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.

   Codes as below:

```
# make a copy
df <- gdp_data
# add a new column as outcome variable
df$GDPWdiff <- ifelse(df$GDPWdiff > 0, 'positive', ifelse(df$GDPWdiff < 0, '
    negative', 'no change'))
# convert integers into factors
df$GDPWdiff <- as.factor(df$GDPWdiff)
df$REG <- as.factor(df$REG)
df$OIL <- as.factor(df$OIL)
# set a reference variable for the outcome
df$GDPWdiff <- relevel(df$GDPWdiff, ref = "no change")
# run a unordered multinomial logistic regression model
mult_log <- multinom(GDPWdiff ~ REG + OIL, data = df)
```

   Summary output:

	Model 1
negative: (Intercept)	3.805***
	(0.271)
negative: REG	1.379
	(0.769)
negative: OIL	4.784
	(6.885)
positive: (Intercept)	4.534***
	(0.269)
positive: REG	1.769*
	(0.767)
positive: OIL	4.576
	(6.885)
AIC	4690.770
BIC	4728.101
Log Likelihood	−2339.385
Deviance	4678.770
Num. obs.	3721
K	3

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 1: unordered multinomial model

   Regression models:

$$\ln\left(\frac{P(GDPWdiff = \text{negative})}{P(GDPWdiff = \text{no change})}\right) = 3.81 + 1.38 \times \text{REG} + 4.78 \times \text{OIL}$$

$$\ln\left(\frac{P(GDPWdiff = \text{positive})}{P(GDPWdiff = \text{no change})}\right) = 4.53 + 1.77 \times \text{REG} + 4.58 \times \text{OIL}$$

Coefficient interpretation:

For the first model (predicting negative GDP change):

1). Intercept -3.81 (cutoff point):
In non-democratic countries where fuel exports are not the dominant exports (i.e., REG=0 and OIL=0), the log odds suggest that the likelihood of GDP decrease compared to no change in GDP is lower (as the coefficient is negative).

2). Coefficient for REG 1.38:
Holding other conditions constant, the log odds of GDP decrease versus no change in GDP is, on average, 1.38 times higher in democratic countries compared to non-democratic countries. However, the non-significant p-value indicates that we do not have sufficient evidence to support that the regression coefficient is significantly different from zero.

3). Coefficient for OIL 4.78:
Holding other conditions constant, for countries where fuel exports dominate, the log odds of GDP decrease is, on average, 4.78 times higher compared to other countries. However, the non-significant p-value indicates that we do not have sufficient evidence to support that the regression coefficient is significantly different from zero.

For the second model (predicting positive GDP change):

1). Intercept -4.53 (cutoff point):
In non-democratic countries where fuel exports are not the dominant exports, the likelihood of GDP growth is lower (as the log odds are negative).

2). Coefficient for REG 1.77:
Holding other conditions constant, the log odds of GDP growth versus no change in GDP is, on average, 1.77 times higher in democratic countries compared to non-democratic countries, implying that democracy may have a stronger positive correlation with GDP growth.

3). Coefficient for OIL 4.58:
Holding other conditions constant, for countries where fuel exports dominate, the log odds of GDP growth versus no change in GDP is, on average, 4.58 times higher compared to other countries. However, the non-significant p-value indicates that we do not have sufficient evidence to support that the regression coefficient is significantly different from zero.

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

Codes as below:

```r
# run an ordered multinomial logistic regression model
df$GDPWdiff <- factor(df$GDPWdiff,
                      levels = c("negative","no change","positive"),
                      ordered = TRUE)

ord_log <- polr(GDPWdiff ~ REG + OIL, data = df, Hess = TRUE)
```

Regression models:

$$\ln\left(\frac{P(GDPWdiff = \text{negative})}{P(GDPWdiff = \text{no change})}\right) = -0.731 + 0.398 \times \text{REG} - 0.199 \times \text{OIL}$$

$$\ln\left(\frac{P(GDPWdiff = \text{no change})}{P(GDPWdiff = \text{positive})}\right) = -0.710 + 0.398 \times \text{REG} - 0.199 \times \text{OIL}$$

Summary output:

	Model 2
REG	0.398^{***}
	(0.075)
OIL	-0.199
	(0.116)
$Negative \| noChange$	-0.731^{***}
	(0.048)
$NoChange \| Positive$	-0.710^{***}
	(0.048)
hline AIC	4695.689
BIC	4720.576
Log Likelihood	-2343.845
Deviance	4687.689
Num. obs.	3721

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: ordered multinomial logit models

Coefficient interpretation:

1). Coefficient for REG 0.398:
Holding other conditions constant, the log odds of GDP decrease versus GDP same (or GDP same versus GDP growth)is, on average, 0.398 times higher in democratic countries compared to non-democratic countries.

2). Coefficient for OIL -0.199:
Holding other conditions constant, the log odds of GDP decrease versus GDP same (or GDP same versus GDP growth)is, on average, 0.199 times lower in countries where fuel exports exceed half of total exports to other else. However, the non-significant p-value indicates that we do not have sufficient evidence to support that the regression coefficient is significantly different from zero.

3) Cutoff point for Negative | No Change -0.731
The cutoff point for comparing the negative and no change categories. A negative coefficient implies that as the log odds increase, the probability of predicting the outcome as no change relative to negative increases.

4) Cutoff point for No Change | Positive -0.710
The cutoff point for comparing the no change and positive categories. A negative coefficient indicates that as the log odds increase, the probability of predicting the outcome as positive relative to no change increases.

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.

Codes as below:

```
# perform poisson regression model
reg.ps <- glm(PAN.visits.06 ~., data = dat, family = poisson)
summary(reg.ps)
# over-dispersion test
install.packages("AER")
library(AER)
dispersiontest(reg.ps)
```

Summary output:

	Model
(Intercept)	−3.810***
	(0.222)
marginality.06	−2.080***
	(0.117)
PAN.governor.06TRUE	−0.312
	(0.167)
competitive.districtTRUE	−0.081
	(0.171)
AIC	1299.213
BIC	1322.357
Log Likelihood	−645.606
Deviance	991.253
Num. obs.	2407

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 3: poisson regression model

As we can see from the table above, the coefficient of the competitive.district dummy variable (where TRUE = 1 represents swing districts) is negative, indicating a negative association between the number of visits and swing districts. Additionally, the p-value of the competitive.district variable is greater than 0.05 (lack of asterisks), suggesting that there is insufficient evidence to conclude that PAN presidential candidates visit swing districts more.

The result of over-dispersion test (P-value = 0.143 > 0.05) shows that there is no significant evidence to reject null hypothesis that the data show no signs of overdispersion.

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

Coefficient interpretation:

1). Coefficient for marginality.06 (2.080)
This coefficient is negative and significant ($p < 0.001$), suggesting that as the measure of poverty increases, the expected count of the PAN presidential candidate's visits decreases.

Specifically, for each unit increase in the marginality index, the log count of visits is expected to decrease by 2.080 on average, holding constant all other variables.

Exponentiate coefficients: an increase of one unit in marginality decreases the expected mean count of visits by a multiplicative factor of exp(-2.080) = 0.125.

2). Coefficient for PAN.governor.06 (TRUE: -0.312)
This negative coefficient suggests that there is a decrease in the expected count of visits by the PAN candidate in districts where the governor is affiliated with PAN, though it's not statistically significant.

The coefficient implies that, compared to when the governor is not from PAN, having a governor from PAN is associated with a decrease of 0.312 in the log count of visits, holding all else equal.

Exponentiate coefficients: having a governor from PAN compared to not from PAN, the expected mean count of visits decrease by a multiplicative factor of exp(-0.312) = 0.732.

(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).

Codes as below:

```r
# c) Estimated mean of visits
coef <- coef(reg.ps)
exp(coef[1] + coef[2]*0 + coef[3]*1 + coef[4]*1)
# Result: 0.0149
```

In this given context, the estimated mean number of visits is about 0.0149.