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 for which 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.

   Wrangling and preparing the data:

```
    # Loading the data
    gdpChange <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2023/main/datasets/gdpChange.csv")

    # Producing three factor variables: Decrease, Increase, and No Change
    gdpChange$GDPWdiff <- as.factor(ifelse(gdpChange$GDPWdiff < 0, "Decrease",
     ifelse(gdpChange$GDPWdiff > 0, "Increase", "No Change")))

```

   Performing the unordered multinomial logit

```
    # Setting the reference level for the outcome
    gdpChange$GDPWdiff <- relevel(gdpChange$GDPWdiff, ref = "No Change")

    multinomial.GDPWdiff <- multinom(GDPWdiff ~ REG + OIL, data = gdpChange)

    summary(multinomial.GDPWdiff)
    exp(coef(multinomial.GDPWdiff))

```

   Calling summary:

```
    summary(multinomial.GDPWdiff)


    Call:
    multinom(formula = GDPWdiff ~ REG + OIL, data = gdpChange)

    Coefficients:
    (Intercept)       REG       OIL
    Decrease    3.805370 1.379282 4.783968
    Increase    4.533759 1.769007 4.576321

    Std. Errors:
    (Intercept)       REG       OIL
    Decrease   0.2706832 0.7686958 6.885366
    Increase   0.2692006 0.7670366 6.885097

```

```
Residual Deviance: 4678.77
AIC: 4690.77

```

Exponentiating the coefficients to get the odds ratios:

```
exp(coef(multinomial.GDPWdiff))

(Intercept)        REG         OIL
Decrease     44.94186  3.972047  119.57794
Increase     93.10789  5.865024   97.15632

```

Estimating the cutoff points:

```
# Finding the cut points for the unordered model
polr(GDPWdiff ~ REG + OIL, data = gdpChange)

Call:
polr(formula = GDPWdiff ~ REG + OIL, data = gdpChange)

Coefficients:
REG          OIL
0.3984834  -0.1987177

Intercepts:
Decrease|No Change No Change|Increase
-0.7311784           -0.7104851

Residual Deviance: 4687.689
AIC: 4695.689

```

Interpreting the unordered multinomial logit: In ceteris paribus, each one-unit increase in `OIL` increases the odds of a decrease in GDP 119.58 times relative to no change. In ceteris paribus, each one-unit increase in `OIL` increases the odds of an increase in GDP 97.16 times relative to no change. In ceteris paribus, each one-unit increase in `REG` increases the odds of a decrease in GDP 3.97 times relative to no change. In ceteris paribus, each one-unit increase in `REG` increases the odds of an increase in GDP 5.87 times relative to no change.

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

   Performing the ordered multinomial logit:

```
# Converting the response variable (GDPWdiff) from and unordered
factor variable to an ordered factor variable
```

```
    gdpChange$GDPWdiff <- factor(gdpChange$GDPWdiff, levels = c("
  Decrease", "No Change", "Increase"), ordered = TRUE)

    # Constructing the ordered/proportional odds logistic model
    ordered.GDPWdiff <- polr(GDPWdiff ~ REG + OIL, data = gdpChange,
  Hess = TRUE)

```

Calling summary:

```
    summary(ordered.GDPWdiff)

    Call:
    polr(formula = GDPWdiff ~ 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
    Decrease|No Change   -0.7312    0.0476    -15.3597
    No Change|Increase   -0.7105    0.0475    -14.9554

    Residual Deviance: 4687.689
    AIC: 4695.689

```

Exponentiating the coefficients to get the odds ratios:

```
    exp(coef(ordered.GDPWdiff))

    REG        OIL
    1.4895639 0.8197813

```

Interpreting the ordered multinomial logit: Holding `OIL` constant,a one-unit increase in `REG` increases the odds of GDP increasing by 1.49. Holding `REG` constant for each one-unit increase in `OIL` decreases the odds of GDP increasing by 0.82.

The cutpoints for this (both unordered and ordered) multinomial model are -0.7312 and -0.7105. This makes sense considering only 16 of datapoints in `GDPWdiff` equate to no change.

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, and running the Poisson regression model

```
# Loading the data
MexicoMuni <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2023/main/datasets/MexicoMuniData.csv")

# Running the Poisson model
poisson_Mexico <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = MexicoMuni, family = "poisson")

```

Calling summary:

```
summary(poisson_Mexico)


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

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.district  -0.08135      0.17069   -0.477
0.6336
```

```
marginality.06         -2.08014     0.11734  -17.728 <0.0000000000000002 ***
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
```

Calling the coefficients:

```
coef(poisson_Mexico)

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

Conducting a Chi-Squared goodness of fit test to determine whether the model fits the data:

```
pchisq(991.25, 2403, lower.tail=FALSE)

[1] 1
```

Running a dispersion test to determine if I need to do a zero-inflated Poisson model:

```
dispersiontest(poisson_Mexico)


Overdispersion test

data:  poisson_Mexico
z = 1.0668, p-value = 0.143
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
2.09834
```

The result of the Chi-Squared goodness of fit test is 1; this indicates the model is a perfect fit for the data.

The overdispersion test's p-value of 0.143 indicates we can't reject the null hypothesis - i.e., we can't refute that the true dispersion isn't greater than 1.

Evidence for whether PAN presidential candidates visit swing districts more: Evidence is insufficient to suggest PAN presidential candidates visit swing districts more (or less for that matter). The log odds for `competitive.district` is -0.08135, and its odds rations (the exponentiated coefficient) is 0.92 - this suggest that competitive districts lead to 8% fewer visits by PAN presidential candidates. `competitive.district` has a test statistics (z-value/score) of -0.477, meaning it's less than half a standard deviation below the sample's mean, and a p-value of 0.6336. These aren't statistically significant, thus we can't reject the null hypothesis stating the predictor variable `competitive.district` and the response variable `PAN.visits.06` aren't significantly associated i.e., for now, we assume they aren't significantly associated.

Interestingly, when I ran an interactive model, the p-value for the `competitive.district` and `marginality.06` (0.01243) was statistically significant at the 99% level, with an odds ratio of 0.48 suggesting the interactive effect between a district being marginal and its poverty rate decreases PAN presidential candidates visiting a district by 52%. However, the three other interactions weren't statistically significant. Thus, we can reject the null hypothesis stating the interaction between predictor variables `competitive.district` and `marginality.06` isn't significantly associated with the response variable `PAN.visits.06` - we can say there's an interactive effect between poverty levels and swing districts on how often PAN presidential visit a district.

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

The log odds for `marginality.06` is -2.08, and the odds ratio (the exponentiated coefficient) is 0.125. This indicates a district's poverty rate led to 82.5% fewer vists by a winning PAN presidential candiate. This difference is statistically signifcant at the 99.99% level (p-value = 0.0000000000000002), allowing us to reject the null hypothesis stating the predictor variable `marginality.06` and the response variable `PAN.visits.06` aren't significantly associated i.e., for now, we assume they are significantly associated.

The log odds for `PAN.governor.06` is -0.31158, and the odds ratio (the exponentiated coefficient) is 0.73. This indicates a district with a PAN-affiliated governor led to 27% fewer visits by a winnin PAN presidential candidate. However, this difference isn't statistically significant (p-value = 0.0617). Thus, we can't reject the null hypothesis asserting the predictor variable `PAN.governor.06` and the response variable `PAN.visits.06` aren't significantly associated - i.e., we assume for now they're not associated significantly.

(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).

```
Mexico.coefs <- coef(poisson_Mexico)

exp(Mexico.coefs[1] + Mexico.coefs[2]*1 + Mexico.coefs[3]*0 +
Mexico.coefs[4]*1)

(Intercept)
0.01494818


# Creating a data frame to check how robust my initial calculation
was
means.Mexico <- data.frame(competitive.district = 1,
marginality.06 = 0,
PAN.governor.06 = 1)

mean_PAN.visits <- predict(poisson_Mexico, means.Mexico, type = "
response")

mean_PAN.visits

1
0.01494818
```

The average number of visits by a winning PAN presidential candidate for a hypothetical competitive district with an average poverty rate of 0 and a PAN-affiliated governor was 0.015.