Problem Set 2

Applied Stats II

Due: February 18, 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 February 18, 2024. No late assignments will be accepted.

We're interested in what types of international environmental agreements or policies people support (Bechtel and Scheve 2013). So, we asked 8,500 individuals whether they support a given policy, and for each participant, we vary the (1) number of countries that participate in the international agreement and (2) sanctions for not following the agreement.

Load in the data labeled `climateSupport.RData` on GitHub, which contains an observational study of 8,500 observations.

- Response variable:
  - `choice`: 1 if the individual agreed with the policy; 0 if the individual did not support the policy
- Explanatory variables:
  - `countries`: Number of participating countries [20 of 192; 80 of 192; 160 of 192]
  - `sanctions`: Sanctions for missing emission reduction targets [None, 5%, 15%, and 20% of the monthly household costs given 2% GDP growth]

Please answer the following questions:

1. Remember, we are interested in predicting the likelihood of an individual supporting a policy based on the number of countries participating and the possible sanctions for non-compliance.

   Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.
   Interpretation:

   Since the dependent variable is binary, the entire question uses a logistic regression model. And use the Likelihood ratio test to test the validity of the coefficients. Global null hypothesis : all coefficients (coefficients of independent variables) are zero, and there is no significant relationship between independent variables and dependent variable , that is, the independent variables in the model have no predictive power on the dependent variable.

	Model
(Intercept)	-0.27^{***}
	(0.05)
countries80 of 192	0.34^{***}
	(0.05)
countries160 of 192	0.65^{***}
	(0.05)
sanctions5%	0.19^{**}
	(0.06)
sanctions15%	-0.13^{*}
	(0.06)
sanctions20%	-0.30^{***}
	(0.06)
AIC	11580.26
BIC	11622.55
Log Likelihood	-5784.13
Deviance	11568.26
Num. obs.	8500

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Model

```
install.packages(”texreg”)
library(texreg)

install.packages(”xtable”)
library(xtable)
```

```r

# load data
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/climateSupport.RData?raw=true"))
load("/Users/daisy/Downloads/climateSupport.RData")
ls()

# Preview data
summary(climateSupport)  # View summary statistics of the data
head(climateSupport)     # View the first few rows of the data

# Convert data format of DV into binary outcome
climateSupport$choice <- ifelse(climateSupport$choice == "Supported", 1, 0)

# Change data format of IVs to unordered factors
climateSupport$countries <- factor(climateSupport$countries, ordered = FALSE)
climateSupport$sanctions <- factor(climateSupport$sanctions, ordered = FALSE)

# Use the relevel function to set the baseline level and fit logistic regression model
climateSupport$countries <- relevel(climateSupport$countries, ref = "20 of 192")
climateSupport$sanctions <- relevel(climateSupport$sanctions, ref = "None")
model <- glm(choice ~ ., family = binomial(link = "logit"), data = climateSupport)

# Output the results of the model
summary(model)

# Create LaTeX table for the model
texreg(model)

## Likelihood ratio test

# H0: beta_1 = beta_2 = ...= beta_5 = 0
# H1: at least one slope is not equal to 0
#  Create a null model
null_model <- glm(choice ~ 1,
                  data = climateSupport,
                  family = "binomial")

#  Run an anova test on the model compared to the null model
anova_result_1 <- anova(null_model, model, test = "LRT")

# print results
print(anova_result_1)

```

```r
# Create LaTeX table for the model

# Create a data frame containing ANOVA results
result_table_1 <- xtable(anova_result_1)

# Convert to xtable format
print(result_table_1, include.rownames = FALSE)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
8499	11783.41			
8494	11568.26	5	215.15	0.0000

From the table, the difference in degrees of freedom is 5, while the difference in bias is 215.15, and the p-value is extremely small (2.2e-16), indicating that the difference in bias between the two models is significant.This suggests that a model that includes the countries and sanctions variables has an advantage in explaining the data relative to the null model.

2. If any of the explanatory variables are significant in this model, then:

   (a) For the policy in which nearly all countries participate [160 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)
   Interpretation:
   Calculate the logarithm of the odds in two different situations, then make the difference, and then exponentialize the difference to get the change in odds ratio.

```r
# (a)

# Fitted logistic regression model is:
# logit^(p/(1-p)) = beta0 + beta_1*countries_80 + beta_2*countries_160 +
# beta_3*sanctions_5% + beta_4*sanctions_15% + beta_5*sanctions_20%

# After substituting the coefficients, get:
# logit^(p/(1-p)) = -0.27266 + 0.33636*countries_80 + 0.64835*countries_160 +
# 0.19186*sanctions_5% - 0.13325*sanctions_15% - 0.30356*sanctions_20%

# according to: logit^(p_5%/(1-p_5%)) =  -0.27266 + 0.64835*countries_160 + 0.19186*sanctions_5%
log_odd_p_5 <- -0.27266 + 0.64835*1 + 0.19186*1

# According to: logit^(p_15%/(1-p_15%)) =  -0.27266 + 0.64835*countries_160 - 0.13325*sanctions_15%
log_odd_p_15 <- -0.27266 + 0.64835*1 - 0.13325*1
```

```r

# Subtract two log odds, get the difference of log odds
delta = log_p_15 - log_p_5

# Indexed results according to the formula: OR = exp^(delta)
exp(delta)

# Print results:0.7224479
```

For the policy in which nearly all countries participate [160 of 192], increasing sanctions from 5% to 15% will increase the odds that an individual will support the policy by 0.7224479.

(b) What is the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions?

```r
# (b)

# After substituting the coefficients, get:
# logit^(p_80/(1-p_80)) =  -0.27266 + 0.33636*countries_80
log_odd_p_80 <- -0.27266 + 0.33636*1

# According to the formula: P = exp(logit^(P/(1-P))/(1+exp(logit^(P/
    (1-P)))
p_80 = exp(log_odd_p_80) / (1 + exp(log_odd_p_80))

# Print the results
print(p_80)

# Results : 0.5159196
```

Interpretation:
The estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions is about 51.6%.

(c) Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?

- Perform a test to see if including an interaction is appropriate.
  Interpretation:

  Use the chi-square test. First fit a logistic regression model containing interaction terms, then use anova() to compare with the original model, and finally analyze the p value in the anova result to see the significance of the coefficient.

```r
# (c)

# Fit a logistic regression model that includes interaction terms
interaction_model <- glm(choice ~ countries * sanctions, family =
    binomial(link = "logit"), data = climateSupport)
```

```

# Use anova() to compare with the addictive model
anova_result_2 <- anova(interaction_model, model, test = "Chi")

# Print results
print(anova_result_2)

# Create LaTeX table for the model

# Create a data frame containing ANOVA results
result_table_2 <- xtable(anova_result_2)

# Convert to xtable format
print(result_table_2, include.rownames = FALSE)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
8488	11561.97			
8494	11568.26	-6	-6.29	0.3912

Since the p value is 0.3912 which is bigger than 0.05, it indicates that the deviation difference between the two models is not significant, that is, the interaction term has no significant impact on the explanatory power of the model.The answers to 2a and 2b would not potentially change if we included the interaction term in this model.