Problem Set 3

Applied Stats II

Due: March 26, 2023 - Marcus Ó Faoláin 16327268

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 26, 2023. 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.

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

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.

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

(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).

Question 1

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.

0.1 Data Wrangling

a. We start off by importing the libraries necessary to run and interpret unordered multinomial regression models. These libraries include MASS, nnet and ggplot2.

```
# Importing libraries
library(MASS)
library(nnet)
library(ggplot2)
```

b. We then read in the GDP change data.

```
gdpChange <- read.csv("gdpChange.csv")
```

c. We create a new column 'Response' in the dataframe. In this column we will convert the GDPWdiff from numerical values into the categorical values "No change", "Positive" and "Negative". We will do this using indexing. Each value where GDPWdiff is equal to, greater than or less than zero will be converted into "No change", "Positive" and "Negative" respectively.

```
gdpChange$Response[gdpChange$GDPWdiff == 0] <- "No change"
gdpChange$Response[gdpChange$GDPWdiff > 0] <- "Positive"
gdpChange$Response[gdpChange$GDPWdiff < 0] <- "Negative"
```

d. In order to run the unordered multinomial regression, we will need to convert the 'Response' column in the dataframe from a 'character' type into a 'factor' type. As we want an unordered regression, we do not in this case include the argument for `ordered` in this code.

```
gdpChange$Response <- factor(gdpChange$Response,
  levels = c("No change", "Positive", "Negative"),
  labels = c("No change", "Positive","Negative")
  )
```

e. We change the reference category of 'Response' to "No change" by using the relevel function.

```
gdpChange$Response <- relevel(gdpChange$Response, ref = "No change")
```

0.2 Running the Unordered Multinomial Model

f. With this pre-processing complete, we can now run the unordered model. We set 'Response' as our outcome variable and 'REG' and 'OIL' as our independent variables, in an additive model.

When we use a summary call on the multinomial regression, we get the following output:

```
mult.log <- multinom(Response ~ REG + OIL, data = gdpChange)
summary(mult.log)
```

```
> summary(mult.log)
Call:
multinom(formula = Response ~ REG + OIL, data = gdpChange)

Coefficients:
         (Intercept)      REG      OIL
Positive    4.533759 1.769007 4.576321
Negative    3.805370 1.379282 4.783968

Std. Errors:
         (Intercept)       REG      OIL
Positive   0.2692006 0.7670366 6.885097
Negative   0.2706832 0.7686958 6.885366

Residual Deviance: 4678.77
AIC: 4690.77
```

We can see here the estimated coefficients of the regression, namely $REG_p = 1.769007$, $OIL_p = 4.576321$, $REG_n = 1.379282$ and $OIL_p = 4.783968$.

We can also see the estimated cut off points of $\beta_p = 4.533759$ and $\beta_n = 3.805370$.

0.3 Interpretation of coefficients

g. Interpretation of coefficients: multinomial regression models take the following form:

$$ln(\frac{p_{i2}}{p_{i1}}) = \beta_{02} + \beta_{12}X_i \tag{1}$$

$$ln(\frac{p_{i3}}{p_{i1}}) = \beta_{03} + \beta_{13}X_i \tag{2}$$

In these examples, our reference category is p_i. This means that our coefficient in the first equation, β_{12}, is the change in the log odds of going from category 1 to category 2 for a one unit increase in x_i, holding all else constant.

β_{02} refers to the intercept of the equation

Similarly, our coefficient β_{13} is the change in the log odds of going from category 1 to category 3 for a one unit increase in x_i, holding all else constant.

β_{03} refers to the intercept of the second equation.

We can apply these equations to the coefficients from our multinomial model.

$$ln(\frac{p_{iPositive}}{p_{iNoChange}}) = \beta_{0NCtoP} + \beta_{R1}X_R + \beta_{O1}X_O \tag{3}$$

Using our coefficients from the previous table, we can form the following

$$ln(\frac{p_{iPositive}}{p_{iNoChange}}) = 4.533759 + 1.769007X_R + 4.576321X_O \tag{4}$$

We can also apply our unordered regression model equation to the second coefficients

$$ln(\frac{p_{iNegative}}{p_{iNoChange}}) = \beta_{0NCtoN} + \beta_{R2}X_R + \beta_{O2}X_O \tag{5}$$

When we input the coefficients from the table, we can get the following equation:

$$ln(\frac{p_{iNegative}}{p_{iNoChange}}) = 3.805370 + 1.379282X_R + 4.783968X_O \tag{6}$$

0.3.1 Interpretation of the intercepts

We can interpret the intercepts of the regressions as follows:

The log odds that a countries GDP growth will go from 'No change' to 'Positive', when the covariate for `REG` (X_{R1}) is equal to zero and the covariate for `OIL` (X_{O1}) is equal to zero is 4.533759.

In real terms, 4.533759 is the log odds of that a non-democracy for whom oil is not a majority of their exports will experience a change from "No change" to "Positive" GDP growth.

The log odds that a countries GDP growth will go from 'No change' to 'Negative', when the covariate for `REG` (X_{R1}) is equal to zero and the covariate for `OIL` (X_{O1}) is equal to zero is 3.805370.

In real terms, 3.805370 is the log odds of that a non-democracy for whom oil is not a majority of their exports will experience a change from "No change" to "Negative" GDP growth.

0.3.2 Interpretation of `REG` coefficients

We can interpret the first `REG` coefficient β_{R1} as follows: a one unit in X_R is associated with a 1.769007 increase the log odds that GDP growth will change from "No change" to "Positive" that year, holding all other factors, such as oil exports, constant.

In real terms, this means that we would expect a 1.769007 increase in the log odds that GDP growth would change from "No change" to "Positive", when looking at democracies versus non-democracies, holding all other factors, such as oil exports constant.

We can interpret the second `REG` coefficient β_{R2} as follows: for a a one unit increase in X_R, we would expect a 1.379282 increase in the log odds that GDP growth for countries will change from "No change" to "Negative", holding all other factors, such as oil exports, constant.

In real terms, this means that we would expect a 1.379282 increase in the log odds that GDP growth will change from "No change" to "Negative", when looking at democracies versus non-democracies, holding all other factors, such as oil exports, constant.

0.3.3 Interpretation of `OIL` coefficients

We can interpret the first `OIL` coefficient β_{O1} as follows: a one unit increase in X_O is associated with a 4.5476321 increase in the log odds that GDP growth for countries will change from "No change" to "Positive", holding all other factors, such as regime type, constant.

In real terms, this means that we would expect a 4.576321 increase in the log odds that GDP growth will change from "No change" to "Positive", when looking a major oil producers versus non-major oil producers, holding all other factors, such as regime type, constant.

We can interpret the second `OIL` coefficient β_{O2} as follows: a one unit increase in X_O is associated with a 4.783968 increase in the log likelihood that GDP growth will go from "No change" to "Negative", holding all other factors, such as regime type, constant.

In real terms, this means that we would expect a 4.783968 increase in the log likelihood that GDP growth will go from "No change" to "Negative", when looking at a major oil producers versus non-major oil producers, holding all other factors, such as regime type, constant.

0.4 Statistical tests of the coefficients

We can calculate z scores by dividing the coefficients by their standard errors. We can then use these z scores to determine the p-values of each of the coefficients. We can use the following code to do so:

```
z <- summary(mult.log)$coefficients/summary(mult.log)$standard.errors
(p <- (1-pnorm(abs(z), 0, 1))*2)
```

We get the following z scores

	Intercept	REG	OIL
Positive	16.84156	2.306287	0.6646706
Negative	14.05839	1.794314	0.6948023

We get the following p-values

	Intercept	REG	OIL
Positive	0	0.02109459	0.5062612
Negative	0	0.07276308	0.4871792

From these tables, at an 0.05 alpha level of significance ($\alpha = 0.05$), we can see that only one of the coefficients, β_{RP}, has a p-value below 0.05 and is statistically significantly different from zero. In the cases of the other coefficients, there is not enough evidence to reject the null hypothesis that the coefficients are not equal to zero.

1 Question 1

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

1.1 Data Wrangling

Much of the necessary wrangling for the data, such as converting from string to factor, has already been completed. However, we still need to do a bit.

(a) We start by converting from unordered categorical data into ordinal data. We use the code similar to that from the previous section, using the `factor()` function with an added `ordered = TRUE` argument, and with 'Negative' as the lowest category..

```
gdpChange$Response <-factor(gdpChange$Response,
  levels = c("Negative",
             "No change",
             "Positive"),
  labels = c("Negative",
             "No change",
             "Positive"),
  ordered = TRUE)
```

1.2 Running the Model

(b) We can run the model with the following code

```
ord.log <- polr(Response ~ REG + OIL,
             data = gdpChange,
             Hess = TRUE)

summary(ord.log)
```

This gives us the following output:

```
> summary(ord.log)
Call:
polr(formula = Response ~ REG + OIL, data = gdpChange, Hess = TRUE)

Coefficients:
      Value Std. Error t value
REG  0.3985    0.07518   5.300
OIL -0.1987    0.11572  -1.717

Intercepts:
                    Value    Std. Error t value
Negative|No change  -0.7312   0.0476   -15.3597
No change|Positive  -0.7105   0.0475   -14.9554

Residual Deviance: 4687.689
AIC: 4695.689
```

We can see in this regression table the estimated coefficients for the regression, namely `REG` = 0.3985 and `OIL` = -0.1987.

We can also see the cutoff points for the regression, namely `Negative|No change` = -0.7312 and `No change|Positive` = -0.7105.

1.3 Statistical tests of the coefficients

Using the following code we can find p-values for our coefficients:

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

Which gives us the following output:

```
> (ctable <- cbind(ctable, "p value" = p))
                        Value Std. Error    t value                                                      p value
REG                 0.3984834 0.07518479   5.300054 0.0000001157686505843804679586172136189492487545749099808745
OIL                -0.1987177 0.11571713  -1.717271 0.0859296745454580790379850441240705549716949462890625000000
Negative|No change -0.7311784 0.04760375 -15.359680 0.0000000000000000000000000000000000000000000000000000305077
No change|Positive -0.7104851 0.04750680 -14.955440 0.0000000000000000000000000000000000000000000000000143529045
```

As we can see, at an alpha significance level of 0.05 ($\alpha = 0.05$), the intercepts for "Negative" to "No change" and "No change" to "Positive", as well as the coefficient for `REG`, we can reject the null hypothesis that these intercepts and the coefficient are equal to zero (since their p values are below 0.05) and accept the alternative hypothesis that there is a relationship between the `REG` predictor and the ordinal outcome variable "Response"

In the case of the coefficient for `OIL`, at alpha significance level of 0.05 ($\alpha = 0.05$), we cannot reject the null hypothesis that there the coefficient is equal to zero. We therefore cannot conclude that the `OIL` variable has an effect on the outcome variable "Response" or leads to a change in the direction of GDP growth.

1.4 Interpretation of the coefficients

When examine the data structure type of the ordinal response variable in R, we get the following output:

```
str(gdpChange$Response)
```

```
Ord.factor w/ 3 levels "Negative"<"No change"<..:  3 1 3 3 3 3 1 3 1 3 ...
```

This indicates to us that R has ordered the data in such a way that "Negative" corresponds with 1, "No change" corresponds with 2 and "Positive" corresponds with 3.

Using the coefficients and cut off points from the ordinal logistic regression table, we can get the estimated model as follows:

We can calculate the odds ratio for the coefficients, which will allow us to interpret the coefficients, as follow:

```
(ci <- confint(ord.log))
exp(cbind(OR = coef(ord.log), ci))
```

This gives us the following output:

```
> exp(cbind(OR = coef(ord.log), ci))
           OR     2.5 %   97.5 %
REG 1.4895639 1.2861520 1.727083
OIL 0.8197813 0.6545844 1.030656
```

We can interpret these odds ratios in the following way: for those countries that democracies, the odds experiencing non-negative growth ("No change" or "Positive") is 1.49 times that of countries that are not democracies, holding all other variables constant

For those countries that are major oil exporters, the odds of that they will experience non-negative GDP growth ("No change" or Positive) is 19% lower than non-major oil exporters, holding all other factors constant.

2 Question 2

(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.

2.1 Theory

The Poisson regression model is a generalised linear model takes the form

$$ln(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} \tag{7}$$

Where

$$\lambda_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}} \tag{8}$$

such that $\ln(\lambda_i)$ refers to the natural log of the Poisson parameter λ_i

2.2 Model 1

(a) Our first step is to load in the municipal data, which we do as shown below:

```
muniData <- read.csv("MexicoMuniData.csv")
```

(b) We then test a model whereby `PAN.visits.06` is the outcome variable and `competitive.district` is our predictor variable.

```
# Running Poisson model 1
poisson.reg <- glm(PAN.visits.06 ~ competitive.district ,
data = muniData,
family = poisson)

summary(poisson.reg)
```

In calling the summary() function, we get the following output.

```
> summary(poisson.reg)

Call:
glm(formula = PAN.visits.06 ~ competitive.district, family = poisson,
    data = muniData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4575  -0.4220  -0.4220  -0.4220  18.6647

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.2571     0.1491 -15.141   <2e-16 ***
competitive.district  -0.1617     0.1670  -0.968    0.333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1473.9  on 2406  degrees of freedom
Residual deviance: 1473.0  on 2405  degrees of freedom
AIC: 1776.9

Number of Fisher Scoring iterations: 6
```

2.2.1 Interpretation

An initial look at this table appears to indicate that the relationship between the outcome variable `PAN.visits.06` and the predictor variable `competitive.district` is not statistically significant from zero. In other words there is not enough evidence to suggest that there is a relationship between the two variables based on this model.

We can further test this by looking at the test statistic and a p-value.

The test statistic, or z value, for the hypothetical coefficient β_{cd} is -0.968.

The p-value for the coefficient is 0.333.

Assuming our alpha level of significance to be 0.05 ($\alpha = 0.05$), there is not enough evidence to conclude that our coefficient for $beta_{cd}$ is statistically significant from zero.

We can therefore cannot conclude that PAN presidential candidates visit swing districts more.

2.3 Model 2

(a) We run another Poisson regression model with `PAN.visits.06` as our outcome variable and `competitive.district`, `marginality.06` and `PAN.governor.06` as our predictor variables.

```
poisson.reg1 <- glm(PAN.visits.06 ~ competitive.district+
  marginality.06 +
  PAN.governor.06,
  data = muniData,
  family = poisson)

summary(poisson.reg1)
```

The summary call gives us the following output:

```
> summary(poisson.reg1)

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

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.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 .
---
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
```

We can see from the output of this model that there appears to be statistically significant non-zero relationship between the predictor `marginality.06` and our outcome variable `PAN.visits.06`.
Our test statistic, or z-score, for the coefficient is z = -17.728.

Our p-value is less than 2e-16.

This means that at an alpha level of significance of 0.05 ($\alpha = 0.05$), we can reject the null hypothesis that there is a non-zero relationship between the `marginality.06` and our outcome variable `PAN.visits.06`.

$$ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{9}$$

$$ln(\lambda) = -3.81023 - 0.08135 X_1 - 2.08014 X_2 - 0.31158 X_3 \tag{10}$$

As we can see from the table, the coefficient `competitive.district` has a test statistic of -0.477 and a p-value of 0.6336. Assuming an alpha significance level of 0.05 ($\alpha = 0.05$), since the `competitive.district` p-value of 0.6336 is greater than 0.05, there is not enough evidence to conclude that PAN presidential candidates visit swing districts more.

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

We can interpret our `PAN.governor.06` coefficient (β_3) by saying that a one unit increase in the covariate (X) of `PAN.governor.06` is associated with e^{β_3} times more visits by the winning PAN presidential candidates, according to our model.

We can calculate e^{β_3} using the following code:

```
exp(cfs[4])
```

This gives us a result of 0.7322898.

This is how many more visits a district with a PAN governor could expect versus a district versus a district without a PAN governor, holding all other factors and assuming that the coefficient is statistically significantly different from 0.

If we were to use an alpha significance level of 0.1 ($\alpha = 0.1$), we can use this interpretation, since our p value of 0.0617 is less than 0.1.

However, if we use an alpha significance level of 0.05 ($\alpha = 0.05$), we cannot conclude that there is a not a non-zero relationship between the outcome variable and `PAN.governor.06`. That is to say, we would not be able to reject the null hypothesis that $\beta_3 = 0$.

We can interpret our `marginality.06` by saying that for a one unit increase in our predictor variable for `marginality.06`, we would expect an e^{β_2} times more visits according t our model.

We can calculate e^{β_2} with the following code:

```
exp(cfs[2])
```

This gives us a result of 0.9218693.

This is how many more visits a district could expect for a one unit increase in `marginality.06`, according to our model and holding all other factors constant.

It is not fully clear whether a positive marginality score refers to a richer than average or a poorer than average district.

Assuming that a positive marginality score refers to richer than average districts, it would mean that we would expect more visits to richer than average areas by the PAN presidential

candidate, holding other factors constant.

Assuming a positive marginality score refers to poorer than average districts, it means that we would expect more visits to poorer than average districts by the PAN presidential candidate, holding other other factors constant.

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

We can use the following code save the coefficients of the poisson regression.

```
cfs <- coef(poisson.reg1)
```

Index 1 of `cfs` corresponds with the intercept of the regression, index 2 corresponds with the coefficient for `competitive.district`, index 3 corresponds with the coefficient for `marginality.06` and index 4 corresponds with the coefficient for `PAN.governor.06`.

With this in mind, we can form a formula to calculate the estimated number of visits with the indices, by multiplying index 2 by 1 (`competitive.district` = 1), multiplying index 3 by 0 (`marginality.06`) and multiplying index 4 by 1 (`PAN.governor.06`).

We can do this by running the following code:

```
exp(cfs[1] + cfs[2]*1 + cfs[3]*0+ cfs[4]*1)
```

With this we get an expected number visits of 0.01494818.