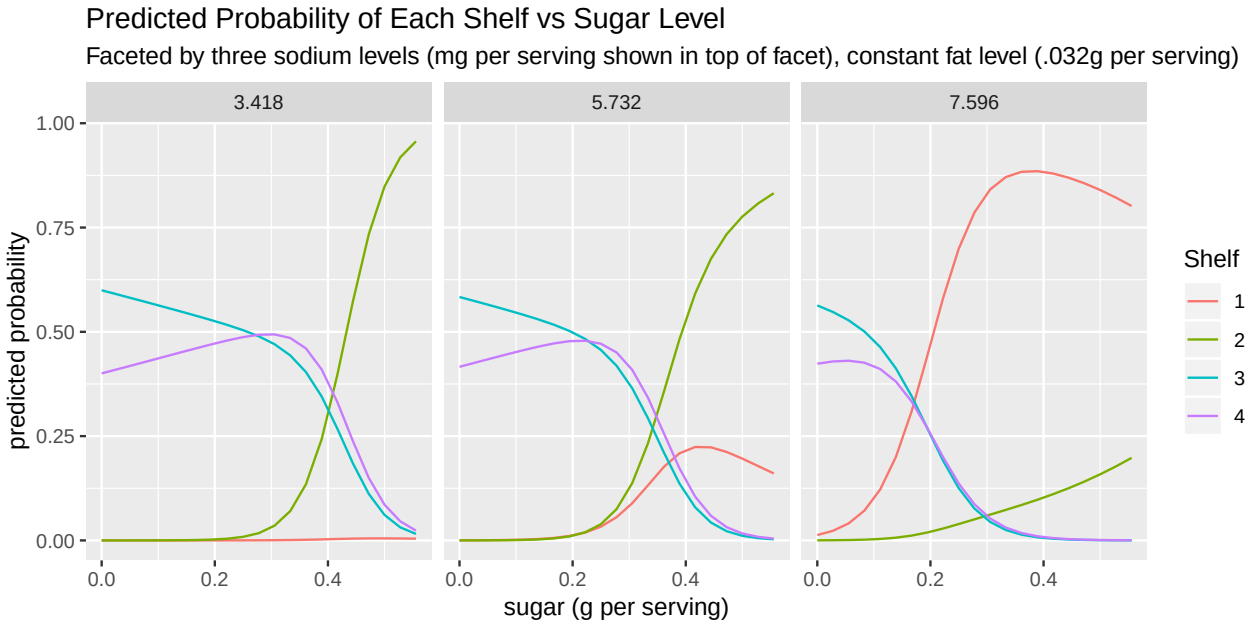
g

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.

To estimate the probabilities, we use a constant, scaled fat level of .347 (the mean value) given fat is not significant, we do not expect the curves to shift significantly at different levels of fat. We facet by scaled sodium values at values of .319, .535, and .709 (the 15th percentile, median, and 85th percentile) to see how much things change for low, high, and middle levels of sodium.

```
predict_data <- data.frame(
  sugar = rep(seq(0,1,by=.05),3),
  fat = .347,
  sodium = rep(c(.319,.535,.709), each=21)
)
```

```r
cbind(predict_data,
      predict(m, type="probs", newdata=predict_data)) %>%
  gather(Shelf, Predicted_Probability,-sugar,-fat,-sodium) %>%
  mutate(
    # note don't need to worry about min because is 0 for all three vars
    sugar_g_per_serving=sugar*max(d$sugar_g_per_serving),
    fat_g_per_serving=fat*max(d$fat_g_per_serving),
    sodium_mg_per_serving=sodium*max(d$sodium_mg_per_serving)
         ) %>%
  ggplot(aes(x=sugar_g_per_serving, y=Predicted_Probability, colour=factor(Shelf))) +
  geom_line() +
  facet_grid(. ~ round(sodium_mg_per_serving,3)) +
  guides(colour=guide_legend(title="Shelf")) +
  labs( title="Predicted Probability of Each Shelf vs Sugar Level",
        subtitle="Faceted by three sodium levels (mg per serving shown in top of facet), consta
        y="predicted probability",
        x="sugar (g per serving)")
```



Low sugar levels indicate that the cereal is on shelf 3 or 4. As the sugar level increases, we see the cereal is more likely to be on shelf 1 or 2. Sodium gives a big boost to shelf 1, such that it has the highest probability even at high values of sugar (though if we could imagine a cereal with a much higher sugar level, shelf 2 would still get the highest prediction), This potentially lends evidence to the hypothesis that children enjoy sugary cereal so sugary cereals are placed low where the children can see them.

h

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

```
# constant shift by .1 times the range (which equals the max bc all mins 0)
round(.1*max(d$fat_g_per_serving),2)
```

```
## [1] 0.01
```

```
round(.1*max(d$sodium_mg_per_serving),2)
```

```
## [1] 1.07
```

```
round(.1*max(d$sugar_g_per_serving),2)
```

```
## [1] 0.06
```

```
# ORs
round(exp(.1*summary(m)$coefficients),2)
```

```
##   (Intercept)  fat sodium sugar
## 2        1.99 1.50   0.17  1.31
## 3        8.74 0.95   0.08  0.29
## 4        8.41 0.92   0.08  0.32
```

```
# Wald confidence intervals
round(exp(.1*confint(m, level=.95)),2)
```

```
## , , 2
##
##             2.5 % 97.5 %
## (Intercept)  0.56   7.11
## fat          0.96   2.36
## sodium       0.04   0.70
## sugar        0.49   3.52
##
## , , 3
##
##             2.5 % 97.5 %
## (Intercept)  2.03  37.65
## fat          0.59   1.52
## sodium       0.02   0.40
## sugar        0.11   0.77
##
## , , 4
##
##             2.5 % 97.5 %
## (Intercept)  1.96  36.09
## fat          0.57   1.47
```

```
## sodium       0.02   0.41
## sugar        0.12   0.83
```

fat

The odds of being on shelf 2 instead of shelf 1 change by 1.50 times (between .96 and 2.36 times with 95% confidence) for an increase of .01g of fat per serving. The odds of being on shelf 3 instead of shelf 1 change by .95 times (between .59 and 1.52 times with 95% confidence) for an increase of .01g of fat per serving. The odds of being on shelf 4 instead of shelf 1 change by .92 times (between .57 and 1.47 times with 95% confidence) for an increase of .01g of fat per serving.

These confidence bands all contain 1 indicating how fat is not significant.

sodium

The odds of being on shelf 2 instead of shelf 1 change by .17 times (between .04 and .70 times with 95% confidence) for an increase of 1.07mg of sodium per serving. The odds of being on shelf 3 instead of shelf 1 change by .08 times (between .02 and .40 times with 95% confidence) for an increase of 1.07mg of sodium per serving. The odds of being on shelf 4 instead of shelf 1 change by .08 times (between .02 and .41 times with 95% confidence) for an increase of 1.07mg of sodium per serving.

All of these confidence intervals are below one showing the signficance of sodium. Also, as we saw in the predict probability plots, when sodium is high, category 1 is most likely (thus odds of all other classes vs 1 go down as sodium increases). The box plot also showed that shelf 1 has much higher sodium levels than all other shelves.

sugar

The odds of being on shelf 2 instead of shelf 1 change by 1.31 times (between .49 and 3.52 times with 95% confidence) for an increase of .06g of sugar per serving. The odds of being on shelf 3 instead of shelf 1 change by .29 times (between .11 and .77 times with 95% confidence) for an increase of .06g of sugar per serving. The odds of being on shelf 4 instead of shelf 1 change by .32 times (between .12 and .83 times with 95% confidence) for an increase of .06g of sugar per serving.

Two of these confidence intervals are below one showing that sugar is a significant differentiator between shelves 1 vs 3 or 4. As we saw in the predicted probability plots, the probability of shelves 3 and 4 decreased as sugar content increased. On the other hand, the predicted probability of shelf 2 and 3 increased to a point, though 2 takes priority (given the OR above 1 between 1 and 2) at higher sugar contents. This also aligns with the box plot showing the shelf 2 clearly has the highest sugar content.