# Problem Set 3

Applied Stats II

Due: March 28, 2022

## 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 class on Monday March 28, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

## 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.

   The R code I used to answer this question was as follows:

```
gdpChange <- read.csv("C:/Users/olivi/OneDrive/Documents/TCD ASDS/Stats II/PS3/gdpChange.csv")

summary(gdpChange)


head(gdpChange)

tail(gdpChange)

######################################
#Assigning levels to make variable categorical
#in the correct way

#if value is no change = 0

#if value is positive 1

#if value is negative 2

gdpChange <- within(gdpChange, {
  GDPWdiff.cat <- NA # need to initialize variable
  GDPWdiff.cat[GDPWdiff < 0] <- "negative"
  GDPWdiff.cat[GDPWdiff == 0] <- "no change"
  GDPWdiff.cat[GDPWdiff > 0] <- "positive"
} )

gdpChange$GDPWdiff.cat <- factor(gdpChange$GDPWdiff.cat, levels = c("no change", "positive", "negative"))

summary(gdpChange$GDPWdiff.cat)

is.factor(gdpChange$GDPWdiff.cat) #this is now a factor variable, as it should be


#Coding categorical variables as factors-
#Links: https://bookdown.org/carillitony/bailey/chp6.html -
#"R uses factor vectors to to represent dummy or categorical data.
#Factors can be ordered or unordered. Factor vectors are built on top of integer vectors and
```

```
#include a unique label for each integer."

#making sure REG and OIL are also factors as they should be:

gdpChange$REG <- as.factor(gdpChange$REG)

gdpChange$OIL <- as.factor(gdpChange$OIL)


#finally able to move onto setting the reference category
#and creating the regression model

#setting reference category
gdpChange$GDPWdiff2 <- relevel(gdpChange$GDPWdiff.cat, ref = "no change")


multinom_gdp <- multinom(gdpChange$GDPWdiff2 ~ REG + OIL, data =
    gdpChange)

summary(multinom_gdp)

#exp(coef(multinom_gdp)[ ,c(1:3)]) #this was giving me crazy values, so I
    commented it out
#and decided not to exponentiate for interpretation in the end




```

Summary output of unordered multinomial model- Table 1.

Interpretation:

Predictor/explanatory variables-

REG- positive- for a unit change in REG i.e. going from 0 to 1, non-democracy to democracy, the log odds that there will be a positive change in GDP from one year to the next increase by 1.769007, when all other variables in the model are held constant and the reference category is "no change".

REG- negative- for a unit change in REG i.e. going from 0 to 1, non-democracy to democracy, the log odds that there will be a negative change in GDP from one year to the next increase by 1.379282, when all other variables in the model are held constant and the reference category is "no change."

OIL- positive- when there is a unit change in the OIL variable, i.e. increasing from 0

Table 1:

| | *Dependent variable:* | |
|---|---|---|
| | positive | negative |
| | (1) | (2) |
| REG1 | 1.769** | 1.379* |
| | (0.767) | (0.769) |
| OIL1 | 4.576 | 4.784 |
| | (6.885) | (6.885) |
| Constant | 4.534*** | 3.805*** |
| | (0.269) | (0.271) |
| Akaike Inf. Crit. | 4,690.770 | 4,690.770 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

to 1, it means the average ratio of fuel exports to total exports in 1984-86 exceeded 50 per cent. Here, when there is a unit change in the oil variable and the average ratio of fuel exports to total exports in 1984-86 exceeded 50 per cent, the log odds that there will be a positive difference in GDP in a country from one year to the next increase by 4.576321, when all other variables in the model are held constant and the reference category is "no change."

OIL- negative- when there is a unit change in the oil variable from 0 to 1, meaning the average ratio of fuel exports to total exports in 1984-86 exceeded 50 per cent, the log odds that there will be a negative difference in GDP in a country from one year to the next increase by 4.783968, when all other variables in the model are held constant and the reference category is "no change."

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

   The R code I used for this part was as follows:

```
ordered_multinom_gdp <- polr(GDPWdiff2 ~ REG + OIL, data = gdpChange,
    Hess = TRUE)

summary(ordered_multinom_gdp)

```

```
#exponentiating for interpretation, as per lecture slides:
exp(cbind(OR=coef(ordered_multinom_gdp), confint(ordered_multinom_gdp)))

```

Summary output of ordered logit- see table 2:

Table 2:

| | *Dependent variable:* |
|---|---|
| | GDPWdiff2 |
| REG1 | $-0.357^{***}$ |
| | (0.075) |
| OIL1 | $0.231^{**}$ |
| | (0.115) |
| Observations | 3,721 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Exponentiated coefficient output- see Table 3-

Table 3:

| | OR | 2.5 % | 97.5 % |
|---|---|---|---|
| REG1 | 0.700 | 0.604 | 0.810 |
| OIL1 | 1.259 | 1.003 | 1.575 |

Interpretation:

Interpreting odds ratios here based on week 8 slides 8-10-

REG - where REG = 1, i.e. a country is a democracy, there is a 0.7000737 times increase in the odds that their GDP will differ from the previous year than non-democratic countries, when all other variables in the model are held constant.

OIL- when OIL =1 , i.e. the average ratio of fuel exports to total exports in a country in 1984-86 exceeded 50 per cent, the odds of a difference in that country's GDP from one year to the next are 1.2593051 times greater than countries with a less than 50 per cent ratio, when all other variables in the model are held constant.

## 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.

Here is the R code from the poisson regression I created:

```r

mexico_data <- read.csv("C:/Users/olivi/OneDrive/Documents/TCD ASDS/Stats II/PS3/MexicoMuniData.csv")


#outcome variable:

mexico_data$PAN.visits.06


#predictor variables- making sure they're factors as they should be as dummy/categorical variables:

mexico_data$PAN.governor.06 = as.factor(mexico_data$PAN.governor.06)

mexico_data$competitive.district = as.factor(mexico_data$competitive.district)

#checking to see how the data looks now
View(mexico_data)


mexico_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, family = poisson (link = "log"), data=mexico_data)
```

```r


summary(mexico_poisson)

exp(mexico_poisson$coefficients)


```

Summary output for poisson regression model- see Table 4.

Table 4:

| | *Dependent variable:* |
|---|---|
| | PAN.visits.06 |
| competitive.district1 | −0.081 |
| | (0.171) |
| marginality.06 | $-2.080^{***}$ |
| | (0.117) |
| PAN.governor.061 | $-0.312^{*}$ |
| | (0.167) |
| Constant | $-3.810^{***}$ |
| | (0.222) |
| Observations | 2,407 |
| Log Likelihood | −645.606 |
| Akaike Inf. Crit. | 1,299.213 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Z value test statistic for competitive.district: -0.477

P value for competitive.district: 0.6336

Exponentiated coefficient output for poisson regression model- see Table 5:

For competitive.district = 1, i.e. a swing district:
beta = -0.08135
exp(beta)= 0.92186932

Table 5:

| (Intercept) | competitive.district1 | marginality.06 | PAN.governor.061 |
|---|---|---|---|
| 0.022 | 0.922 | 0.125 | 0.732 |

The coefficient is negative, so the expected count is 0.92186932 times smaller than when X= 0, with all else held constant.

This expected count refers to visits by PAN presidential candidates to swing districts, i.e. the count of visits would be 0.92 times smaller, with all else held constant.

Thus, PAN presidential candidiates are 0.92 times less likely to visit swing districts than a "safe seat."

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

marginality.06-
beta = -2.08014361
exp(beta) = 0.12491227

The coefficient is negative, so...

When there is a unit increase in marginality.06, the expected count decreases by 0.12491227 times.

PAN.governor.06-
beta = -0.31157887
exp(beta) = 0.73228985

The PAN.governor.06 coefficient is negative, so when there is a unit increase in PAN.governor.06, i.e. when PAN.governor.06 = 1 and the state has a PAN-affiliated governor, expected count decreases by 0.73228985 times.

(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).

```r

#Using method on week 9 slide 14:

coeff <- mexico_poisson$coefficients

est_mean <- exp(coeff[1] + coeff[2]*1 + coeff[3]*0 + coeff[4]*1)

est_mean #this seems to be such a low value because there's a lot of zeros in the PAN.visits.06 column


```

Estimated mean output found: (Intercept) 0.01494818 visits