Problem Set 3

Samanta Nedzinskaite

Due: March 26, 2023

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub in `.pdf` form.
- This problem set is due before 23:59 on Sunday March 26, 2023. No late assignments will be accepted.

Question 1

We are interested in how governments' management of public resources impacts economic prosperity. Our data come from Alvarez, Cheibub, Limongi, and Przeworski (1996) and is labelled `gdpChange.csv` on GitHub. The dataset covers 135 countries observed between 1950 or the year of independence or the first year forwhich data on economic growth are available ("entry year"), and 1990 or the last year for which data on economic growth are available ("exit year"). The unit of analysis is a particular country during a particular year, for a total $>$ 3,500 observations.

- Response variable:
  - `GDPWdiff`: Difference in GDP between year t and $t-1$. Possible categories include: "positive", "negative", or "no change"
- Explanatory variables:
  - `REG`: 1=Democracy; 0=Non-Democracy
  - `OIL`: 1=if the average ratio of fuel exports to total exports in 1984-86 exceeded 50%; 0= otherwise

Please answer the following questions:

1. Construct and interpret an unordered multinomial logit with `GDPWdiff` as the output and "no change" as the reference category, including the estimated cutoff points and coefficients.

```r
data <- read_csv("~/Documents/GitHub/StatsII_Spring2023/datasets/gdpChange.csv")

##response variable to negative/positive/no change
data$GDPwdiff <- ifelse(data$GDPWdiff<0, "negative",
                 ifelse(data$GDPWdiff>0, "positive",
                        "no change"))

#'no change' is reference category
data$GDPwdiff <- factor(data$GDPwdiff, levels = c("no change", "negative", "positive"))

#unordered multinomial model
multinom_model1 <- multinom(GDPwdiff ~ REG + OIL, data = data)
summary(multinom_model1)
```

Coefficients:

```
Call:
multinom(formula = GDPwdiff ~ REG + OIL, data = data)

Coefficients:
         (Intercept)      REG      OIL
negative    3.805370 1.379282 4.783968
positive    4.533759 1.769007 4.576321

Std. Errors:
         (Intercept)       REG      OIL
negative   0.2706832 0.7686958 6.885366
positive   0.2692006 0.7670366 6.885097

Residual Deviance: 4678.77
AIC: 4690.77
```

```r
#get cut-off points using predicted probabilities
pred_probs <- predict(multinom_model1, type = "probs")
cutoff_points <- t(apply(pred_probs, 2, function(x) quantile(x, probs = 0.5)))
#print cutoff points
colnames(cutoff_points) <- c("no change", "negative", "positive")
```

```r
rownames(cutoff_points) <- names(multinom_model1$coefficients)
print(cutoff_points)
```

Cut-off-points:

```
        no change negative  positive
[1,] 0.007191671 0.323207 0.6696013
```

Interpretation:
In an unordered multinomial model, the coefficients represent the change in the log-odds of moving from one category to a specific reference category. The unordered multinomial logistic regression model aims to predict the outcome variable 'GDPwdiff', which has three (unordered) categories - 'no change', 'negative', and 'positive'. The coefficients in the summary output represent the estimated log odds of being in each category, relative to the reference category - which in this case is the 'no change' category.

For the 'negative' category, the intercept is 3.805370, meaning that the log odds of being in the negative category, relative to the reference category ('no change'), are 3.805370 when both REG and OIL are equal to 0. Further, for one-unit increase in REG, the log odds of being in the negative category increase by 1.379282, and for one-unit increase in OIL, the log odds of being in the negative category increase by 4.783968.

The same interpretation applies to the positive category and its coefficients.

The cutoff points represent the thresholds for the predicted probabilities that determine which outcome category is most likely given the predictor values. For example, for the first predictor variable, if the predicted probability of "no change" is less than 0.007, the model predicts that the outcome will be "negative". If the predicted probability of "no change" is greater than 0.669, the model predicts that the outcome will be "positive". If the predicted probability of "no change" falls between these two cutoff points, the model predicts that the outcome will be "no change". These cutoff points are useful for interpreting the model's predictions based on the model's outputs.

2. Construct and interpret an ordered multinomial logit with `GDPWdiff` as the outcome variable, including the estimated cutoff points and coefficients.

```r
#ordered multinomial model
ord_log <- polr(GDPwdiff ~ REG + OIL, data = data, Hess = TRUE)
summary(ord_log)

# Calculate a p value
ctable <- coef(summary(ord_log))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
```

Coefficients:

```
Call:
polr(formula = GDPwdiff ~ REG + OIL, data = data, Hess = TRUE)

Coefficients:
      Value Std. Error t value
REG  0.4102    0.07518   5.456
OIL -0.1788    0.11546  -1.549

Intercepts:
                    Value    Std. Error t value
no change|negative  -5.3199   0.2523   -21.0865
negative|positive   -0.7036   0.0476   -14.7932

Residual Deviance: 4686.606
AIC: 4694.606
```

```r
#get cut-off points using predicted probabilities
pred_probs2 <- predict(ord_log, type = "probs")
cutoff_points2 <- t(apply(pred_probs2, 2, function(x) quantile(x, probs =
    0.5)))
#print cutoff points
colnames(cutoff_points2) <- c("no change", "negative", "positive")
rownames(cutoff_points2) <- names(multinom_model1$coefficients)
print(cutoff_points2)
```

Cut-off-points:

```
       no change  negative  positive
[1,] 0.004869417 0.3261451 0.6689855
```

Interpretation:

In an ordered multinomial model, the coefficients represent the change in the log-odds of moving from one category to the next higher category. For example, the coefficient for REG in the 'no change' category represents the change in the log-odds of moving from 'no change' to 'negative' category.

This model is trying to predict whether GDPwdiff will be 'no change', 'negative', or 'positive' based on two variables called REG and OIL. The coefficient for REG is 0.4102, which means that for each one-unit increase in REG, the likelihood of GDPwdiff being in the 'no change' category increases by 0.4102. The coefficient for OIL is -0.1788, which means that for each one-unit increase in OIL, the likelihood of GDPwdiff being in the 'no change' category decreases by 0.1788.

The intercepts represent the starting point for each category of GDPwdiff. The intercept for the 'no change' category is -5.3199, which means that if REG and OIL are

both 0, the likelihood of GDPwdiff being in the 'no change' category is very high. The intercept for the 'negative' category is -0.7036, which means that if REG and OIL are both 0, the likelihood of GDPwdiff being in the 'negative' category is relatively high.

In an ordered multinomial logistic regression model, the cutoff points are the thresholds for the predicted probabilities that determine which outcome category is most likely given the predicted cut-off points. For example, if the predicted probability of 'no change' is less than 0.004869417, the model predicts that the outcome will be 'negative'. If the predicted probability of 'no change' is greater than 0.6689855, the model predicts that the outcome will be 'positive'. If the predicted probability of 'no change' falls between these two cutoff points, the model predicts that the outcome will be 'no change'.

Question 2

Consider the data set `MexicoMuniData.csv`, which includes municipal-level information from Mexico. The outcome of interest is the number of times the winning PAN presidential candidate in 2006 (`PAN.visits.06`) visited a district leading up to the 2009 federal elections, which is a count. Our main predictor of interest is whether the district was highly contested, or whether it was not (the PAN or their opponents have electoral security) in the previous federal elections during 2000 (`competitive.district`), which is binary (1=close/swing district, 0="safe seat"). We also include `marginality.06` (a measure of poverty) and `PAN.governor.06` (a dummy for whether the state has a PAN-affiliated governor) as additional control variables.

(a) Run a Poisson regression because the outcome is a count variable. Is there evidence that PAN presidential candidates visit swing districts more? Provide a test statistic and p-value.

```
with(df,
     list(mean(PAN.visits.06), var(PAN.visits.06)))

# The variance is MUCH greater than the mean...
# This suggests that we will have over-dispersion in the model.

mod_ps <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN
    .governor.06, data = df, family = poisson)
summary(mod_ps)
```

```
Call:
glm(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
    PAN.governor.06, family = poisson, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2309  -0.3748  -0.1804  -0.0804  15.2669

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.81023    0.22209 -17.156   <2e-16 ***
competitive.district -0.08135    0.17069  -0.477   0.6336
marginality.06       -2.08014    0.11734 -17.728   <2e-16 ***
PAN.governor.06      -0.31158    0.16673  -1.869   0.0617 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1473.87  on 2406  degrees of freedom
Residual deviance:  991.25  on 2403  degrees of freedom
AIC: 1299.2

Number of Fisher Scoring iterations: 7
```

Based on the output of the Poisson regression, we can see that the coefficient estimate for competitive.district is negative (-0.08135) and not statistically significant (p-value = 0.6336), which suggests that there is no evidence to support the claim that PAN presidential candidates visit swing districts more. The test statistic for the Poisson model is the z-value. The z-value is given for each coefficient in the "Coefficients" table under the column "z value".

(b) Interpret the `marginality.06` and `PAN.governor.06` coefficients.

The coefficient for marginality.06 in the Poisson regression model is -2.08014. This indicates that for a one-unit increase in marginality, holding all other variables constant, the expected log count of PAN presidential candidate visits decreases by 2.08014.

The coefficient for PAN.governor.06 is -0.31158. This means that for a one-unit increase in the presence of a PAN governor in a district, holding all other variables constant, the expected log count of PAN presidential candidate visits decreases by 0.31158.

(c) Provide the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive (`competitive.district`=1), had an average poverty level (`marginality.06` = 0), and a PAN governor (`PAN.governor.06`=1).

```
mean_visits <- exp(-3.81023 - 0.08135*1 - 2.08014*0 - 0.31158*1)
mean_visits
```

```
[1] 0.01494827
```

This means that on average, the winning PAN presidential candidate is estimated to make 0.0149 visits to a hypothetical district with the given characteristics.