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
if (!require(nnet)) install.packages(”nnet”)
library(nnet)

# Set working directory
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# Load the GDP change data from the provided URL
gdp_data <- read.csv(”/Users/iseli/Downloads/gdpChange.csv”, stringsAsFactors = F)

# Inspect the structure of the data
str(gdp_data)

# View the first few rows of the data
head(gdp_data)

# Create a categorical variable 'GDPWdiff_cat' with levels ”positive”, ”negative”, ”no change”
gdp_data$GDPWdiff_cat <- factor(ifelse(gdp_data$GDPWdiff > 0, ”positive”,
                                        ifelse(gdp_data$GDPWdiff < 0, ”negative”, ”no change”)))

# Check the levels of the new factor variable to ensure they are correct
levels(gdp_data$GDPWdiff_cat)

# Set ”no change” as the reference level for the factor
gdp_data$GDPWdiff_cat <- relevel(gdp_data$GDPWdiff_cat, ref = ”no change”)

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

# Summarize the model to view the estimated coefficients and other statistics
summary(model)

# Interpret the coefficients
# The coefficients for REG and OIL represent the log odds of observing a ”positive” or ”negative” GDPWdiff
# relative to ”no change”, holding all other variables constant. A positive coefficient indicates that
# an increase in the predictor variable is associated with higher odds of the corresponding category of GDPWdiff,
# while a negative coefficient indicates lower odds.
```

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

```r
# Remove objects from the workspace and detach all non-default packages
rm(list=ls())
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0) for (package in package.list) detach(package, character.only=TRUE)
}
detachAllPackages()

# Load necessary libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}
lapply(c("nnet", "MASS"), pkgTest)

setwd("/Users/iseli/Downloads/gdpChange.csv")

# Load the GDP change data
gdp_data <- read.csv("/Users/iseli/Downloads/gdpChange.csv", stringsAsFactors = F)

# Inspect the data structure
str(gdp_data)

# View the first few rows of the dataset
head(gdp_data)

# Create an ordered factor variable for GDPWdiff
gdp_data$GDPWdiff_cat <- factor(ifelse(gdp_data$GDPWdiff > 0, "positive",
                                       ifelse(gdp_data$GDPWdiff < 0, "negative", "no change")),
                                levels = c("negative", "no change", "positive"), ordered = TRUE)

# Fit the ordered multinomial logit model using polr from the MASS package
model_ordered <- polr(GDPWdiff_cat ~ REG + OIL, data = gdp_data, Hess=TRUE)

# Display the summary of the model
summary(model_ordered)
```

```
# Interpretation
# Coefficients indicate the change in log odds of moving to a higher category
# Thresholds (cutpoints) separate the categories and are part of the model output
```

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 data
mexico_elections <- read.csv("/Users/iseli/Documents/GitHub/StatsII_Spring2024/datasets/MexicoMuniData.csv")



# Load the MexicoMuniData.csv data
mexico_elections <- read.csv("/Users/iseli/Documents/GitHub/StatsII_Spring2024/datasets/MexicoMuniData.csv")

# Inspect the data structure
str(mexico_elections)

# View the first few rows of the dataset
head(mexico_elections)

# Run a Poisson regression
# PAN.visits.06 is the outcome variable
# competitive.district, marginality.06, and PAN.governor.06 are the predictors
poisson_model <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06,
                     data = mexico_elections, family = poisson())

# Display the summary of the model to view coefficients, test statistics, and p-values
```

```r
summary(poisson_model)

# Extract and display the specific coefficient, test statistic, and p-value for competitive.district
coef_summary <- summary(poisson_model)$coefficients["competitive.district", ]
cat("Coefficient for competitive.district:", coef_summary["Estimate"], "\n")
cat("Test statistic for competitive.district:", coef_summary["z value"], "\n")
cat("P-value for competitive.district:", coef_summary["Pr(>|z|)"], "\n")
```

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

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 freedomAIC: 1299.2

Number of Fisher Scoring iterations: 7

> coef_summary <- summary(poisson_model)$coefficients["competitive.district", ]> ca
Coefficient for competitive.district: -0.08135181
> cat("Test statistic for competitive.district:", coef_summary["z value"], "\n")
Test statistic for competitive.district: -0.4766106
> cat("P-value for competitive.district:", coef_summary["Pr(>|z|)"], "\n")
P-value for competitive.district: 0.6336394
```

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

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

# Display the summary of the model
model_summary <- summary(poisson_model)

# Interpret the coefficients for marginality.06 and PAN.governor.06
marginality_coef <- model_summary$coefficients["marginality.06", ]
pan_governor_coef <- model_summary$coefficients["PAN.governor.06", ]

cat("Interpretation of marginality.06 coefficient:\n")
cat("Coefficient estimate:", marginality_coef["Estimate"], "\n")
cat("Standard error:", marginality_coef["Std. Error"], "\n")
cat("z-value:", marginality_coef["z value"], "\n")
cat("P-value:", marginality_coef["Pr(>|z|)"], "\n\n")

cat("Interpretation of PAN.governor.06 coefficient:\n")
cat("Coefficient estimate:", pan_governor_coef["Estimate"], "\n")
cat("Standard error:", pan_governor_coef["Std. Error"], "\n")
cat("z-value:", pan_governor_coef["z value"], "\n")
cat("P-value:", pan_governor_coef["Pr(>|z|)"], "\n")
```

```
> cat("Interpretation of marginality.06 coefficient:\n")
Interpretation of marginality.06 coefficient:
> cat("Coefficient estimate:", marginality_coef["Estimate"], "\n")
Coefficient estimate: -2.080144
> cat("Standard error:", marginality_coef["Std. Error"], "\n")
Standard error: 0.1173386
> cat("z-value:", marginality_coef["z value"], "\n")
z-value: -17.72771
> cat("P-value:", marginality_coef["Pr(>|z|)"], "\n\n")
P-value: 2.562806e-70 >
> cat("Interpretation of PAN.governor.06 coefficient:\n")
Interpretation of PAN.governor.06 coefficient:
> cat("Coefficient estimate:", pan_governor_coef["Estimate"], "\n")
Coefficient estimate: -0.3115789
> cat("Standard error:", pan_governor_coef["Std. Error"], "\n")
Standard error: 0.1667306
> cat("z-value:", pan_governor_coef["z value"], "\n")
z-value: -1.868757
> cat("P-value:", pan_governor_coef["Pr(>|z|)"], "\n")
P-value: 0.06165665
```

(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
# Create a new data frame for the hypothetical district
hypothetical_district <- data.frame(competitive.district = 1,
                                    marginality.06 = 0,
                                    PAN.governor.06 = 1)

# Predict the expected count (mean number of visits) for the hypothetical district
predicted_visits <- predict(poisson_model, newdata = hypothetical_district, type = "response")

# Print the estimated mean number of visits
cat("Estimated mean number of visits for the hypothetical district:", predicted_visits, "\n")
```

```
Estimated mean number of visits for the hypothetical district: 0.01494818
```