

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 for which 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.

```
1
2 # load data
3 gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/
  StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
4
5 # Check the structure of the dataset
6 str(gdp_data)
7
8 # Fit the unordered multinomial logistic regression model
9 model <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)
10
11 #Because It seems I have too many levels in one of my categorical
  variables
12 #I will proceed to inspect the levels of GDPWdiff
13 table(gdp_data$GDPWdiff)
14
15 # Define breakpoints for binning
16 breakpoints <- quantile(gdp_data$GDPWdiff, probs = seq(0, 1, by = 0.1))
17
18 # Create a new categorical variable by binning GDPWdiff
19 gdp_data$GDPWdiff_group <- cut(gdp_data$GDPWdiff, breaks = breakpoints,
  labels = FALSE)
20
21 # Fit the multinomial logistic regression model using the binned variable
22 model <- multinom(GDPWdiff_group ~ REG + OIL, data = gdp_data)
23
24 # Summarize the model
25 summary(model)
26
```

```
multinom(formula = GDPWdiff_group ~ REG + OIL, data = gdp_data)
```

Coefficients:

	(Intercept)	REG	OIL
2	0.44787599	-0.6300308	-1.5567870
3	0.65516574	-1.3224053	-1.6480791
4	0.66889824	-1.5204747	-1.5707004
5	0.56041172	-1.0167280	-1.3985742
6	0.48915063	-0.7418814	-1.6696137
7	0.23433147	-0.2256526	-0.8533760
8	-0.01306567	0.2774703	-0.8571151
9	-0.16667708	0.5803102	-1.0127366
10	-0.42147924	0.8224895	-0.2648593

Std. Errors:

	(Intercept)	REG	OIL
2	0.1044569	0.1544339	0.2487385
3	0.1006293	0.1675130	0.2457301
4	0.1004054	0.1748111	0.2399128
5	0.1022532	0.1606005	0.2316745
6	0.1036795	0.1559679	0.2562922
7	0.1085848	0.1511739	0.2099432
8	0.1147097	0.1508068	0.2176762
9	0.1193782	0.1524712	0.2319998
10	0.1258310	0.1551094	0.1993666

Residual Deviance: 16601.56

AIC: 16655.56

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

```
1 # Load necessary library
2 library(MASS)
3
4 # Convert GDPWdiff_group to factor
5 gdp_data$GDPWdiff_group <- factor(gdp_data$GDPWdiff_group)
6
7 # Fit ordered multinomial logit model
8 ordered_model <- polr(GDPWdiff_group ~ REG + OIL, data = gdp_data, Hess =
  TRUE)
9
10 # Summarize the model
11 summary(ordered_model)
```

```
polr(formula = GDPWdiff_group ~ REG + OIL, data = gdp_data, Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
REG	0.87554	0.06161	14.2098
OIL	0.04442	0.10459	0.4247

Intercepts:

	Value	Std. Error	t value
1 2	-1.9272	0.0584	-32.9823
2 3	-1.1215	0.0459	-24.4249
3 4	-0.5717	0.0416	-13.7413
4 5	-0.1193	0.0404	-2.9559
5 6	0.3099	0.0409	7.5838
6 7	0.7394	0.0428	17.2858
7 8	1.2114	0.0462	26.2210
8 9	1.7803	0.0519	34.3080
9 10	2.6294	0.0647	40.6400

Residual Deviance: 16924.32

AIC: 16946.32

(1 observation deleted due to missingness)

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.

```
1 # load data
2 mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/
  StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
3
4 # Fit Poisson regression model
5 poisson_model <- glm(PAN.visits.06 ~ competitive.district, data =
  mexico_elections, family = poisson)
6
7 # Summary of the model
8 summary(poisson_model)
```

```
glm(formula = PAN.visits.06 ~ competitive.district, family = poisson,
data = mexico_elections)
```

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.2571 0.1491 -15.141 <2e-16 ***

competitive.district -0.1617 0.1670 -0.968 0.333

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1473.9 on 2406 degrees of freedom

Residual deviance: 1473.0 on 2405 degrees of freedom

AIC: 1776.9

Number of Fisher Scoring iterations: 6

- (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`).