Problem Set 3

Applied Stats II

Due: March 26, 2023

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.

Question 1a

```
data <- gdpChange
###make character first
ndaf <-as.character(data)
ndaf <- factor(ifelse(data$GDPWdiff < 0, "negative",
                      ifelse(data$GDPWdiff == 0, "neutral", "positive
    ")),
               levels = c("negative", "neutral", "positive"))
##replacing old measure
data$GDPWdiff <- ndaf
#########################ordered
model1 <- multinom(GDPWdiff ~ REG + OIL, data)
summary(model1)
#intereptation
exp(coef(model1))
# get p values
z <- summary(model1)$coefficients/summary(model1)$standard.errors
(p <- (1 - pnorm(abs(z), 0, 1)) * 2)
```

```
summary of model
Call:
multinom(formula = GDPWdiff ~ REG + OIL, data = data)

Coefficients:
(Intercept)        REG         OIL
neutral    -3.8011902 -1.351703 -7.9240683
positive    0.7284081  0.389905 -0.2076511

Std. Errors:
(Intercept)        REG         OIL
neutral    0.27014596 0.75825317 32.9772055
positive   0.04789662 0.07552484  0.1158094

Residual Deviance: 4678.728
AIC: 4690.728

Exp coeff of Model
          (Intercept)       REG          OIL
neutral    0.02234416 0.2587991 0.0003619269
positive   2.07177984 1.4768404 0.8124904479

P-values
          (Intercept)          REG         OIL
neutral             0 7.464265e-02 0.81010602
positive            0 2.435359e-07 0.07296613
```

- The p-values indicate that the estimated coefficients for REG and OIL are statistically significant for predicting the response variable (GDPWdiff) at a 5 percent significance level.

  The results suggest that non-democratic countries are more likely to experience a positive GDPWdiff compared to democratic countries, and countries with a lower ratio of fuel exports to total exports are more likely to experience a positive GDPWdiff compared to countries with a higher ratio.

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

Question 1b

```r
data$GDPWdiff <- factor(data$GDPWdiff, ordered = TRUE,
                        levels = c("negative", "no change", "positive"))

model_ordered <- polr(GDPWdiff ~ REG + OIL, data = data, Hess = TRUE)

summary(model_ordered)

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

# Calculate confidence intervals
(ci <- confint(model_ordered))

# convert to odds ratio
```

```
Summary
Call:
polr(formula = GDPWdiff ~ REG + OIL, data = data, Hess = TRUE)

Coefficients:
Value Std. Error t value
REG   0.3901     0.07553    5.165
OIL  -0.2080     0.11581   -1.796

Intercepts:
Value     Std. Error t value
negative|no change   -0.7285    0.0479    -15.2097
no change|positive   -0.7285    0.0479    -15.2091

Residual Deviance: 4482.588
AIC: 4490.588
(16 observations deleted due to missingness)
P value
                        Value Std. Error     t value       p value       p
  value
```

```
REG                  0.3901136  0.07552803    5.165150  2.402462e-07
  2.402462e-07
OIL                 -0.2080302  0.11580601   -1.796368  7.243602e-02
  7.243602e-02
negative|no change  -0.7285030  0.04789736  -15.209670  3.050777e-52
  3.050777e-52
no change|positive  -0.7284712  0.04789721  -15.209054  3.079625e-52
  3.079625e-52
Odds ratio
            OR    2.5 %    97.5 %
REG 1.4771486 1.273332 1.714715
OIL 0.8121825 0.647811 1.023041
```

- he results show that for REG, the odds ratio is 1.477 with a 95 percent confidence interval of 1.273 to 1.715, which means that for every one unit increase in REG, the odds of being in a higher GDPWdiff category (positive or negative) are 1.477 times higher, with a range of 1.273 to 1.715.

  For OIL, the odds ratio is 0.812 with a 95 percent confidence interval of 0.648 to 1.023. This means that for every one unit increase in OIL, the odds of being in a higher GDPWdiff category are 0.812 times lower, with a range of 0.648 to 1.023. Note that the confidence interval for OIL includes 1, which suggests that there is not enough evidence to conclude that OIL has a significant effect on GDPWdiff.

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.

lstinputlisting

```r
glimpse(dat)

model <- glm(PAN.visits.06 ~ competitive.district + marginality.06 +
    PAN.governor.06, data = dat, family = poisson)
summary(model)
#interpreting outputs
```

```r
cfs <- coef(model)

# calculate pseudo R squared
1 - (model$deviance/model$null.deviance)

# calculate RMSE
sqrt(mean((model$model$PAN.visits.06 - model$fitted.values)^2))
```

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

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 .
---

(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 output shows that all the independent variables are statistically significant at the 5 percent level, except for "competitive.district". The model also reports the null and residual deviances, and the AIC (Akaike Information Criterion) value, which is used to compare the relative goodness-of-fit of different models.

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

```
Model$coeffecients
   (Intercept) competitive.district       marginality.06
PAN.governor.06
     -3.81023498          -0.08135181          -2.08014361
-0.31157887

```

(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).

lstinputlisting

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

mean_visits <- exp(predict(model, newdata, type = "response"))
mean_visits
```

```
       1
1.01506

```