Problem Set 3

Applied Stats II

Due: March 26, 2023

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 26, 2023. 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.*

   In order to carry out an unordered multinomial logit the `GDPWdiff` variable first had to be changed to indicate an increase, decrease, or no change. This was done using the following code:

```
#data wrangling
head(data_copy)

data_copy$GDPWdiff <- cut(data_copy$GDPWdiff,
                          breaks = c(-Inf,-.9, .9, Inf),
                          labels = c("NEG", "NONE", "POS"),
                          right = FALSE,
                          ordered_result = FALSE)
```

   An unordered multinomial logit was then constructed using the following code, which gave the output as shown in Table 1:

```
# set a reference level for the outcome
data_copy$GDPWdiff <- relevel(data_copy$GDPWdiff, ref = "NONE")


#Run unordered multinomial model
multinom_model1 <- multinom(GDPWdiff ~ REG + OIL, data = data_copy)
summary(multinom_model1)
exp(coef(multinom_model1))
```

   In Table 1, we can see that only one predictor, whether or not a country was a democracy was significant in determining the log odds of an increase in GDP with respect to no change in GDP. Holding all other variables constant, if a country is a democracy, it increases the log odds by 1.769.

   We can also exponentiate these coefficients to get the multiplicative factors of the covariates, as shown in Table 2.

Table 1: Unordered Multinomial Logit

	Dependent variable:	
	NEG	POS
	(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	

Table 2: Exponentiated Coefficients

	(Intercept)	REG	OIL
NEG	44.942	3.972	119.578
POS	93.108	5.865	97.156

For the democracy covariate in the positive equation, this means that if the country is a democracy, the odds of the difference in GDP being positive with respect to no change in GDP multiply by a factor of 5.865. However, as can be shown by the following table, we can see that the model does not perform well in predicting the correct category of the outcome variable, suggesting that the model does not fit the data well:

```
#prediction accuracy
addmargins(table(data_copy$GDPWdiff, predict(multinom_model1, type = "
    class")))
```

	NONE	NEG	POS	Sum
NONE	0	0	16	16
NEG	0	0	1105	1105
POS	0	0	2600	2600
Sum	0	0	3721	3721

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

   An unordered multinomial logit was then constructed using the following code:

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

   This gave the following output:

Table 3:

	Dependent variable:
	GDPWdiff
REG	0.410^{***}
	(0.075)
OIL	-0.179
	(0.115)
Observations	3,721
Note:	*p<0.1; **p<0.05; ***p<0.01

In interpreting the table, as with the positive equation in the previous model, the `REG` variable was the only significant predictor in the model. For the coefficient, this means that holding all other variables constant, if a country is a democracy, it increases the log odds of the GDP difference being positive with respect to there being a change in

GDP with respect to no change by 0.4.

We can then calculate the odds ratios and confidence intervals:

Table 4:

	OR	2.5 %	97.5 %
REG	1.507	1.301	1.747
OIL	0.836	0.668	1.051

In terms of the `REG` variable, this means that the odds of there being a change in GDP is increased by a multiplicative factor of 1.5. However, we also need to check whether the parallel lines assumption holds. To do this, we can run individual logistic regressions for each category. This produces the following table:

Table 5:

	Dependent variable:		
	GDPWdiff		
	(POS)	(NEG)	(NONE)
REG	0.083^{***}	-0.078^{***}	-0.005^{**}
	(0.015)	(0.015)	(0.002)
OIL	-0.042^{*}	0.047^{*}	-0.006
	(0.025)	(0.025)	(0.004)
Constant	0.669^{***}	0.323^{***}	0.007^{***}
	(0.010)	(0.010)	(0.001)
Observations	3,721	3,721	3,721
Log Likelihood	−2,364.514	−2,350.415	4,869.173
Akaike Inf. Crit.	4,735.028	4,706.831	−9,732.345
Note:			*p<0.1; **p<0.05; ***p<0.01

As can be seen from the table, while the first two logistic regression models are similar, the no change model does not have a similar slope to the other two lines. This may indicate that the categories are not meant to be ordered categories.

Regardless, the table displays all of the predicted probabilities with all combinations of covariates:

Table 6:

	REG	OIL	Level	Probability
1	0	1	NONE	0.006
2	1	1	NONE	0.004
3	0	0	NONE	0.005
4	1	0	NONE	0.003
5	0	1	NEG	0.366
6	1	1	NEG	0.278
7	0	0	NEG	0.326
8	1	0	NEG	0.244
9	0	1	POS	0.628
10	1	1	POS	0.718
11	0	0	POS	0.669
12	1	0	POS	0.753

Therefore, by taking the means of each of the levels' predicted probabilities, the predicted cutoffs are 0 - 0.3035 for the NONE category, 0.3035 - 0.692 for the NEG category, and 0.692 - 1 for POS.

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.*

A Poisson regression was run using the following code:

```
# Conduct poisson regression
mexico_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality
    .06 + PAN.governor.06, data=mexico, family="poisson")
summary(mexico_poisson)
```

As shown in Table 7, the competitive district variable is negative, meaning that the number of times the winning PAN presidential candidate visited a district decreases if the district was a "swing district" by a multiplicative factor of $\exp^{-0.081} = 0.922$. The coefficient did also not have a significant non-zero effect on the outcome variable. This can be displayed through a t-test, where we can get a t statistic by using the following formula:

$$T = \frac{\beta_1}{se_{\beta_1}} = \frac{-0.081}{0.171} = -.0474$$

Using R, we can then find the p-value of this statistic:

```
pt(-0.474, 2406, lower.tail = TRUE)
[1] 0.3177715
```

As $p > 0.05$, we fail to reject the null hypothesis that $\beta_1 = 0$

A pseudo R-squared test was also carried out using the following code:

```
#Pseudo R Squared
1 - (mexico_poisson$deviance/mexico_poisson$null.deviance)
```

Table 7:

	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

This gave a pseudo R-squared value of 0.32745, meaning that the model only explains 33% of the variance in the outcome variable, being the number of visits made by the winning PAN presidential candidate.

A dispersion test was carried out to make sure the assumption of equal variances held true. The test revealed a p-value $> .05$, meaning a zero-inflated Poisson was not carried out.

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

Both the `marginality.06` and `PAN.governor.06` had significant coefficients, meaning they had non-zero effects on the number of times the winning PAN presidential candidate visited the district. The `marginality.06` coefficient indicates that for every unit increase in the poverty measure, the amount of times the winning PAN presidential candidate visited the districted was lowered by a multiplicative factor of $\exp^{-2.080} = 0.125$, while holding the other variables constant.

The `PAN.governor.06` coefficient indicates that if the district had a PAN-affiliated governor, the amount of times the winning PAN presidential candidate visited the districted was lowered by a multiplicative factor of $\exp^{-0.312} = 0.732$, while holding the other variables constant.

(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).*

The estimated mean number of visits from the winning PAN presidential candidate for the above mentioned hypothetical district was obtained from the code below:

```
#estimate
lambda <- exp(-3.810 - 0.081 - 0.312)
lambda
```

```
[1] 0.01495066
```

When looking at the data, this makes sense as the winning PAN presidential candidate only visited 135 out of 2407 districts, meaning they only visited around 6% of districts. The district they visited the most (35 times) also had no PAN-affiliated governor and was a swing district with -1.505 on the marginality score. The next most visited district (5 times) also had no PAN-affiliated governor and was a swing district with -1.831 on the marginality score. Therefore, the model predicts that the winning PAN presidential candidate would most likely not visit the above hypothetical district