Problem Set 3

Yajie Dong

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.

   Step 1: Let’s load the dataset from R:

```r
# Step 1:load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
```

   Step 2: Transform the outcome variable into an unordered categorical format, with ”no change” designated as the baseline category. Let’s do that by R:

```r
# Step 2: Create a factor variable for GDP change outcome
# Assign "no change" to rows where GDP difference is 0
gdp_data$GDPcat[gdp_data$GDPWdiff == 0] <- "no change"
# Assign "positive" to rows where GDP difference is greater than 0
gdp_data$GDPcat[gdp_data$GDPWdiff > 0] <- "positive"
# Assign "negative" to rows where GDP difference is less than 0
gdp_data$GDPcat[gdp_data$GDPWdiff < 0] <- "negative"
# Convert the 'GDPcat' column to a factor and set "no change" as the reference level
gdp_data$GDPcat <- factor(gdp_data$GDPcat, levels = c("no change", "positive", "negative"))
```

   Step 3: Now we can run unordered multinomial logistic regression,let do that by R:

```r
# Step 3:Run unordered multinomial logistic regression
library(nnet)
unordered_logit <- multinom(GDPcat ~ REG + OIL, data = gdp_data)
# Display the summary of the multinomial logistic regression model
summary(unordered_logit)
```

   The output we got from running the unordered multinomial logistic regression model in R as following:

Table 1: Coefficients and Standard Errors of the Multinomial Logistic Regression

	Coefficient		**Standard Error**	
	Positive	**Negative**	**Positive**	**Negative**
Intercept	4.533759	3.805370	0.2692006	0.2706832
REG	1.769007	1.379282	0.7670366	0.7686958
OIL	4.576321	4.783968	6.885097	6.885366

Residual Deviance: 4678.77
AIC: 4690.77

Now let’s interpret the output: The model predicts the category of GDP change (‘GDPcat‘), which can be ”positive” or ”negative”, based on two independent variables: ‘REG‘ and ‘OIL‘ . The reference category is for ‘GDPcat‘ . Here’s an interpretation of the key components of the output:

Coefficients: The coefficients represent the log odds of being in the specified category (”positive” or ”negative”) relative to the reference category (”no change”), for a one-unit increase in the predictor variable, holding all other variables constant.

(Intercept) for Positive: The intercept for ”positive” is 4.533759, indicating the log odds of observing a ”positive” GDP change when ‘REG‘ and ‘OIL‘ are 0.

REG for Positive: The coefficient for ‘REG‘ in predicting a ”positive” outcome is 1.769007, suggesting that a one-unit increase in ‘REG‘ is associated with an increase in the log odds of a ”positive” GDP change by approximately 1.77, holding ‘OIL‘ constant.

OIL for Positive: Similarly, a one-unit increase in ‘OIL‘ is associated with an increase in the log odds of a ”positive” GDP change by approximately 4.58, holding ‘REG‘ constant.

- (Intercept), REG, and OIL for Negative: These coefficients can be interpreted similarly but are specific to the odds of a ”negative” GDP change relative to ”no change”.

Standard Errors: The standard errors measure the variability or uncertainty in the coefficient estimates. Smaller standard errors indicate more precision in the estimation of the coefficients.

(Intercept), REG, and OIL for Positive and Negative: The standard errors for the intercepts and coefficients of ‘REG‘ and ‘OIL‘ are provided, indicating the precision of these estimates. For example, the standard error for the ‘REG‘ coefficient in predicting a ”positive” outcome is 0.7670366.

Residual Deviance and AIC:

Residual Deviance: The residual deviance is 4678.77. In general, lower deviance indicates a better fit of the model to the data. However, without a baseline comparison, it’s hard to evaluate this number on its own.

AIC (Akaike Information Criterion): The AIC is 4690.77, a measure of the relative quality of statistical models for a given set of data. Among models, the one with the lowest AIC is generally preferred.

Interpretation Summary:

This model examines how regional effects (‘REG‘) and oil production/consumption impacts (‘OIL‘) influence the likelihood of a country’s GDP changing positively or negatively compared to no change. The positive coefficients for ‘REG‘ and ‘OIL‘ across both ”positive” and ”negative” outcomes suggest that increases in these variables are associated with greater odds of observing both positive and negative changes in GDP compared to no change. This might reflect the complex and significant roles these factors play in economic dynamics, where they can contribute to both growth and decline under different circumstances.

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

   Step 1: Reorganize the outcome variable to ensure that the categories are arranged in ascending order: starting with ”negative,” followed by ”no change,” and ending with ”positive.”Let’ do that by R:

```r
#Step 1:  Prepare for ordered logistic regression.Adjust the levels of the factor variable to establish a specific sequence.
gdp_data$GDPcat <- factor(gdp_data$GDPcat, levels = c("negative", "no change", "positive"))
```

   Step 2: Prior to executing our ordered logistic regression model, it’s necessary to adjust the categorization of our outcome variable to reflect this specified sequence.Let’ do that by R:

```r
# Step 2:Run the ordered logistic regression
ordered_logit <- polr(GDPcat ~ REG + OIL, data = gdp_data)

# Display the summary of the ordered logistic regression model
summary(ordered_logit)
```

   The output from R:

Table 2: Ordinal Logistic Regression (polr) Output

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

Now let’s interpret the output: The output comes from an ordered logistic regression .The dependent variable ‘GDPcat‘ has three ordered categories: ”negative,” ”no change,” and ”positive.” The independent variables in the model are ‘REG‘ and ‘OIL‘.

Coefficients:

REG: The coefficient for ‘REG‘ is 0.3985 with a standard error of 0.07518, resulting in a t-value of 5.300. This positive coefficient suggests that an increase in ‘REG‘ is associated with moving to a more positive category of ‘GDPcat‘, from ”negative” to ”no change” or from ”no change” to ”positive.” The statistical significance (implied by the high t-value) indicates a strong association.

OIL: The coefficient for ‘OIL‘ is -0.1987 with a standard error of 0.11572 and a t-value of -1.717. This negative value implies that an increase in ‘OIL‘ is associated with a shift towards a more negative category of ‘GDPcat‘. However, the lower t-value compared to ‘REG‘ suggests this association might not be as strong or statistically significant.

Intercepts (Cutpoints):

The intercepts in ordered logistic regression are also known as thresholds or cutpoints. They indicate the latent (unobserved) variable’s value at which the probability of being in one category or the next is equal.

negative—no change: The intercept for moving from ”negative” to ”no change” is -0.7312 with a standard error of 0.0476 and a very significant t-value of -15.3597. This shows a strong distinction between these categories.

no change—positive: The intercept for moving from ”no change” to ”positive” is -0.7105 with a standard error of 0.0475 and a t-value of -14.9554, also indicating a significant distinction between these categories.

Model Fit:

Residual Deviance: The residual deviance is 4687.689. In the context of logistic regression, lower residual deviance indicates a better fit of the model to the data.

AIC (Akaike Information Criterion): The AIC value is 4695.689. AIC is a measure of the relative quality of a statistical model for a given set of data. When comparing models, the one with the lowest AIC is preferred.

In summary, the ‘REG‘ variable has a statistically significant positive effect on the probability of being in a higher category of ‘GDPcat‘. The ‘OIL‘ variable's effect is negative, suggesting it decreases the likelihood of being in a higher category, but its impact is less statistically significant compared to ‘REG‘. The significant cutpoints indicate clear distinctions between the categories of the dependent variable. The overall model fit, as shown by the residual deviance and AIC, provides a basis for evaluating the model's performance relative to alternative models.

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.

Step 1: let's load the dataset by R:

```
# Step1:load data
mexico_elections <- read.csv(”https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv”)
```

Step 2: Now let's run a Poisson regression by R:

```
# Step 2:Fit the Poisson regression model
mex_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06,
                          data = mexico_elections, family = poisson())

# Display the summary of the Poisson model
summary(mex_poisson)
```

The output of Poisson regression model from R:

Table 3: Summary of the generalized linear model (GLM) for predicting PAN visits in 2006 using Poisson regression.

Coefficient	Estimate	Std. Error	z value	Pr(¿—z—)
Intercept	-3.81023	0.22209	-17.156	¡2e-16 ***
competitive.district	-0.08135	0.17069	-0.477	0.6336
marginality.06	-2.08014	0.11734	-17.728	¡2e-16 ***
PAN.governor.06	-0.31158	0.16673	-1.869	0.0617 .
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				
Number of Fisher Scoring iterations: 7				

The output from the Poisson regression model **does not show evidence** that PAN presidential candidates visit swing districts more frequently. The coefficient for competitive.district is -0.08135 with a standard error of 0.17069, resulting in a z value of -0.477 and a p-value of 0.6336. This p-value is much greater than the conventional threshold of 0.05, indicating that the relationship between being a swing district (competitive.district) and the number of visits from PAN presidential candidates is not statistically significant. In practical terms, this means we do not have sufficient evidence to conclude that swing districts receive more visits from PAN presidential candidates compared to non-swing districts based on this model.

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

Now let's plot he relationship between ‘marginality.06‘ and the predicted number of PAN visits by R:

```
#(b) plotting of Predicted PAN Visits vs. Marginality
# Create a new data frame for predictions
new_data <- data.frame(
  competitive.district = mean(mexico_elections$competitive.district, na.rm = TRUE),
  marginality.06 = seq(min(mexico_elections$marginality.06, na.rm = TRUE),
                       max(mexico_elections$marginality.06, na.rm = TRUE), length.out = 100),
  PAN.governor.06 = mean(mexico_elections$PAN.governor.06, na.rm = TRUE)
)

# Predict the number of visits using the new data
new_data$predicted_visits <- predict(mex_poisson, newdata = new_data, type = "response")

# Plotting
library(ggplot2)
ggplot(new_data, aes(x = marginality.06, y = predicted_visits)) +
  geom_line(color = "blue") +
  labs(title = "Predicted PAN Visits vs. Marginality",
       x = "Marginality in 2006",
       y = "Predicted Number of PAN Visits") +
  theme_minimal()
```

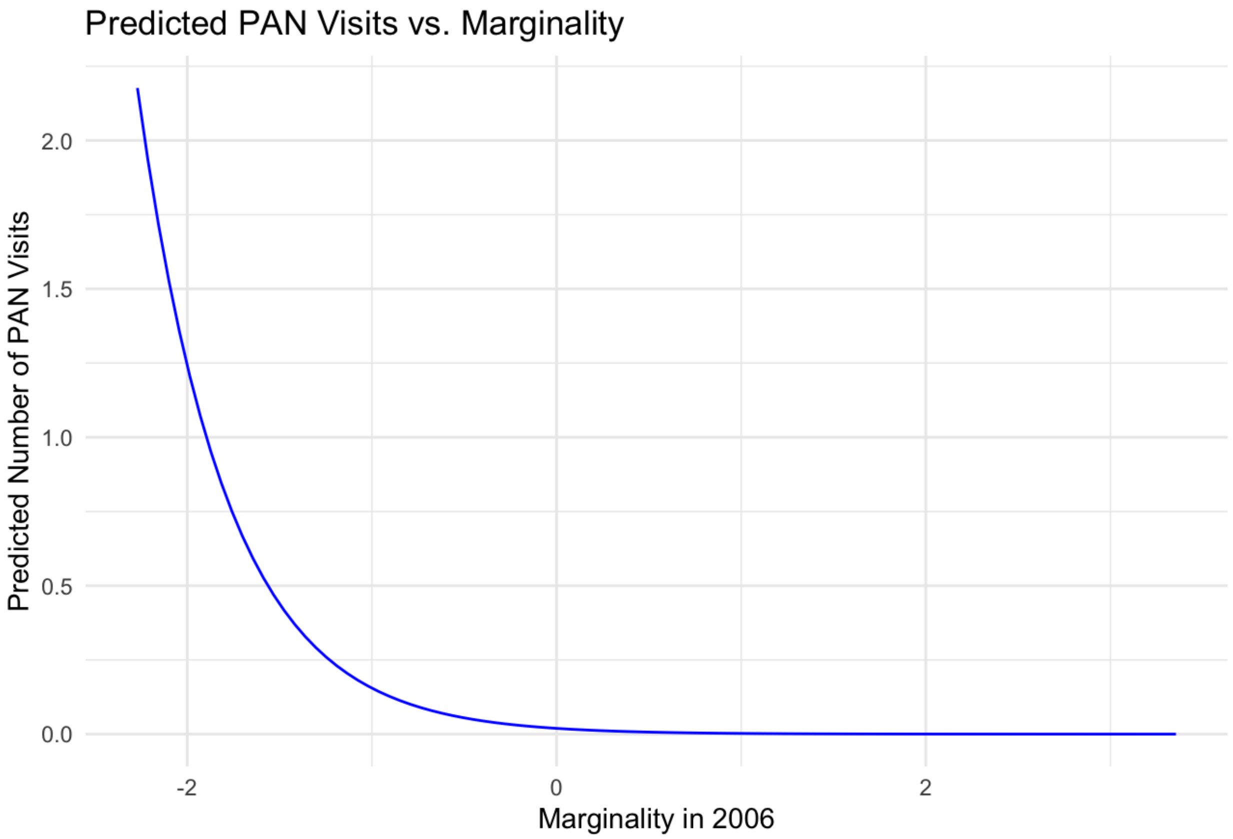

Figure 1: Predicted PAN Visits vs. Marginality

The plot shows the relationship between ‘marginality.06‘ and the predicted number of PAN visits. It’s a descending curve that flattens as marginality increases, suggesting a higher number of predicted visits in areas with lower levels of marginality in 2006, and fewer predicted visits as marginality increases.

Now, let’s interpret the coefficients from the Poisson regression model related to ‘marginality.06‘ and ‘PAN.governor.06‘:

‘marginality.06‘: The coefficient for ‘marginality.06‘ is -2.08014 with a highly significant p-value (¡2e-16), which implies that increases in marginality (poverty) are strongly associated with a decrease in the expected number of visits by the PAN candidate. In practical terms, this means that for each unit increase in marginality, the expected number of visits multiplies by ‘exp(-2.08014)‘, which is a small fraction, thus significantly reducing the number of visits.

‘PAN.governor.06‘: The coefficient for ‘PAN.governor.06‘ is -0.31158, and it is marginally significant (p-value = 0.0617). This indicates that having a PAN-affiliated governor is associated with a decrease in the expected number of visits by the PAN candidate, though the evidence is not strong enough to be considered significant at the conventional 0.05 level. The number of expected visits multiplies by ‘exp(-0.31158)‘ for districts with a PAN governor, which implies a moderate decrease.

The plot and the coefficients together suggest that while poverty strongly and negatively impacts the number of PAN candidate visits, the presence of a PAN governor also seems to reduce the visits, but not as dramatically or certainly as marginality does.

(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
# Predict expected PAN visits for a competitive district with average marginality and a PAN governor.
predict(mex_poisson, newdata = data.frame(competitive.district = 1, marginality.06 = 0, PAN.governor.06 = 1), type = "response")
```

The result form R:0.01494818

Based on the output from the R ‘predict‘ function and the specified model conditions, the estimated mean number of visits from the winning PAN presidential candidate to a hypothetical district that is competitive (with ‘competitive.district=1‘), has an average poverty level (‘marginality.06 = 0‘), and is governed by a PAN-affiliated governor (‘PAN.governor.06=1‘) is approximately 0.015. This figure is derived from the predictive analysis using the fitted Poisson regression model, reflecting the expected frequency of visits under these particular conditions.