Problem Set 3

Doireanna Craven

Due: March 28, 2022

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.

   First, wrangle the data to keep only `GDPWdiff`, `REG`, and `OIL`; bin `GDPWdiff` and cast as factors; and remove NaN.

```
# Drop unnecessary rows
gdpChange <- gdpChange[,c("GDPWdiff", "REG", "OIL")]

```

```r
# Just check if there are any zero values for "no change" category:
length(which(gdpChange$GDPWdiff==0))
# There are 16 zero entries for CGDWdiff

# Alter GDPWdiff to indicate "positive", "negative" or "no change".
# Bin GDPWdiff, will assign factors
gdpChange$GDPWdiff <- cut(gdpChange$GDPWdiff,
                          breaks=c(min(gdpChange$GDPWdiff), -1, 0,
                                   max(gdpChange$GDPWdiff)),
                          labels=c("negative","no change","positive"))

# Check for N/A
unique(gdpChange$GDPWdiff)
# There is an N/A - find it:
which(is.na(gdpChange$GDPWdiff))
# Line 2221, remove it:
gdpChange <- gdpChange[-2221,]
```

Now run the multinomial logit regression, with 'no change' as the reference category.

```r
# Set a reference level for the outcome
gdpChange$GDPWdiff <- relevel(gdpChange$GDPWdiff, ref = "no change")

# Run multinomial logit model
mult.log <- multinom(GDPWdiff ~ ., data = gdpChange)
sum_mult.log <- summary(mult.log)

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

# Create tables for 'negative' and 'positive' compared to 'no change'
tab_neg_mult.log <- rbind(sum_mult.log$coefficients[1,],
                          sum_mult.log$standard.errors[1,], z[1,], p[1,])
rownames(tab_neg_mult.log) <- c("Coeff","Std. Errors","z stat","p value")


tab_pos_mult.log <- rbind(sum_mult.log$coefficients[2,],
                          sum_mult.log$standard.errors[2,], z[2,], p[2,])
rownames(tab_pos_mult.log) <- c("Coeff","Std. Errors","z stat","p value")
```

This table shows the estimated coefficients for the probability of showing a NEGATIVE change in GDP compared to the baseline category of no change:

	(Intercept)	REG	OIL
Coefficient	3.805	1.380	4.758
Std. Errors	0.271	0.769	6.823
z stat	14.058	1.795	0.697
p value	0	0.073	0.486

This table shows the estimated coefficients for the probability of showing a POSITIVE change in GDP compared to the baseline category of no change:

	(Intercept)	REG	OIL
Coefficient	4.534	1.769	4.557
Std. Errors	0.269	0.767	6.823
z stat	16.842	2.306	0.668
p value	0	0.021	0.504

The logit model can be written as:

$$ln(\frac{P(diff = negative)}{P(diff = nochange)}) = 3.805 + 1.380(REG) + 4.758(OIL)$$

$$ln(\frac{P(diff = positive)}{P(diff = nochange)}) = 4.534 + 1.769(REG) + 4.557(OIL)$$

To interpret the coefficients it can be more useful to use exponentiated coefficients:

```
# Interpret the coefficients:
exp(coef(mult.log))
```

	(Intercept)	REG	OIL
negative	44.934	3.976	116.492
positive	93.118	5.867	95.344

- There is an increase in the baseline odds that a country will see a decrease in GDP when that country is a democracy, by a factor of 3.976.
- There is an increase in the baseline odds that a country will see a decrease in GDP when that country's exports are more than 50% oil, by a factor of 116.492.
- There is an increase in the baseline odds that a country will see an increase in GDP when that country is a democracy, by a factor of 5.867.
- There is an increase in the baseline odds that a country will see an increase in GDP when that country's exports are more than 50% oil, by a factor of 95.344.

Let's look at the mean probabilities within each category of GDP change, first for `REG`:

```
predicted_values2$REG: 0
  no change    negative    positive
0.003631204 0.347012860 0.649355936
-------------------------------------
-----------------------------
predicted_values2$REG: 1
   no change     negative     positive
0.0006956138 0.2658060611 0.7334983250
```

For a non-democratic country, the probability of an increase in GDP is 0.65, the probability of a decrease in GDP is 0.35, with 'no change' negligible.
For a democratic country, the probability of an increase in GDP is 0.73, the probability of a decrease in GDP is 0.27, with 'no change' negligible.

And now for `OIL`:

```
predicted_values2$OIL: 0
  no change    negative    positive
0.004284532 0.284628012 0.711087456
-------------------------------------
-----------------------------
predicted_values2$OIL: 1
   no change     negative     positive
4.228551e-05 3.281909e-01 6.717668e-01
```

For a non-major oil exporting country, the probability of an increase in GDP is 0.71, the probability of a decrease in GDP is 0.28, with 'no change' negligible.

For a major oil exporting country, the probability of an increase in GDP is 0.67, the probability of a decrease in GDP is 0.33, with 'no change' negligible.

We can calcuate cut-points using fitted values.

```
pp <- data.frame(fitted(mult.log))
head(data.frame(GDPWdiff = gdpChange$GDPWdiff,
                no.change = pp$no.change,
                negative = pp$negative,
                positive = pp$positive))
```

This returns cut-points of 37.08% for negative—no change, and 37.09% for no change—positive.

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

   The ordered proportional-odds logistic regression was run using `polr`; then p-values and confidence intervals were calculated. An exponentiated table was finally produced.

```
# Construct an ordered multilogit
# First, re-level factors
gdpChange$GDPWdiff <- relevel(gdpChange$GDPWdiff, "negative")

# Run ordered (proportional odds) logistic regression
ord.log <- polr(GDPWdiff ~ ., data = gdpChange, Hess = TRUE)
(sum_ord.log <- summary(ord.log))

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

# Calculate confidence intervals
(ci <- confint(ord.log))

# Convert to odds ratio
(final_table <- exp(cbind(OR = coef(ord.log), ci)))
```

	OR	2.5 %	97.5 %
REG	1.488	1.285	1.726
OIL	0.826	0.659	1.039

A unit change in REG increases the odds of GDP change achieving the leap into a higher category by a factor of $e^{0.3977}$=1.488, with a 95% confidence interval of [1.285, 1.726].
A unit change in OIL increases the odds of GDP change achieving the leap into a higher category by a factor of $e^{-0.1914}$=0.826, with a 95% confidence interval of [0.659, 1.039].

Predicted probabilities for each category can be calculated using fake data.

```
# Generate data for prediction purposes:
predict_data2 <- data.frame(REG = rep(c(0,1), each = 2), OIL = rep(c(0,1)
    , 2))
```

```
# Predict values
plot_data <- melt(cbind(predict_data2, predict(ord.log, predict_data2,
                type = "probs")), id.vars = c("REG", "OIL"),
                variable.name = "Level" , value.name = "Probability")
```

	REG	OIL	variable	value
1	0	0	negative	0.324873191
2	0	1	negative	0.368171106
3	1	0	negative	0.244306800
4	1	1	negative	0.281341737
5	0	0	no change	0.004557724
6	0	1	no change	0.004829558
7	1	0	no change	0.003842886
8	1	1	no change	0.004205314
9	0	0	positive	0.670569086
10	0	1	positive	0.626999336
11	1	0	positive	0.751850314
12	1	1	positive	0.714452949

The cut points reflect the predicted cumulative probabilities at covariate values of zero. These are returned by the `polr` function.

```
(cut_points <- exp(ord.log$zeta))
```

negative|	no change	no change|	positive
0.481		0.491	

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.

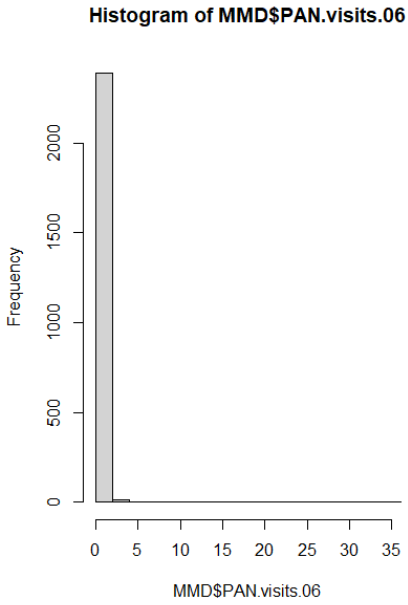
First look at a histogram to check if a zero-inflated model is required:

```
MMD <- read_csv("./MexicoMuniData.csv")

# Remove unnecessary columns
MMD <- MMD[,c("PAN.visits.06", "competitive.district", "marginality.06",
              "PAN.governor.06")]

# Check for N/A
which(is.na(MMD))
# There are none

# Run Poisson regression
# First check if we need a zero-inflated model.
hist(MMD$PAN.visits.06)
```



A zero-inflated model should be used here. We can confirm by running a dispersion test.

```
MMD_poisson <- glm(MMD$PAN.visits.06 ~ ., data = MMD, family = poisson)
summary(MMD_poisson)

dt <- dispersiontest(MMD_poisson)
dt
```

This returns a dispersion of 2.098, which is significanty more than 1, which confirms that a zero-inflated model would be appropriate here.

```
mod.zip <- zeroinfl(PAN.visits.06 ~ ., data = MMD, dist = "poisson")
summary(mod.zip)
```

	Count	Zip
(Intercept)	-1.914	1.272
competitive.district	0.402	0.900
marginality.06	-1.240	0.872
PAN.governor.06	-0.470	-0.175

The ZIP model pertains to the case where the count is zero, of the form binomial with logit link. The baseline odds of a district not having a PAN candidate visit is $e^{1.2719}$ = 3.568. Probability = odds/(1 + odds) = 3.568/(1 + 3.568) = 0.781. The odds of a visit in this case is increased for a swing seat by a factor of $e^{0.900}$ =2.460, when other variables are held constant.
The count model pertains to those cases which are non-zero, of the form poisson with log link. For those districts having a PAN visit, the baseline number of visits is $e^{-1.915}$ = 0.147. For a swing seat, this baseline number is multipied by $e^{0.4024}$ = 1.495, when other variables are held constant.
For both ZIP and count models the odds of a PAN visit are increased for swing seats.

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

The baseline odds of not receiving a visit = 3.568 are multiplied by $e^{0.872}$ = 2.391 for each unit increase in the `marginality.06` measure of poverty, when other variables are held constant.

For those districts having a PAN visit, the baseline number of visits = 0.147 is multiplied by $e^{-1.240}$ = 0.289, for each unit increase in the `marginality.06` measure of poverty, when other variables are held constant.

The baseline odds of not receiving a visit = 3.568 are multiplied by $e^{-0.175}$ = 0.840 where the district has a PAN affiliated governor, when other variables are held constant.

For those districts having a PAN visit, the baseline number of visits = 0.147 is multiplied by $e^{-0.470}$ = 0.625 where the district has a PAN affiliated governor, when other variables are held 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).

```
data <- data.frame(PAN.visits.06 = 1,
                   competitive.district = 1,
                   marginality.06 = 0,
                   PAN.governor.06 = 1)

predict(mod.zip, newdata = data, type = "zero", se = TRUE)
predict(mod.zip, newdata = data, type = "count", se = TRUE)
```

The model predicts that the district will have a 0.880 probability of not receiving a visit.
In the event that there is a visit, the expected count of visits will be 0.138.