Problem Set 3

Applied Stats II
Dan Zhang 23335541

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, let's load the data. Then, we need to convert `GDPWdiff` into factor. Next, run an unodered multinomial logit model with ”no change” as the reference category.

```
# load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)

# (1)
# Convert GDPWdiff to a factor
gdp_data$GDPWdiff <- as.factor(ifelse(gdp_data$GDPWdiff > 0, "positive",
                                  ifelse(gdp_data$GDPWdiff < 0, "negative", "no change")))
# Set the reference level
gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")

# Run an unordered multinomial logit model
multi_unordered <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data, reflevel = "no change")
multi_unordered
```

Here are the results:

Table 1:

	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	

Convert the coefficients into exponents:

```r
# Exponentiate the coefficients
exp(coef(multi_unordered))
```

Table 2:

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

Interpretation:
`Intercept:`

- ”negative”: The constant term of the ”negative” category is 3.805, and the significance level is p<0.01, indicating that for a non-democracy country and its average ratio of fuel exports to total exports in 1984-86 no more than 50%, the log odds of having a negative GDP difference versus no change is 3.805.
- ”positive”: The constant term of the ”positive” category is 4.534 with a p<0.01, suggesting that for a non-democracy country and its average ratio of fuel exports to total exports in 1984-86 no more than 50%, the odds of having a positive GDP difference versus no change is 93.108.

`Coefficents of REG:`

- ”negative”: The coefficient is 1.379 with a p<0.1, indicating a statistically significant relationship at the 10% level. This suggests that being a democracy (REG = 1) increases the log odds of having a negative GDP difference compared to no change by 1.379, when all other variables are held constant.
- ”positive”: The coefficient is 1.769 with a p<0.05, indicating a statistically significant relationship at the 5% level. This implies that democracies have a positive GDP difference compared to no change, with the odds multiplying by a factor of 5.865 when all other variables are constant.

`Coefficients of OIL:`

- The coefficients for OIL are not statistically significant at the conventional levels (p>0.1). This suggests that the ratio of fuel exports to total exports does not significantly influence the likelihood of experiencing either a positive or negative GDP difference compared to no change.

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

First, let's convert `GDPWdiff` into an ordered variable. Then, fit an ordered multinomial logit model.

```r
# Convert GDPWdiff to an ordered factor
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, levels = c("negative", "no change", "positive"), ordered = TRUE)

# Run an ordered multinomial logit model
multi_ordered <- polr(GDPWdiff ~ REG + OIL, data = gdp_data, Hess = TRUE)
```

Here are the results:

Table 3:

	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

Let's calculate a p value for the model:

```r
# Calculate a  p value
ctable <- coef(summary(multi_ordered))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
stargazer((ctable <- cbind(ctable, "p value" = p)))
```

Table 4:

	Value	Std. Error	t value	p value	p value.1
REG	0.398	0.075	5.300	0.00000	0.00000
OIL	-0.199	0.116	-1.717	0.086	0.086
negative|no change	-0.731	0.048	-15.360	0	0
no change|positive	-0.710	0.048	-14.955	0	0

Convert the coeffiecients into exponents and add CIs:

```r
# Get odds ratios and CIS
exp_coefs <- exp(cbind(OR = coef(multi_ordered), confint(multi_ordered)))
```

Table 5:

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

Interpretation:
Check Table 3, we can see the coefficients of explanatory varibles are statistically significant. Let's interpret them respectively.

`Coefficents of REG:` The coefficient for REG is 0.398 with a p-value < 0.01, indicating a statistically significant positive association between being a democracy and the ordered level of GDP change. A positive coefficient suggests that as a country changes from non-democracy to democracy, the log odds of being in a higher category of difference of GDP also increase. The odds ratio for REG suggests that being a democracy increases the odds of experiencing a positive GDP change by 49% compared to a non-democracy, holding all else constant. The confidence interval (1.286 to 1.727) does not include 1, confirming the statistical significance of this finding.

`Coefficients of OIL:` The coefficient for OIL is -0.199 with a p-value < 0.1, suggesting a statistically significant but negative relationship at the 10% level. This implies that countries where fuel exports exceed 50% of total exports are less likely to experience higher levels of GDP change. The presence of significant oil exports decreases the log odds of moving to a more positive category of GDP change. The odds ratio for OIL indicates that countries with significant oil exports have 18% lower odds of moving to a more positive GDP change category compared to countries without significant oil exports. The confidence interval (0.655 to 1.031) is close to including 1, indicating that this result is less robust and should be interpreted with caution.

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.

First, let's load the data:

```r
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/
    StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
```

Having a look at the data dispersion.

```r
hist(mexico_elections$PAN.visits.06)
ggplot(mexico_elections, aes(x = PAN.visits.06, y = factor(competitive.
    district), color = marginality.06, shape = factor(PAN.governor.06))) +
  geom_jitter(alpha = 0.5) +
  labs(x = "Number of PAN Visits in 2006", y = "Competitive District",
    color = "Marginality", shape = "PAN Governor in 2006") +
  ggtitle("Visualization of PAN Visits by District Characteristics")
```

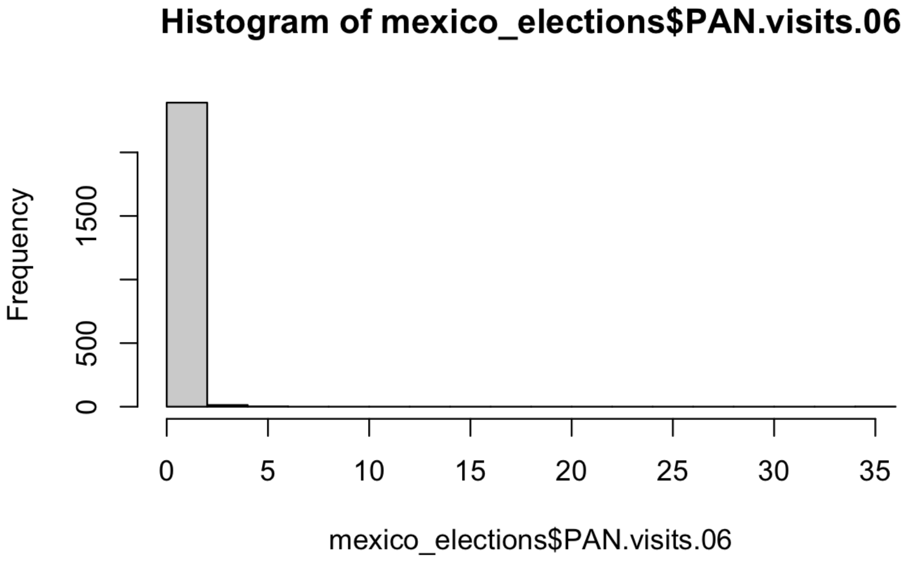

Figure 1: Outcome

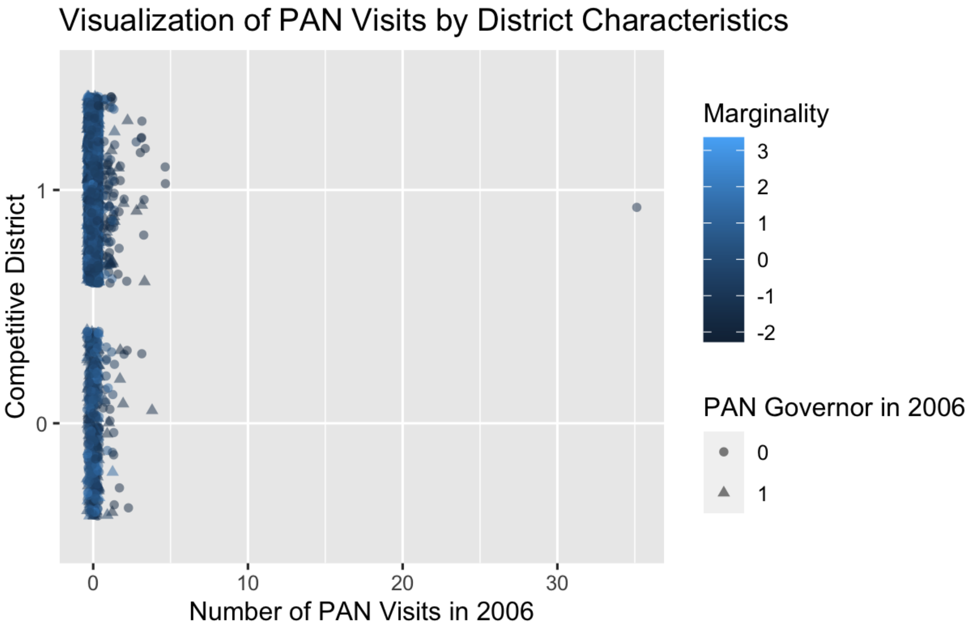

Figure 2: Variables distribution

By checking these pictures, we can see there are many zero values. In order to prevent over-dispersion problem, we need to run a dispersion test for poisson model. Let's run a poisson regression model.

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

Run a dispersion test.

```r
dispersiontest(model_poisson)
```

Table 6: Overdispersion Test Results

Test Detail	**Result**
Data	model_poisson
Z-value	1.0668
P-value	0.143
Alternative Hypothesis	True dispersion is greater than 1
Sample Estimates: Dispersion	2.09834

As the p-value = 0.143, the results are not statistically significant. There is not enough evidence to prove that there is a over-dispersion problem. Therefore, according to this test, he Poisson regression model is applicable for this dataset.

Here are the test statistics and p values:

```r
# Extract p-value and test statistic
p_value <- coef(summary(model_poisson))[, 4]
test_statistic <- coef(summary(model_poisson))[, 3]
stargazer(cbind(test_statistic, p_value))
```

Table 7:

	test_statistic	p_value
(Intercept)	-17.156	0
competitive.district	-0.477	0.634
marginality.06	-17.728	0
PAN.governor.06	-1.869	0.062

As we can see from the Table 6, the test-statistic coefficient for "competitive.district" is -0.477 with a p-value of 0.634. This suggests that, controlling all other variables constant in the model, being a competitive district does not have a statistically significant effect on the number of visits by PAN presidential candidates. The high p-value indicates that we fail to reject the null hypothesis of no effect. Besides, the coefficient of ”competitive.district” is negative, which indicate that if this variable were statistically significant, it would indicate that competitive districts would have fewer PAN presidential candidate visits than non-competitive districts. However, due to the high p-value for this result, we cannot make this conclusion.

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

Here are the coefficients of the model:

Table 8:

	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

- `marginality.06`: The coefficient of `marginality.06` is -2.08 which is significantly negative ($p < 0.01$). This indicates that as marginality (a measure of poverty) increases, the number of visits by PAN presidential candidates in 2006 decreases. Specifically, for a one-unit increase in marginality, the expected log count of PAN visits decreases by 2.080. This may suggest that PAN candidates were less likely to visit municipalities with higher levels of poverty in 2006.
- `PAN.governor.06`: The coefficient of `PAN.governor.06` is -0.312 which is also negative and statistically significant at the $p < 0.1$ level. This suggests that the presence of a PAN-affiliated governor in the state is associated with a decrease in the number of visits by PAN presidential candidates. For districts within states governed by PAN in 2006, the expected log count of visits from PAN presidential

candidates decreases by 0.312. This might imply that PAN candidates probably allocated their campaign efforts away from states already governed by their party, maybe focus on regions where gaining or maintaining political support was deemed more critical

(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).

Use the predict function to check.

```r
# Create a new dataframe for hypothetical district
new_data <- data.frame(competitive.district = 1, marginality.06 = 0, PAN.governor.06 = 1)
# Predict the mean number of visits from the winning PAN presidential candidate
pred_visits <- predict(model_poisson, newdata = new_data, type = ”response”)
exp(pred_visits)
```

The estimated mean number of visits would be 1.015 and its exponent value is 0.015. This means that the estimated mean number for a winning PAN presidential candidate to visit a competitive district which had an average poverty level and a PAN-affiliated governor is 1.015.