Problem Set 3

Applied Stats II

Due: March 26, 2023

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 26, 2023. 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.

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

   MY ANSWER FOR QUESTION 1

```
#####################################################################
# Question 1
#####################################################################

# Part 1
# Required to construct and interpret an unordered multinomial model
# Given - the response variable GDPWdiff is difference between year t and
    t-1
# possible catagories "pos" and "neg"
# and "no change" for the reference category

# expanding on this and to make the variables categorical let:
# no change = 0
# pos = 1
# neg = 2

# note that within() evaluates the expression and creates a copy of the
   original data frame ps3Data
ps3Data <- within(ps3Data, {
  # this initializes the variable
  GDPWdiff.cat <- NA
  # for the neg value
  GDPWdiff.cat[GDPWdiff < 0] <- "neg"
  # for the no change value
  GDPWdiff.cat[GDPWdiff == 0] <- "no change"
  # for the pos value
  GDPWdiff.cat[GDPWdiff > 0] <- "pos"}
)

# check that everything looks okay
print(head(ps3Data,n=5))

# X COUNTRY CTYNAME YEAR GDPW OIL REG    EDT GDPWlag GDPWdiff GDPWdifflag
   GDPWdifflag2 GDPWdiff.cat
# 1 1        1 Algeria 1965 6620   1   0  1.45    6502      118
   419         1071           pos
# 2 2        1 Algeria 1966 6612   1   0  1.56    6620       -8
   118          419           neg
# 3 3        1 Algeria 1967 6982   1   0 1.675    6612      370
   -8           118           pos
```

```r
# 4 4        1  Algeria  1968  7848    1    0  1.805     6982       866      370            -8            pos
# 5 5        1  Algeria  1969  8378    1    0   1.95     7848       530      866           370            pos

# the next step is to turn the data into factors or convert numeric to a factor
# can use either as.factor() or factor() which is wrapper for factor but allows quick return if the input factor
# is already a factor
ps3Data$GDPWdiff.cat <- factor(ps3Data$GDPWdiff.cat, levels = c("no change", "pos", "neg"))
# ps3Data$GDPWdiff.cat <- as.factor(ps3Data$GDPWdiff.cat)

# check that everything looks okay
print(summary(ps3Data$GDPWdiff.cat))

# no change      pos        neg
#        16       2600       1105

# REG and OIL should also be factors
ps3Data$REG <- as.factor(ps3Data$REG)
ps3Data$OIL <- as.factor(ps3Data$OIL)

# the next step is to address the reference category and create the p regression model
# the relevel() function doesn't afffect the original dataset
# lm(x ~ y + relevel(b, ref = "3")) ... just working out how to implement it
ps3Data$GDPWdiff2 <- relevel(ps3Data$GDPWdiff.cat, ref = "no change")

# returns the following
# weights:  12 (6 variable)
# initial  value 4087.936326
# iter  10 value 2340.076844
# final  value 2339.385155
# converged

ps3Multinomial <- multinom(ps3Data$GDPWdiff2 ~ REG + OIL, data = ps3Data)

# check that everything looks okay
print(summary(ps3Multinomial))

# returns the following
# Call:
#  multinom(formula = ps3Data$GDPWdiff2 ~ REG + OIL, data = ps3Data)

# Coefficients:
#       (Intercept)      REG1      OIL1
# pos     4.533759 1.769007 4.576321
# neg     3.805370 1.379282 4.783968
```

```
# Std. Errors:
#        (Intercept)       REG1      OIL1
# pos    0.2692006 0.7670366 6.885097
# neg    0.2706832 0.7686958 6.885366

# Residual Deviance: 4678.77
# AIC: 4690.77

## Add this stargazer table to the TeXShop / latex document file
stargazer(ps3Multinomial, title = "unordered multinominal logit")


############ now we can visualize the data with a jitter plot
ggplot(ps3Data, aes(x = OIL , y = GDPWdiff.cat)) +
  geom_jitter(alpha = .5, color="purple") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1)) +
  theme(legend.position="bottom")


ggplot(ps3Data, aes(x = REG, y = GDPWdiff.cat)) +
  geom_jitter(alpha = .5, color="purple") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1)) +
  theme(legend.position="bottom")


### get the P-values and the estimated cutoff points and the coefficients

# Interpretation
## I got a little bit confused here. Referencing the lecture slides (Week
   8: Multinomial Logit Regression)
## and Slide 11. replicating exp(coef(multinom_model) [,c(1:5)])
## to exp(coef(ps3Multinomial)[ ,c(1:3)]) this was giving some unusual
   values
## removing the exp or exponent and the indexing was not returning unusal
   values
# coef(ps3Multinomial) ??

ps3table <- coef(summary(ps3Multinomial))
ps3table

# this returns
#            (Intercept)      REG1      OIL1
# pos            4.533759 1.769007 4.576321
# neg            3.805370 1.379282 4.783968


#### Interpretations continued
#### REG1  'pos'
# for REG1 'pos' = for a unit change in REG going from 0 to 1, non
   democracy to a democracy, the log-odds are that there will
```

```
# be a pos change in the GDP from one year to the next increase by (1.769007) when all other variables in the multinom_model are
# held constant and the ref category is "no change"

#### REG1 'neg'
# for REG1 'neg' = for a unit change in REG1 going from 0 to 1, non democracy to a democracy, the log-odds are that there
# will be a neg change in the GDP from one year to the next increase by (1.379282) when all other variables in the multinom_model
# are held constant and the ref category is "no change"

#### OIL1 'pos'
# OIL1 'pos' when there is a unit change in the OIL variable, where it increases from 0 to 1
# this indicates that the average ratio of fuel exports to total exports in 1984-86 exceeded by 50%
# the log-odds here means  that there will be a pos difference in the total GDP in a COUNTRY from one year to the next year
# and ths results in an increase of (4.576321) and the other variables in the model are held constant


### OIL 'neg'
# for OIL 'neg' when there is a unit change in the OIL variable from going from 0 to 1, then the average ratio of fuel exports
# to total exports in 1984-86 exceeded by 50%, while the log-odds results in a neg difference in the total GDP of a country
# from one year to the next and increases by (4.783968)  while all all other variables are held constant


########## Ordered multinominal logit - Q1 - Part 2
# running the ordered logit
# referencing lecture slide 45
# Hess=TRUE to have the model return the observed information matrix from optimization (called the Hessian)
# which is used to get standard errors.
# ref. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/

ps3_ordered_multi <- polr(GDPWdiff2 ~ REG + OIL, data = ps3Data, Hess = T)

## referencing lecture slide 46
test <- exp(cbind(OR = coef(ps3_ordered_multi), confint(ps3_ordered_multi))) # odds ratio
test
### This returns
# OR              2.5 %       97.5 %
# REG1 0.7000737  0.6042257   0.8102918
# OIL1 1.2593051  1.0029960   1.5754005

```

```


summary(ps3_ordered_multi)

## this returns

#Coefficients:
#      Value Std.     Error     t value
#REG1 -0.3566     0.07485   -4.764
#OIL1  0.2306     0.11510    2.003

#Intercepts:
#                   Value     Std. Error  t value
#no change|pos   -5.5846    0.2534    -22.0376
#pos|neg          0.7491    0.0479     15.6475

#Residual Deviance: 4692.109
#AIC: 4700.109
```

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.

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

(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).

MY ANSWER FOR QUESTION 2

```



##################################################################
# Question 2
```

```r
##############################################################

# import the data MexicaMuniData.csv
mexData <- read.csv("/Users/marklikeman/desktop/ASDS-2023/applied-stats-2-2023/problemset03/MexicoMuniData.csv",
                    stringsAsFactors = FALSE)

# inspecting the data
head(mexData, n=5)
tail(mexData)
str(mexData)
summary(mexData)
stargazer(mexData, title = "MexicoMuniData")

# a quick look at the loaded data
# MunicipCode pan.vote.09 marginality.06 PAN.governor.06 PAN.visits.06 competitive.district
# 1    1001        0.283          -1.831                 0                5            1
# 2    1002        0.352          -0.620                 0                0            1
# 3    1003        0.359          -0.875                 0                0            1
# 4    1004        0.238          -0.747                 0                0            1
# 5    1005        0.378          -1.234                 0                0            1

#### looking at the question 2 details the variables outcome Pan.visits.06 looks to be of interest here
#### the main predictor of interest is whether the district was highly contested, plus measure of property
#### and PAN.governor.06   ...
#### PAN.governor.06 and competitive.district are the binary variables, (1 = close/swing district, 0 = 'safe seat')


# note that within() evaluates the expression and creates a copy of the original dataset
mexData <- within(mexData, {
  PAN.governor.06 <- as.logical(PAN.governor.06) # binary
  competitive.district <- as.logical(competitive.district)} # binary
)

############################################### Part a
## running the poisson regression
pm  <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexData, family = poisson)
summary(pm)

## this returns
```

```
###################################################################
# Deviance Residuals:
# Min        1Q    Median       3Q      Max
# -2.2309  -0.3748  -0.1804  -0.0804  15.2669

# Coefficients:
#                          Estimate Std. Error z value Pr(>|z|)
# (Intercept)             -3.81023    0.22209 -17.156   <2e-16 ***
#  competitive.districtTRUE -0.08135    0.17069  -0.477   0.6336
# marginality.06          -2.08014    0.11734 -17.728   <2e-16 ***
# PAN.governor.06TRUE     -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
###################################################################


# looking at the poisson model it appears to be the case that when
  changing from a safe swing seat, this decreases
# the log-odds that there will be a PAN presidential candidates visit
  while holding all the other visits contant


## for the TeXShop template
stargazer(pm, title = "Poisson Model")


#### visualizing the data
ggplot(data = NULL, aes(x = pm$fitted.values, y = mexData$PAN.visits.06))
   +
  geom_jitter(alpha = 0.5) +
  geom_abline(color = "purple") +
  theme(legend.position="bottom")




############################################### Part b

### the coef marginality.06 is -> marginality.06    -2.08014
### this indicates that for a unit one increase in a measure of property,
   the log-odds of a PAN presidential candidates visit
```

```
### will decrease by factor 2.080 which indicates that poorer districts have a low probility of getting a visit from
### a AN presidential candidate


############################################ Part c


#### referring to the ouput from poisson regression above

# Coefficients:
#                             Estimate Std. Error z value Pr(>|z|)
# (Intercept)               -3.81023    0.22209 -17.156   <2e-16 ***
#  competitive.districtTRUE -0.08135    0.17069  -0.477   0.6336
# marginality.06            -2.08014    0.11734 -17.728   <2e-16 ***
# PAN.governor.06TRUE       -0.31158    0.16673  -1.869   0.0617 .

### this used to estimate the mean

# (intercept*1)  + (competitive.districtTRUE*1) + (marginality.06*0) + (PAN.governor.06TRUE*1)
part_c <- exp((-3.81023*1) + (-0.08135*1) + (2.08014*0) + (-0.31158*1))
part_c
## this gives
## 0.01494827

## interpretation
# the estimated mean for the amount of times that a PAN presidential candidate winning in 2006 is 0.01494827
```