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.

```r
# wrangling
gdp_data_sub$OIL <- as.factor (gdp_data_sub$OIL)
gdp_data_sub$REG <- as.factor (gdp_data_sub$REG)
gdp_data_sub$COUNTRY <- as.factor (gdp_data_sub$COUNTRY)

gdp_data_sub$GDPWdiff_category <- ifelse(gdp_data_sub$GDPWdiff > 0,
                                         "positive",
                                       ifelse(gdp_data_sub$GDPWdiff < 0,
                                              "negative", "no change"))
gdp_data_sub$GDPWdiff_category_unordered <- factor(gdp_data_sub$GDPWdiff_
    category, ordered = FALSE)
gdp_data_sub$GDPWdiff_category_unordered <- relevel(gdp_data_sub$GDPWdiff
    _category_unordered, ref = "no change")

# summarize the original data again
summary(gdp_data_sub)

# run unordered multinomial model
gdp_unorderd <- multinom(gdp_data_sub$GDPWdiff_category_unordered ~ OIL +
    REG, data = gdp_data_sub)
summary(gdp_unorderd)
```

Table 1: Multinomial Regression Results

	Negative	**Positive**
(Intercept)	3.805 (0.271)	4.534 (0.269)
OIL1	4.784 (6.885)	4.576 (6.885)
REG1	1.379 (0.769)	1.769 (0.767)
Residual Deviance	4678.77	
AIC	4690.77	

(a) **The unordered multinomial logit with GDPWdiff as the output and "no change" as the reference category is as follows:**

$$\ln\left(\frac{p(\text{negative})}{p(\text{no change})}\right) = 3.805370 + \text{OIL1} \times 4.783968 + \text{REG1} \times 1.379282$$

Coefficient for OIL1: Holding REG1 constant, for every one unit increase in OIL1, the odds of Y= "negative" vs. Y= "no change" increase by exp(4.783968)

Coefficient for REG1: Holding OIL1 constant, for every one unit increase in REG1, the odds of Y= "negative" vs. Y= "no change" increase by exp(1.379282)

Intercept: When OIL1 and REG1 both equal to 0, the odds of Y= "negative" vs. Y= "no change" equals to exp(3.805370)

(b) **The unordered multinomial logit with GDPWdiff as the output and "no change" as the reference category is as follows:**

$$\ln\left(\frac{p(\text{negative})}{p(\text{no change})}\right) = 4.533759 + \text{OIL1} \times 4.576321 + \text{REG1} \times 1.769007$$

Coefficient for OIL1: Holding REG1 constant, for every one unit increase in OIL1, the odds of Y= "positive" vs. Y= "no change" increase by exp(4.576321)

Coefficient for REG1: Holding OIL1 constant, for every one unit increase in REG1, the odds of Y= "positive" vs. Y= "no change" increase by exp(1.769007)

Intercept: When OIL1 and REG1 both equal to 0, the odds of Y= "positive" compared to Y= "no change" equals to exp(4.533759)

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

```
""
# (2) ordered multinomial
gdp_data_sub$GDPWdiff_category_ordered <- factor(gdp_data_sub$GDPWdiff_
    category, ordered = TRUE, levels = c("positive", "no change", "
    negative"))
gdp_orderd <- polr(gdp_data_sub$GDPWdiff_category_ordered ~ OIL + REG,
    data = gdp_data_sub, Hess = TRUE)
summary(gdp_orderd)
```

(a) **The ordered logistic regression model is represented as follows:**

$$\ln\left(\frac{p(\text{GDPWdiff\_category\_ordered} + 1)}{p(\text{GDPWdiff\_category\_ordered})}\right) = 0.1987 \times \text{OIL1} + (-0.3985 \times \text{REG1})$$

Coefficient for OIL1: Holding REG1 constant, for every one unit increase in OIL1, the odds of GDP becoming from positive change to no change or from no change to positive change increased by exp(0.1987) times.

Table 2: Ordered Logistic Regression Results

	Value	Std. Error	t value
OIL1	0.1987	0.11572	1.717
REG1	-0.3985	0.07518	-5.300
Intercepts:			
positive|no change	0.7105	0.0475	14.9554
no change|negative	0.7312	0.0476	15.3597
Residual Deviance		4687.689	
AIC		4695.689	

Coefficient for REG1: Holding OIL1 constant, for every one unit increase in REG1, the odds of GDP becoming from positive change to no change or from no change to positive change decreased by exp(0.3985) times.

(b) **The estimated cutoffs for odds between Y= positive and Y= no change is 0.7105; The estimated cutoffs for odds between Y= no change and Y= negative is 0.7312.**

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. data wrangling

```
"" ↑
# Subset the dataset
mexico_elections_subset <- mexico_elections[,c("PAN.visits.06","
    competitive.district","marginality.06","PAN.governor.06")]

# identify if there are na in the df
rows_with_missing_mexico <- which(apply(mexico_elections_subset, 1,
    function(x) any(is.na(x))))
```

```r
print(rows_with_missing_mexico) #there are no na

# summarize the data
summary(mexico_elections_subset)

# wrangling
mexico_elections_subset$competitive.district <- as.factor (mexico_
    elections_subset$competitive.district)
mexico_elections_subset$PAN.governor.06 <- as.factor (mexico_elections_
    subset$PAN.governor.06 )

# summarize the original data again
summary(mexico_elections_subset)

# run poisson regression
mexico_poisson <- glm(formula = PAN.visits.06 ~ ., family = poisson, data
     = mexico_elections_subset)
summary(mexico_poisson)
```

Table 3: Poisson Regression Results

	Estimate	**Std. Error**	**z value**	**Pr($>|z|$)**
(Intercept)	-3.81023	0.22209	-17.156	$< 2 \times 10^{-16}$ ***
competitive.district1	-0.08135	0.17069	-0.477	0.6336
marginality.06	-2.08014	0.11734	-17.728	$< 2 \times 10^{-16}$ ***
PAN.governor.061	-0.31158	0.16673	-1.869	0.0617 .
Signif. codes: *** < 0.001, ** < 0.01, * < 0.05, . < 0.1, ' ' < 1				
Dispersion parameter for poisson family taken to be 1				
Null deviance	1473.87 on 2406 degrees of freedom			
Residual deviance	991.25 on 2403 degrees of freedom			
AIC	1299.2			

```r
"" ↑ """""
# check equal variance assumption
dispersiontest( mexico_poisson )
```

A over-dispersion test (as below) was run. The p-value of the test is 0.143 which is bigger than 0.05. So we cannot reject the null hypothesis that true dispersion is smaller or equal to 1. Therefore, the zero-inflated model is not considered. According to the regression, the equation is as follows:

$$\begin{aligned}\ln(\text{visit}) = &-3.81023 - 0.08135 \times \text{competitive} \\ &- 2.08014 \times \text{marginality} - 0.31158 \times \text{governor}\end{aligned}$$

The equation suggests that holding other variables constant, PAN visited wing district about 8% ($\exp(-0.08135) = 0.92186932$) less than "saft seat" district.

Table 4: Poisson Regression Results

	Estimate	**Std. Error**	**z value**
(Intercept)	-3.81023	0.22209	-17.156
competitive.district1	-0.08135	0.17069	-0.477
marginality.06	-2.08014	0.11734	-17.728
PAN.governor.061	-0.31158	0.16673	-1.869
Signif. codes: *** < 0.001, ** < 0.01, * < 0.05, . < 0.1, ' ' < 1			
Dispersion parameter for poisson family taken to be 1			
Null deviance	1473.87 on 2406 degrees of freedom		
Residual deviance	991.25 on 2403 degrees of freedom		
AIC	1299.2		

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

- **The marginality.06 coefficients:** Holding other variables constant, an unit increase in marginality is expected to decrease the number of PAN visits by a multiplicative factor of $e^{-2.08014} \approx 0.12491227$.
- **The PAN.governor.06 coefficients:** Holding other variables constant, PAN visits the states with a PAN-affiliated governor about 27% ($\exp(-0.31158) \approx 0.73228985$) less than the states with a non-PAN-affiliated governor.

(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).

```
"" ↑"""""" ^ ^ ^ ^^
#extract coefficient
cfs <- coef(mexico_poisson)
# the estimated mean number
exp(cfs[1] + cfs[2]*1 + cfs[3]*0 + cfs[4]*1)#0.01494818
```

Therefore, the estimated number is 0.01494818.