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.

   First Create new column and fill it with catagories. Make it a factor and set ”no change” as the reference. Then run the mode (see table one)

```
# Initialize GDP_cat as a character vector with "NA" as placeholder
gdp_data$GDP_cat <- rep("NA", nrow(gdp_data))

# Create category
for (i in 1:nrow(gdp_data)){
  if (gdp_data$GDPWdiff[i] > 0) {
    gdp_data$GDP_cat[i] = "positive"
  }
  else if (gdp_data$GDPWdiff[i] < 0){
    gdp_data$GDP_cat[i] = "negative"
  }
  else gdp_data$GDP_cat[i] = "no change"
}
# Set the ref category to no change
gdp_data$GDP_cat <- relevel(as.factor(gdp_data$GDP_cat), ref = "no change")

#Run regression
mult.log <- multinom(GDP_cat ~ REG + OIL, data = gdp_data)
summary(mult.log)
```

   Intercept negative: When the regime is not a democracy and the fuel export ratio is below 50 percent, the log odds of changing from ”no change” to ”negative” is approximately 3.80. This value is not statistically significant.

   Intercept positive: When fuel export is below 50 percent and the regime is a non-democracy, the estimated log odds of moving from ”no change” to ”positive” is approximately 4.53. This value is not statistically significant.

   Holding OIL constant, a change from not democracy to democracy is associated with an estimated log odds of moving from ”no change” to ”negative” of approximately 1.38. This value is not statistically significant.

   Holding OIL constant, a change from not democracy to democracy is associated with an estimated log odds of moving from ”no change” to ”positive” of approximately 1.78. This value is not statistically significant.

Table 1:

	Dependent variable:	
	negitive	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	

Holding REG constant, a change in the fuel export ratio from below 50 percent to above 50 percent is associated with an estimated log odds of moving from ”no change” to ”negative” of approximately 4.73. This value is not statistically significant.

Holding REG constant, a change in the fuel export ratio from below 50 percent to above 50 percent is associated with an estimated log odds of moving from ”no change” to ”positive” by approximately 4.58. This value is not statistically significant.

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

```
## Ordered Logistic Regression (change the lowest category to reference)
gdp_data$GDP_cat <- relevel(as.factor(gdp_data$GDP_cat), ref = ”negative”
    )
ord.log <- polr(GDP_cat ~ REG + OIL, data = gdp_data, Hess = TRUE)
summary(ord.log)
```

   Holding OIL constant, a change from not democracy to democracy is associated with an estimated log odds of moving up one category (negative -¿ no change -¿ positive) by approximately 0.40.

Holding REG constant, a change in the fuel export ratio from below 5050 percent is associated with an estimated log odds of moving one category (negative -¿ no change -¿ positive) by approximately -0.1987.

Intercepts or Cut off points negative—no change: -0.7312 It is the log odds of being in ”negative” compared to ”no change” or ”positive” when all coefficients are at zero.

no change—positive -0.7105: It is the log odds of being in ”negative” or ”no change” compared to being in ”positive” when all coefficients are at zero.

BONUS
I'll try to take it a step further and test the parallel lines regression assumption. We need to ensure that the relationship between our predictors is the same between each category. I can use this function. As can be seen, the parallel regression assumption holds!

```r
install.packages("brant")
library(brant)
brant(ord.log)
```

Test for X2 df probability

Omnibus 2.64 2 0.27REG 2.64 1 0.1OIL 0 1 0.99

H0: Parallel Regression Assumption holds

Test for	X^2	df	probability
Omnibus	2.64	2	0.27
REG	2.64	1	0.1
OIL	0	1	0.99

Table 2: Test of Parallel Regression Assumption

H_0: Parallel Regression Assumption holds

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

We do not find statistically significant evidence to support the alternate hypothesis that candidates visit swing districts more when the district is competitive (we fail to reject the NULL). The p-value for "competitive.district" variable is 0.6336 (p is not greater than .05) and the test statistic is -0.477.

Table 3:

	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

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

Marginality.06: Holding all other variables constant, a one-unit increase in marginality is associated with an estimated decrease in the expected number of visits by a multiplicative factor of exp(-2.08014) = 0.1249127. This value has high statistical significance.

PAN.governor.06: Holding all other variables constant, a change from PAN governor from 0 to 1 is associated with an estimated decrease in the expected number of visits by a multiplicative factor of exp(-0.31158) = 0.732289. This value is not statistically significant at alpha = 0.05.

(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).

```r
# Estimated mean number of visits.
exp(cfs[1]*1 + cfs[2]*1 + cfs[3]*0 + cfs[4]*1) # 0.01494818
# Above is a linear equation which includes the coefficients at their given values including the intercept.
```

Above is a linear equation that includes the coefficients at their given values, including the intercept. It is estimated that he will make 0.01494818 visits under the given circumstances, so he probably won't visit... so sad.

BONUS

An assumption of Poisson regression is the mean and the variance are the same or close to the same. Let's take a quick look to see if the mean equals the variance in the response variable.

```r
with(mexico_elections,
     list(mean(PAN.visits.06), var(PAN.visits.06))) # do we meet assumptions for Poisson?
```

The mean 0.09181554 and variance is 0.6436861
We need to do further test to make sure but it doesn't like the mean and variance are close enough.
Another problem in Poisson regression is having too many zeros, lets check it out.

```r
table(mexico_elections$PAN.visits.06)
```

```
   0     1    2    3   4   5  35
2272   102   17   12   1   2   1
```

As you can see we have A LOT of zeros, I would popbaly want to do a zero inflated test and check that out.

Thanks for reading