Problem Set 3

Applied Stats II

Zhuo Zhang/23346227

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.

```r
# load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)
# do some wrangling
gdp_data$REG <- factor(gdp_data$REG,
                                levels = c(0,1),
                                labels = c("Non-Democracy", "Democracy"))
gdp_data$OIL <- factor(gdp_data$OIL,
                       levels = c(0,1),
                       labels = c("Below50%","Beyond50%"))


gdp_data$GDPWdiff <- factor(
  ifelse(gdp_data$GDPWdiff > 0, "positive",
         ifelse(gdp_data$GDPWdiff == 0, "no change", "negative")),
  levels = c( "negative","positive", "no change"),
  labels = c( "negative","positive", "no change")
)

# Set "no change" as the reference category
gdp_data$GDPWdiff <- relevel(gdp_data$GDPWdiff, ref = "no change")
# Set "Democracy" and "Beyond50%" as the reference categories
gdp_data$REG <- relevel(gdp_data$REG, ref = "Non-Democracy")
gdp_data$OIL <- relevel(gdp_data$OIL, ref = "Below50%")

# Constructing an unordered multiple logistic regression model
model1 <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)

# View Model Summary
summary(model1)
```

Result:

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

Coefficients:
         (Intercept) REGDemocracy OILBeyond50%
negative    3.805370     1.379282     4.783968
positive    4.533759     1.769007     4.576321

Std. Errors:
         (Intercept) REGDemocracy OILBeyond50%
negative   0.2706832    0.7686958     6.885366
positive   0.2692006    0.7670366     6.885097

Residual Deviance: 4678.77
AIC: 4690.77
```

Intercept coefficient: For GDPWdiff, the reference category is "negative" and the intercept coefficient is 3.805370. This indicates that, with other predictor variables remaining unchanged, the log adds of GDPWdiff in the negative category are approximately 3.805370 relative to the reference category.

REGDemocracy coefficient: For the case where the REG variable belongs to the "Democracy" category, the REGDemocracy coefficient is 1.379282 relative to the "Non Democracy" category of the REG variable. This indicates that, while keeping other predictor variables constant, the logarithmic probability of GDPWdiff in the "positive" category is approximately 1.379282 compared to the reference category when REG is "Democracy".

OILBeyond50% coefficient: For cases where the OIL variable is in the "Beyond 50%" category, the coefficient for OILBeyond50% is 4.783968 relative to the "Below 50%" category of the OIL variable. This indicates that, while keeping other predictor variables constant, the logarithmic probability of GDPWdiff in the negative category is approximately 4.783968 compared to the reference category when OIL is "Beyond 50%".
The standard error provides information about the accuracy of coefficient estimation. A smaller standard error indicates a more accurate estimation.
The residual deviation measurement model measures the overall goodness of fit of the data. A smaller residual deviation value indicates a better fit with the observed data. AIC is a measure of the goodness of fit of a model, adjusted based on the number of predictive factors. The smaller the AIC value, the more suitable the model is.

```r
# Extract estimated values of intercept and coefficients
intercepts <- coef(model1)
# View estimated values of intercept and coefficients
print("Intercepts:")
print(intercepts)
```

Result:

```
         (Intercept) REGDemocracy OILBeyond50%
negative    3.805370     1.379282     4.783968
positive    4.533759     1.769007     4.576321
```

For negative level GDPWdiff, the Intercept at the cut-off point is 3.81, and the coefficient of REGDemography is 1.379282
The coefficient of OILBeyond50% is 4.783968 For the positive level of GDPWdiff, the Intercept at the cut-off point is 4.533759, the coefficient for REGDemocracy is 1.769007, and the coefficient for OILBeyond50% is 4.576321

```r
#exponentiate coefficients
exp(coef(model1)[,c(1:3)])
```

Result:

```
          (Intercept) REGDemocracy OILBeyond50%
negative     44.94186     3.972047    119.57794
positive     93.10789     5.865024     97.15632
```

1.Intercept (Negative to Positive): The exponentiated intercept of 44.94 indicates that when all other predictors are held constant, the odds of transitioning from the negative level to the positive level of GDPWdiff are approximately 44.94 times higher than remaining in the negative level, for observations with the reference categories of REG (Non-Democracy) and OIL (Below50%).

2.REGDemocracy (Negative to Positive): The exponentiated coefficient of 3.97 for REGDemocracy suggests that, holding OIL constant, the odds of transitioning from the negative level to the positive level of GDPWdiff are approximately 3.97 times higher for observations with a regime categorized as Democracy compared to Non-Democracy.

3.OILBeyond50% (Negative to Positive): The exponentiated coefficient of 119.58 for OILBeyond50% implies that, when controlling for REG, the odds of transitioning from the negative level to the positive level of GDPWdiff are approximately 119.58 times higher for observations with oil production categorized as Beyond50% compared to Below50%.

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

```r
# Encode GDPWdiff as an ordered factor variable GDPWdiff
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff,
                            levels = c("no change", "negative","positive"),
                            labels = c("no change", "negative","positive"),
                            ordered = TRUE
)

# Constructing an ordered multiple logistic regression model
model2 <- polr(GDPWdiff ~ REG + OIL, data = gdp_data, Hess=T)

# View Model Summary
summary(model2)
```

Result:

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

Coefficients:
               Value Std. Error t value
REGDemocracy  0.4102    0.07518   5.456
OILBeyond50% -0.1788    0.11546  -1.549

Intercepts:
                    Value    Std. Error t value
no change|negative  -5.3199   0.2523   -21.0865
negative|positive   -0.7036   0.0476   -14.7932

Residual Deviance: 4686.606
AIC: 4694.606
```

In the results of this ordered multiple logistic regression model, the REGDemocracy coefficient is 0.4102, the standard error is 0.07518, and the t-value is 5.456. This indicates that when the REG variable changes from "Non Democracy" to "Democracy", the logarithmic probability of transitioning from lower GDPWdiff categories to higher GDPWdiff categories increases by 0.4102 units. A t-value of 5.456 indicates that this change is significant, as its absolute value is much greater than 1.96 (at a 95% confidence level).

The OILBeyond 50% coefficient is -0.1788, the standard error is 0.11546, and the t-value is -1.549. When the OIL variable changed from "Below50%" to "Beyond 50%", the log adds for transitioning from higher GDPWdiff categories to lower GDPWdiff categories decreased by 0.1788 units. Although the t-value is -1.549, it is not significant as its absolute value is less than 1.96.

The intercept represents the log adds for converting from one GDPWdiff class to another. The intercept from no change to negative category is -5.3199, which means that the log addresses for converting from no change to negative category are -5.3199. The intercept from negative to positive category is -0.7036, indicating that the log adds for transitioning from negative to positive category are -0.7036.

```r
# Extract estimated values of intercept and coefficients
intercepts <- coef(model2)
# View estimated values of intercept and coefficients
print("Intercepts:")
print(intercepts)
```

Result:

```
REGDemocracy OILBeyond50%
   0.4101691   -0.1788330
```

The estimated value of the REGDemocracy coefficient is 0.4101691. This indicates that when the REG variable changes from "Non Democracy" to "Democracy", the log additions for transitioning from lower GDPWdiff categories to higher GDPWdiff categories

increase by 0.4101691 units. The estimated value of the OILBeyond 50% coefficient is -0.1788330. This indicates that when the OIL variable changes from "Below50%" to "Beyond 50%", the log adds for transitioning from higher GDPWdiff categories to lower GDPWdiff categories decrease by 0.1788330 units.

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.

```r
# load data
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv")
model3 <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexico_elections, family = "poisson")
summary(model3)
```

Result:

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

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
```

According to the Poisson regression results above, the coefficient of the competitive.

logistic variable is -0.08135, with a p-value of 0.6336. This means that, while keeping other variables constant, the relationship between competitive constituencies and PAN.visits.06 is not significant. Therefore, according to this model, there is no evidence to suggest that PAN presidential candidates are more inclined to visit competitive constituencies. The test statistic is -0.477.

(b) Interpret the `marginality.06` and `PAN.governor.06` coefficients.
Result:
The margin. 06 coefficient is -2.08014, and its p-value is less than 2e-16, which is very significant. This indicates a significant negative correlation between marginalization (poverty level) and PAN.visits.06, while keeping other variables constant. The coefficient of PAN.governor.06 is -0.31158, with a p-value of 0.0617, which is close to the significance level of 0.05. This suggests that the relationship between the coefficient of whether the governor is a PAN party and PAN. visits.06 may be significant, while other variables remain unchanged, but further confirmation is needed.

(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).

```
hypothetical_district <- data.frame(competitive.district = 1, marginality
    .06 = 0, PAN.governor.06 = 1)
estimated_mean_visits <- predict(model3, newdata = hypothetical_district,
     type = "response")
estimated_mean_visits
```

Result:

```
          1
0.01494818
```

For hypothetical regions with competitive constituencies (competitive. district=1), average poverty level (marginal. 06=0), and PAN party governors (PAN. governor. 06=1), the average predicted number of visits by PAN presidential candidates is 0.0149.