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 I examined the data set to understand the variables on interest and to make sure that the different variables were in the right format, such as the outcome of interest being categorical. This following table was produced:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	3,721	1,861.000	1,074.305	1	931	2,791	3,721
COUNTRY	3,721	70.420	37.535	1	39	103	135
YEAR	3,721	1,974.786	9.847	1,954	1,967	1,983	1,990
GDPW	3,721	9,276.381	8,198.898	509	2,566	13,470	37,903
GDPWlag	3,721	9,090.429	8,066.179	509	2,533	13,167	37,089
GDPWdiff	3,721	185.952	590.095	−9,257	−24	415	7,867
GDPWdifflag	3,721	189.683	583.573	−9,257	−20	415	7,867
GDPWdifflag2	3,721	189.926	582.219	−9,257	−19	405	7,867

Table 1 above shows that our outcome variable GDPWdiff has been input as numerical, therefore data wrangling will need to be preformed to make this into categorical data. The following code was used:

```
gdp_data <- within(gdp_data, {
  GDPWdiff.cat <- NA # need to initialize variable
  GDPWdiff.cat[GDPWdiff < 0] <- ”negative”
  GDPWdiff.cat[GDPWdiff == 0 ] <- ”no change”
  GDPWdiff.cat[GDPWdiff > 0] <- ”positive”
} )

#Next we need to make this data into factors
gdp_data$GDPWdiff.cat <- as.factor(gdp_data$GDPWdiff.cat)
```

Following this the explanatory variables were examined, as both REG and OIL were binary these were encoded with the correct labels using the following code:

```
gdp_data$REG <- as.integer(as.logical(gdp_data$REG))
gdp_data$REG <- factor(gdp_data$REG,
                          levels = c(0,1),
                          labels = c(”Non Democracy”, ”Democracy”))

gdp_data$OIL <- as.integer(as.logical(gdp_data$OIL))
gdp_data$OIL <- factor(gdp_data$OIL,
                         levels = c(0,1),
                         labels = c(”Otherwise”, ”Ratio exceeded 50%”))
```

Following this, the following graphs were produced to further understand the data:

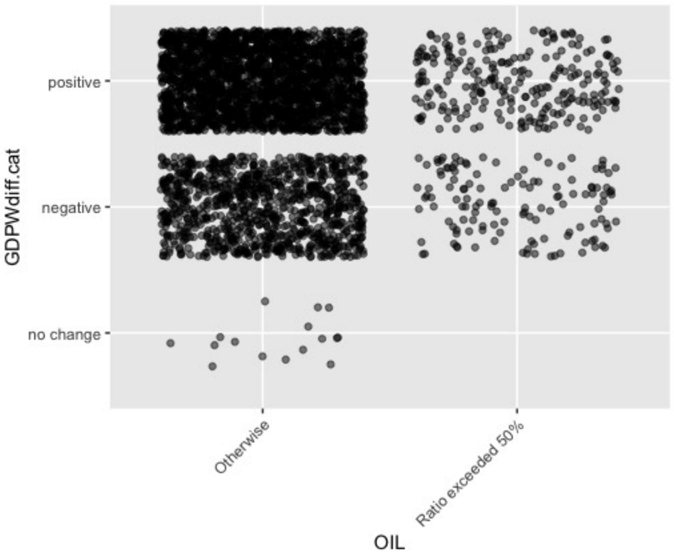

Figure 1: OIL

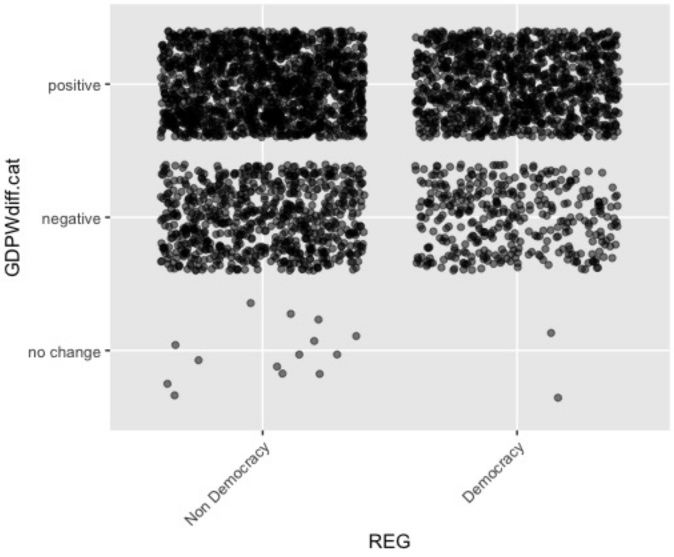

Figure 2: REG

Both of these plots show the spread of the data.

Next the reference category was set to ”no change” in the outcome variable GDPWdiff:

```
gdp_data$GDPWdiff.cat <- relevel(gdp_data$GDPWdiff.cat, ref = ”no change”
    )
```

Next an Unordered Multinomial Logit was preformed using the following code:

```
model <- multinom(GDPWdiff.cat ~ REG + OIL, data = gdp_data)
```

This produced the following output:

Table 2: Unordered Multinomial Logit

	Dependent variable:	
	negative	positive
	(1)	(2)
REGDemocracy	1.379*	1.769**
	(0.769)	(0.767)
OILRatio exceeded 50%	4.784	4.576
	(6.885)	(6.885)
Constant	3.805***	4.534***
	(0.271)	(0.269)
Akaike Inf. Crit.	4,690.770	4,690.770
Note:	*p<0.1; **p<0.05; ***p<0.01	

Interpretation: For every one unit increase in X, the log odds of Y= j vs Y = 1 increase by beta1

REG:

Non Democracy = 0, Democracy = 1

For REG, in a given country, there is an increase in the log odds that GDP difference will be positive by 1.769, when the country is democratic.

For REG, in a given country, there is an increase in the log odds that GDP difference will be negative by 1.379, when the country is democratic.

OIL:

Otherwise = 0, average ratio of fuel exports to total exports in 1984-86 exceeded 50% = 1

For OIL, in a given country, there is an increase in the log odds that GDP difference will be positive, by 4.784, when the fuel ration exceeds 50%

For OIL, in a given country, there is an increase in the log odds that GDP difference will be negative, by 4.576, when the fuel ration exceeds 50%

However, the p value for these coefficients is not significant, and therefor we fail to reject the null hypothesis that the slope is different to zero.

The table above shows which of the coefficients in the model are significant, but the following code was used to get the actual p values: The p value and confidence intervals were produced using the following code:

```
#for a coefficient table
ctable <- coef(summary(model))
#To get the p value
z <- summary(model)$coefficients/summary(model)$standard.errors
(p <- (1 - pnorm(abs(z), 0, 1)) * 2)
#Bind together the p value with the coefficients
(ctable <- cbind(ctable, ”p value” = p))
#For the confidence intervals
(ci <- confint(model))
```

This produced the following: Firstly the p value and coefficients, the first half of the table shows the coefficients and the second half shows the p values:

	(Intercept)	REG	OIL	(Intercept)	REG	OIL
negative	3.805370	1.379282	4.783968	0	0.07276308	0.4871792
positive	4.533750	1.769007	4.576321	0	0.02109459	0.5062612

Following this the below results show the confidence intervals for the model:

Negative	2.5%	97.5%
(Intercept)	3.2748404	4.335899
REG	-0.1273345	2.885898
OIL	-8.7111015	18.279038

Positive	2.5%	97.5%
(Intercept)	4.0061354	5.061383
REG	0.2656425	3.272371
OIL	-8.9182207	18.070863

The confidence interval shows that there is 95% confidence that the parameter will fall between the values in the table above.

We can see that for OIL there is a very large confidence interval, in both the negative and positive categories.

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

   To construct an ordered multinomial logit the data wrangling from above was preformed again to fit the data into categories and binary variables when needed, but the reference category line was removed.

   The following code was then used to produce the ordered multinomial logit:

```r
ordinal_model <- polr(GDPWdiff.cat ~ REG + OIL, data = gdp_data, Hess = TRUE)
```

   The following output was produced:

Table 3: Ordered Multinomial Logit

	Dependent variable:
	GDPWdiff.cat
REGDemocracy	0.398***
	(0.075)
OILRatio exceeded 50%	−0.199*
	(0.116)
Observations	3,721
Note:	*p<0.1; **p<0.05; ***p<0.01

   Interpretation: For every one unit increase in X, the log odds of y = j vs y = j + one increase in beta.

   REG

   For countries which are democratic, the log odds of having a GDP difference that is positive is 0.398 times higher than non democratic countries, holding constant all other variables.

   OIL

   For countries which have the oil ratio exceeding 50%, the log odds of having a GDP difference that is positive is 0.199 times lower than countries who's oil ratio is other than 50%, holding constant all other variables.

   The following code was used to produce the cut of points and the coefficients:

```r
ctable <- coef(summary(ordinal_model))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
(ci <- confint(ordinal_model))
```

Table 4: Confidence intervals for the Ordered Model

	2.5 %	97.5 %
REGDemocracy	0.252	0.546
OILRatio exceeded 50%	-0.424	0.030

The table above shows the confidence intervals for the ordered model, this shows that OIL ratio has larger interval

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.**

Firstly, I examined the data to see if any columns would need to changed into binary categories:

Table 5:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
MunicipCode	2,407	19,505.170	7,308.898	1,001	14,107.5	24,039.5	32,057
pan.vote.09	2,407	0.272	0.454	0.005	0.135	0.360	17.000
marginality.06	2,407	−0.001	0.983	−2.270	−0.746	0.629	3.355
PAN.governor.06	2,407	0.215	0.411	0	0	0	1
PAN.visits.06	2,407	0.092	0.802	0	0	0	35
competitive.district	2,407	0.821	0.383	0	1	1	1

We can see that PAN.governor.06 columns is being treated as numerical data even though it it a binary classifier, and the competitive.district column as well. Following this the following code was used to wrangle the data:

```r
mexico_data <- within(mexico_data, {
  PAN.governor.06 <- as.logical(PAN.governor.06)
   competitive.district <- as.logical(competitive.district)
})
```

Once this was completed a Poisson regression was ran with (`PAN.visits.06`) as the outcome variable. The following code was used:

```r
poisson_model  <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexico_data, family = poisson)
```

This produced the following results:

Table 6: Poisson Model

	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

The following R output includes the test statistics and the p values.

```
Call:
glm(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
    PAN.governor.06, family = poisson, data = mexico_data)

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 <0.0000000000000002
    ***
competitive.districtTRUE  -0.08135    0.17069  -0.477              0.6336
marginality.06            -2.08014    0.11734 -17.728 <0.0000000000000002
    ***
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
```

The Poisson models shows that changing from a swing district to a safe seat , decreases the log odds that there will be a PAN presidential visit, while holding all else constant. This variable had a test statistic of -0.477. This suggests that there would be a higher chance of a PAN president visiting a swing district, however, this figure was not statistically significant in the model, with a p value of 0.6336, which is over the alpha threshold of 0.05, which is needed to reject the null hypothesis that there is no relationship. This suggests that there is not enough evidence to suggest that PAN presidential candidates visit swing districts more.

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

The coefficient for `marginality.06` is -2.080. For every one unit increase in poverty score, the log odds of a presidential candidate visit decreases by a multiplicative factor of 2.080. This suggests that poorer district were less likely to get a visit from a PAN presidential candidate.

The coefficient for `PAN.governor.06` is -0.312. Changing from a state that has a PAN-affiliated governor to a non-PAN-affiliated governor, the log odds of a presidential candidate visit decreases by a multiplicative factor of 0.312, while holding all else constant. Suggesting that if the district has a PAN affiliated governor, the log odds of a presidential visit increase.

(c) **Provide the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive (`competitive.district=` had an average poverty level (`marginality.06` = 0), and a PAN governor (`PAN.governor.06`=1).**

The following code was used to calculate the estimated mean for the number of visits from the winning PAN presidential candidate, with the above conditions:

```

#                              Estimate      Question
#(Intercept)                   -4.20317
#Competitive.districtTRUE        -0.08135         1
#marginality.06                  -2.08014         0
#PAN.governor.06TRUE             -0.31158         1

lamda <- exp((-4.20317*1) + (-0.8135*1) + (-2.08014*0) + (-0.31158*1))
```

This means the estimated mean for the number of times the winning PAN presidential candidate in 2006 is 0.004852555.