Problem Set 3

Applied Stats II

Maiia Skrypnyk 23371609

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.

   Loading the data:

```
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
```

   Preparing the data:

```
#Categorising the output variable into three categories
gdp_data$GDPWdiff1 <- ifelse(gdp_data$GDPWdiff > 0, "positive",
                      ifelse(gdp_data$GDPWdiff == 0, "no change", "negative"))

#Relevelling so the reference category is "no change"
gdp_data$GDPWdiff1 <- relevel(factor(gdp_data$GDPWdiff1), ref = "no change")
```

   Essentially, by manually categorising the numerical observations of the outcome variable into three groups, we are embedding the cutoff points of < 0, 0, > 0 into the model estimation process.

   Constructing an unordered multinomial logit model:

```
model1 <- multinom(GDPWdiff1 ~ REG + OIL,
                   data = gdp_data)
summary(model1) #Coefficients indicate the change in log odds
```

(Please see the summary table on the next page)

Table 1: Unordered Multinomial Logit Model

	Dependent variable:	
	negative	positive
	(1)	(2)
REG	1.379*	1.769**
	(0.769)	(0.767)
OIL	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	

Let's exponentiate the coefficients and add confidence intervals for a more intuitive interpretation to follow:

```
exp_model1 <-exp(cbind(OR = coef(model1), confint(model1)))
exp_model1
```

Table 2: Unordered Multinomial Logit Model (OR)

	(Intercept)	REG	OIL
negative	44.942	3.972	119.578
positive	93.108	5.865	97.156

- **Intercept (negative), 44.942:** On average, when a country is non-democratic and does not produce any oil, the baseline odds of going from having no change in GDP difference to a negative GDP difference ≈ 44.942.
- **Intercept (positive), 93.108:** On average, when a country is non-democratic and does not produce any oil, the baseline odds of going from having no change in GDP difference to a positive GDP difference ≈ 93.108.

  Both intercepts are statistically differentiable from zero at the α-level = 0.05.
- **REG(negative), 3.972**: On average and holding all the other predictors constant, a shift from being a non-democratic to a democratic country is associated

with an increase by a multiplicative factor of 3.972 in the odds of going from having no change in GDP difference to a negative GDP difference.

- **REG(positive), 5.865**: On average and holding all the other predictors constant, a shift from being a non-democratic to a democratic country is associated with an increase by a multiplicative factor of 5.865 in the odds of going from having no change in GDP difference to a positive GDP difference. *The coefficient is statistically differentiable from zero at the α-level = 0.05.*
- **OIL(negative), 119.578**: On average and holding all the other predictors constant, a shift from not producing oil to producing oil for a country is associated with an increase by a multiplicative factor of 119.578 in the odds of going from having no change in GDP difference to a negative GDP difference.
- **OIL(positive), 97.156**: On average and holding all the other predictors constant, a shift from not producing oil to producing oil for a country is associated with an increase by a multiplicative factor of 97.156 in the odds of going from having no change in GDP difference to a positive GDP difference.

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

Preparing the data:

```
#Relevelling to ordered
gdp_data$GDPWdiff1 <- factor(gdp_data$GDPWdiff1, levels = c("negative", "
    no change", "positive", ordered = TRUE))
```

Constructing an ordered multinomial logit model:

```
model2 <- polr(GDPWdiff1 ~ REG + OIL,
               data = gdp_data, Hess=T)
summary(model2)
```

(Please see the summary table on the next page)

Table 3: Ordered Multinomial Logit Model

	Dependent Variable: GDPWdiff1		
	Value	Std. Error	t-value
REG	0.398***	0.075	5.300
OIL	-0.199*	0.116	-1.717
Intercepts:			
	Value	Std. Error	t-value
negative—no change	-0.731	0.048	-15.360
no change—positive	-0.711	0.048	-14.955
positive—TRUE	1232.909	0.048	25952.256
Observations: 3,721			
Residual Deviance: 4687.689			
AIC: 4697.689			
Note:	*p<0.1; **p<0.05; ***p<0.01		

Let's exponentiate the coefficients and add confidence intervals for a more intuitive interpretation to follow:

```
exp_model2 <-exp(cbind(OR = coef(model2), confint(model2)))
exp_model2
```

Table 4: Ordered Multinomial Logit Model (OR)

	OR	2.5%	97.5%
REG	1.490	1.286	1.727
OIL	0.820	0.655	1.031

- **Intercept: negative-no change, $e^{-0.731} = 2.07$ + Intercept: no-change-positive, $e^{-0.711} = 2.03$**:
  - The intercepts are very similar to each other. Therefore, we assume that the "proportional odds"/"parallel regression assumption" holds for the ordered model, which makes sense: the categories we imposed are inherently ordinal (negative → no change → positive). Thus, also, the effect of the predictors on the outcome variable is the same across all of its categories.
  - While the intercepts for the ordered model do not exactly have an intuitive meaning, or an interpretation alike to the ordered model, we might interpret them as the cutoff points for shifting from the lower category of the outcome variable to the higher one.

- **REG, 0.398:** On average and holding all the other predictors constant, a shift from being a non-democratic to a democratic country is associated an increase by a multiplicative factor of 1.490 in the odds of observing a next/higher category (any one of them, "proportional odds"/"parallel regression assumption" again) of the GDP difference. (Increase in odds and probability). *This coefficient is statistically differentiable from zero at the α-level = 0.05.*

- **OIL, 0.820**: On average and holding all the other predictors constant, a shift from not producing oil to producing oil for a country is associated with a change in odds of observing a next/higher category of the GDP difference by a multiplicative factor of 0.820. (Decrease in odds and probability).

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.

Loading the data:

```
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
```

Constructing a Poisson model:

```
mexico_poisson <- glm(PAN.visits.06 ~ factor(competitive.district) + marginality.06 + factor(PAN.governor.06), data = mexico_elections, family = poisson)
summary(mexico_poisson) #Coefficients represent change in the log count
```

Let's exponentiate the coefficients and add confidence intervals for a more intuitive interpretation to follow:

```
exp_mexico_poisson <-exp(cbind(OR = coef(mexico_poisson), confint(mexico_poisson)))
exp_mexico_poisson
```

Table 5: Poisson Model

	Dependent variable:			
	PAN.visits.06			
	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	−3.810	0.222	−17.156	<2e-16***
factor(competitive.district)1	−0.081	0.171	−0.477	0.634
marginality.06	−2.080	0.117	−17.728	<2e-16***
factor(PAN.governor.06)1	−0.312	0.167	−1.869	0.062
Observations	2,407			
Log Likelihood	−645.606			
Akaike Inf. Crit.	1,299.213			
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Poisson Model (exp.)

	OR	2.5 %	97.5 %
(Intercept)	0.022	0.014	0.034
factor(competitive.district)1	0.922	0.666	1.302
marginality.06	0.125	0.099	0.156
factor(PAN.governor.06)1	0.732	0.523	1.007

The coefficient for `competitive.district` would mean that for a given district, shifting from being a 'safe bet' to a 'swing' district, on average and holding other predictors constant, is associated with a change in the number of PAN presidential candidate visits by a multiplicative factor of 0.922 (which is essentially a decrease). Anyways, this coefficient is **not** statistically differentiable from zero at the α-level = 0.05. Therefore, there is no sufficient evidence to reject the null hypothesis that PAN presidential candidates do not visit swing districts more often. In other words, there is no sufficient evidence to say that the candidates visit the 'swing' districts more often than the 'safe' ones.

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

- `marginality.06`, **0.125**: For a given district, shifting from having an average poverty level (0) to a marginal poverty level (1) is, on average and holding all the other predictors constant, associated with a change in the number of PAN presidential candidate visits by a multiplicative factor of 0.125 (which means a decrease in the number of visits).

- `PAN.governor.06`, **0.732**: For a given district, shifting from not having a PAN-affiliated governor (0) to having one (1) is, on average and holding all the other predictors constant, associated with a change in the number of PAN presidential candidate visits by a multiplicative factor of 0.732 (which means a decrease in the number of visits).

(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).

Just in terms of conducting basic exploratory data analysis, to get an overall idea of the visits frequency, let's calculate the mean number of visits of the winning PAN presidential candidate:

```
mean(mexico_elections$PAN.visits.06)
```

```
[1] 0.09181554
```

Poisson regression equation (general):

$$\ln(\widehat{\texttt{PAN.visits.06}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \texttt{competitive.district}_i + \hat{\beta}_2 \texttt{marginality.06}_i + \hat{\beta}_3 \texttt{PAN.governor.06}_i$$

$$\ln(\widehat{\texttt{PAN.visits.06}}_i) = -3.810 - 0.081 \cdot \texttt{competitive.district}_i - 2.080 \cdot \texttt{marginality.06}_i - 0.312 \cdot \texttt{PAN.governor.06}_i$$

$$\widehat{\texttt{PAN.visits.06}}_i = e^{(-3.810 - 0.081 \cdot \texttt{competitive.district}_i - 2.080 \cdot \texttt{marginality.06}_i - 0.312 \cdot \texttt{PAN.governor.06}_i)}$$

Estimating mean number of visits from the winning PAN presidential candidate:

Given `competitive.district` = 1, `marginality.06 = 0`, `PAN.governor.06` = 1:

$$\widehat{\texttt{PAN.visits.06}}_i = e^{(-3.810 - (0.081 \cdot 1) - (2.080 \cdot 0) - (0.312 \cdot 1)}$$

```
coeff <- mexico_poisson$coeff
lambda <- exp(coeff[1]+coeff[2]+coeff[4])
lambda
```

```
0.01494818
```

Therefore, the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive, had an average poverty level, and a PAN governor, ≈ 0.015.