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.

```r
table(gdp_data$GDPWdiff_category)
```

```
> table(gdp_data$GDPWdiff_category)
 negative no change  positive
     1105        16      2600
```

```r
## Transform the REG (Regime) from binary (0;1) into the Categorical 0=Non-Democracy; 1=Democracy
gdp_data$REG <- factor(gdp_data$REG, levels=c(0, 1), labels=c("Non-Democracy", "Democracy"))
table(gdp_data$REG)
```

```
Non-Democracy     Democracy
         2227          1494
```

```r
#Transform the variable OIL from binary to Categorical Variable:
gdp_data$OIL <- factor(gdp_data$OIL, levels=c(0, 1), labels=c("Not Exceed 50%","otherwise"))
table(gdp_data$OIL)
```

```
Not Exceed 50%      otherwise
          3347            374
```

```r
ftable(xtabs(~REG + GDPWdiff_category + OIL, data=gdp_data))
```

```
 OIL Not Exceed 50% otherwise
REG           GDPWdiff_category
Non-Democracy negative                               641         93
              no change                               14          0
              positive                              1284        195
Democracy     negative                               332         39
              no change                                2          0
              positive                              1074         47
```

```
gdp_data$GDPWdiff_category <- factor(gdp_data$GDPWdiff_category, levels=c("negative","no change","positive"),
                                     labels=c("negative","no change","positive"))
####################
###################################
#PROBLEM SET III. Question 1: Part 1.
# Fitting an unordered multinomial logit with as the output and setting a reference category "no change".
##
gdp_data$GDPWdiff_category <- relevel(gdp_data$GDPWdiff_category , ref="no change")
# Run the Model:
multinom_model_unordered <- multinom(GDPWdiff_category ~ REG + OIL, data = gdp_data)
summary(multinom_model_unordered) # # Summary of the model
```

```
Call:
multinom(formula = GDPWdiff_category ~ REG + OIL, data = gdp_data)

Coefficients:
         (Intercept) REGDemocracy OILotherwise
negative    3.805370     1.379282     4.783968
positive    4.533759     1.769007     4.576321

Std. Errors:
         (Intercept) REGDemocracy OILotherwise
negative   0.2706832    0.7686958     6.885366
positive   0.2692006    0.7670366     6.885097

Residual Deviance: 4678.77
AIC: 4690.77
```

```
# ln(GDPWdiif_negative/DGPWdiff_nochange)=3.805370 + 1.379282*REGDemocracy + 4.783968*OILotherwise
#(GDPWdiif_negative/DGPWdiff_nochange) = exp(3.805370 + 1.379282*REGDemocracy + 4.783968*OILotherwise)
```

$$\ln\left(\frac{GDPWdiif_{\text{negative}}}{DGPWdiff_{\text{nochange}}}\right) = 3.805370 + 1.379282 \times \text{REGDemocracy} + 4.783968 \times \text{OILotherwise}$$

```
#(ii) For GDPWdiff_positive and the reference category is DGPWdiff_nochange:

# ln(GDPWdiif_positive/DGPWdiff_nochange)=4.533759 + 1.769007*REGDemocracy + 4.576321*OILotherwise
#(GDPWdiif_positivetive/DGPWdiff_nochange) = exp(4.533759 + 1.769007*REGDemocracy + 4.576321*OILotherwise)
```

```

exp(coef(multinom_model_unordered)) # Convert the coefficients to odds
    ratio
# Answer/Output:
#              (Intercept) REGDemocracy OILotherwise
#negative     44.94186      3.972047    119.57794
#positive     93.10789      5.865024     97.15632
#####
# Calculate the p-values:
z <- summary(multinom_model_unordered)$coefficients/summary(multinom_
    model_unordered)$standard.errors
(p <- (1 - pnorm(abs(z), 0, 1))*2)
```

exp(coef(multinom$_{m}odel_{u}nordered)thatconvertthecoefficientstooddsratioisgivenby$:

article graphicx booktabs

Table 1: The Coefficients of Odds Ratio

	(Intercept)	REGDemocracy	OILotherwise
negative	44.94186	3.972047	119.57794
positive	93.10789	5.865024	97.15632

Interpretations: Intercept (for "negative"): $\hat{\beta}_0 = 3.80537$, it indicates that when all predictor variables (REGDemocracy and OILotherwise) are zero, the log-odds of observing a "negative" outcome are 3.805.
REGDemocracy (for "negative"):A positive coefficient (1.379282) suggests that as REGDemocracy increases by one unit, the log-odds of the outcome being "negative" versus "no change" increase by approximately 1.379.
OILotherwise (for "negative"): Similarly, this coefficient (4.783968) indicates the change in log-odds of the outcome being "negative" versus "no change" for a one-unit increase in the OILotherwise predictor variable, holding all other variables constant. A higher coefficient suggests a larger effect on the log-odds.
Interpretation of Odds Ratio:
Intercept(Negative Category):For every one unit increase in the odds of observing a "negative" change in GDPWdiff, the odds of observing "no change" in GDPWdiff decrease by a factor of approximately 44.94, holding all other variables constant.
For every one unit increase in the odds of observing a "negative" change in GDPWdiff, the odds of observing a "positive" change in GDPWdiff decrease by a factor of approximately 93.11, holding all other variables constant.

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

```
multinom_model_ordered <- polr(GDPWdiff_category ~ REG + OIL, data = gdp_
    data, Hess = TRUE)
summary(multinom_model_ordered) # # Summary of the model
```

```
#Call:
#   polr(formula = GDPWdiff_category ~ REG + OIL, data = gdp_data,
#        Hess = TRUE)
```

Table 2: Estimated Coefficients for Ordered Multinomial Logit Reg

	Value	Std. Error	t value
Coefficients			
REGDemocracy	0.3985	0.07518	5.300
OILNot Exceed 50%	0.1987	0.11572	1.717
Intercepts			
negative—no change	-0.5325	0.1097	-4.8544
no change—positive	-0.5118	0.1097	-4.6671

Residual Deviance: 4687.689
AIC: 4695.689

```
## Calculating the p-value
ctable1 <- coef(summary(multinom_model_ordered)) # Extract coefficient
    summary
p <- 2 * (1 - pnorm(abs(ctable1[, "t value"]))) # Calculate the p-value
ctable1 <- cbind(ctable1, "p-value" = p) ## Combine coefficient summary
    and p-values
print(ctable1) # Print the results
## Answer:
```

Table 3: Calculating the p-values

	Value	Std. Error	t value	p-value
REGDemocracy	0.3984828	0.07518478	5.300046	1.157735e-07
OILNot Exceed 50%	0.1987196	0.11571711	1.717288	8.592653e-02
negative—no change	-0.5324600	0.10968546	-4.854426	1.207358e-06
no change—positive	-0.5117652	0.10965270	-4.667147	3.054110e-06

```
# Calculating  95% confidence intervals:
(ci <- confint(multinom_model_ordered))
## Answer:
```

Table 4: Calculating 95 percent Confidence Intervals

	2.5%	97.5%
REGDemocracy	0.25165482	0.5464341
OILNot Exceed 50%	-0.03019571	0.4237548

```
# Converting to odds ratio:
exp(cbind(OR=coef(multinom_model_ordered), ci))
## Answer:
```

Table 5: Odds Ratios and Confidence Intervals

	OR	2.5%	97.5%
REGDemocracy	1.489563	1.2861520	1.727083
OILNot Exceed 50%	1.219840	0.9702556	1.527687

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
mexico_elections <- read.csv("C:/NewGithubFolder/StatsII_Spring2024/datasets/MexicoMuniData.csv")
#
head(mexico_elections)
names(mexico_elections)#"MunicipCode" ; "pan.vote.09"; "marginality.06"; "PAN.governor.06"; "PAN.visits.06"; "competitive.district"
dim(mexico_elections) # Rows/Observations=2407; Columns/Variables=6
str(mexico_elections)
# ################################################################################################
# ################################################################################################
##############
# Outcome variable: "PAN.visits.06"
###
# Predictors Variables of the interest:
# "competitive.district: 1=close/swing district; 0="safe seat") "
#PAN.governor.06"
#"marginality.06"
#"PAN.governor.06"
####
table(mexico_elections$PAN.visits.06)
# Answer:
#
```

```
# 0 1 2 3 4 5 35
# 2272 102 17 12 1 2 1
###############
table(mexico_elections$"marginality.06")
# (a) Answers:
poisson_reg.model <- glm(PAN.visits.06 ~ competitive.district +
    marginality.06 +
                          PAN.governor.06 , data=mexico_elections ,
    family=poisson)
summary(poisson_reg.model)
```

(a) 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.

Table 6: Model Coefficients Estimated of Poisson Regression

	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 .

Interpretation:
Intercept
:
The estimated intercept is approximately -3.81023, indicating the expected log count of PAN visits when all other predictors are zero. The associated standard error is 0.22209. The z-value is -17.156, and the p-value is less than 2e-16, indicating that the intercept is statistically significant.
competitive.district:
The coefficient estimate for competitive.district is approximately -0.08135. This suggests that for a one-unit increase in the competitive.district variable (indicating a swing district), the log count of PAN visits decreases by 0.08135 units, holding all other variables constant. However, the associated p-value is 0.6336, indicating that this coefficient is not statistically significant at conventional levels.
Over-dispersion test - check equal variance assumption:

```r
install.packages("AER")
library("AER")
#
dispersiontest(poisson_reg.model)
## Answer:
#Overdispersion test
```

Table 7: Overdispersion Test-Check Equal Variance Assumption

Data	poisson_reg.model
z	1.0668
p-value	0.143
Alternative hypothesis	true dispersion is greater than 1
Sample estimates	
Dispersion	2.09834

Note that one of the common cause of over-dispersion is excess zeros, which in turn are generated by an additional data generating process. In this situation, zero-inflated model should be considered/applied

```r
#Slide 22  Zip Model in R
# R contributed package "pscl" contains the function zeroinfl:
install.packages("pscl")
library("pscl")
###
zeroinfl_poisson_1 <- zeroinfl(PAN.visits.06 ~ competitive.district +
    marginality.06 +
                                       PAN.governor.06 , data=mexico_
    elections, dist="poisson")
summary(zeroinfl_poisson_1)
## Answer:
```

(b)

```r
#(b)
exp(coef(zeroinfl_poisson_1))
## Answer:
```

Interpretation:

For the Model with Zero-Inflation Component (zeroinfl$_{p}oisson_{1}$) :

count$_{m}arginality.06$:

The exponentiated coefficient is approximately 0.2894. This suggests that for a one-unit increase in the marginality.06 variable (which likely represents a measure of poverty), the expected count of PAN visits decreases by a factor of 0.2894, holding all other variables constant. In other words, districts with higher poverty levels tend to have fewer PAN visits.

count$_{P}AN.governor.06$:

Table 8: Model Coefficients and Pearson Residuals

	Estimate	**Std. Error**	**z value**	**Pr(>—z—)**
Pearson Residuals:				
Min	-0.95323			
1Q	-0.24006			
Median	-0.12842			
3Q	-0.06045			
Max	37.56115			
Count Model Coefficients (poisson with log link):				
(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):				
(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

Table 9: Count and Zero-Inflation Model Coefficients

	Count	**Zero**
(Intercept)	0.1474155	3.5675098
competitive.district	1.4953556	2.4596673
marginality.06	0.2894367	2.3906532
PAN.governor.06	0.6247883	0.8395139

The exponentiated coefficient is approximately 0.6248. This suggests that districts with a PAN-affiliated governor in 2006 have a count of PAN visits that is approximately 0.6248 times the count in districts without such a governor, holding all other variables constant. It indicates that the presence of a PAN-affiliated governor is associated with a decrease in the expected count of PAN visits, although the effect is not as strong as poverty.

(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).

```
##(c)
# Coefficients from the Poisson regression model
coefficients <- coef(zeroinfl_poisson_1)

# Calculating the linear predictor (eta) using the coefficients and values
Est_mean.visitors <- data.frame(competitive.district=1, marginality.06=0, PAN.governor.06=1)
exp(predict(zeroinfl_poisson_1, Est_mean.visitors)) # Answer: 1.016598
```

commentary of the outcome result:
This means that, on average, the winning PAN presidential candidate is expected to visit the hypothetical district 1.016598 times during the specified time period (e.g., leading up to the 2009 federal elections).