STAA 552: Final Exam Part 3

Matthew Stoebe

Honor Code from Part 1 applies here, too.

Chocolate (Q14 - Q19)

We consider Ratings (1-5 scale) for n = 432 different chocolate bars. We consider two predictors: Cocoa (%) and Country. **Rating is the response variable for all models.**

The data is available from Canvas as Chocolate.csv.

The data includes the following variables:

- Company, Bar_Name and Ref_ID: should NOT be used for model fitting
- Cocoa (%)
- Country: Canada, Ecuador, France or U.K.
- Rating: 1 = Disappointing, 2 = Passable, 3 = Satisfactory, 4 = Praiseworthy, 5 = Premium

Note: Some of the results may be surprising.

Q14 (2 pts)

Create a summary table of **Country** and Rating. For each country, this table should give the proportion of bars with each rating. For example, for Canadian chocolate bars what proportion have rating 1, 2, etc?

```
## # A tibble: 20 x 3
## # Groups:   Country [4]
##    Country Rating   prop
##    <fct>    <int>  <dbl>
##  1 Canada       1 0.016
##  2 Canada       2 0.144
##  3 Canada       3 0.36
##  4 Canada       4 0.4
##  5 Canada       5 0.08
##  6 Ecuador      1 0.0909
##  7 Ecuador      2 0.273
##  8 Ecuador      3 0.345
##  9 Ecuador      4 0.236
## 10 Ecuador      5 0.0545
## 11 France       1 0.0577
## 12 France       2 0.167
## 13 France       3 0.301
```

```
## 14 France       4 0.327
## 15 France       5 0.147
## 16 U.K.         1 0.0625
## 17 U.K.         2 0.333
## 18 U.K.         3 0.312
## 19 U.K.         4 0.25
## 20 U.K.         5 0.0417
```

Q15 (2 pts)

Create a summary table giving the number of observations and the mean(**Cocoa**), for each value of Rating.

```
## # A tibble: 5 x 3
##   Rating count mean_cocoa
##    <int> <int>      <dbl>
## 1      1    22       84.8
## 2      2    91       73.6
## 3      3   141       72.4
## 4      4   138       70.8
## 5      5    40       70.9
```

Q16

Fit an appropriate model including only additive effects (no interaction). Show the coefficients table (including coefficient estimates and Wald test p-values). For consistency, use vlgm() to fit the model.

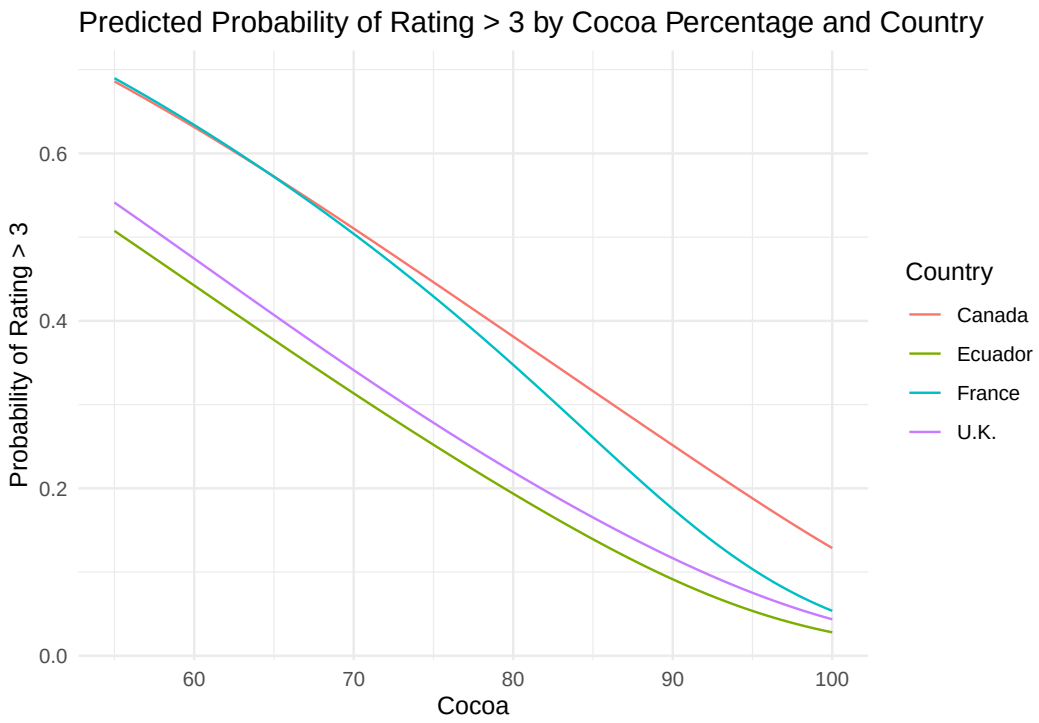
```
## Warning: package 'VGAM' was built under R version 4.4.2

## Call:
## vglm(formula = Rating ~ Cocoa + Country, family = multinomial(refLevel = 1),
##     data = choc)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    11.61392    2.31864   5.009 5.47e-07 ***
## (Intercept):2    14.08166    2.34476   6.006 1.91e-09 ***
## (Intercept):3    17.42773    2.52150   6.912 4.79e-12 ***
## (Intercept):4    15.12898    3.02294   5.005 5.59e-07 ***
## Cocoa:1          -0.11920    0.02614  -4.561 5.10e-06 ***
## Cocoa:2          -0.14046    0.02645  -5.311 1.09e-07 ***
## Cocoa:3          -0.18554    0.02946  -6.298 3.01e-10 ***
## Cocoa:4          -0.17590    0.03741  -4.702 2.58e-06 ***
```

```
## CountryEcuador:1 -1.06337    0.98028  -1.085   0.2780
## CountryEcuador:2 -1.75041    0.96468  -1.815   0.0696 .
## CountryEcuador:3 -2.33217    0.99123  -2.353   0.0186 *
## CountryEcuador:4 -2.15788    1.14775  -1.880   0.0601 .
## CountryFrance:1  -1.53136    0.91917  -1.666   0.0957 .
## CountryFrance:2  -1.87354    0.89680  -2.089   0.0367 *
## CountryFrance:3  -1.93567    0.90093  -2.149   0.0317 *
## CountryFrance:4  -1.11311    0.95573  -1.165   0.2442
## CountryU.K.:1    -0.29708    0.94342  -0.315   0.7528
## CountryU.K.:2    -1.25469    0.93312  -1.345   0.1787
## CountryU.K.:3    -1.54702    0.94520  -1.637   0.1017
## CountryU.K.:4    -1.73594    1.08604  -1.598   0.1100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]),
## log(mu[,4]/mu[,1]), log(mu[,5]/mu[,1])
##
## Residual deviance: 1152.505 on 1708 degrees of freedom
##
## Log-likelihood: -576.2524 on 1708 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## 'Cocoa:3', 'Cocoa:4'
##
##
## Reference group is level  1  of the response
```

Q17 (6 pts)

Using your model, create a graph of the (smooth) fitted curves representing the predicted probability of "Praiseworthy" or better ($P(Y \geq 4)$) over the range of Cocoa. There should be 4 fitted curves, color coded by Country.



Q18

Using your model, provide a detailed one sentence **interpretation** of the coefficient corresponding to **Cocoa** in context.

Response

For every one unit increase in cocoa concentration, the log odds of being rated at 2,3,4,5 instead of 1 decrease by between 11% and 18%. (depending on the comparison being made)

Q19 (6 pts)

After controlling for Cocoa, is there evidence of differences in chocolate ratings between **Countries**? Run an appropriate test. Use the test results and the graph from Q17 to briefly discuss the Countries.

```
## [1] 3.808885e-08
```

Discussion: There is still a significant difference after controlling for Cocoa. this is evident above where the full model is signifficantly better than the model with just cocoa.

You can also see this in the plot. The trend is consistent for cooca between the countries but there is still a seperation between them that seems to be driven by country. ******

Appendix

```r
#Retain this code chunk!!!
library(knitr)
library(tidyverse)
library(ggplot2)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#Q14
choc <- read.csv("Chocolate.csv")
choc$Country <- factor(choc$Country, levels = c("Canada", "Ecuador", "France", "U.K."))


prop_table <- choc %>%
  group_by(Country, Rating) %>%
  summarize(n = n()) %>%
  mutate(prop = n / sum(n)) %>%
  select(Country, Rating, prop)

prop_table

#Q15
summary_by_rating <- choc %>%
  group_by(Rating) %>%
  summarize(
    count = n(),
    mean_cocoa = mean(Cocoa, na.rm = TRUE)
  )

summary_by_rating
#Q16
library(VGAM)
model <- vglm(Rating ~ Cocoa + Country,
              family = multinomial(refLevel = 1),
              data = choc)

summary(model)
#Q17

cocoa_seq <- data.frame(Cocoa = seq(min(choc$Cocoa), max(choc$Cocoa), length.out = 100))
plot_data <- expand.grid(Cocoa = cocoa_seq$Cocoa, Country = levels(choc$Country))
pred_probs <- predict(model, newdata = plot_data, type = "response")


plot_data$P_ge_4 <- pred_probs[,4] + pred_probs[,5]


ggplot(plot_data, aes(x = Cocoa, y = P_ge_4, color = Country)) +
  geom_line() +
  labs(
    title = "Predicted Probability of Rating > 3 by Cocoa Percentage and Country",
    x = "Cocoa",
    y = "Probability of Rating > 3"
```

```
  ) +
  theme_minimal()

# Q19
model_reduced <- vglm(Rating ~ Cocoa,
                      family = multinomial(refLevel = 1),
                      data = choc)

model_full <- model

dev_full <- deviance(model_full)
dev_reduced <- deviance(model_reduced)
LRT <- dev_reduced - dev_full
p_value <- pchisq(LRT, 1, lower.tail = FALSE)
p_value
```