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.

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

Question 1: Part 1

```r
# load data
gdp <- read.csv("gdpChange.csv")

# View data structure
summary(gdp)
str(gdp)

# Convert values to character strings
gdp2 <- gdp
gdp2$GDPWdiff2 <- gdp$GDPWdiff

gdp2$GDPWdiff2[gdp2$GDPWdiff == 0] <- "no change"
gdp2$GDPWdiff2[gdp2$GDPWdiff < 0] <- "negative"
gdp2$GDPWdiff2[gdp2$GDPWdiff > 0] <- "positive"

# Remove order from factors
#gdp2$GDPWdiff2 <- factor(gdp2$GDPWdiff2, ordered = FALSE)

# Set reference level for factors
gdp2$GDPWdiff2 <- relevel(as.factor(gdp2$GDPWdiff2), ref = "no change")

# Remove additional column data
gdp2 <- subset(gdp2, select = c(REG, OIL, GDPWdiff2))

summary(gdp2)
str(gdp2)

# Unordered multinomial logit
multinom_1 <- multinom(GDPWdiff2 ~ REG + OIL, data = gdp2)

# Getting multinom_1 information, including coefficients
multinom_1
```

Table 1: Multinom_1 Results

	Dependent variable:	
	negative	positive
	(1)	(2)
REG	1.379*	1.769**
	(0.769)	(0.767)
OIL	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	

```r
# Specifically coefficients
coef_1 <- exp(coef(multinom_1))
```

Table 2: Multinom_1 Coefficients

	(Intercept)	REG	OIL
negative	44.942	3.972	119.578
positive	93.108	5.865	97.156

```r
# Cutoff points
confint_exp_m1 <- exp(confint(multinom_1))
```

Table 3: Multinom_1 Confidence Intervals: Negative

	2.5%	97.5%
(Intercept)	$2.643900e + 01$	$7.639360e + 01$
REG	$8.804391e - 01$	$1.791965e + 01$
OIL	$1.647467e - 04$	$8.679315e + 07$

Table 4: Multinom_1 Confidence Intervals: Positive

	2.5%	97.5%
(Intercept)	$5.493416e + 01$	$1.578085e + 02$
REG	$1.304269e + 00$	$2.637379e + 01$
OIL	$1.339263e - 04$	$7.048166e + 07$

Interpretation of coefficients and cutoff points

So in any particular country, there is a 5.865 times increase in the baseline odds that there will be positive GDP growth if that country is a democracy, holding everything else at its mean. The 95% confidence intervals for this 5.865 value are 1.304 26.373.

Question 1: Part 2

```r
#### 2. Ordered Multinomial Logit ####

gdp3 <- gdp
gdp3$GDPWdiff3 <- gdp$GDPWdiff

gdp3$GDPWdiff3[gdp3$GDPWdiff == 0] <- "no change"
gdp3$GDPWdiff3[gdp3$GDPWdiff < 0] <- "negative"
gdp3$GDPWdiff3[gdp3$GDPWdiff > 0] <- "positive"

str(gdp3)

# Remove additional column data
gdp3 <- subset(gdp3, select = c(REG, OIL, GDPWdiff3))

# Check new data structure
summary(gdp3)

#Check nature of ordering for gdp3$GDPWdiff3
is.ordered(gdp3$GDPWdiff3) # returns FALSE

# Make it ordered
as.ordered(gdp3$GDPWdiff3) # returns "Levels: negative < no change < positive"
```

```r

summary(gdp3)

# Ordered multinomial logit
multinom_2 <- multinom(GDPWdiff3 ~ REG + OIL, data = gdp3)

# Getting multinom_2 information, including coefficients
multinom_2
```

Table 5: Multinom_2 Results

	Dependent variable:	
	no change	positive
	(1)	(2)
REG	-1.352^{*}	0.390^{***}
	(0.758)	(0.076)
OIL	-7.924	-0.208^{*}
	(32.977)	(0.116)
Constant	-3.801^{***}	0.728^{***}
	(0.270)	(0.048)
Akaike Inf. Crit.	4,690.728	4,690.728
Note:	*p<0.1; **p<0.05; ***p<0.01	

```r
# Specifically coefficients
exp(coef(multinom_2))
```

Table 6: Multinom_2 Coefficients

	(Intercept)	REG	OIL
no change	0.02234416	0.2587991	0.0003619269
positive	2.07177984	1.4768404	0.8124904479

```r
# And cutoff points
exp(confint(multinom_2))
```

Table 7: Multinom_2 Confidence Intervals: No Change

	2.5%	97.5%
(Intercept)	$1.315877e-02$	$3.794135e-02$
REG	$5.855130e-02$	$1.143903e+00$
OIL	$3.078737e-32$	$4.254701e+24$

Table 8: Multinom_2 Confidence Intervals: Positive

	2.5%	97.5%
(Intercept)	$5.493416e+01$	$1.578085e+02$
REG	$1.304269e+00$	$2.637379e+01$
OIL	$1.339263e-04$	$7.048166e+07$

Interpretation of coefficients and cutoff points for ordered multinomial
In our ordered multonomial logit with "negative" set as our reference there is a 1.476 times increase in the baseline odds that there will be positive GDP growth if that country is a democracy, holding everything else at its mean. The 95% confidence intervals for this 1.476 value are 1.273 1.712.

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).

Question 2: Part A

```r
# Load data
data <- read.csv("MexicoMuniData.csv")
View(data)
str(data)

# Convert variables to factors
as.factor(data$PAN.governor.06)
as.factor(data$competitive.district)

# Poisson Regression Model
data_poisson <- glm (PAN.visits.06 ~ competitive.district + marginality.06 +
    PAN.governor.06, data=data, family=poisson)
```

Table 9: Poisson Regression

	Dependent variable:
	PAN.visits.06
competitive.district	−0.081
	(0.171)
marginality.06	−2.080***
	(0.117)
PAN.governor.06	−0.312*
	(0.167)
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

From these results it looks like being a competitive district is not is a significant predictor of increased visits.

```r
coeffs <- coefficients(data_poisson)
```

Table 10: Coefficients

(Intercept)	competitive.district	marginality.06	PAN.governor.06
-3.810	-0.081	-2.080	-0.312

```r
coeffs_exp <- exp(coeffs)
```

Table 11: Exponentiated Coefficients

(Intercept)	competitive.district	marginality.06	PAN.governor.06
0.02214298	0.92186932	0.12491227	0.73228985

A 1 unit increase in competitiveness (from 0 to 1, safe to close / swing district) decreases our expected number of visits by a multiplicative factor of 0.92 (the exponent value of -0.081, indeed the log odds standard error of this value is 0.171, which means the true value could be 0).

```r
# Let's look at a histogram of visits
hist(data$PAN.visits.06,
     main = paste("Histogram of visits"),
     xlab = "Number of visits",
     ylab = "Frequency")
```

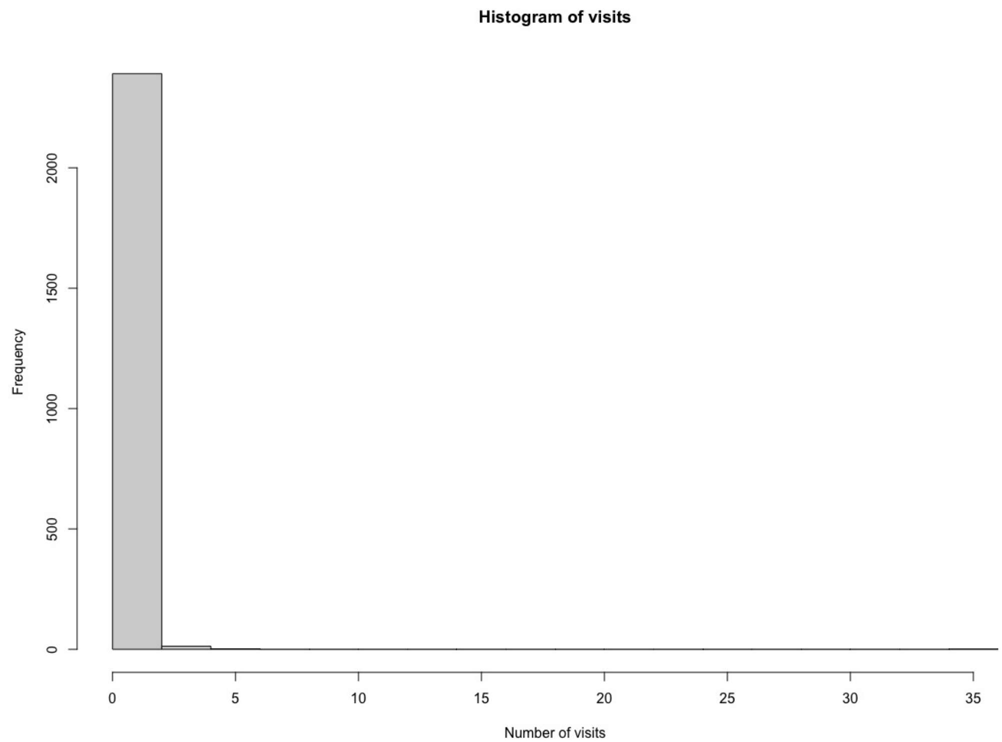

Figure 1: Number of visits to different districts

The large number of '0' values in visits may indicate that a Poisson regression may not be a good fit. To confirm this let's run a dispersion test.

```
disp_test <- dispersiontest(data_poisson)
```

The dispersion test reveals a z-score of 1.067 with a p-value of 0.143.
Let's run a ZIP regression and compare to be sure.

```
zeroinfly_poisson <- zeroinfl(PAN.visits.06 ~ competitive.district +
    marginality.06 + PAN.governor.06, data=data, dist='poisson')
```

Now to compare which is better I'll use a Vuong test of the two models.

```
vuong(data_poisson, zeroinfly_poisson)
```

With the null hypothesis being that the two models are indistinguishable, the results of this voung comparison have p-values of more than 0.10 suggesting that our ordinary Poisson regression model is sufficient. See table 12.

Table 12: Vuong Results

	Vuong z-statistic	H_A p-value
Raw	1.1844529 model2 >	model1 0.11812
AIC-corrected	−1.0796817 model2 >	model1 0.14014
BIC-corrected	−0.7765712 model2 >	model1 0.21871

Question 2: Part B

From table 9, it looks like marginality is a significant negative predictor of visits.

Looking at the exponentiated coefficients in table 11 tells us that a 1 unit increase in marginality decreases our expected number of visits by a multiplicative factor of 0.12 (the exponent value of -2.08).

From the same results having a PAN-affiliate governor is also a significant predictor of visits. The exponentiated coefficients tells us that a 1 unit increase in PAN-affiliated governor (from not being affiliate to being affiliated) decreases our expected number of visits by a multiplicative factor of 0.7322 (the exponent value of -.312).

Question 2: Part C - Estimated mean number of visits.

Going back to the coefficients of the original model (reproduced from table 10)...

Table 13: Coefficients

(Intercept)	competitive.district	marginality.06	PAN.governor.06
-3.810	-0.081	-2.080	-0.312

The predicted mean number of visits for a district that was competitive(1), had an average poverty level(0), and a PAN governor (1) is...

```
exp(coeffs[1] + coeffs[2]*1 + coeffs[3]*0 + coeffs[4]*1)
```

This gives us an estimated number of visits of 0.0149.