Problem Set 3

Applied Stats II

Due: March 24, 2024

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.

   Steps:
   1. Data Check
   2. Handling Missing Values
   3. Model Construction
   4. Output Model Results

```r
# load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/gdpChange.csv", stringsAsFactors = F)

# Import necessary packages
library(dplyr)

# Check the data
# View the structure and first few rows
head(gdp_data)
```

```
X COUNTRY CTYNAME YEAR GDPW OIL REG   EDT GDPWlag GDPWdiff GDPWdifflag GDPWdifflag2
1       1 Algeria 1965 6620   1   0  1.45    6502      118         419         1071
2       1 Algeria 1966 6612   1   0  1.56    6620       -8         118          419
3       1 Algeria 1967 6982   1   0 1.675    6612      370          -8          118
4       1 Algeria 1968 7848   1   0 1.805    6982      866         370           -8
5       1 Algeria 1969 8378   1   0  1.95    7848      530         866          370
6       1 Algeria 1970 8536   1   0   2.1    8378      158         530          866
```

```r
# Use the str function to view the data structure
str(gdp_data)
```

```
  'data.frame': 3721 obs. of  12 variables:
 $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ COUNTRY     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ CTYNAME     : chr  "Algeria" "Algeria" "Algeria" "Algeria" ...
 $ YEA-R        : int  1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 ...
 $ GDPW        : int  6620 6612 6982 7848 8378 8536 7816 9372 9361 10480 ...
 $ OIL         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ REG         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ EDT         : chr  "1.45" "1.56" "1.675" "1.805" ...
 $ GDPWlag     : int  6502 6620 6612 6982 7848 8378 8536 7816 9372 9361 ...
 $ GDPWdiff    : Ord.factor w/ 3 levels "negative"<"no change"<..: 2 1 3 3 3 3 1 3
 $ GDPWdifflag : int  419 118 -8 370 866 530 158 -720 1556 -11 ...
 $ GDPWdifflag2: int  1071 419 118 -8 370 866 530 158 -720 1556 ...
```

```r
# Use the summary function to view the data summary
summary(gdp_data)
```

```
#   X             COUNTRY          CTYNAME                YEAR            GDPW
#   Min.   :   1   Min.   :  1.00   Length:3721         Min.   :1954   Min.   :  509
#   1st Qu.: 931   1st Qu.: 39.00   Class :character   1st Qu.:1967   1st Qu.: 2566
#   Median :1861   Median : 71.00   Mode  :character   Median :1976   Median : 6425
#   Mean   :1861   Mean   : 70.42                      Mean   :1975   Mean   : 9276
#   3rd Qu.:2791   3rd Qu.:103.00                      3rd Qu.:1983   3rd Qu.:13470
#   Max.   :3721   Max.   :135.00                      Max.   :1990   Max.   :37903
#   OIL               REG               EDT                GDPWlag          GDPWdiff
#   Min.   :0.0000   Min.   :0.0000   Length:3721         Min.   :  509   negative :1
#   1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Qu.: 2533   no change:
#   Median :0.0000   Median :0.0000   Mode  :character   Median : 6245   positive :2
#   Mean   :0.1005   Mean   :0.4015                      Mean   : 9090
#   3rd Qu.:0.0000   3rd Qu.:1.0000                      3rd Qu.:13167
#   Max.   :1.0000   Max.   :1.0000                      Max.   :37089
#   GDPWdifflag       GDPWdifflag2
#   Min.   :-9257.0   Min.   :-9257.0
#   1st Qu.:  -20.0   1st Qu.:  -19.0
#   Median :  117.0   Median :  116.0
#   Mean   :  189.7   Mean   :  189.9
#   3rd Qu.:  415.0   3rd Qu.:  405.0
#   Max.   : 7867.0   Max.   : 7867.0
```

```r
# Check the number of observations in each category
table(gdp_data$GDPWdiff)
```

```
#   negative no change  positive
#   1177         1      2543
```

```r
# Check the number of observations in each category
table(gdp_data$GDPWdiff)
#   negative no change  positive
#   1177         1      2543


# Convert ”GDPWdiff” to a factor variable and set ”no change” as the reference level
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, levels = c(”negative”, ”no change”, ”positive”))
unique(gdp_data$GDPWdiff)
#  [1] no change negative  positive
#  Levels: negative < no change < positive


# Recalculate missing values and convert them to factor variables
# Converts missing values to the ”no change” category and ensures the horizontal order of factor variables
gdp_data$GDPWdiff[is.na(gdp_data$GDPWdiff)] <- ”no change”
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, levels = c(”negative”, ”no change”, ”positive”))


# Check the modified unique value
# This output is the same as the previous one, reconfirming the unique value and factor horizontal order in the response variable GDPWdiff.
unique(gdp_data$GDPWdiff)
#  [1] no change negative  positive
#  Levels: negative < no change < positive

# Build unordered multinomial Logit model
# A disordered multinomial Logit model is constructed using multinom function to explain the relationship between response variable GDPWdiff and explanatory variables REG and OIL.
model_unordered <- multinom(GDPWdiff ~ REG + OIL, data = gdp_data)
#  # weights:  12 (6 variable)
#  initial  value 4087.936326
#  iter  10 value 2316.550625
#  iter  20 value 2313.050603
#  final  value 2312.908537
#  converged
```

This output displays the process of building an unordered multinomial logit model using the multinom function. ”Weights: 12 (6 variable)” indicates that the model estimates 12 parameters, including intercept and coefficients of the explanatory variables REG and OIL, totaling 6 variables. ”Initial value” and ”final value” represent the initial and final estimates of the model, ”iter” indicates the number of iterations, and ”converged” indicates whether the model converged. This output provides relevant information about the model fitting process, including initial values, iteration process, and final convergence status.

```
# Output model results
summary(model_unordered)
#  Call:
#  multinom(formula = GDPWdiff ~ REG + OIL, data = gdp_data)

#  Coefficients:
```

```
#    (Intercept)          REG          OIL
#  no change  -12.5609032  -8.3458842   7.8854448
#  positive     0.6324751   0.3907864  -0.1162928

#  Std. Errors:
#    (Intercept)           REG          OIL
#  no change 20.58809041 134.79565119 20.6125497
#  positive   0.04700425   0.07399261   0.1156075

#  Residual Deviance: 4625.817
#  AIC: 4637.817
```

This unordered multinomial Logit model is used to explain the relationship between national economic growth (GDPWdiff) and two explanatory variables (REG and OIL) Model Results

Coefficients:

For the ‘no change‘ category:

Intercept: -12.5609032. This means that when REG and OIL are both 0 (i.e., when REG is unrestricted and OIL is also unrestricted), the expected log probability of the ‘no change‘ category is -12.5609032.

REG coefficient: -8.3458842. This means that when REG increases by one unit, the expected log probability of the ‘no change‘ category decreases by approximately 8.35 units.

OIL coefficient: 7.8854448. This means that when OIL increases by one unit, the expected log probability of the ‘no change‘ category increases by approximately 7.89 units.

For the ‘positive‘ category:

Intercept: 0.6324751. This means that when REG and OIL are both 0, the expected log probability of the ‘positive‘ category is 0.6324751.

REG coefficient: 0.3907864. This means that when REG increases by one unit, the expected log probability of the ‘positive‘ category increases by approximately 0.39 units.

OIL coefficient: -0.1162928. This means that when OIL increases by one unit, the expected log probability of the ‘positive‘ category decreases by approximately 0.12 units.

Residual Deviance: 4625.817. This indicates the degree of unexplained variability in the data.

AIC (Akaike Information Criterion): 4637.817. A smaller AIC value indicates better performance of the model in explaining the data.

Overall, this model tells us that changes in economic growth are associated with the region (REG) and oil production (OIL). Specifically, changes in REG and OIL are related to the probability of economic growth categories ‘no change‘ and ‘positive‘.

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

```r
# Ordered multinomial Logit model
# Converts "GDPWdiff" to an ordered factor, in which "negative" is considered the lowest level and "positive" is considered the highest.
gdp_data$GDPWdiff <- factor(gdp_data$GDPWdiff, levels = c("negative", "no change", "positive"), ordered = TRUE)

# Build an ordered multinomial Logit model where GDPWdiff is the result variable and REG and OIL are the explanatory variables.
model_ordered <- polr(GDPWdiff ~ REG + OIL, data = gdp_data, Hess = TRUE)

# Output model results
summary(model_ordered)
#  Call:
#  polr(formula = GDPWdiff ~ REG + OIL, data = gdp_data, Hess = TRUE)

#  Coefficients:
#     Value Std. Error t values
#  REG  0.3912    0.07399   5.287
#  OIL -0.1192    0.11529  -1.034

#  Intercepts:
#     Value    Std. Error t value
#  negative|no change   -0.6332   0.0470   -13.4694
#  no change|positive   -0.6319   0.0470   -13.4442

#  Residual Deviance: 4630.758
#  AIC: 4638.758
```

This ordered multinomial logistic regression model is used to explain GDPWdiff (GDP change) as the outcome variable.

Intercepts:
negative—no change: This intercept represents the transition from "negative GDP change" to "no change in GDP change". The estimated intercept value is -0.6332. This means the likelihood of transitioning from negative GDP change to no change in

GDP change when the other explanatory variables (REG and OIL) are 0.

no change—positive: This intercept represents the transition from ”no change in GDP change” to ”positive GDP change”. The estimated intercept value is -0.6319. This means the likelihood of transitioning from no change in GDP change to positive GDP change when the other explanatory variables (REG and OIL) are 0.

Coefficients:
REG: The coefficient estimate for REG (government management index) is 0.3912. This indicates that when the value of REG increases by one unit, the log odds of transitioning from a lower GDP change to a higher GDP change increase by 0.3912. The standard error of the coefficient is 0.07399, indicating the uncertainty in the coefficient estimate. In this case, with a t-value of 5.287, the coefficient is statistically significant.

OIL: The coefficient estimate for OIL (oil export index) is -0.1192. This indicates that when the value of OIL increases by one unit, the log odds of transitioning from a lower GDP change to a higher GDP change decrease by 0.1192. The standard error of the coefficient is 0.11529, with a t-value of -1.034, indicating that the coefficient is not statistically significant.

Residual Deviance:
The residual deviance measures the degree of variability unexplained by the model. In this case, the residual deviance is 4630.758.

AIC (Akaike Information Criterion):
AIC is used to measure the goodness of fit and complexity of the model. In this model, the AIC is 4638.758. A lower AIC value indicates a better model fit.

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.

```
# Step 1. Build assumptions
H0: Competitive and non-competitive regions have the same average number of visits.
H1: The average number of visits is different between competitive and non-competitive regions.


#  load data
mexico_elections <- read.csv(”https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2024/main/datasets/MexicoMuniData.csv”)

# View the data structure and the first few lines
head(mexico_elections)
#  MunicipCode pan.vote.09 marginality.06 PAN.governor.06 PAN.visits.06
#  1        1001       0.283         -1.831               0             5
#  2        1002       0.352         -0.620               0             0
#  3        1003       0.359         -0.875               0             0
#  4        1004       0.238         -0.747               0             0
#  5        1005       0.378         -1.234               0             0
#  6        1006       0.145         -1.306               0             0
#  competitive.district
#  1                  1
#  2                  1
#  3                  1
#  4                  1
#  5                  1
#  6                  1

# Step 2. Run the Poisson regression model
model_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06,
                      data = mexico_elections, family = poisson)

# Step 3. Explanatory coefficient
summary(model_poisson)

#  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
```

4. Explaining the Poisson Regression Model Results:

4.1 Model Explanation:
The objective of the model is to explain the number of visits by PAN presidential candidates to regions before the 2009 federal elections by considering the following predictor variables:

‘competitive.district‘: Whether the district is competitive
‘marginality.06‘: Level of poverty
‘PAN.governor.06‘: Whether the state has a governor affiliated with the PAN party

4.2 Coefficient Interpretation:
The coefficient estimate for ‘competitive.district‘ is -0.08135 with a p-value of 0.6336. This suggests that in the presence of other variables, there is no significant association between competitive districts and the number of visits by PAN presidential candidates.
The coefficient estimate for ‘marginality.06‘ is -2.08014, with a p-value very close to 0, indicating a significant negative correlation between areas with higher poverty levels and the number of visits by PAN presidential candidates.
The coefficient estimate for ‘PAN.governor.06‘ is -0.31158 with a p-value of 0.0617, close to the significance level of 0.05. This suggests that in the presence of other variables, having a governor affiliated with the PAN party may be associated with fewer visits by PAN presidential candidates, but this relationship is not very significant.

4.3 Model Fit:
The dispersion of residuals is 1, the Null deviance is 1473.87, the Residual deviance is 991.25, and the AIC is 1299.2. These metrics are used to evaluate the models fit, with smaller residual dispersion and AIC values indicating better model fit.

```
# 5. Predict average visits
# Create virtual data
new_data <- data.frame(competitive.district = 1, marginality.06 = 0, PAN.
    governor.06 = 1)
# Predict average number of visits
predicted_visits <- predict(model_poisson, newdata = new_data, type = ”
    response”)
```

```r

# Output result
predicted_visits
#          1
# 0.01494818

#  According to the model's prediction,
#  for a hypothetical electoral district with the following characteristics:
#  `competitive.district` equals 1 (highly competitive district)
#  `marginality.06` equals 0 (average poverty level)
#  `PAN.governor.06` equals 1 (PAN-affiliated governor)

# The estimated average number of visits by PAN presidential candidates to this district is 0.0149 times. This suggests that, on average, PAN presidential candidates visit this type of district approximately 0.015 times.


# In the above code, we have used the Poisson regression model and obtained the summary information of the model by summary(model_poisson).
# We can extract information about the competitive region coefficient from the model summary and calculate the t-statistic and p-value.
# Extract coefficients and standard errors for competitive regions
coef_competitive <- coef(summary(model_poisson))["competitive.district", "Estimate"]
se_competitive <- coef(summary(model_poisson))["competitive.district", "Std. Error"]

# Calculate the t statistic
t_statistic <- coef_competitive / se_competitive

# Calculate the p-value
p_value <- 2 * pt(-abs(t_statistic), df = model_poisson$df.residual)

# Output test statistics and p-values
cat("t_statistic:", t_statistic, "\n")
# t_statistic: -0.4766106

cat("p_value:", p_value, "\n")
# p_value: 0.6336828
```

Based on the results of the Poisson regression model, we obtained the following test statistics and p-values:

Test Statistic: In this case, the test statistic is the estimated coefficient of the competitive district variable in the Poisson regression model. Specifically, it represents the coefficient estimate for the ‘competitive. district‘ variable, indicating the relationship between competitive districts and the number of visits by PAN presidential candidates, holding other explanatory variables constant. The value is -0.4766106, suggesting a potential decrease in the average visitation frequency in competitive districts.

p-value: In this scenario, the p-value is 0.6336828, indicating whether there is a significant difference in the average visitation frequency between competitive and non-competitive districts. With a p-value greater than the typical significance level (usually 0.05), we fail to reject the null hypothesis, suggesting no significant difference in the average visitation frequency between competitive and non-competitive districts.

Predicted Average Visitation: According to the models predictions, for districts characterized by competitive districts, average poverty levels, and PAN-affiliated governors, the estimated average visitation frequency by PAN presidential candidates is approximately 0.0149 times.

In summary, based on the results of the Poisson regression model and the test statistics, we find that the relationship between competitive districts and the visitation frequency of PAN presidential candidates is not significant in this model (test statistic = -0.4766106, p-value = 0.6336828). Additionally, the predicted average visitation frequency for districts with these characteristics is approximately 0.015 times. Therefore, we cannot conclude that PAN presidential candidates visit competitive districts more frequently.

Thus, based on the results of this Poisson regression model, we fail to reject the null hypothesis that there is no significant difference in the visitation frequency of PAN presidential candidates between swing and non-swing districts.

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

Explain the meaning of the coefficients' marginality.06 'and' PAN.governor.06 'as follows:

1. Coefficient for 'marginality.06':
The coefficient estimate is -2.08014, indicating the effect of a one-unit change in 'marginality.06' on the number of visits by PAN presidential candidates, holding other explanatory variables constant.
Here, 'marginality.06' is a measure of poverty level, and its negative coefficient suggests that an increase in poverty level (i.e., in more impoverished areas) is associated with a decrease in the number of visits by PAN presidential candidates. In other words, areas with higher levels of poverty may experience fewer visits by PAN presidential candidates.

2. Coefficient for 'PAN.governor.06':
The coefficient estimate is -0.31158, representing the effect of having a PAN-affiliated

governor on the number of visits by PAN presidential candidates, holding other explanatory variables constant.
‘PAN.governor.06‘ is a dummy variable, and its negative coefficient suggests a negative correlation between having a PAN-affiliated governor and the number of visits by PAN presidential candidates. In other words, in states with PAN-affiliated governors, the number of visits by PAN presidential candidates may be fewer.

In summary, the interpretation of the coefficients for ‘marginality.06‘ and ‘PAN.governor.06‘ suggests that higher levels of poverty and the presence of a PAN-affiliated governor may lead to a decrease in the number of visits by PAN presidential candidates.

(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).

```r
# To define the characteristics of the hypothetical district:

#  competitive.district is 1 (highly competitive district).
#  marginality.06 is 0 (average poverty level).
#  PAN.governor.06 is 1 (PAN-affiliated governor).


# Create virtual data points
new_data <- data.frame(competitive.district = 1,
                        marginality.06 = 0,
                       PAN.governor.06 = 1)

# Predict average number of visits
predicted_visits <- predict(model_poisson, newdata = new_data, type = "
   response")

# Output result
predicted_visits
#  1
#  0.01494818
```

The result indicates that the hypothetical district with these characteristics (highly competitive, average poverty level, PAN-affiliated governor) is estimated to be visited by the PAN presidential candidate approximately 0.015 (0.01494818) times on average.