Problem Set 3

Yuanyuan Liu

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

ftable(xtabs(~OIL+REG+GDPWdiff,data = gdp_data))

#do some wrangling
gdp_data$GDPWdiff <- ifelse(gdp_data$GDPWdiff > 0, 'positive',
                            ifelse(gdp_data$GDPWdiff < 0, 'negative', 'no change'))
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff,
                            levels = c('positive', 'negative','no change'),
                            labels = c('positive', 'negative','no change'))
gdp_data$REG <- factor(gdp_data$REG,
                       levels = c(0,1),
                       labels = c('Non-Democracy','Democracy'))
gdp_data$OIL <- factor(gdp_data$OIL,
                       levels = c(0,1),
                       labels = c('Otherwise','exceeded'))

ftable(xtabs(~OIL+REG+GDPWdiff,data = gdp_data))

#set a reference level for the outcome
gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")

#Constract a unordered multinomial logit
unorder_model <- multinom(GDPWdiff~OIL+REG, data = gdp_data,)
summary(unorder_model)
texreg(list(unorder_model), digits=3)

# get p values
z <- summary(unorder_model)$coefficients/summary(unorder_model)$standard.errors
(p <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

	Model 1
positive: (Intercept)	4.534^{***}
	(0.269)
positive: OILexceeded	4.576
	(6.885)
positive: REGDemocracy	1.769^{*}
	(0.767)
negative: (Intercept)	3.805^{***}
	(0.271)
negative: OILexceeded	4.784
	(6.885)
negative: REGDemocracy	1.379
	(0.769)
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: Statistical models

interpret the cutoff points:
Positive: The intercept of GDPWdiff is 4.534. This is the log-odds of being a ”positive” category in difference in GDP between year t and t-1 compared to ”no change” when all the predictor variables are the reference level.
Negative: The intercept of GDPWdiff is 3.805. This is the log-odds of being a ”negative” category in difference in GDP between year t and t-1 compared to ”no change” by 3.805 when all the predictor variables are the reference level.

Interpret the coefficients:
OILexceeded:
Positive: The coefficient is 4.576. This suggests that for countries where the average ratio of fuel exports to total exports exceeded 50%, the log-odds of having a ”positive” change in difference in GDP between year t and t-1 (as opposed to ”no change”) is increased by 4.576 compared to otherwise.
Negative: The coefficient is 4.784. This suggests that for countries where the average ratio of fuel exports to total exports exceeded 50%, the log-odds of having a ”negative” change in difference in GDP between year t and t-1 (as opposed to ”no change”) is increased by 4.576 compared to otherwise.

REGDemocracy:
Positive: The coefficient is 1.769. This indicates that for democracies, the log-odds of having a ”positive” change in difference in GDP between year t and t-1 (as opposed to ”no change”) are increased by 1.769 compared to non-democracies.
Negative: The coefficient is 1.379. For democracies, the log-odds of having a ”negative” change in difference in GDP between year t and t-1 (as opposed to ”no change”) are increased by 1.379 compared to non-democracies.

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

```
#Perform an ordered (proportional odds) logistic regression
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff,
                            levels = c( 'negative','no change','positive'))
levels(gdp_data$GDPWdiff)
order_model <- polr(GDPWdiff~OIL+REG, data = gdp_data, Hess = TRUE)
summary(order_model)
texreg(list(order_model), digits=3)

#Calculate P value
ctable <- coef(summary(order_model))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
```

	Model 1
OILexceeded	-0.199
	(0.116)
REGDemocracy	0.398^{***}
	(0.075)
negative—no change	-0.731^{***}
	(0.048)
no change—positive	-0.710^{***}
	(0.048)
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: Statistical models

Coefficients

OILexceeded: The coefficient is -0.199. This suggests that for countries where the average ratio of fuel exports to total exports exceeded 50%, the log-odds of being a higher category in difference in GDP between year t and t-1 (such as going from 'negative' to 'no change', or from 'no change' to 'positive') is decreased by -0.199 compared to otherwise. and the t value of -1.717 suggests that this effect might not be statistically significant.

REGDemocracy: The coefficient is 0.398. Being a democracy is associated with increase log-odds of being a higher category of in difference in GDP between year t and t-1 by 0.398 compared to non-democracies. And the t value of 5.300 indicates this is likely a statistically significant predictor.

Intercepts (Thresholds or Cutoff Points):

no change—positive: The intercept is -0731. The log-odds of being in the "no change" category in difference in GDP between year t and t-1 from "negative" is -07312 when OILexceeded and REGDemocracy are the refernece level. The value of t value is -15.3597, indicating that this intercept is statistically significant.

positive—negative: The intercept is -0.7105. The log-odds of being in the "positive" category in difference in GDP between year t and t-1 from "no change" is -0.711 when OILexceeded and REGDemocracy are the refernece level. The value of t value is -14.955, which also means that this intercept is statistically significant.

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.

```
#run a poisson regression
# load data
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
# EDA
str(mexico_elections)
summary(mexico_elections)

with(
  mexico_elections,
  list(mean(PAN.visits.06), var(PAN.visits.06))
) # we can meet assumptions for Poisson


#Run a poisson regression model
model.pois <- glm(PAN.visits.06 ~competitive.district+marginality.06+PAN.governor.06, data = mexico_elections, family = poisson)
summary(model.pois)
texreg(list(model.pois), digits=3)
```

	Model 1
hline(Intercept)	-3.810^{***}
	(0.222)
competitive.district	-0.081
	(0.171)
marginality.06	-2.080^{***}
	(0.117)
PAN.governor.06	-0.312
	(0.167)
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: Statistical models

The coefficient for competitive.district is -0.081, with a standard error of 0.171 and a z value of -0.477. The associated p-value is 0.6336. The p-value is greater than common significance levels (0.05, 0.01, etc.), indicating that we can refuse the null hypothese, there is not evidence that PAN presidential candidates visit swing districts more.

(b) Interpret the `marginality.06` and `PAN.governor.06` coefficients.
The coefficient for marginality.06 is -2.080. One unit increase of marginality.06 (a measure of poverty), excepted to decrease log count of the number of visits by the PAN presidential candidate by 2.080. And p-value is <2e16, it can prove that the marginality has a significant impact on competitive.district on the significant level of 0.001.
The coefficient for PAN.governor.06 is -0.312. This coefficient reflects the expected change in the log count of the number of visits by the PAN presidential candidate if the district has a PAN-affiliated governor (1) versus compared having one (0), with other variables held constant. And p-value is 0.062, it can prove that the marginality has a significant impact on competitive.district on the significant level of 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).

```
# predicted the mean number by R
pred <- data.frame(
  competitive.district=1,
  marginality.06 = 0,
  PAN.governor.06=1
)

# check with predict() function
predict(model.pois, newdata = pred, type = "response")

# calculate pseudo R squared
1 - (model.pois$deviance / model.pois$null.deviance)

# c) Over-dispersion?
install.packages("AER")
library(AER)

dispersiontest(model.pois)
```

Table 4: Model Prediction Results

Observation	**Predicted Probability**
1	0.01494818

The value of estimated mean number of visits from the winning PAN presidential candidate is 0.0149.

Table 5: Overdispersion Test Results

Statistic	**Value**
Test Statistic (z)	1.0668
p-value	0.143
Alternative Hypothesis	true dispersion is greater than 1
Sample Estimate	dispersion: 2.09834

Although a dispersion estimate is 2.0983, greater than 1 indicates evidence of overdispersion, a p-value greater than 0.05 indicates that this level of dispersion is not statistically significant. That is, according to this test, the dispersion of the data is not significantly higher than the hypothesis of a Poisson distribution.