Problem Set 3

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.

```r
data <- read.csv("C:/Users/crowl/Documents/Trinity/ASDS Course/POP77003_Applied_Statistical_Analysis_2/ProblemSet/PS3/template/GDPchange.csv")


library(foreign)
library(nnet)
library(stargazer)
library(jtools)
library(MASS)
library(nnet)
library(ggplot2)
install.packages("pscl")
library(pscl)

summary(data)
```

Firslty loading in my data and inspecting it.

```r
data$GDPWdiff <- cut(x = data$GDPWdiff, breaks = c(-9257, -0.9, 0.1, 7867)) #creating bounds of values for the levels , as they are real numbers
levels(data$GDPWdiff) <- c("negative", "no_change", "positive")

data$GDPWdiff


data$GDPWdiff <- factor(data$GDPWdiff , ordered = FALSE ) #how to make a table un-ordered

table(data$GDPWdiff)


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

Firstly I inspect my data and see that I need to create a bounds of values for the 3 catagories of "negative", "no change", "positive". I then create a contingency table to help run a regression. Then I set the reference category to “no change” as specified

```r
data$GDPWdiff = relevel(data$GDPWdiff, ref = "no_change") #setting the reference level to 0

# run model
mult.log <- multinom(GDPWdiff ~ OIL + REG, data = data)
summary(mult.log)
expc <- exp(coef(mult.log))

```

```
stargazer(expc, type="latex")

# get p values
z <- summary(mult.log)$coefficients/summary(mult.log)$standard.errors
z
p <- ((1 - pnorm(abs(z), 0, 1)) * 2) #2 tailed z test
p

stargazer(p, type="latex")

library(stargazer)
stargazer(mult.log, type="latex") #gives the P value and coefficents

head(fitted(mult.log)) # checking the fitted values


##############
```

In this I set my reference catagory and run the regression, and obtain the P values and coefficents.

Table 1:

	(Intercept)	OIL	REG
negative	44.934	116.492	3.976
positive	93.118	95.344	5.867

These are the logit coefficients relative to the reference category of no change. For example, under 'REG', the 1.38 suggests that for one unit increase in 'REG' score, the logit coefficient for 'negative' relative to 'no change' will go increase by 1.38. In other words, if your REG score increases one unit, your chances of staying in the noc hange GDPWdiff category are higher compared to staying in low GDPWdiff. When the coefficients are expontiated keeping all other variables constant, if the oil score increases one unit, you are 116.492 times more likely to stay in the negative category as compared to the no change category. The coefficient, however, is not significant. keeping all other variables constant, if the oil score increases one unit, you are 95.344 times more likely to stay in the positive category as compared to the no change category. The coefficient, however, is not significant. The p values are not statistically significant and we fail to reject the null hypothesis of our model

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

```
ord.log <- polr(GDPWdiff ~ OIL + REG, data = data, Hess = TRUE)
summary(ord.log)

```

Table 2:

	Dependent variable:	
	negative	positive
	(1)	(2)
OIL	4.758	4.557
	(6.823)	(6.823)
REG	1.380*	1.769**
	(0.769)	(0.767)
Constant	3.805***	4.534***
	(0.271)	(0.269)
Akaike Inf. Crit.	4,688.792	4,688.792
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
stargazer(ord.log, type=”latex”)

# Calculate a p value
ctable <- coef(summary(ord.log))
p <- pnorm(abs(ctable[, ”t value”]), lower.tail = FALSE) * 2
ctable <- cbind(ctable, ”p value” = p)
ctable

stargazer(ctable, type=”latex”)

# Calculate confidence intervals
ci <- confint(ord.log)
ci


# convert to odds ratio
exp(cbind(OR = coef(ord.log), ci))
```

For nations with REG(1) ie. Democracies, the odds of being more likely (i.e., very or somewhat likely versus unlikely) to have positive change in GDP is 105 percent more than undemocratic nations, holding constant all other variables. For nations with Oil(1) ie. 50 percent + exports, the odds of being more likely (i.e., very or somewhat likely versus unlikely) to have positive change in GDP is 84 percent more than non 50 percent oil exporting nations, holding constant all other variables.

Table 3:

	Dependent variable:
	GDPWdiff
OIL	−0.172
	(0.116)
REG	0.409***
	(0.075)
Observations	3,720
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4:

	Value	Std. Error	t value	p value
OIL	-0.172	0.116	-1.483	0.138
REG	0.409	0.075	5.446	0.00000
no_change|negative	-5.319	0.252	-21.084	0
negative|positive	-0.704	0.048	-14.798	0

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
Mex_data <- read.csv("C:/Users/crowl/Documents/Trinity/ASDS Course/POP77003_Applied_Statistical_Analysis_2/ProblemSet/PS3/template/MexicoMuniData.csv")

summary(Mex_data)
str(Mex_data)

with(Mex_data, #Testing to see if our response variable meets the assumptions for a Poisson test.
     list(mean(PAN.visits.06), var(PAN.visits.06))) #The outcome would suggest it does as the expected mean and expected  are approx equal
```

loading in the data. As the mean is 0.09181554 and the variance is 0.6436861, which show they are approximately the same, we progress with running a Poisson regression.

```r
Mex_data <- within(Mex_data, {
  PAN.governor.06 <- as.logical(PAN.governor.06)
  competitive.district <- as.logical(competitive.district)
})

mod.pos <- glm(PAN.visits.06 ~ ., data = Mex_data, family = poisson(link = "log"))
summary(mod.pos)


table(Mex_data$competitive.district) #this shows that FALSE is the reference category

                                                        #I want to change this to TRUE to easily compare my categories
Mex_data$competitive.district <- factor(Mex_data$competitive.district , ordered = FALSE )

```

```r
Mex_data$competitive.district2 <- relevel(Mex_data$competitive.district,
    ref= "TRUE")

table(Mex_data$competitive.district2) #shows the referennce has now
    changed

mod.pos1 <- glm(PAN.visits.06 ~ MunicipCode + pan.vote.09 + marginality
    .06 + PAN.governor.06 +competitive.district2, data = Mex_data, family
    = poisson)
summary(mod.pos1)
```

As competitive.district and PAN.governor.06 are binary (0/1) responses, I will turn them into logical (true/false) responses. I then run the model.

Table 5:

	Dependent variable:
	PAN.visits.06
MunicipCode	−0.00001
	(0.00001)
pan.vote.09	0.232***
	(0.052)
marginality.06	−2.060***
	(0.120)
PAN.governor.06	−0.269
	(0.167)
competitive.district2FALSE	−0.068
	(0.182)
Constant	−3.761***
	(0.244)
Observations	2,407
Log Likelihood	−639.535
Akaike Inf. Crit.	1,291.070
Note:	*p<0.1; **p<0.05; ***p<0.01

In our model when the district is non competitive the intercept is -3.829, with all other factors held at zero. And -3.761 when the district is competitive, both of these

Table 6:

	Dependent variable:
	PAN.visits.06
MunicipCode	−0.00001
	(0.00001)
pan.vote.09	0.232^{***}
	(0.052)
marginality.06	-2.060^{***}
	(0.120)
PAN.governor.06	−0.269
	(0.167)
competitive.district	0.068
	(0.182)
Constant	-3.829^{***}
	(0.302)
Observations	2,407
Log Likelihood	−639.535
Akaike Inf. Crit.	1,291.070
Note:	*p<0.1; **p<0.05; ***p<0.01

intercepts have p values that denote significance. When we change the model from non competitive to competitive we see a 0.64 difference in the coefficient intercept.

The coefficient for competitive.district is 0.6849. This means that the expected log count for a one-unit increase in competitive.district is 0.6849. However the p value of competitive.district (0.182) is not significant and therefore we must conclude that a districts competitiveness does not significantly influence the vistits of a presidential candidate.

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

```
cofs <- coef(mod.pos)
cofs

stargazer(cofs, type=”latex”)

confint(mod.pos) #confidence interval set at 95%

exp(coef(mod.pos)) #exponential model coefficients
exp(confint(mod.pos)) #CI for the exponential model coefficients


sjPlot::tab_model(mod.pos , show.intercept = TRUE, #shows the P value
                  show.se = FALSE, dv.labels = ”Predictors of presidental
    Visit”, auto.label = TRUE, show.re.var = FALSE, show.icc =FALSE,
                  show.r2 = FALSE, show.ngroups = FALSE, show.obs = FALSE
    )
```

marginality.06 has a coefficient of -2.06 and a p value of 0.12 which is statistically significant. PAN.governor.06 has a coefficient of -0.269 and a p value of 0.167 which is not statifcally significant. After this I then created a 95percent confidence interval and exponentiated the model's coefficients. This then produced a table showing each of the variables exponentiated coefficients.

there appears to be a significant association between marginality.06 and PAN.visits.06 IRR = 0.13 P value = ¡0.001 , 95percent CI = 0.10 – 0.16

there appears to be a non-significant association between PAN.governor.06 and PAN.visits.06 IRR = 0.76 P value = 0.107 , 95 percent CI = 0.55 – 1.05

Table 7:

(Intercept)	MunicipCode	pan.vote.09	marginality.06	PAN.governor.06TRUE	competitive.dist
-3.829	-0.00001	0.232	-2.060	-0.269	0.068

(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).

```
exp(cofs[1] + cofs[2] + cofs[3] + cofs[4]*0 + cofs[5]*1 + cofs[6]*1)

pred <- data.frame(Mex_data$MunicipCode,
                   Mex_data$pan.vote.09,
                   marginality.06 = 0,
                   PAN.governor.06 = TRUE,
                   competitive.district = TRUE)

colnames(pred)

names(pred)[1] <- "MunicipCode" #had to rename the columns, or they would not be recognised
names(pred)[2] <- "pan.vote.09"

colnames(pred)


# check with predict() function
predict(mod.pos, newdata = pred, type = "response")

mod.zip <- zeroinfl(PAN.visits.06 ~ ., data = Mex_data, dist = "poisson") #i get an error message when I attempt to run this, it a non finite
summary(mod.zip) # vlaue has been supplied, I have not been able to reslove it.
```

When the model is exponentiated the expected intercept is 0.02172546 When the model is exponentiated and the variables are selected the expected intercept is 0.02243164