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.

```r
######################
# Problem 1
######################
library(nnet)

### 1
# load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
head(gdp_data)

# Assign negative numbers, 0 and positive numbers to different strings respectively
gdp_data$GDPWdiff_category <- ifelse(gdp_data$GDPWdiff < 0, "negative",
                                     ifelse(gdp_data$GDPWdiff == 0, "no_change", "positive"))

# Convert required columns to factor variables
gdp_data$GDPWdiff_category <- factor(gdp_data$GDPWdiff_category, ordered = FALSE )
gdp_data$REG <- as.factor(gdp_data$REG)
gdp_data$OIL <- as.factor(gdp_data$OIL)

# Check the transformed dataframe
head(gdp_data)

# Set "no change" to the reference category
gdp_data$GDPWdiff_category <- relevel(gdp_data$GDPWdiff_category, ref = "no_change")

# run the muiltinomial regression model
multinom_model <- multinom(GDPWdiff_category ~ REG + OIL, data = gdp_data)

# Check the information of the model
summary(multinom_model)

# so,the unordered multinomial logit models are :
# for negative: ln(P_negative/P_no change) = 3.805370 + 1.379282*X_REG + 4.783968*X_OIL
# for positive: ln(P_positive/P_no change) = 4.533759 + 1.769007*X_REG + 4.576321*X_OIL

# Exponentiate coefficients
```

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

Table 1: Results

	Dependent variable:	
	negative	positive
	(1)	(2)
REG1	1.379*	1.769**
	(0.769)	(0.767)
OIL1	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	

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.

```r
### 2

# Run ordered logit
ordered_model <- polr(GDPWdiff_category ~ REG + OIL, data = gdp_data,
    Hess = T)

# Check the information of the model
summary(ordered_model)

# so,the ordered multinomial logit models are :
# ln( P(Y <= no change)) = -5.3199 + 0.4102*REG -0.1788*OIL
# ln( P(Y <= negative)) = -0.7036 + 0.4102*REG -0.1788*OIL


# Get odds ratios and CIs
exp(cbind(OR = coef(ordered_model), confint(ordered_model)))
```

Table 3: Results

	Dependent variable:
	GDPWdiff_category
REG1	0.410***
	(0.075)
OIL1	−0.179
	(0.115)
Observations	3,721
Note:	*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 Yi = no change
if 5.3199 is less than P is less than -0.7036, then predict Yi = negative
if P is bigger than -0.7036, then predict Yi 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.

```r
### 1

# load data
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
head(mexico_elections)
View(mexico_elections)

# Convert required columns to factor variables
mexico_elections$PAN.governor.06 <- as.factor(mexico_elections$PAN.governor.06)
```

```r
mexico_elections$competitive.district <- as.factor(mexico_elections$
    competitive.district)

# Run poisson regression model
poisson_model <- glm( PAN.visits.06 ~ PAN.governor.06 + competitive.
    district + marginality.06, data = mexico_elections, family = poisson(
    link = "log"))

# Check the information of the model
summary(poisson_model)

# so the poisson regression model is :
# ln(lambda) = -3.81023 - 0.31158*X_PAN.governor.06 - 0.08135*X_
    competitive.district - 2.08014*X_marginality.06
```

Table 5: Results

	Dependent variable:
	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
Note:	*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).

```r
### 3

# Extract coefficients
coeffs <- coefficients(poisson_model)

# Substitute values into the model, so we can get the estimated mean
    number of visits
lambda <- exp(coeffs[1] + coeffs[2]*1 + coeffs[3]*1 + coeffs[4]*0)

# Print the results
print(lambda)
```

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