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.

   I initiate by inspecting the data:

```r
str(gdp_data) # Call structure of the data
summary(gdp_data) # Call summary statistics for each column

head(gdp_data, 5) # Call first 5 observations
tail(gdp_data, 5) # Call final 5 observations
colnames(gdp_data) # Call column names
lapply(gdp_data, typeof) # Call datatypes for each column
View(gdp_data) # View whole dataset

print(unique(gdp_data$CTYNAME)) # Call names of countries in dataset
length(unique(gdp_data$CTYNAME)) # Call total number of countries
```

   Note I prepare the response variable, coercing into a factor and then releveling the variable for the purposes of analysis.

```r
# Check response variable. Note is numeric.
print(gdp_data$GDPWdiff)
# Transform variable into factor with levels "negative", "no change", and "positive"
gdp_data$GDPWdiff <- as.factor(ifelse(gdp_data$GDPWdiff < 0, "negative",
                                      ifelse(gdp_data$GDPWdiff == 0, "no change",
                                             "positive")))
# Confirm now a factor
is.factor(gdp_data$GDPWdiff)
# Check levels are updated
levels(gdp_data$GDPWdiff)
# Relevel so "no change" is reference category
gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")
# Confirm releveling is successful
levels(gdp_data$GDPWdiff)
```

   I also note that the predictor/explanatory variables are both sparse, with x almost 60% sparse, and y about 90% sparse (each past the 50% baseline). This may be problematic for analysis.

```r
print(gdp_data$REG)
cat("Note that the data is", round((1 - sum(gdp_data$REG)/length(gdp_data$REG))*100, 2),
    "% 'empty' or zeros, meaning it is considered sparse")

print(gdp_data$OIL)
cat("Note that the data is", round((1 - sum(gdp_data$OIL)/length(gdp_data$OIL))*100, 2),
    "% 'empty' or zeros, meaning it is considered sparse")
```

Fit unordered multinominal logistic regression

```r
# Fit unordered multinomial logistic regression model
model1 <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)
# Print regression table
summary(model1)
# Print odds ratios
model1_ORs <- exp(coef(model1))
```

Interpretation:

The intercept represents the average baseline **log-odds** of being in the reference category (in this case, ”no change” for GDPWdiff), when all other predictor variables are zero.

The rest of the coefficients represent the average change in the log-odds of membership in each outcome category *compared to* the reference category, for a 1-unit increase in the corresponding predictor variable, holding all other variables constant.

The exponents of these log-odds values are the odds ratio (likelihood) of membership with the given category with a 1-unit change across the corresponding label, all other predictors holding constant.

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

   Fit ordered multinominal logistic regression

   Here, the model assumes a relative increasing magnitude across the values of `GDPWdiff` AND the values of `REG` (”Non-Democracy” to ”Democracy”).

```r
# Fit ordered multinomial logistic regression model
model2 <- polr(GDPWdiff ~ REG + OIL, data = gdp_data, Hess = TRUE)
# Print regression table
summary(model2)
# Print odds ratios
model2_ORs <- exp(coef(model2))
```

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.

I initiate by inspecting and preparing the data - I coerve the binary variables as logicals for analysis.

```r
# Inspect data
str(mexico_elections) # Call the structure of the data
summary(mexico_elections) # Call summary statistics for each column

head(mexico_elections, 5) # Call first 5 observations
tail(mexico_elections, 5) # Call final 5 observations
colnames(mexico_elections) # Call column names
lapply(mexico_elections, typeof) # Call datatypes for each column - note all numeric
summary(mexico_elections) # Call summary statistics for each column
View(mexico_elections) # Call the data in a dedicated window view

# Note binary variables must be coerced into logicals for analysis
is.logical(mexico_elections$competitive.district)
is.logical(mexico_elections$PAN.governor.06)

# Coerce within the dataframe's environment
mexico_elections <- within(mexico_elections, {
  competitive.district <- as.logical(competitive.district)
  mexico_elections$PAN.governor.06 <- as.logical(mexico_elections$PAN.governor.06)
})

# Note binary variables now return TRUE as logicals
is.logical(mexico_elections$competitive.district)
is.logical(mexico_elections$PAN.governor.06)
```

I also check the assumption for the Poisson distribution that the mean and the variance of the outcome variable are approximately the same. This assumption does not hold as they are significantly divergent. However, for the purposes of the assignment, I continue.

```r
with(mexico_elections,
     list(mean(PAN.visits.06), var(PAN.visits.06)))

# Visualise with a histogram - shows very strong right-skew...
ggplot(data = mexico_elections, aes(PAN.visits.06)) +
```

First, run Poisson regressions with only `competitive.district` as the predictor, and then with `competitive.district`, `marginality.06` and `PAN.governor.06` as predictors.

```r
# Run Poisson regression...
# ... with only competitive.district as predictor
mod.ps1 <- glm(PAN.visits.06 ~ competitive.district,
                 data = mexico_elections,
                 family = poisson)
summary(mod.ps1)

# ... with competitive.district, marginality.06 and PAN.governor.06 as predictors
mod.ps2 <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06,
                 data = mexico_elections,
                 family = poisson)
summary(mod.ps2)
```

```r
pos_exp1 <- exp(mod.ps1$coefficients)
pos_exp2 <- exp(mod.ps2$coefficients)
```

Note the coefficients of a Poisson regression represent the **average change** in the **expected log-count** of the outcome/response variable with a **one-unit change** in the corresponding predictor, holding all other predictors constant. Therefore, the corresponding **exponentiated value** represents the **average change in the expected count** of the outcome variable **in percentage terms** for a 1-unit increase in the corresponding predictor, holding all others constant.

Note the first model does not return a statistically-significant coefficient for `competitive.district` (i.e. $p-value > 0.05$), failing to reject the null hypothesis; that there is no difference in the number of visits between 'swing' and 'safe' states by PAN presidential candidates.

If it were statistically significant, the regression would indicate an average change of -0.162 in the log-count of PAN presidential candidate visits were a state to change from 'safe' to 'swing' - in clearer terms, the average expected count in visits to a 'swing' state would be 85% of the average expected count for a 'safe' state.

Note the second model also fails to return a statistically significant coefficient for `competitive.district`, and therefore also fails to give evidence to reject the null hypothesis (stated above).

If it were statistically significant, the regression coefficient would indicate an average change of -0.081 in the log-count of PAN presidential candidate visits were a state to change from 'safe' to 'swing', all other predictors held constant - in clearer terms, the average expected count in visits to a 'swing' state would be 92% of the average expected count for a 'safe' state, all other predictors held constant.

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

The average expected change in the log-count for a 1-unit increase in the poverty measure is -2.08, all other predictors held constant, which is statistically significant ($p-value < 0.05$) - therefore, this result gives evidence to reject the null hypothesis that the average expected number of visits to a state is the same across state-level poverty levels, holding all other predictors constant; and gives evidence for the alternative

Table 1: Poisson Regression: One Predictor

	Dependent variable:
	PAN.visits.06
competitive.district	−0.162
	(0.167)
Constant	-2.257^{***}
	(0.149)
Observations	2,407
Log Likelihood	−886.462
Akaike Inf. Crit.	1,776.925
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Poisson Regression 2: Three Predictors

	Dependent variable:
	PAN.visits.06
competitive.district	−0.081
	(0.171)
marginality.06	-2.080^{***}
	(0.117)
PAN.governor.06	-0.312^{*}
	(0.167)
Constant	-3.810^{***}
	(0.222)
Observations	2,407
Log Likelihood	−645.606
Akaike Inf. Crit.	1,299.213
Note:	*p<0.1; **p<0.05; ***p<0.01

hypothesis that the average number of visits is different across poverty levels, holding all other predictors constant.

In other terms, a 1-unit increase in the poverty measure indicates an average expected number of visits to be 12% that which would otherwise be expected, holding all other predictors constant.

The average expected change in the log-count for a state without or with a PAN-affiliated governor is -0.312, all other predictors held constant, which not statistically significant ($p - value > 0.05$) - therefore, this result fails to give evidence to reject the null hypothesis that the average expected number of visits to a state is the same for states with a pre-existing PAN-affiliated governor and states without, holding all other predictors constant.

If this coefficient was statistically significant, the average expected number of visits to states with a PAN-affiliated governor would be 73% that of a state without an affiliated governor, holding all other predictors constant.

(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).

The predicted mean number of visits for the given hypothetical district is 0.015 visits.

```r
# Create hypothetical district
pred <- data.frame(MunicipCode = 0,
                   pan.vote.09 = 0,
                   marginality.06 = 0,
                   PAN.governor.06 = 1,
                   competitive.district = 1)

pred_hyp <- predict(mod.ps2, newdata = pred, type = ”response”)
pred_hyp <- cbind(pred_hyp, ”Exp” = exp(pred_hyp))
```