GLMs for Categorical Data Exercise

```
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
```

```
library(tidyverse)
library(effects)
library(VGAM)
library(nnet)
library(lubridate)
```

Part I

Load the dataset `squirrel_census.csv`. This file includes a number of variables describing squirrel behavior and physical features during squirrel sightings in Central Park. We are interested in finding whether we are able to successfully use this information to predict a squirrel's primary fur color.

What type of categorical scale best describes the response variable here, and what would be the most appropriate model for this task?

The primary fur color is a categorical variable on the nominal scale. The baseline-category model would likely be the most appropriate model for the given modeling outcome.

Fit the model mentioned above, and provide an interpretation for the model coefficients.

When we load the dataset, we also reformat many of the variable names and discard some of the variables that have inconsistent values.

```
squirrels <- read_csv("squirrel_census.csv") %>%
  rename(
    id = "Unique Squirrel ID",
    hectare = "Hectare",
    shift = "Shift",
```

```
  date = "Date",
  hectare_squirrel = "Hectare Squirrel Number",
  age = "Age",
  primary_fur = "Primary Fur Color",
  highlight_fur = "Highlight Fur Color",
  primary_highlight_fur = "Combination of Primary and Highlight Color",
  color_notes = "Color notes",
  location = "Location",
  sighter_measurement = "Above Ground Sighter Measurement",
  specific_location = "Specific Location",
  running = Running,
  chasing = Chasing,
  climbing = Climbing,
  eating = Eating,
  foraging = Foraging,
  other_activities = "Other Activities",
  kuks = Kuks,
  quaas = Quaas,
  moans = Moans,
  flags = "Tail flags",
  twitches = "Tail twitches",
  approaches = Approaches,
  indifferent = Indifferent,
  runs_from = "Runs from",
  interactions = "Other Interactions",
  lat_long = "Lat/Long"
) %>%
select(
  X,
  Y,
  id,
  hectare,
  shift,
  date,
  age,
  primary_fur,
  highlight_fur,
  primary_highlight_fur,
  location,
  sighter_measurement,
  running,
```

```
    chasing,
    climbing,
    eating,
    foraging,
    kuks,
    quaas,
    moans,
    flags,
    twitches,
    approaches,
    indifferent,
    runs_from
  ) %>%
  mutate(
    date = mdy(date)
  )
```

When fitting the model, we include most of the available predictors. Because all data collection takes place within Central Park, the location variables are likely irrelevant, so they are excluded, as are variables with a large number of missing values. From the model summary, we see that the squirrels with black fur are used as the baseline category, so the estimated coefficients describe the differences between the baseline group and the groups with gray and cinnamon fur, respectively.

```
bc_mod <- multinom(
  primary_fur ~ shift +
    age +
    location +
    running +
    chasing +
    climbing +
    eating +
    foraging +
    kuks +
    quaas +
    moans +
    flags +
    twitches +
    approaches +
    indifferent +
    runs_from,
  data = squirrels
```

```
  )
```

```
# weights:  57 (36 variable)
initial  value 3102.481103
iter  10 value 1556.060877
iter  20 value 1500.941617
iter  30 value 1493.740140
iter  40 value 1493.475223
final  value 1493.466178
converged
```

```
  summary(bc_mod)
```

```
Call:
multinom(formula = primary_fur ~ shift + age + location + running +
    chasing + climbing + eating + foraging + kuks + quaas + moans +
    flags + twitches + approaches + indifferent + runs_from,
    data = squirrels)

Coefficients:
         (Intercept)   shiftPM  ageAdult ageJuvenile locationGround Plane
Cinnamon   -5.591596 0.1397004  6.397133    7.170776          -0.06692869
Gray       11.853451 0.2255340 -8.847345   -8.497004          -0.19114618
         runningTRUE chasingTRUE climbingTRUE eatingTRUE foragingTRUE  kuksTRUE
Cinnamon   0.3638170   0.5164466   0.07359606  0.3056715    0.4908310 0.2254169
Gray       0.1533483   0.6615701  -0.08158615  0.1451103    0.3454597 0.4499831
         quaasTRUE moansTRUE  flagsTRUE twitchesTRUE approachesTRUE
Cinnamon -1.409787 -5.747410  0.2284257   0.07491175      0.6195083
Gray     -1.493057  8.991143 -0.1701263  -0.28280684     -0.1753301
         indifferentTRUE runs_fromTRUE
Cinnamon     0.000737242    -0.3775194
Gray         0.102508352    -0.3103793

Std. Errors:
         (Intercept)   shiftPM  ageAdult ageJuvenile locationGround Plane
Cinnamon   0.2843044 0.2303325 0.2003653   0.2819596            0.3452386
Gray       0.2592116 0.2087515 0.1846167   0.2655712            0.3123951
         runningTRUE chasingTRUE climbingTRUE eatingTRUE foragingTRUE  kuksTRUE
Cinnamon   0.2845245   0.4830348    0.3460046  0.2767739    0.2695153 0.7077605
Gray       0.2604177   0.4450385    0.3128648  0.2546084    0.2459071 0.6363207
```

```
         quaasTRUE    moansTRUE flagsTRUE twitchesTRUE approachesTRUE
Cinnamon 0.6784857 4.193847e-10 0.4742414    0.3004620      0.4763197
Gray     0.5263668 3.413211e-07 0.4373953    0.2759829      0.4533780
         indifferentTRUE runs_fromTRUE
Cinnamon       0.2797034     0.3036993
Gray           0.2545063     0.2726369

Residual Deviance: 2986.932
AIC: 3058.932
```

We can interpret each coefficient as the increase in log-odds of falling into one category versus another as a result of a one-unit increase in the applicable explanatory variable. For example, with all else equal, the log-odds of a squirrel having gray fur versus black fur increases by 0.276 if it is seen running instead of not running, meaning that all else equal, a squirrel seen running is 1.32 times more likely to have gray fur instead of black fur than a squirrel not seen running. Other coefficients can be interpreted analogously.

Part II

Use the `effects` package to load the `WVS` dataset. This file includes information from the world values survey. More information can be found with `?WVS`. We are interested in the outcome of `poverty`, an evaluation of whether the government's efforts to help those in poverty are too little, about right, or too much. What type of data is this, and what type of model would be appropriate for this situation?

```
data(WVS)
```

This data is categorical on an ordinal scale. Because of this, a cumulative logit model or proportional odds model would likely be appropriate.

Fit a cumulative logit model and a proportional odds model. Does the proportional odds assumption apply in this setting? Provide an interpretation of the model coefficients for the proportional odds model.

First, we begin by fitting the cumulative logit model.

```
mod1 <- vglm(
  poverty ~ religion + degree + country + age + gender,
  data = WVS,
  family = cumulative
)
```

```r
summary(mod1)
```

```
Call:
vglm(formula = poverty ~ religion + degree + country + age +
    gender, family = cumulative, data = WVS)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept):1    0.700879   0.109198   6.418 1.38e-10 ***
(Intercept):2    2.412286   0.152479  15.820  < 2e-16 ***
religionyes:1   -0.104622   0.080699  -1.296 0.194822
religionyes:2   -0.286809   0.108066  -2.654 0.007954 **
degreeyes:1     -0.180002   0.070303  -2.560 0.010456 *
degreeyes:2     -0.007199   0.102278  -0.070 0.943889
countryNorway:1  0.125588   0.078091   1.608 0.107786
countryNorway:2  1.765905   0.182010   9.702  < 2e-16 ***
countrySweden:1  0.444542   0.082808   5.368 7.95e-08 ***
countrySweden:2  2.053499   0.213638   9.612  < 2e-16 ***
countryUSA:1    -0.357866   0.073421  -4.874 1.09e-06 ***
countryUSA:2    -0.894826   0.087845 -10.186  < 2e-16 ***
age:1           -0.010672   0.001649  -6.472 9.68e-11 ***
age:2           -0.010292   0.002217  -4.643 3.43e-06 ***
gendermale:1    -0.197816   0.055611  -3.557 0.000375 ***
gendermale:2    -0.104302   0.077804  -1.341 0.180061
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 10031.68 on 10746 degrees of freedom

Log-likelihood: -5015.84 on 10746 degrees of freedom

Number of Fisher scoring iterations: 6

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
  religionyes:1   religionyes:2     degreeyes:1     degreeyes:2 countryNorway:1
      0.9006653       0.7506554       0.8352683       0.9928272       1.1338150
```

```
countryNorway:2 countrySweden:1 countrySweden:2    countryUSA:1    countryUSA:2
      5.8468608       1.5597749       7.7951280       0.6991671       0.4086787
          age:1           age:2    gendermale:1    gendermale:2
      0.9893843       0.9897605       0.8205208       0.9009535
```

At this point, looking at the model summary provides one important piece of information: the residual deviance, which is a measure of model fit. We will use this value to compare the cumulative logit model and the proportional odds model that we fit below.

```r
mod2 <- vglm(
  poverty ~ religion + degree + country + age + gender,
  data = WVS,
  family = cumulative(parallel = TRUE)
)

summary(mod2)
```

```
Call:
vglm(formula = poverty ~ religion + degree + country + age +
    gender, family = cumulative(parallel = TRUE), data = WVS)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.729769   0.104044   7.014 2.32e-12 ***
(Intercept):2  2.532483   0.110154  22.990  < 2e-16 ***
religionyes   -0.179733   0.076565  -2.347 0.018902 *
degreeyes     -0.140918   0.066714  -2.112 0.034663 *
countryNorway  0.322353   0.075444   4.273 1.93e-05 ***
countrySweden  0.603300   0.080895   7.458 8.79e-14 ***
countryUSA    -0.617773   0.068391  -9.033  < 2e-16 ***
age           -0.011141   0.001557  -7.157 8.24e-13 ***
gendermale    -0.176370   0.052877  -3.335 0.000851 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 10402.59 on 10753 degrees of freedom

Log-likelihood: -5201.296 on 10753 degrees of freedom
```

```
Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
  religionyes     degreeyes countryNorway countrySweden    countryUSA
    0.8354929     0.8685603     1.3803723     1.8281418     0.5391437
          age    gendermale
    0.9889208     0.8383078
```

This model fit shows higher deviance than in the cumulative logit model, which is expected given that the proportional odds model is less flexible than the more general cumulative logit model, and will thus show worse fit to the data.

```r
pchisq(
  deviance(mod2) - deviance(mod1),
  df = df.residual(mod2) - df.residual(mod1),
  lower.tail = FALSE
)
```

```
[1] 4.096967e-76
```

The p-value from the χ^2 distribution shown above indicates that there is significant evidence to show that the proportional odds assumption does not hold in this case. Examining the coefficients in the cumulative logit model, we can see substantial differences in coefficients between response levels, particularly for the country coefficients. This is not surprising given that whether someone feels that the government is providing enough support to those in poverty will depend largely on the policies of the government in question. To address this concern, we may consider a partial proportional odds assumption, rather than enforcing a complete proportional odds assumption. Under this assumption, we will assume that the country coefficients will very by response level, but the other coefficients will stay constant. WE fit this model as follows:

```r
mod3 <- vglm(
  poverty ~ religion + degree + country + age + gender,
  data = WVS,
  family = cumulative(parallel = FALSE~country)
)

summary(mod3)
```

```
Call:
vglm(formula = poverty ~ religion + degree + country + age +
    gender, family = cumulative(parallel = FALSE ~ country),
    data = WVS)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept):1    0.717842   0.104683   6.857 7.02e-12 ***
(Intercept):2    2.360861   0.115250  20.485  < 2e-16 ***
religionyes     -0.149106   0.076443  -1.951  0.05111 .
degreeyes       -0.141428   0.066883  -2.115  0.03447 *
countryNorway:1  0.122710   0.077846   1.576  0.11495
countryNorway:2  1.781195   0.181065   9.837  < 2e-16 ***
countrySweden:1  0.444902   0.082441   5.397 6.79e-08 ***
countrySweden:2  2.068669   0.213359   9.696  < 2e-16 ***
countryUSA:1    -0.362552   0.073344  -4.943 7.69e-07 ***
countryUSA:2    -0.872748   0.086634 -10.074  < 2e-16 ***
age             -0.010605   0.001559  -6.804 1.02e-11 ***
gendermale      -0.173844   0.052931  -3.284  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 10040.25 on 10750 degrees of freedom

Log-likelihood: -5020.123 on 10750 degrees of freedom

Number of Fisher scoring iterations: 6

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
    religionyes       degreeyes countryNorway:1 countryNorway:2 countrySweden:1
      0.8614775       0.8681179       1.1305563       5.9369444       1.5603377
countrySweden:2    countryUSA:1    countryUSA:2             age      gendermale
      7.9142848       0.6958980       0.4178017       0.9894508       0.8404278
```

```r
pchisq(
  deviance(mod3) - deviance(mod1),
  df = df.residual(mod3) - df.residual(mod1),
  lower.tail = FALSE
)
```

```
[1] 0.07293254
```

When running the same test of the proportional odds assumption, we again find evidence that it is not the most accurate, but the p-value is not significant at $\alpha = 0.05$, and at this point the convenience of the assumption likely outweighs any loss in performance resulting from the less flexible model specification.

Looking at the exponentiated coefficients shows that, for example, the odds of men feeling like government is doing too little to support those in poverty compared to doing about the right amount or too much is approximately 0.84 times the odds of women feeling like that. Because this is a (partial) proportional odds model, this coefficient also applies to the odds of feeling like government is doing either too little or about the right amount compared with too much to support those in poverty. Additionally, we see that for each increase in age of one year, the estimated odds of feeling like government is doing too little to support those in poverty compared with doing about the right amount or too much decreases by about one percent.