Problem Set 3

Student: Shekhar Kedia (23351315)
Applied Stats II

Due: March 24, 2024

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 24, 2024. 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.

   Solution:
   Firstly using `R`, we load the dataset. Then, create levels for the response variable (including a reference category) by running the following codes:

```r
# Loading the dataset to the environment
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)

# Creating levels for factor variable and setting the reference (base) category
gdp_data$GDPWdiff_l <- ifelse(gdp_data$GDPWdiff > 0, "positive",
                 ifelse(gdp_data$GDPWdiff < 0, "negative", "no change"))
gdp_data$GDPWdiff_l <- relevel(factor(gdp_data$GDPWdiff_l, ordered=F),
                               ref="no change")
```

   Then, we run the unordered multinomial logit model with `GDPWdiff` as the outcome variable and "no change" as the reference category and create the summary output using the following codes:

```r
# Running the unordered multinomial model
multinom_unor <- multinom(GDPWdiff_l ~ REG + OIL, data = gdp_data)
summary(multinom_unor)
stargazer(multinom_unor) #Exporting results to LaTeX
```

Table 1:

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

   The result shows that we have two different regression models, one where we estimate shift from "no change" category to "negative" category and the second where we estimate shift from "no change" category to "positive" category for the response variable.

For the "negative" model (i.e., shift from "no change" to "negative"), we can make the following interpretations:

- The intercept value refers to the increase in estimated log odds on average for `GDPWdiff` outcome variable moving from "no change" to "negative" when the `REG` and `OIL` values are both 0.
- We see that change of `REG` from 0 to 1 or "non-democracy" to "democracy" is associated with an increase of 1.379 log odds on average for `GDPWdiff` outcome variable moving from "no change" to "negative" keeping `OIL` value constant. However, the coefficient is significant at 10% error probability.
- The coefficient for `OIL` is not statistically reliable.

For the "positive" model (i.e., shift from "no change" to "positive"), we can make the following interpretations:

- The intercept value refers to the increase in estimated log odds on average for `GDPWdiff` outcome variable moving from "no change" to "positive" when the `REG` and `OIL` values are both 0.
- We see that change of `REG` from 0 to 1 or "non-democracy" to "democracy" is associated with an increase of 1.769 log odds on average for `GDPWdiff` outcome variable moving from "no change" to "positive" keeping `OIL` value constant.
- The coefficient for `OIL` is not statistically reliable.

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

Solution:
I reorder the levels for the response variable and set "negative" as the reference category by running the following codes:

```
# Setting the order of the factor variable
gdp_data$GDPWdiff_1 <- relevel(gdp_data$GDPWdiff_1, ref="negative")
levels(gdp_data$GDPWdiff_1) #Checking the levels are in order: negative,
    no change, positive
```

Then, I run the ordered multinomial logit model with `GDPWdiff` as the outcome variable and create the summary output using the following codes:
N.B.: Note that the P-value was calculated separately and added to the table

```
# Running the ordered multinomial model
multinom_or <- polr(GDPWdiff_1 ~ REG + OIL, data = gdp_data, Hess = T)
summary(multinom_or)

# Calculating the p-value
ctable <- coef(summary(multinom_or))
```

```r
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
```

```
                        Value   Std. Error    t value      p value
REG                 0.3984834 0.07518479   5.300054  1.157687e-07
OIL                -0.1987177 0.11571713  -1.717271  8.592967e-02
negative|no change -0.7311784 0.04760375 -15.359680  3.050770e-53
no change|positive -0.7104851 0.04750680 -14.955440  1.435290e-50
```

We can see from the results that change of `REG` from 0 to 1 or "non-democracy" to "democracy" is associated with an increase of 0.398 log odds on average for a one step change of `GDPWdiff` from base value i.e., "negative" to "no change" and from "no change" to "positive" keeping `OIL` value constant.

Also, change of `OIL` from 0 to 1 is associated with a decrease of 0.198 log odds on average for a one step change of `GDPWdiff` from base value i.e., "negative" to "no change" and from "no change" to "positive" keeping `REG` value constant. However, the coefficient is significant at 10% error probability.

The cutoff points refer to the shift in category from "negative" to "no change" when the value is -0.73 and shift in category from "no change" to "positive" when the value is -0.71.

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.

Solution:
Firstly using `R`, we load the dataset. Then, we run the Poisson regression model as the outcome (`PAN.visits.06`) is a count variable using the following codes:

```r
# Loading the dataset to the environment
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")

# Running the poisson regression model
mexico_poisson <- glm(PAN.visits.06 ~ competitive.district + PAN.governor.06 + marginality.06, data = mexico_elections, family = poisson)
summary(mexico_poisson)
```

It's better to run a over-dispersion test to see if a zero-inflated model is required. We run the over-dispersion test in `R` using the `dispersiontest()` function from the "`AER`" package using the following codes:

```r
# Running the overdispersion test to check over-dispersion
dispersiontest(mexico_poisson)
```

```
Overdispersion test:

data:  mexico_poisson
z = 1.0668, p-value = 0.143
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
2.09834
```

We see that we don't have sufficient evidence (p-value > 0.05) to reject the null hypothesis that the true dispersion is not greater than 1. In other words, we don't find evidence for over-dispersion and so, we don't need a zero-inflated model and can use the base model.

The output from the base model is as follows:

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

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
PAN.governor.06      -0.31158    0.16673  -1.869   0.0617 .
marginality.06       -2.08014    0.11734 -17.728   <2e-16 ***
---
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
```

From the result we see that the coefficient for `competitive.district` is not statistically reliable i.e., the test-statistics (-0.477) is not large enough to reject the null hypothesis that competitive.district has no effect on PAN presidential candidates visit and the p-value is greater than 0.05.
Therefore, we don't find enough evidence to say that PAN presidential candidates visit swing districts more.

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

Solution:
From the above regression output, we can interpret that a one unit increase in `marginality.06` value i.e., measure of poverty is associated with a decrease in 2.08 log count PAN presidential candidates visit on average, keeping `PAN.governor.06` and `competitive.district` value constant.

Also, we can interpret that a one unit increase in `PAN.governor.06` i.e., moving from state not having a PAN-affiliated governor to state having a PAN-affiliated governor is associated with a decrease in 0.31 log count PAN presidential candidates visit on average, keeping `marginality.06` and `competitive.district` value constant.

(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).

Solution:
We use the following codes to make the prediction:

```r
# Making the prediction
predict(mexico_poisson, newdata = data.frame(competitive.district = 1,
    marginality.06 = 0, PAN.governor.06 = 1), type = "response")
```

We see that the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive, had an average poverty level and a PAN governor is found to be 0.01494818.