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.

```

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

# Check the structure of the dataset
str(gdp_data)

# Fit the unordered multinomial logistic regression model
model <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)

#Because It seems I have too many levels in one of my categorical
    variables
#I will proceed to inspect the levels of GDPwdiff
table(gdp_data$GDPWdiff)

# Define breakpoints for binning
breakpoints <- quantile(gdp_data$GDPWdiff, probs = seq(0, 1, by = 0.1))

# Create a new categorical variable by binning GDPWdiff
gdp_data$GDPWdiff_group <- cut(gdp_data$GDPWdiff, breaks = breakpoints,
    labels = FALSE)

# Fit the multinomial logistic regression model using the binned variable
model <- multinom(GDPWdiff_group ~ REG + OIL, data = gdp_data)

# Summarize the model
summary(model)

```

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

Coefficients:
(Intercept)         REG        OIL
2   0.44787599 -0.6300308 -1.5567870
3   0.65516574 -1.3224053 -1.6480791
4   0.66889824 -1.5204747 -1.5707004
5   0.56041172 -1.0167280 -1.3985742
6   0.48915063 -0.7418814 -1.6696137
7   0.23433147 -0.2256526 -0.8533760
8  -0.01306567  0.2774703 -0.8571151
9  -0.16667708  0.5803102 -1.0127366
10 -0.42147924  0.8224895 -0.2648593

Std. Errors:
(Intercept)        REG       OIL
2    0.1044569 0.1544339 0.2487385
3    0.1006293 0.1675130 0.2457301
4    0.1004054 0.1748111 0.2399128
5    0.1022532 0.1606005 0.2316745
6    0.1036795 0.1559679 0.2562922
7    0.1085848 0.1511739 0.2099432
8    0.1147097 0.1508068 0.2176762
9    0.1193782 0.1524712 0.2319998
10   0.1258310 0.1551094 0.1993666

Residual Deviance: 16601.56
AIC: 16655.56
```

Coefficientes; Each row corresponds to a different outcome category of the response variable They represent the change in log odds for a one-unit increase in the respective predictor variable, holding all other variables constant.

REG (Region): The coefficient for each level of the region variable represents the change in the log odds of being in that category compared to the reference category for a one-unit increase in the REG variable while holding other variables constant. For example:

For category 2, an increase of one unit in REG leads to a decrease of approximately 0.63 in the log odds of being in category 2 compared to category 1. Similarly, for category 3, an increase of one unit in REG leads to a decrease of approximately 1.32

in the log odds of being in category 3 compared to category 1, and so on for the other categories.

OIL (Presence of Oil Resources): The coefficient for the presence of oil resources variable represents the change in the log odds of being in each category for a one-unit increase in the OIL variable while holding other variables constant. For example:

For category 2, the presence of oil resources leads to a decrease of approximately 1.56 in the log odds of being in category 2 compared to category 1. Similarly, for category 3, the presence of oil resources leads to a decrease of approximately 1.65 in the log odds of being in category 3 compared to category 1, and so on for the other categories.

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

```
# Load necessary library
library(MASS)

# Convert GDPWdiff_group to factor
gdp_data$GDPWdiff_group <- factor(gdp_data$GDPWdiff_group)

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

# Summarize the model
summary(ordered_model)
```

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

Coefficients:
Value Std. Error t value
REG 0.87554    0.06161 14.2098
OIL 0.04442    0.10459  0.4247

Intercepts:
Value    Std. Error t value
1|2   -1.9272   0.0584   -32.9823
2|3   -1.1215   0.0459   -24.4249
3|4   -0.5717   0.0416   -13.7413
4|5   -0.1193   0.0404    -2.9559
5|6    0.3099   0.0409     7.5838
6|7    0.7394   0.0428    17.2858
7|8    1.2114   0.0462    26.2210
8|9    1.7803   0.0519    34.3080
9|10   2.6294   0.0647    40.6400
```

```
Residual Deviance: 16924.32
AIC: 16946.32
(1 observation deleted due to missingness)
```

Coefficients:

REG (Region): The coefficient represents the change in the log odds of being in a higher GDP change category for a one-unit increase in the REG variable, holding other variables constant. The coefficient for REG is 0.87554, indicating that as the region variable increases by one unit, the log odds of being in a higher GDP change category increase by 0.87554 units.

OIL (Presence of Oil Resources): The coefficient for the presence of oil resources variable represents the change in the log odds of being in a higher GDP change category for a one-unit increase in the OIL variable, holding other variables constant. The coefficient for OIL is 0.04442, indicating that as the presence of oil resources increases by one unit, the log odds of being in a higher GDP change category increase by 0.04442 units.

Intercepts:

Each intercept represents the threshold between adjacent categories of the dependent variable. For example, the intercept between categories 1 and 2 is -1.9272, indicating the log odds of being in category 1 versus category 2.

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(”https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv”)

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

# Summary of the model
summary(poisson_model)
```

```
glm(formula = PAN.visits.06 ~ competitive.district, family = poisson,
data = mexico_elections)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.2571     0.1491 -15.141   <2e-16 ***
competitive.district -0.1617     0.1670  -0.968    0.333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1473.9  on 2406  degrees of freedom
Residual deviance: 1473.0  on 2405  degrees of freedom
AIC: 1776.9

Number of Fisher Scoring iterations: 6
```

H0 (Null Hypothesis): There is no difference in the number of visits made by PAN presidential candidates between swing districts and safe districts.

H1 (Alternative Hypothesis): PAN presidential candidates visit swing districts more frequently than safe districts.

Test Statistic = -0.1617 / 0.1670 = -0.968

P value for competitive district is 0.333, greater than the typical significance value at 0.05, so, we fail to reject the null hypothesis. Therefore, based on this analysis, there is no statistically significant evidence to suggest that PAN presidential candidates visit swing districts more frequently compared to safe, after controlling for other factors included in the model.

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

```
poisson_modelB <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06,
data = mexico_elections, family = poisson)

# Summary of the model
summary(poisson_modelB)
```

```
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 coefficient for marginality.06 is -2.08014. This indicates that, holding all other variables constant, for each one-unit increase in marginality.06, the expected log count of PAN visits decreases by 2.08014 units. This effect is statistically significant (p-value ¡ 0.001).

The coefficient for PAN.governor.06 is -0.31158. This suggests that, holding all other variables constant, having a PAN governor in 2006 decreases the expected log count of PAN visits by 0.31158 units. However, this effect is marginally significant (p-value = 0.0617), meaning it doesn't quite reach conventional levels of significance (e.g., p ¡ 0.05), but it's close.

(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).

Intercept: -3.81023

competitive.district: -0.08135

marginality.06: -2.08014

PAN.governor.06: -0.31158

log(count)=B0+B1×competitive.district+B2×marginality.06+B3×PAN.governor.

competitive.district = 1

marginality.06 = 0

PAN.governor.06 = 1

We substitute these values into the equation:

log (count) = + (3.81023) + (0.08135) x 1 + (2.08014) x 0 + (0.31158) x 1

log (count) = + (3.81023) + (0.08135) + (0.31158)

log (count) = (4.20316)

count = e(4.20316)

count = 0.0147

So, the estimated mean number of visits from the winning PAN presidential candidate for the hypothetical district is approximately 0.0147.