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.

   Codes as below:

```
# convert variables into unordered factor variables
climateSupport$sanctions_fac <- as.factor(as.integer(climateSupport$sanctions))
climateSupport$countries_fac <- as.factor(as.integer(climateSupport$countries))

# convert response variable into logical variable
climateSupport$choice <- as.logical(ifelse(climateSupport$choice == 'Supported', 1, 0))

# 1. run an additive model
mod <- glm(choice ~ countries_fac + relevel(sanctions_fac, ref = '1'),
          data = climateSupport[, !names(climateSupport) %in% c('countries', 'sanctions')],
          family = binomial(link = 'logit'))

# print the summary
summary(mod)
```

Summary output:

Table 1: Logistic regression model wIth 20 countries and 5% sanction as reference level

	Addition Model
(Intercept)	-0.081
	(0.053)
80 of 192 countries	0.336^{***}
	(0.054)
160 of 192 countries	0.648^{***}
	(0.054)
No sanctions	-0.192^{**}
	(0.062)
15% sanctions	-0.325^{***}
	(0.062)
20% sanctions	-0.495^{***}
	(0.062)
AIC	11580.260
BIC	11622.547
Log Likelihood	-5784.130
Deviance	11568.260
Num. obs.	8500

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

Global Hypothesis:

$$H_0 : \text{all slopes} = 0$$

$$H_A : \text{at least one } \beta_j \neq 0$$

As we can see from above summary output, expect for intercept, all coefficients of explained variables are showed statistically significant. We can check using R language as well, coding as below:

```r
# check if coefficients' p valuesa re less than 0,05 alpha level
p_values <- summary(mod)$coefficients[, "Pr(>|z|)"]
print(p_values)
p_values <= 0.05
```

Table 2: Coefficients' p-values

(Intercept)	80 countries	160 countries	No sanction	15% sanction	20% sanction
0.12848	0	0	0.00203	0.0000002	0

Conclusion:

- The intercept (β_0) of -0.081 suggests that when number of countries is 20 of 192 and the sanction is 5% (considered as the reference category), the estimated log odds of the individual agrees with the policy is -0.081 on average.

- The coefficient for *80 of 192 participating countries* (β_1) is 0.336, indicating that increasing the number of countries from 20 to 80 increases the estimated log odds of the individual agreeing with the policy by 0.336 on average.

- The coefficient for *160 of 192 participating countries* (β_2) is 0.648, indicating that increasing the number of countries from 20 to 160 increases the estimated log odds of the individual agreeing with the policy by 0.648 on average.

- The coefficient for *No sanction* (β_3) is -0.192, indicating that compared to 5% sanction selection, no sanction decreases the estimated log odds of the individual agreeing with the policy by 0.192 on average.

- The coefficient for *15% sanction* (β_4) is -0.325, indicating that compared to 5% sanction selection, choosing 15% sanction decreases the estimated log odds of the individual agrees with the policy by 0.352 on average.

- The coefficient for *20% sanction* (β_5) is -0.495, indicating that compared to 5% sanction selection, choosing 20% sanction decreases the estimated log odds of the individual agrees with the policy by 0.495 on average.

Based on these results, we can conclude the following:

1. Country number matters: Increasing participating countries associated with higher odds individual support climate policy. This implies that more countries to participate can improve the likelihood of people choosing support the policy.

2. Sanction choices impact personal support: no sanction, 15% sanction and 20% sanction, these three selections all are associated with lower odds of if individual support this policy. This suggests that 5% sanction may be a more effective scheme for promoting people' support.

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)

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

   (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.

   (a) Answer:

   The coefficient for 15% sanctions is -0.325, and the coefficient for 160 participating countries is 0.648. Holding the number of countries constant at 160 (setting its value to 1), choosing 15% sanctions rather than 5% sanctions will increase the odds of an individual supporting the policy by a multiplicative factor, namely 1.275, or approximately 27.5% on average.

   Mathematically, the expression is :

$$\text{Odds increase} = \exp(-0.008 + 0.648 - 0.325) = \exp(0.243) \approx 1.275$$

   Computing as below:

```
# 2.(a) increasing sanctions from 5% to 15%
log_odd <- -0.08 + 0.336 * 0 + 0.648 * 1 - 0.192 * 0 - 0.325 * 1 -
    0.495 * 0
round(exp(log_odd),3)
```

   (b) Answer: With 80 countries and no sanction, we can set corresponding values and compute the estimated log odds is around 0.063. Then we use logistic function to convert the logit value into a estimated probability Finally, the probability is rounded as 0.516. This means that given 80 countries without sanctions, the probability of individual support policy is about 51.6%.

   Computing as below:

```
# 2.(b) estimated prob given that 80 of 192 countries with no
    sanctions
# compute logit value
logit <- -0.081 + 0.336 * 1 + 0.648 * 0 - 0.192 * 1 - 0.325 * 0 -
    0.495 * 0

# Use the logistic function for converting
esti_prob <- plogis(logit)

# print the result
round(esti_prob, 3)
```

(c) Answer:

The answers to questions (a) and (b) may not change even when considering the interaction term. This is because not all the coefficients of the interaction term is not statistically significant (evident showed on the summary output table). Additionally, the p-value from the Likelihood Ratio Test (can be seen from the below ANOVA test result), is 0.3912, much higher than 0.05.

Therefore, the interaction between the level of sanctions and the number of participating countries may not contribute significantly to explaining the variation in the dependent variable.

In general, if we cannot reject the null hypothesis in an anova test, it may indicate that interaction terms for the logistic regression model are not significant predictor, and it would be more appropriate for the model selection not to include the interaction item. So I lean towards not change the answers to above two sub-questions.

Coding as below:

```r
# 2.c) additive model vs interation term model
mod_int <- glm(choice ~ countries_fac * relevel(sanctions_fac, ref = '2'),
               data = climateSupport[, !names(climateSupport) %in% c('countries', 'sanctions')],
               family = binomial(link = 'logit'))

# print the model output
summary(mod_int)
stargazer(mod_int)

# perform an likelihood ratio test
anova(mod, mod_int, test = "LRT")
```

Anova Test Output:

```
Analysis of Deviance Table

Model 1: choice ~ countries_fac + relevel(sanctions_fac, ref = "2")
Model 2: choice ~ countries_fac * relevel(sanctions_fac, ref = "2")

Resid. Df   Resid. Dev   Df    Deviance   Pr(>Chi)
1      8494       11568
2      8488       11562        6       6.2928    0.3912
```

Model Summary Output:(next page)

Table 3: Logistic regression model wIth interaction term

	Dependent variable:
	choice
80 of 192 countries	0.470^{***}
	(0.109)
160 of 192 countries	0.743^{***}
	(0.106)
No sanction	-0.122
	(0.105)
15% sanction	-0.219^{**}
	(0.107)
20% sanction	-0.374^{***}
	(0.107)
80 countries : No sanction	-0.095
	(0.152)
160 countries : No sanction	-0.130
	(0.151)
80 countries : 15% sanction	-0.147
	(0.154)
160 countries : 15% sanction	-0.182
	(0.151)
80 countries : 20% sanction	-0.292^{*}
	(0.153)
160 countries: 20% sanction	-0.073
	(0.152)
Constant	-0.153^{**}
	(0.073)
Observations	8,500
Log Likelihood	−5,780.983
Akaike Inf. Crit.	11,585.970

*p<0.1; **p<0.05; ***p<0.01

Table 4: Top 30 Words Metrics per Topic

Topic	Top 30 Words Weight (%)	Top 30 Words in Total Vocab (%)	Multiplier
1	22.07	0.30	73.59
2	19.17	0.30	63.93
3	19.60	0.30	65.34
4	17.77	0.30	59.26
5	13.53	0.30	45.11
6	22.61	0.30	75.39
7	26.01	0.30	86.74
8	24.23	0.30	80.79
9	17.62	0.30	58.74
10	12.84	0.30	42.82
11	14.63	0.30	48.77
12	29.94	0.30	99.84
13	17.70	0.30	59.03
14	21.44	0.30	71.48
15	26.93	0.30	89.79
16	9.43	0.30	31.44
17	16.89	0.30	56.31
18	14.42	0.30	48.07
19	18.03	0.30	60.11
20	25.61	0.30	85.41
21	16.54	0.30	55.14
22	26.26	0.30	87.57
23	19.91	0.30	66.40
24	23.28	0.30	77.63
25	24.76	0.30	82.57
26	14.05	0.30	46.84
27	20.38	0.30	67.94
28	29.77	0.30	99.26
29	22.45	0.30	74.86
30	14.49	0.30	48.31