

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 for which 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.

```
1 #####
2 # Problem 1
3 #####
4 library(nnet)
5
6 ### 1
7 # load data
8 gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
9 head(gdp_data)
10
11 # Assign negative numbers, 0 and positive numbers to different strings
    respectively
12 gdp_data$GDPWdiff_category <- ifelse(gdp_data$GDPWdiff < 0, "negative",
13                                     ifelse(gdp_data$GDPWdiff == 0, "no_
        change", "positive"))
14
15 # Convert required columns to factor variables
16 gdp_data$GDPWdiff_category <- factor(gdp_data$GDPWdiff_category, ordered
    = FALSE )
17 gdp_data$REG <- as.factor(gdp_data$REG)
18 gdp_data$OIL <- as.factor(gdp_data$OIL)
19
20 # Check the transformed dataframe
21 head(gdp_data)
22
23 # Set "no change" to the reference category
24 gdp_data$GDPWdiff_category <- relevel(gdp_data$GDPWdiff_category, ref = "
    no_change")
25
26 # run the multinomial regression model
27 multinom_model <- multinom(GDPWdiff_category ~ REG + OIL, data = gdp_data
    )
28
29 # Check the information of the model
30 summary(multinom_model)
31
32 # so, the unordered multinomial logit models are :
33 # for negative:  $\ln(P_{\text{negative}}/P_{\text{no change}}) = 3.805370 + 1.379282 \cdot X_{\text{REG}} + 4.783968 \cdot X_{\text{OIL}}$ 
34 # for positive:  $\ln(P_{\text{positive}}/P_{\text{no change}}) = 4.533759 + 1.769007 \cdot X_{\text{REG}} + 4.576321 \cdot X_{\text{OIL}}$ 
35
36 # Exponentiate coefficients
```

37 `exp(coef(multinom_model)[,c(1:3)])`

Table 1: Results

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

Table 2: Exponentiated Coefficients

	(Intercept)	REG1	OIL1
negative	44.94	3.97	119.58
positive	93.11	5.87	97.16

for negative:

cutoff point is 3.805370;

For every one unit increase in REG, the log-odds of Y = no change vs. Y = negative increase by 1.379282

For every one unit increase in OIL, the log-odds of Y = no change vs. Y = negative increase by 4.783968

for positive:

cutoff point is 4.533759;

For every one unit increase in REG, the log-odds of Y = no change vs. Y = positive increase by 1.769007

For every one unit increase in OIL, the log-odds of Y = no change vs. Y = positive increase by 4.576321

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

```

1 ### 2
2
3 # Run ordered logit
4 ordered_model <- polr(GDPWdiff_category ~ REG + OIL, data = gdp_data,
   Hess = T)
5
6 # Check the information of the model
7 summary(ordered_model)
8
9 # so, the ordered multinomial logit models are :
10 # ln( P(Y <= no change)) = -5.3199 + 0.4102*REG -0.1788*OIL
11 # ln( P(Y <= negative)) = -0.7036 + 0.4102*REG -0.1788*OIL
12
13
14 # Get odds ratios and CIs
15 exp(cbind(OR = coef(ordered_model), confint(ordered_model)))

```

Table 3: Results

	<i>Dependent variable:</i>
	GDPWdiff_category
REG1	0.410*** (0.075)
OIL1	-0.179 (0.115)
Observations	3,721
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Table 4: Odds Ratios and 95% Confidence Intervals

	OR	2.5%	97.5%
REG1	1.51	1.30	1.75
OIL1	0.84	0.67	1.05

Cutoff points are -5.3199 and -0.7036 which are intercepts (I do not know why they disappeared here)

if P is less than -5.3199, then predict Y_i = no change
 if -5.3199 is less than P is less than -0.7036, then predict Y_i = negative
 if P is bigger than -0.7036, then predict Y_i is positive;

for $P(Y$ is less than or equal to no change):

For every one unit increase in REG, the log-odds of Y = no change vs. Y = negative increase by 0.4102;

For every one unit increase in OIL, the log-odds of Y is no change vs. Y is negative increase by -0.1788;

for $P(Y$ is less than or equal to negative):

For every one unit increase in REG, the log-odds of Y = negative vs. Y = positive increase by 0.4102;

For every one unit increase in OIL, the log-odds of Y = negative vs. Y = positive increase by -0.1788;

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.

```
1 ### 1
2
3 # load data
4 mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/
  StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
5 head(mexico_elections)
6 View(mexico_elections)
7
8 # Convert required columns to factor variables
9 mexico_elections$PAN.governor.06 <- as.factor(mexico_elections$PAN.
  governor.06)
```

```

10 mexico_elections$competitive.district <- as.factor(mexico_elections$
    competitive.district)
11
12 # Run poisson regression model
13 poisson_model <- glm( PAN.visits.06 ~ PAN.governor.06 + competitive.
    district + marginality.06, data = mexico_elections, family = poisson(
    link = "log"))
14
15 # Check the information of the model
16 summary(poisson_model)
17
18 # so the poisson regression model is :
19 # ln(lambda) = -3.81023 - 0.31158*X_PAN.governor.06 - 0.08135*X_
    competitive.district - 2.08014*X_marginality.06

```

Table 5: Results

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

After checking the results of regression, competitive.district's test statistic is -0.477, P value is 0.6336 which is not significant, so there is no evidence that PAN presidential candidates visit swing districts more.

- (b) Interpret the `marginality.06` and `PAN.governor.06` coefficients.
For the coefficient of `marginality.06` which is -2.08014, it means increasing marginal-

ity.06 by 1 unit has a multiplicative effect on the mean of Poisson by $\exp(-2.08014)$;

For the coefficient of PAN.governor.06 which is -0.31158, it means increasing PAN.governor.06 by 1 unit has a multiplicative effect on the mean of Poisson by $\exp(-0.31158)$;

- (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`).

```
1 ### 3
2
3 # Extract coefficients
4 coeffs <- coefficients(poisson_model)
5
6 # Substitute values into the model, so we can get the estimated mean
  number of visits
7 lambda <- exp(coeffs[1] + coeffs[2]*1 + coeffs[3]*1 + coeffs[4]*0)
8
9 # Print the results
10 print(lambda)
```

After calculating, the estimated mean number of visits is 0.01494818, almost 0.