Problem Set 3

Applied Stats II

Due: March 28, 2022

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 class on Monday March 28, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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, run descriptive statistics on the dataset using the summary() function. Then create a table consisting of the X variables OIL and REG to get a more immediate understanding on this data. As you can see below, the vast majority of fuel exports do not exceed 50 percent. By comparison, there is a similar sample of democratic and nn-democratic countries.

```r
# Libraries
library(MASS)
library(nnet)
library(ggplot2)

data <- read.csv('https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2022/main/datasets/gdpChange.csv')
data

# Run descriptive stats on dataset
summary(data)
ftable(xtabs(~ OIL + REG, data = data))
# Vast majority of fuel exports do not exceed 50%; by comparsion there is a similar sample of
# democratic and non-democratic countries.
```

   Now, manipulate the outcome variable to make relevant character categories out of the numeric values. Then, turn it into a factor and set 'no change' as the reference catgeory. With this complete, create two dummy variables out of both the democratic and OIL predictors. After this, create a new smaller dataset containing the cleaned outcome and predictor variables. This dataset is then ready for unordered multinomil logit.

```r
# Manipulate GDPWdiff variable to make categories
# Assign all negative values to 'negative' category, all positives to 'positive', and all
# zeroes to 'no change' category
data$GDPWdiff[data$GDPWdiff > 0] <- 'positive'
data$GDPWdiff[data$GDPWdiff < 0] <- 'negative'
data$GDPWdiff[data$GDPWdiff == 0] <- 'no change'
data
# Set reference category to no change
data$GDPWdiff <- as.factor(data$GDPWdiff)
data$GDPWdiff <- relevel(data$GDPWdiff, ref = 'no change')

# Create two dummy variables out of categories for REG
# Name them Democracy and Non-Democracy
democracy <- ifelse(data$REG == 1, 1, 0)
```

```r
non_democracy <- ifelse(data$REG == 0, 1, 0)

# Create two dummy variables out of categories for OIL exports
# Name them large_oilex and small_oilex
large_oilex <- ifelse(data$OIL == 1, 1, 0)
small_oilex <- ifelse(data$OIL == 0, 1, 0)

final_data <- data.frame(GDPWdiff = data$GDPWdiff,
                         large_oilex = large_oilex,
                         small_oilex = small_oilex,
                         democracy = democracy,
                         non_democracy = non_democracy)
final_data
# Run multinomial unorderd logit
unorder_reg <- multinom(GDPWdiff ~ large_oilex + small_oilex + democracy
   + non_democracy, data = final_data)
summary(unorder_reg)
```

Table 1: The Unordered Multinomial Regression Output

Coefficients:

	(Intercept)	large_oilex	small_oilex	democracy	non_democracy
negative	5.145122	8.385181	-3.240059	3.248796	1.896326
positive	5.554894	8.486243	-2.931349	3.648636	1.906258

Std. Errors:

	(Intercept)	large_oilex	small_oilex	democracy	non_democracy
negative	0.1522334	0.0350539	0.1534428	0.4389436	0.3265688
positive	0.1519580	0.0350539	0.1529560	0.4380432	0.3257614

Residual Deviance: 4678.726
AIC: 4690.726

Now, interpret the coefficients and the intercepts.

1) The Coefficients

A one-unit increase in the variable large_oilex is associated with a large increas

A one-unit increase in the variable large_oilex is associated with a large increase

A one-unit increase in the variable small_oilex is associated with a decrease in lo

A one-unit increase in the variable small_oilex is associated with a decrease in lo

A one-unit increase in the variable democracy is associated with an increase in log

A one-unit increase in the variable democracy is associated with an increase in lo
in the amount of 3.649

A one-unit increase in the variable non_democracy is associated with a slight incr

A one-unit increase in the variable non_democracy is associated with a slight incr

2) The Intercepts:

The intercept for negative GDP diff: when all predictors are at 0, the log odds of

The intercept for positive GDP diff: when all predictors are at 0, the log odds of

Now, run a z-test and get a p-value to look for significance. Drop any predictor from model that is not signficant. Call this model the final model. Then run a final unordered multinomial regression and reinterpret the output.

```
# Run Z-test to get p-value.
# This is needed to see whether these coefficents are statistically significant within
# the model.
z <- summary(unorder_reg)$coefficients/summary(unorder_reg)$standard.errors
z
# Pvalue
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p

#          (Intercept) large_oilex small_oilex    democracy non_democracy
# negative           0           0           0 1.347811e-13  6.367129e-09
# positive           0           0           0 0.000000e+00  4.865686e-09

# From this table, we can see that the coefficients for oil exports (large and small) and one coefficient for democracy are
# statistically significant
# We can also infer that the coefficients for non_democracy appear to not be.
# The intercept is also below 0.05 at 0. It is therefore statistically significant
# Therefore, under this model, only the variables democracy, large_oilex and small_oilex should be kept for further analyses
```

```r

# We can therefore run a final more refined model
final_model <- multinom(GDPWdiff ~ large_oilex + small_oilex + democracy
    - non_democracy, data = final_data) # subtract non_democracy from
    model
summary(final_model)
# Covert coefficients to odds through exponation to make further
    interpretation on non_democracy
# and oil predictors
exp(coef(final_model))

## Interpret Coefficients

# For Democracy Variable:
# There is an increase in the reference category odds that there will be
    a negative GDP difference year-on-year by 3.867 times when a country
    is a democracy
# There is an increase in the reference category odds that there will be
    a positive GDP difference year-on-year by 5.710 times when a country
    is a democracy

# For large_oilex variable:
# There is an increase in the reference category odds that there will be
    a negative GDP difference year-on-year by 3518.108 times when fuel
    exports
# exceed more than 50% of total exports
# There is an increase in the reference category odds that there will be
    a positive GDP difference year-on-year by 3905.215 times when fuel
    exports
# exceed more than 50% of total exports

# For large_oilex variable:
# There is an increase in the reference category odds that there will be
    a negative GDP difference year-on-year by 0.113 times when fuel
    exports
# are less than 50% of total exports
# There is an increase in the reference category odds that there will be
    a positive GDP difference year-on-year by 0.154 times when fuel
    exports
# are less than 50% of total exports
```

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

   Order the outcome variable GDPWdiff as an ordered factor by levels ranging from positive to no change, to negative. Then, run an ordered logit regression using the polyr() function and summarise the output.

```r
# Order GDPWdiff by levels positive, then no change and finally negative
data$GDPWdiff <- factor(data$GDPWdiff, levels = c('positive', 'no change'
    , 'negative'),
```

```
                                    labels = c('postive', 'no change', '
    negative'))
data$GDPWdiff

## Run ordered logit regression using polr() function from MAAS package
order_reg <- polr(GDPWdiff ~ ., data = final_data, Hess = TRUE)
summary(order_reg)
```

```
              Ordinal Regression Output
Call:
polr(formula = GDPWdiff ~ ., data = final_data, Hess = TRUE)

Coefficients:
Value   Std. Error t value
OIL           -0.1788    0.11546  -1.549
REG           0.4102    0.07518   5.456

Intercepts:
Value    Std. Error  t value
no change|negative  -5.3199   0.2523   -21.0865
negative|positive   -0.7036   0.0476   -14.7932

Residual Deviance: 4686.606
AIC: 4694.606
```

The output of the model is visualised above. The coefficient for the OIL variable is -0.1788 and the coefficient for the REG variable is 0.4102. Next, a significance test can be run on each coefficient to validate these values via a p value. This is shown below.

```
# Significance Test: Calculate a p value
coef_table <- coef(summary(order_reg))
pval <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

(coef_table <- cbind(coef_table, "p value" = pval))
### P values for both coefficients are significant at 1.214075e-01 and
# 4.875321e-08
```

Therefore, both coefficients appear significant. Now we can interpret these coefficients in English by exponating them to get the odds ratio. This is shown in code below.

```
### CONVERT TO ODDS RATIO TO INTERPRET COEFFICIENTS
# exponate coefficients
exp((order_reg))
large_oilex    democracy
0.8362455    1.5070726

```

```
# For countries that export oil exceeding 50% of GDP, the odds of there being a positive GDP-differenceversus negative
# year-on-year is 0.84 times that of those whose oil exports do not exceed 50%, holding other variables constant.

# For countries whose regime (REG) is democratic, the odds of there being a positive GDP-difference versus negative
# year-on-year is 1.5 times that of undemocratic regimes, holding other variables constant.
```

Therefore, parts 1 and 2 to this Problem Set have run and interpreted the cutoff points and coefficients for an unordered multinomial and an ordered regression model.

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.

Load in the dataset and run the poisson regression. Summarise the output. The intercept and the coefficient for marginality appear highly significant.The coefficient of interest for the competitive district variable is -0.08135.

```
dataset <- read.csv('https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2022/main/datasets/MexicoMuniData.csv')


dataset <- data.frame(PAN.visits.06 = dataset$PAN.visits.06,
                      competitive = dataset$competitive.district,
                      marginality = dataset$marginality.06,
                    governor = dataset$PAN.governor.06)

##### RUN POISSION REGRESSION
poisson_reg <-glm(formula = PAN.visits.06 ~ ., data = dataset,
              family = poisson)
summary(poisson_reg)
```

```
Summary of Poisson Regression Model

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 -0.08135    0.17069  -0.477   0.6336
marginality -2.08014    0.11734 -17.728   <2e-16 ***
governor    -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
```

Finally, exponate the coefficients to make a final interpretation on the competitive district coefficient. This can determine whether PAN candidates visit competitive districts more often than safe ones. Result: Exponated coefficient for districts (swing or safe) is 0.922

Therefore the expected difference in log count between swing districts compared to safe districts is 0.922

Therefore, it does appear that PAN candidates visit swing districts more often than safe districts.

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

For a one unit increase in marginality, the difference in the logs of expected counts is expected to change by -2.08014, given the other predictors are held constant.

PAN.governor.06 coefficient only slightly negatively correlated with outcome variable (-0.31)

Overall, both coefficients are negative indicting a negative relation. However, unlike PAN.governor.06, the coefficient for marginality is highly significant with a p-value of ¿0.001

(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
competitive <- dataset$competitive
marginality <- dataset$marginality
governor <- dataset$governor
# Create new data
newdata <- data.frame(competitive == 1,
                      governor == 1,
                      marginality == 0)
newdata
# Make the prediction on newdata using predict() function
pred <- predict(poisson_reg, newdata = newdata, type = "response")
mean(pred) # [1] 0.09181554

# Mean visits from a winning candidate for competitive district with an
   avg. poverty level
# and a PAN governor is 0.092.
```