Problem Set 3

Yana Konshyna

March 22, 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.

   Uploading data from GitHub, then convert output variable `GDPWdiff` into categorical variable. Setting reference category as "no change".
   Running unordered multinominal logit regression `unordered_logit`.

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

# Creating factor variable for outcome
gdp_data$GDPWdiff_category <- factor(
  with(gdp_data, ifelse(GDPWdiff == 0, "no change",
                        ifelse(GDPWdiff > 0, "positive", "negative"))),
  levels = c("no change", "positive", "negative"))

# Setting reference category
gdp_data$GDPWdiff_category <- relevel(as.factor(gdp_data$GDPWdiff_category), ref="no change")

# Running unordered multinominal logit regression
unordered_logit <- multinom(GDPWdiff_category ~ OIL + REG, data=gdp_data)
summary(unordered_logit)
```

   Output:

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

Std. Errors:
         (Intercept)      OIL       REG
negative   0.2706832 6.885366 0.7686958
positive   0.2692006 6.885097 0.7670366

Residual Deviance: 4678.77
AIC: 4690.77
```

   Interpretation of the Estimated Coefficient (Negative Outcome):

   - **Intercept:** When `REG` and `OIL` both equal 0, the estimated log odds of going from "no change" to "negative" is approximately 3.8.

- **OIL:** On average, a one unit change in `OIL` (from 0 to 1) is associated with a change in log odds of going from "no change" to "negative" by approximately 4.78, holding `REG` constant.
- **REG:** On average, a one unit change in `REG` (from 0 to 1) is associated with a change in log odds of going from "no change" to "negative" by approximately 1.38, holding `OIL` constant.

Interpretation of the Estimated Coefficient (Positive Outcome):

- **Intercept:** When `REG` and `OIL` both equal 0, the estimated log odds of going from "no change" to "positive" is approximately 4.53.
- **OIL:** On average, a one unit change in `OIL` (from 0 to 1) is associated with a change in log odds of going from "no change" to "positive" by approximately 4.57, holding `REG` constant.
- **REG:** On average, a one unit change in `REG` (from 0 to 1) is associated with a change in log odds of going from "no change" to "positive" by approximately 1.77, holding `OIL` constant.

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

   Reordering outcome variable so that the categories are in increasing order from "negative" to "no change" and then to "positive". Running ordered multinomial logit regression `ordered_logit`.

```r
# Creating factor to have ordering
gdp_data$GDPWdiff_category <- relevel(gdp_data$GDPWdiff_category, ref="
    negative")

# Running ordered multinominal logit regression
ordered_logit <- polr(GDPWdiff_category ~ OIL + REG, data=gdp_data)
summary(ordered_logit)
```

Output:

```
Coefficients:
      Value   Std. Error t value
OIL -0.1987    0.11572  -1.717
REG  0.3985    0.07518   5.300

Intercepts:
                    Value  Std. Error   t value
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
```

Interpretation of the Estimated Coefficients:

- **OIL:** On average, a one unit change in `OIL` is associated with a change in the log odds of going from "negative" to "no change" and from "no change" to "positive" by approximately -0.1987, holding `REG` constant.
- **REG:** On average, a one unit change in `REG` is associated with a change in the log odds of going from "negative" to "no change" and from "no change" to "positive" by approximately 0.3985, holding `OIL` constant.

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.

Uploading data from GitHub, then running a Poisson regression `mexico_poisson`.

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

# Estimating Poisson model
mexico_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexico_elections, family=poisson)
summary(mexico_poisson)
```

Output:

```
Coefficients:
                   Estimate Std. Error z value            Pr(>|z|)
(Intercept)         -3.81023    0.22209 -17.156 <0.0000000000000002 ***
competitive.district -0.08135    0.17069  -0.477              0.6336
marginality.06      -2.08014    0.11734 -17.728 <0.0000000000000002 ***
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
```

For variable `competitive.district` z-value is -0.477, so I don't have a large enough test statistic to reject the null hypothesis that the estimated association between competitive.district and the number of visits from the winning PAN presidential candidate in 2006 is zero (p-value $= 0.63$ that is > 0.05).

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

Interpretation of the Estimated Coefficients:

- **marginality.06:** On average, a one unit increase in poverty is associated with a change in the expected number of visits by a multiplicative factor of $e^{-2.08014} \approx 0.125$, holding other variables constant.
- **PAN.governor.06:** On average, having a PAN governor, in comparing to having a non-PAN governor, is associated with a change in the expected number of visits by a multiplicative factor of $e^{-0.31158} \approx 0.73$, holding other variables constant.

(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).

Calculating the expected number of visits using predict() function

```r
pred <- data.frame(competitive.district = 1,
                   marginality.06 = 0,
                   PAN.governor.06 = 1)
predict(mexico_poisson, newdata = pred, type="response")
```

Output:

```
0.01494818

```

The expected number of visits from the winning PAN candidate is approximately 0.0149 given that a district is competitive (`competitive.district`=1), had an average poverty level (`marginality.06` $= 0$), and a PAN governor (`PAN.governor.06`=1).