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.

   The R script for model fitting of unordered multinomial logit model is:

```r

# 1.1

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

# creating a factor variable
gdp_data$category <- ifelse(gdp_data$GDPWdiff == 0, "no change",
                            ifelse(gdp_data$GDPWdiff > 0, "positive change", "negative change"))


# conevrting category variable into a factor, keeping "no change" as reference level
gdp_data$category <- relevel(as.factor(gdp_data$category), ref="no change")

# using ordinal variables from the dataset to fit an unordered multinomial logit model
un_logit <- multinom(gdp_data$category ~ gdp_data$OIL+gdp_data$REG)
summary(un_logit)
```

Coefficients:
(Intercept) gdp_data$OIL gdp_data$REG

negative change 3.805370 4.783968 1.379282
positive change 4.533759 4.576321 1.769007

Std. Errors:
(Intercept) gdp_data$OIL gdp_data$REG

negative change 0.2706832 6.885366 0.7686958
positive change 0.2692006 6.885097 0.7670366

Residual Deviance: 4678.77
AIC: 4690.77

The log odds when there is no change in both the ordinal variables is 3.8 while when the first variable OIL is changed to negative the log odds become 4.79 and when REG is changed to negative keeping the other variables constant, becomes 1.38.

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

The R script for model fitting of ordered multinomial logit model is:

```
# 1.2

# create factor through relevel for increasing order level creation
gdp_data$category <- relevel(gdp_data$category, ref = "negative change")

# using ordinal variables from the dataset to fit an ordered multinomial
    logit model
ord_logit <- polr(gdp_data$category ~ gdp_data$OIL+gdp_data$REG)

summary(ord_logit)
```

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

Intercepts:
Value Std. Error t value
negative change—no change -0.7312 0.0476 -15.3597
no change—positive change -0.7105 0.0475 -14.9554

Residual Deviance: 4687.689
AIC: 4695.689

The log odds in ordered multinomial logit model of OIL when the other variable is constant is -0.19 while the log odds of REG when OIL is constant is 0.39. This is because of chnage in category of GDP from negative to positive.

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.

The R script for poisson regression is:

```r
# 2.1

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



# Fit Poisson regression model
poisson <- glm(mexico_elections$PAN.visits.06 ~
                 mexico_elections$competitive.district +
                 mexico_elections$marginality.06 +
                 mexico_elections$PAN.governor.06,
               family = poisson())

# Summarize the model
summary(poisson)
```

Coefficients:
Estimate Std. Error z value Pr(¿—z—)
(Intercept) -3.81023 0.22209 -17.156 ¡2e-16 ***
mexico_elections$competitive.district -0.08135 0.17069 -0.477 0.6336
mexico_elections$marginality.06 -2.08014 0.11734 -17.728 ¡2e-16 ***
mexico_elections$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

The p value of visit to competitive districts is 0.6336 which is less than the benchmark 0.05. As the null hypothesis represents status quo, therefore we cannot accept the alternative hypothesis of more visits to competitive districts.

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

The coefficient of marginality is -2.08 which means that when there is 2 units poverty the presidential candidate visited the district. while the coefficient of PAN governor is -0.31 which predicts that the number of visits by a presential candidate is -0.31.

(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).

The R script for poisson regression is:

```
1
2 # 2.3
3
4 # create a new data frame with specific characteristics
5 hyp_data <- data.frame(competitive.district = 1,
6                         marginality.06 = 0,
7                         PAN.governor.06 = 1)
8
9 # predict mean number of visits for hypothetical data
10 predicted_mean <- predict(poisson, data = hyp_data, type = "response")
11
12 estimated_mean <- mean(predicted_mean)
13
14 # print the predicted mean
15 print(estimated_mean)
```

The estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district is 0.91.