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
# Load necessary libraries
library(nnet)
library(dplyr)

# Let's load the dataset
data <- read.csv("gdpChange.csv")

# We should first check the few rows of the dataset
head(data)

# We can now create a new factor column for GDPWdiff categories
data$GDPWdiff_category <- cut(data$GDPWdiff, breaks = c(-Inf, 0, Inf),
    labels = c("negative", "positive"))

# Convert the GDPWdiff_category variable to a factor
data$GDPWdiff_category <- factor(data$GDPWdiff_category, levels = c("
    negative", "positive"))

# Add "no change" category for observations where GDPWdiff is zero
data$GDPWdiff_category <- factor(data$GDPWdiff_category, levels = c("no
    change", "positive", "negative"))

data$GDPWdiff_category[data$GDPWdiff == 0] <- "no change"

# Set "no change" as the reference category
data$GDPWdiff_category <- relevel(data$GDPWdiff_category, ref = "no
    change")

# Fit the unordered multinomial logit model
model <- multinom(GDPWdiff_category ~ REG + OIL, data = data)

# View the model summary
summary(model)
```

The results of the multinomial logit model are as follows:

- **Call**:

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

- **Coefficients**:

```
          (Intercept)      REG      OIL
positive    4.533759 1.769007 4.576321
negative    3.805370 1.379282 4.783968
```

- **Std. Errors**:

```
          (Intercept)       REG      OIL
positive   0.2692006 0.7670366 6.885097
negative   0.2706832 0.7686958 6.885366
```

- **Residual Deviance**: 4678.77
- **AIC**: 4690.77

Our multinomial logit model results show that compared to the reference category of "no change" countries with democracy (1) have significantly higher odds of experiencing positive GDP changes (2) compared to countries with no change, with a coefficient of 1.769 (3), indicating that the odds of observing positive GDP changes increase by approximately 1.769 for every unit increase in the democracy score, holding other variables constant. Additionally, countries with a higher ratio of fuel exports to total exports have significantly higher odds of experiencing both positive and negative GDP changes compared to no change. However, it is important to note the wide standard errors for the coefficient of oil, indicating uncertainty in the estimates of the effect of this variable on GDP variations.

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

```r
# Load necessary library
library(MASS)

# Fit the ordered multinomial logit model
ordered_model <- polr(GDPWdiff_category ~ REG + OIL, data = data, Hess =
    TRUE)

# View the model summary
summary(ordered_model)
```

The results of the ordered multinomial logit model are as follows:

- **Call**:

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

- **Coefficients**:

```
             Value  Std. Error  t value
REG       -0.3566    0.07485   -4.764
OIL        0.2306    0.11510    2.003
```

- **Intercepts**:

```
                              Value    Std. Error    t value
no change|positive    -5.5846     0.2534        -22.0376
positive|negative       0.7491     0.0479         15.6475
```

- **Residual Deviance**: 4692.109
- **AIC**: 4700.109

The ordered multinomial logit model indicates that higher democracy scores are associated with lower probabilities of transitioning from the "no change" category to "positive" or "negative" GDP changes (REG: coefficient = -0.3566, p ¡ 0.001). However, greater dependence on fuel exports increases the odds of transitioning from "no change" to either category.

The intercept values highlight significant impacts on transitions: the probabilities of transitioning from "no change" to "positive" decrease markedly, while those from "positive" to "negative" increase substantially.

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
# Load the dataset
data <- read.csv("MexicoMuniData.csv")

# Check the structure of the data
str(data)

# Run Poisson regression
poisson_model <- glm(PAN.visits.06 ~ competitive.district +
    marginality.06 + PAN.governor.06,
                     data = data,
                     family = poisson)

# Summarize the model
summary(poisson_model)
```

```
Call:
glm(formula = PAN.visits.06 ~ competitive.district + marginality.06
    +
    PAN.governor.06, family = poisson, data = data)

Coefficients:
                   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 .
---
Signif. codes:  0    ***    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

Number of Fisher Scoring iterations: 7
```

The results of the Poisson regression analysis indicate that marginality in 2006 is a significant predictor of the number of visits by PAN presidential candidates to a district ($p < 0.001$). However, the analysis did not find significant evidence ($p = 0.6336$) that PAN candidates visit swing districts more compared to safe districts.

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

The coefficients for the predictors in the Poisson regression model are as follows:

- **marginality.06**: The coefficient for marginality.06 is -2.08014 with a standard error of 0.11734. This coefficient indicates that for every one-unit increase in marginality in 2006, the expected log count of PAN presidential candidate visits to a district decreases by approximately 2.08, holding other variables constant. Since this coefficient is negative and statistically significant ($p < 0.001$), it suggests that districts with higher levels of poverty (higher marginality) in 2006 are associated with fewer visits by PAN presidential candidates.
- **PAN.governor.06**: The coefficient for PAN.governor.06 is -0.31158 with a standard error of 0.16673. This coefficient indicates that districts in states with a PAN-affiliated governor in 2006 have a slightly lower expected log count of PAN presidential candidate visits compared to districts in states without a PAN-affiliated governor, holding other variables constant. However, this coefficient is not statistically significant at the conventional level ($p = 0.0617$).

(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
# Hypothetical values
competitive_district <- 1
marginality_06 <- 0
PAN_governor_06 <- 1

# Create a data frame with the hypothetical values
hypothetical_data <- data.frame(
  competitive.district = competitive_district,
  marginality.06 = marginality_06,
  PAN.governor.06 = PAN_governor_06
)

# Predict the mean number of visits
mean_visits <- predict(poisson_model, newdata = hypothetical_data,
    type = "response")

# Print the estimated mean number of visits
print(mean_visits)
```

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`) is approximately 0.01495 visits.