Problem Set 3

Applied Stats II

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.

Answer:

```r
# Change GDPWdiff from numeric to categorical, and then factor with reference.

gdp_data$GDPWdiff <- ifelse(gdp_data$GDPWdiff == 0, "no change", ifelse(gdp_data$GDPWdiff > 0, "positive", "negative"))

gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, ordered = FALSE)

gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")

model <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)
```

Table 1: the estimated coefficients of the unordered multinomial logistic regression

	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	

The constant (or cutoff point) 3.805 is the log odds that GDPWdiff will be negative relative to the baseline category of no change when both REG and OIL are coded as zero. The constant (or cutoff point) 4.534 is the log odds that GDPWdiff will be positive relative the baseline category of no change when both REG and OIL are coded as zero.

For a one unit increase in REG, that being when a country goes from being coded as 0 (non-democracy) to 1(democracy), there is a 1.379 unit increase in the log odds of GDPWdiff being negative vs the baseline category of no change. There is also a 1.769 unit increase in the log odds of GDPWdiff being positive vs the baseline category.

For a one unit increase in OIL, that being when a county goes from being coded as one in which the ratio of fuel exports to total exports does not exceed 50% (0), to one in which it does (1), there is a 4.784 unit increase in the log odds of GDPWdiff being negative vs the baseline category of no change. There is also a 4.576 unit increase in the log odds of GDPWdiff being positive vs the baseline category.

Table 2: the exponents of the coefficients (the odds) from the multinominal logistic regression model

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

44.942 represents the multiplicative change in the odds that GDPWdiff will be negative vs the baseline category of no change when both REG and OIL are coded as zero 93.108 represents the multiplicative change in the odds that GDPWdiff will be positive vs the baseline category of no change when both REG and OIL are coded as zero.

3.972 represents the multiplicative change in the odds that GDPWdiff will be negative vs the baseline category of no change when REG changes from 0 to 1. 5.856 represents the multiplicative change in the odds that GDPWdiff will be positive vs the baseline category of no change when REG changes from 0 to 1.

119.578 represents the multiplicative change in the odds that GDPWdiff will be negative vs the baseline category of no change when OIL changes from 0 to 1. 97.156 represents the multiplicative change in the odds that GDPWdiff will be positive vs the baseline category of no change when OIL changes from 0 to 1.

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

Answer:

```
# Order the factors.
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, ordered = TRUE, levels =
  c("negative", "no change", "positive"))

model2 <- polr(GDPWdiff ~ REG + OIL, data = gdp_data)
```

Table 3: the estimated coefficients of the ordinal multinomial logistic regression model2

	Dependent variable:
	GDPWdiff
REG	0.398^{***}
	(0.075)
OIL	-0.199^{*}
	(0.116)
Observations	3,721
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: the estimated odds ratios and their confidence intervals of the ordinal multinominal logistic regression model2.

	OR	2.5 %	97.5 %
REG	1.490	1.286	1.727
OIL	0.820	0.655	1.031

For every one unit increase in REG there is an approximate 0.398 increase in the log odds of moving from a lower category of GDPWdiff (e.g., negative or no change) to a higher category (e.g., no change or positive), holding the other variables constant.

For every one unit increase in OIL there is an approximate 0.199 decrease in the log odds of moving from a lower category of GDPWdiff to a higher category, holding other variables constant.

Table 5: Intercepts (cutoff points) of the ordinal multinomial regression model2

	Value	Std. Error	t value
negative—no change	-0.7312	0.0476	-15.3597
no change—positive	-0.7105	0.0475	-14.9554

-0.7312 represents the log odds of being in the negative category of GDPWdiff relative to no change category, when both OIL and REG are at zero, it is the baseline log odds of the negative category relative to the no change category.

-0.7105 represents the log odds of being in the no change category of GDPWdiff relative to the positive category, when both OIL and REG are at zero, it is the baseline log odds of the no change category relative to positive category.

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.

Answer:

The Poisson regression can be run with the following code.

```
model3 <- glm(PAN.visits.06 ~ competitive.district
+ marginality.06 + PAN.governor.06,
family = "poisson", data = mexico_elections)
```

Table 6: the estimated coefficients of the 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

Table 7: the odds ratios and their confidence intervals for model3

	OR	2.5 %	97.5 %
(Intercept)	0.022	0.014	0.034
competitive.district	0.922	0.666	1.302
marginality.06	0.125	0.099	0.156
PAN.governor.06	0.732	0.523	1.007

Table 8: the estimated coefficients along with their test statistics and p-values for model3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.8102	0.2221	-17.16	0.0000
competitive.district	-0.0814	0.1707	-0.48	0.6336
marginality.06	-2.0801	0.1173	-17.73	0.0000
PAN.governor.06	-0.3116	0.1667	-1.87	0.0617

Null Hypothesis (H0) = There is not enough evidence to support the claim that PAN presidential candidates visit swing districts more.

Alternative Hypothesis (HA) = There is sufficient evidence to support the claim that PAN presidential candidates visit swing districts more.

Given the p-value for the coefficient competitive.district is equal to 0.6336 and this is greater than the level of significance ($\alpha = 0.05$) **we fail to reject the null hypothesis, there is not enough evidence to support the claim that PAN presidential candidates visit swing districts more.**

Given the low explanatory capability of the model in this case, it is worth trying a zero inflated Poisson model on the data to see if they are zero inflated.

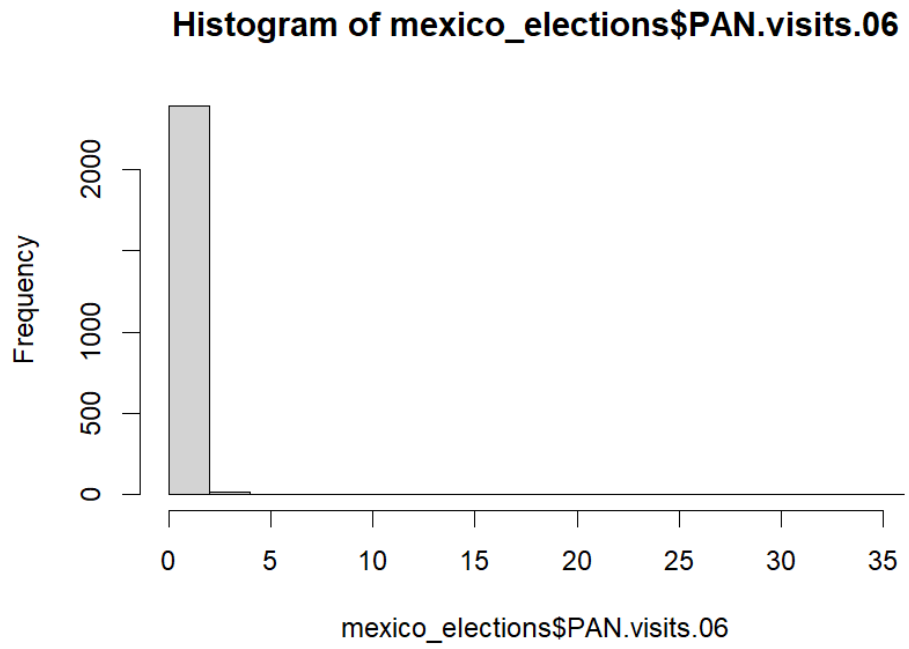

Figure 1: Frequency of number of PAN visits

This histogram of the frequency of each number of PAN visits would suggest that zeros are over represented in the data.

```
library(pscl)
zip_model3 <- zeroinfl(PAN.visits.06 ~ competitive.district
+ marginality.06 + PAN.governor.06,
data = mexico_elections, dist = "poisson")

# Test the models.

library(AER)
dispersiontest(model3)
AIC (model3, zip_model3)
```

After running and testing the zero inflated model is seems the model could be somewhat zero inflated but the high p-value **0.143** of the dispersion test suggests that there is not sufficient evidence to make that claim. Also looking at the coefficients and p-values of the zero inflated model it does not appear to offer more useful insights than the regular Poisson model, e.g. the coefficient p-values are similar. Refer to R file for more details.

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

Answer:

marginality.06: for every one unit increase in marginality.06 there is approximately a 2.080 decrease in the log count of PAN presidential visits, holding all other variables constant.

PAN.governor.06: for a one unit increase in PAN.governor.06, that being when a hypothetical district goes from not having a PAN affiliated governor to having a PAN affiliated governor, there is an approximate 0.312 decrease in the log count of PAN presidential visits, holding all other variables constant. Although considering the high p-value of the PAN.governor.06 coefficient, $0.0617 > \alpha = 0.05$, the effect of this coefficient is not statistically significant.

(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).

Answer:

The following R code is used to do the calculation:

```
exp(model3_cfs[1] + model3_cfs[2]*1 + model3_cfs[3]*0 + model3_cfs[4]*
  1)
```

It does the calculation by getting the exponent of the regression equation for the regression model3, getting the exponent works as a link function in the case of Poisson regression.

The calculation provides a result **= 0.0149** which means that the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive, had an average poverty level, and a PAN governor was 0.0149.

Bibliography:

- Zhang, Z. (no date) 'Estimating The Optimal Cutoff Point For Logistic Regression'.
- Kwak, C. and Clayton-Matthews, A. (2002) 'Multinomial Logistic Regression', Nursing Research, 51(6), p. 404.
- O'Halloran, S. (no date) 'Lecture 10: Logistical Regression II— Multinomial Data'.
- Mokhtar, K.I., Wan Nor Arifin, Tengku Muhammad Hanis Tengku (no date) Chapter 9 Multinomial Logistic Regression — Data Analysis in Medicine and Health using R. (Accessed: 23 March 2024).
- Faraway, J.J. (2016) Extending the Linear Regression Model with R, second edition.
- Dr. Jeffrey Ziegler's lecture material.