Problem Set 3 Claire Mooney

Applied Stats II

Due: March 28, 2022

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 class on Monday March 28, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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

1 Question 1 part 1

data2 ¡- read.csv(”https://raw.githubusercontent.com/
ASDS-TCD/StatsII$_{spring}2022/main/datasets/MexicoMuniData.csv$”)

data1 ¡- read.csv(”https://raw.githubusercontent.com/
ASDS-TCD/StatsII$_{spring}2022/main/datasets/gdpChange.csv$”)

Question 1 part 1

```
library(tidyverse)
library(ggplot2)
set.seed(1234)
install.packages(”MASS”)
library(MASS)
library(nnet)
library(ggplot2)
install.packages(”AER”)
library(AER)
install.packages(”dplyr”)
library(dplyr)


dispersiontest(mod.ps)

install.packages(”pscl”)
library(pscl)

summary(data1)

ftable(xtabs(  GDPWdiff + REG + OIL, data = data1))

x ¡- as.numeric(data1(dollar)GDPWdiff)
```

I was running into issues with the levels so i thought of using a for loop
it assigns all positive values the value of 1
and all negative values -1
it keeps 0 at 0 and this then allowed me to factor

```
for (y in c(1:length(x)))  if (x[y] ¿ 0)  x[y] ¡- 1  else if (x[y] == 0)  x[y] ¡- 0

for (y in c(1:length(x)))  if (x[y] != 1  x[y] != 0)  x[y] ¡- as.numeric(-1)

data1(dollar)GDPWdiff ¡- x

data1(dollar)GDPWdiff ¡- as.factor(data1(dollar)GDPWdiff)

data1(dollar)GDPWdiff ¡- factor(data1(dollar)GDPWdiff,
levels = c(1,-1, 0),
labels = c(”Positive”,
”Negetive”,
”No Change”)
```

I ran into issues with the above code and leveling it did not seems to matter whether or not I factorised the data that was I decided to run the for loop

```
data1(dollar)REG ¡- as.factor(data1(dollar)REG)
data1(dollar)REG ¡- factor(data1(dollar)REG, levels = c(0,1), labels = c(”Non-Democracy”,
”Democracy”))

data1(dollar)OIL ¡- as.factor(data1(dollar)OIL)
data1(dollar)OIL ¡- factor(data1(dollar)OIL, levels = c(0,1), labels = c(”Above 50(percent)”,
”Below 50(percent”))

levels(data1(dollar)OIL)
levels(data1(dollar)GDPWdiff)

ftable(xtabs(  GDPWdiff + REG + OIL, data = data1))
OIL Above 50 Below 50
GDPWdiff REG
Positive Non-Democracy 1284 195
Democracy 1074 47
Negetive Non-Democracy 641 93
Democracy 332 39
No Change Non-Democracy 14 0
```

Democracy 2 0

b) fit a multinomial logit model
set a reference level for the outcome

```
data1(dollar)GDPWdiff ¡- relevel(data1(dollar)GDPWdiff, ref = ”No Change”)
```

run model

```
mult(underscore)log ¡- multinom(data1(dollar)GDPWdiff ., data = data1, MaxNWts = 5200)
```

```
weights: 5199 (3464 variable)
initial value 4087.936326
iter 10 value 1481.481086
iter 20 value 863.921586
iter 30 value 321.785585
iter 40 value 121.129018
iter 50 value 8.226435
iter 60 value 0.196041
iter 70 value 0.000226
final value 0.000056
converged
```

having run a foreloop I was initially unsure of the outcome

```
summary(mult(underscore)log)
exp(coef(mult(underscore)log))
```

```
z ¡- summary(mult(underscore)log)(dollar)coefficients/summary(mult(underscore)log)(dollar)standard.erro
(p ¡- (1 - pnorm(abs(z), 0, 1)) * 2)
```

this would not run on my machine but
I would have done this to get a p value

```
summary(z)
```

```
exp(cbind(OR = coef(mult(underscore)log), confint(mult(underscore)log)))
```

gives odds ratio and confidence interval

For the purposes of this PDF I have changed dollar signs and underscore symbols to their titles.
My original code did run these symbols.

2 Question 1 part 2

```
ord.log ¡- polr(GDPWdiff  ., data = data1, Hess = TRUE) summary(ord.log)
```

again I was having huge issues with r so I couldn't get the output of this model but this is the code I would have theoretically ran

whilst I am unsure if this is correct I wanted to try and see the difference between ordered and unordered multinomial logit

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.

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

(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).

3 Question 2 part 1

```
summary(data2)

data2(dollar)competitive.district ¡-
factor(data2(dollar)competitive.district,
levels = c(0,1),
labels = c(”swing”, ”close”))

mod.ps ¡- glm(PAN.visits.06  ., data = data2, family = poisson)

summary(mod.ps)
```

this returns a value of 2.325 for voter if you take the exponent of 2.325
returns a value of 10.22668009 therefore visiting districts increases

voteshare and visiting swing districts would have a positive impact on the number of votes a canidate recieves.
the expected voteshare increases as the number of visits increases

```
lambda ¡- exp(mod.ps(dollar)coefficients[1] +
mod.ps(dollar)coefficients[2])
lambda
0.02172524
```

```
dispersiontest(mod.ps)
z = 1.0651, p = 0.1434
alternative hypothesis: true dispersion is greater than 1 sample estimates:
dispersion = 2.033988
```

this means we do not have to fit ZIP model

```
z ¡- summary(mod.ps)(dollar)coefficients/summary(mod.ps)(dollar)standard.errors
(p ¡- (1 - pnorm(abs(z), 0, 1)) * 2)
```

again this wouldn't run on my machine but this is what I would have done to get the p values

4 Question 2 part 2

marginality returned a value of -2.060 when you take the exponent returns

0.1274539699 this is not a hugely siginicient figure but does suggest

that visiting people does have a marginiality effect on the marginiality of the votes

pan governor returned a value of -2.690 0.06788093937 this suggests although
it is a positive value above 0 that the govenor did have an affect but a very
small one not a significant impact of them being an influencing factor
as to whether canidates visited swing districts more

```
dispersiontest(mod2.ps)
data: mod2.ps
z = -13.094, p-value = 1
```

alternative hypothesis: true dispersion is greater than 1 sample estimates:
dispersion = 0.7847943

lambda ¡- exp(mod2.ps$coefficients[1] +$
$mod2.ps$coefficients[2])
lambda
0.398053

5 Question 2 part 3

data2(dollar)competitive.district ¡-
factor(data2(dollar)competitive.district,
levels = c(0,1), labels = c(”swing”, ”close”))

counter ¡- 0

Filtering data2 for competetive.district == 1 (later I would have done marginality.06 == 0 and PAN.governor.06 == 1) however I did not have enough time
my R was not running data efficiently I am unsure if it is due to the large
nature of the data or my laptop or R not being able to run the data quickly or sometimes at all
for (x in data2(dollar)competitive.district)
if (x == ”swing”)
data2 ¡- data2[-c(counter),]
print(”Found”)
else
print(”True”)

counter ¡- counter + 1

Need to filter data2 for competetive district == 1, PAN governor == 1 and marginality == 0
Then I can run a model to find the mean for PAN.visits.06 on this filtered data

mod3.ps ¡- glm(PAN.visits.06 ., data = data2, family = poisson) dispersiontest(mod3.ps)