Problem Set 3

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.

   Run the `R` code:

```
##### Problem 1a. the un-ordered regression model

# first step, we need to recode the GDPWdiff variable with ifelse command:
# if < 0 - "negative"
# if = 0 - "no change"
# if > 0 - "positive"
gdp_data$GDPWdiff <- ifelse(gdp_data$GDPWdiff > 0, "positive",
                         ifelse(gdp_data$GDPWdiff == 0, "no change",
                                "negative"))
# transfer GDPWdiff into factor variable,
# this step is to make sure we can put GDPWdiff into response variable
gdp_data$GDPWdiff <- as.factor(gdp_data$GDPWdiff)
gdp_data$GDPWdiff # check if we successful
# the results show:
# Levels: negative no change positive. so we succeed.

# second step, let's fit the model, hu-rrray!
# we have these 2 predictors:
# - REG: 1 for Democracy, 0 for Non-Democracy
# - OIL: 1 for the average ratio of fuel exports in 1984-86 exceeded 50%;
#       0 for otherwise
# and before we fit the model, we should remember to regard "no change"
# as our reference level
gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")
# fit the model
p1a_model <- multinom(GDPWdiff ~ REG + OIL,
                    data = gdp_data)
summary(p1a_model)
stargazer(p1a_model)
# third step, exponentiation the coefficient to interpret
exp(coef(p1a_model)[,c(1:3)])
```

   And we will get the regression results from `R`:

Table 1: Outcome variable is GDPWdiff and the explanatory variables are REG and OIL

	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	

So, we can write our regression formula:

$$ln(\frac{P_{negative}}{P_{nochange}}) = 3.805 + 1.379\texttt{REG} + 4.784\texttt{OIL} \tag{1}$$

$$ln(\frac{P_{positive}}{P_{nochange}}) = 4.534 + 1.769\texttt{REG} + 4.576\texttt{OIL} \tag{2}$$

We can interpret the coefficient that:

For regression formula (1), we can know:

- The constant (3.805) means the log odds of increase in estimated on average for GDPWdiff moving from "no change" to "negative" when REG and OIL are both 0.
- The coefficient of REG (1.379) refers that when OIL is stable, REG change from "non-democracy" (0) to "democracy" (1) will associated with an increase of 1.379 log odds on average for GDPWdiff moving from "no change" to "negative". But we should notice that the signifficant of coefficient of REG is during 0.05 to 0.1.
- The coefficient of OIL (4.784) is not signifficant.

For regression formula (2) we can know:

- The constant (4.534) means the log odds of increase in estimated on average for GDPWdiff moving from "no change" to "positive" when REG and OIL are both 0.
- The coefficient of REG (1.769) refers that when OIL is stable, REG change from "non-democracy" (0) to "democracy" (1) will associated with an increase of 1.769 log odds on average for GDPWdiff moving from "no change" to "negative".
- The coefficient of OIL (4.576) is not signifficant.

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

   We can write `R` code to fit the model and calculate the p-value:

```
##### Problem 1b. the ordered regression model

# because in Problem 1a., we have set "no change" as our reference level.
# but in Problem 1b., we don't need this anymore.
# and if we keep the reference level,
# we will make a wrong order (no change - negative - positive)
# so we need to re-factor variable GDPWdiff, to assign the right order, which is
# (negative - no change - positive)
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff,
                            ordered = TRUE,
                            levels = c("negative", "no change", "positive"))
# then fit the model
p1b_model <- polr(GDPWdiff ~ REG + OIL,
                  data = gdp_data)
summary(p1b_model)

# next step, we wish to calculate p-value and combine p-value into current table
ctable <- coef(summary(p1b_model))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
stargazer(ctable)
```

   And we will get output from `R`:

Table 2: The outcome variable is `GDPWdiff` and explanatory variables are `REG` and `OIL`

	Value	Std. Error	t value	p value
REG	0.398	0.075	5.300	0.00000
OIL	-0.199	0.116	-1.717	0.086
negative|no change	-0.731	0.048	-15.360	0
no change|positive	-0.710	0.048	-14.955	0

So, we can write our regression formula:

$$ln(\frac{P_{negative}}{P_{nochange}}) = -0.7312 + 0.398\texttt{REG} - 0.199\texttt{OIL} \tag{3}$$

$$ln(\frac{P_{positive}}{P_{nochange}}) = -0.7105 + 0.398\texttt{REG} - 0.199\texttt{OIL} \tag{4}$$

We can interpret the coefficient that:

- The cutoff points refers to the shift from "negative" to "no change" and "no change" to "positive" is -0.7312 and -0.7105 respectively.
- The coefficient of `REG` (0.398) refers when `OIL` is stable, `REG` change from "non-democracy" (0) to "democracy" (1) will associated with an increase of 0.398 log odds on average for `GDPWdiff` moving from one step to next one i.e. "negative" to "no change" and from "no change" to "positive".
- The coefficient of `OIL` (-0.199) refers when `REG` is stable, `OIL` change from fuel exports below 50% (0) to exceeded 50% (1) will ssociated with an increase of -0.199 log odds on average for `GDPWdiff` moving from one step to next one i.e. "negative" to "no change" and from "no change" to "positive". But we should notice that the signifficant of coefficient of `OIL` is during 0.05 to 0.1.

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.

Table 3:

	Dependent variable:
	PAN.visits.06
competitive.district	-0.081
	(0.171)
marginality.06	-2.080^{***}
	(0.117)
PAN.governor.06	-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

We can write the poisson regression formula:

$$ln(\lambda) = -3.810 - 0.081 competitive.district - 2.080 marginality.06 - 0.312 PAN.governor.06 \tag{5}$$

To draw a conclusion about if PAN presidential candidate visit swing district more, we need to check the signifficant of coefficient of `competitive.district`.

The Z value of `competitive.district` is -0.477, and the p value of `competitive.district` is 0.6336, so it is not significant. This means we don't find enough evidence to say the coefficient of `competitive.district` is differ from 0. So we can conclude that there is not evidence that PAN presidential candidate visit swing districts more.

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

- The coefficient of `marginality.06` (-2.080) refers when `PAN.governor.06` and `competitive.district` are stable, 1 unit increase in `marginality.06` is associated with an increase of -2.080 log odds on average for the mean value of `PAN.visits.06`.
- The coefficient of `PAN.governor.06` (-0.312) refers when `marginality.06` and `competitive.district` are stable, `PAN.governor.06` change from no PAN-affiliated governor (0) to have PAN-affiliated governor (1) will associated with an increase of -0.312 log odds on average for the mean value of `PAN.visits.06`. But we should notice that the signifficant of coefficient of `PAN.governor` is during 0.05 to 0.1.

(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).

According to the regression formula (5), we can know when `competitive.district` = 1, `marginality.06` = 0 and `PAN.governor` = 1, the formula is:

$ln(\lambda) = -3.810 - 0.081 * 1 - 2.080 * 0 - 0.312 * 0 = -3.891$

So the estimated log odds mean number of `PAN.visits.06` is -3.891 under this hypothetical district.