Problem Set 3

Applied Stats II

Due: March 24, 2024 /// Wei Tang 23362496

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.

   Firstly,we can load, read and clean the dataset, and transform it into the format which is easier for us to analyse. Then we can factorize the response variable and explanatory variables we need.

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

# create a new column called GDP_change, including categories:"increase","decrease","no_change".
# GDP_change is depending on GDPWdiff.
for (i in 1:length(gdp_data$GDPWdiff)) {
  if (gdp_data$GDPWdiff[i] > 0) {
    gdp_data$GDP_change[i] <- "increase"
  } else if (gdp_data$GDPWdiff[i] < 0) {
    gdp_data$GDP_change[i] <- "decrease"
  } else {
    gdp_data$GDP_change[i] <- "no_change"
  }
}

# factorize the response variables and explanatory variables, put their levels and labels in the same order.
gdp_data$GDP_change <- factor(gdp_data$GDP_change, levels = c("decrease","no_change","increase"),labels = c("decrease","no_change","increase"))
gdp_data$OIL <- factor(gdp_data$OIL, levels = c(0,1),labels = c("otherwise","Exceed 50%"))
gdp_data$REG <- factor(gdp_data$REG, levels = c(0,1),labels = c("Non-Democracy","Democracy"))
```

   Now we get the clean data, then we start to run the unordered model.

```r
# unordered model
# relevel GDP_change and set "no_change" as the reference category.
gdp_data$GDP_change <- relevel(gdp_data$GDP_change,ref = "no_change")
# run a unordered model
unordered_mod <- multinom(GDP_change ~ REG + OIL, data = gdp_data)
# get the summary of the model to get coefficients
summary(unordered_mod)
# exponent e by coefficients
exp(coef(unordered_mod))
```

   Then we check the summary of the model:

```
> summary(unordered_mod)
```

```
Coefficients:
          (Intercept) REGDemocracy OILExceed 50%
decrease    3.805370     1.379282      4.783968
increase    4.533759     1.769007      4.576321

Std. Errors:
          (Intercept) REGDemocracy OILExceed 50%
decrease   0.2706832    0.7686958      6.885366
increase   0.2692006    0.7670366      6.885097

Residual Deviance: 4678.77
AIC: 4690.77

> exp(coef(unordered_mod))
          (Intercept) REGDemocracy OILExceed 50%
decrease    44.94186     3.972047     119.57794
increase    93.10789     5.865024      97.15632
```

Formula:

$$\ln\left(\frac{P_{\text{decrease}}}{P_{\text{no\_change}}}\right) = 3.805370 + 1.379282 \times \text{REG} + 4.783968 \times \text{OIL}$$

$$\ln\left(\frac{P_{\text{increase}}}{P_{\text{no\_change}}}\right) = 4.533759 + 1.769007 \times \text{REG} + 4.576321 \times \text{OIL}$$

Interpretation:
Intercept decrease:
When the REG and OIL are both 0, the log odds of GDP_change="decrease" vs. GDP_change="no_change" is 3.805370 Intercept increase:
When the REG and OIL are both 0, the log odds of GDP_change="increase" vs. GDP_change="no_change" is 4.533759 REG decrease:
For every 1 unit increase of REG, the log odds of
GDP_change="decrease" vs. GDP_change="no_change" will increase 1.379282 OIL increase:
For every 1 unit increase of OIL, the log odds of
GDP_change="decrease" vs. GDP_change="no_change" will increase 4.783968

REG increase:
For every 1 unit increase of REG, the log odds of
GDP_change="increase" vs. GDP_change="no_change" will increase 1.769007 OIL increase:
For every 1 unit increase of OIL, the log odds of
GDP_change="increase" vs. GDP_change="no_change" will increase 4.576321

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

```r
# ordered model
# run an ordered model
ordered_mod <- polr(GDP_change ~ REG + OIL, data = gdp_data)
# get the summary of the model to get coefficients
summary(ordered_mod)
# exponent e by coefficients
exp(coef(ordered_mod))
# get the confidence interval of the coefficients
exp(cbind(OR=coef(ordered_mod),confint(ordered_mod)))
```

```
> summary(ordered_mod)
Coefficients:
                Value Std. Error t value
REGDemocracy   0.3985    0.07518   5.300
OILExceed 50% -0.1987    0.11572  -1.717

Intercepts:
                    Value    Std. Error t value
decrease|no_change  -0.7312   0.0476   -15.3597
no_change|increase  -0.7105   0.0475   -14.9554

Residual Deviance: 4687.689
AIC: 4695.689
```

Interpretation:
Coefficients:
REG:
For every 1 unit increase of REG, the log odds of
GDP_change="increase" vs. GDP_change="no_change" will increase 0.3985

OIL:
For every 1 unit increase of OIL, the log odds of
GDP_change="increase" vs. GDP_change="no_change" will decrease 0.1987

Intercepts:
the Value of `decrease|no_change` is -0.7312, means the estimated cutoff point for GDP_Change between "decrease" and "no_change" is -0.7312.
the Value of `no_change|increase` is -0.7105, means the estimated cutoff point for GDP_Change between "no_change" and "increase" is -0.7105.

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 and read data
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
# run Poisson regression and build model
mod.ps <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexico_elections, family = poisson)
# check the summary of the model
summary(mod.ps)
```

Firstly, we can run a Poisson regression to build a model.

```
> summary(mod.ps)

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 freedom
AIC: 1299.2

Number of Fisher Scoring iterations: 7
```

Then, we can do a dispersion test.

```r
# do dispersion test
dispersiontest(mod.ps)
```

```
> dispersiontest(mod.ps)
Overdispersion test
data:  mod.ps
z = 1.0668, p-value = 0.143
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
2.09834
```

According to the dispersion test, the p-value is bigger than 0.05, so we should build a zero inflation model.

```r
# build a zero-inflation model
mod.zip <- zeroinfl(PAN.visits.06 ~ competitive.district + marginality.06
    + PAN.governor.06, data = mexico_elections, dist = "poisson")  #

# check the summary of the model
summary(mod.zip)
```

```
> summary(mod.zip)

Call:
zeroinfl(formula = PAN.visits.06 ~ competitive.district + marginality.06 + PAN.gove

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.95323 -0.24006 -0.12842 -0.06045 37.56115

Count model coefficients (poisson with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.9145     0.4982  -3.843 0.000122 ***
competitive.district  0.4024     0.3119   1.290 0.197028
marginality.06       -1.2398     0.2610  -4.750 2.03e-06 ***
PAN.governor.06      -0.4703     0.2707  -1.737 0.082341 .
```

```
Zero-inflation model coefficients (binomial with logit link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.2719     0.6753   1.883  0.05966 .
competitive.district   0.9000     0.5106   1.763  0.07794 .
marginality.06         0.8716     0.3021   2.885  0.00392 **
PAN.governor.06       -0.1749     0.4119  -0.425  0.67106
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 28
Log-likelihood: -600.4 on 8 Df
```

According to the summary of the model, we can see that the p-value of "competitive.district" is not less than 0.05(z-score = 1.763, p-value = 0.07794), so there is no significant evidence that PAN presidential candidates visit swing districts more.

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

Interpretation:
Intercept :
When the competitive.district, marginality and PAN.governor are all 0, the estimated PAN.visits.06 is 1.2719 competitive.district:
For every 1 unit increase of competitive.district, the estimated PAN.visits.06 will increase 0.9000 marginality.06:
For every 1 unit increase of marginality.06, the he estimated PAN.visits.06 will increase 0.8716 PAN.governor.06:
For every 1 unit increase of PAN.governor.06, the he estimated PAN.visits.06 will decrease 0.1749.

(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
# withdraw coefficients
cfs <- coef(mod.zip)
# predict the d mean number of visits
pre_data <- data.frame(competitive.district = 1, marginality.06 = 0, PAN.
    governor.06 = 1)
# get the estimate from original Poisson regression model
exp(predict(mod.ps, newdata = pre_data))
# get the estimate from zero-inflation model
```

Here is the out put:

```
> exp(predict(mod.zip, newdata = pre_data))
```

```
    1
1.016598
```

So 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 1.06598 according to the model.