Problem Set 2

Samanta Nedzinskaite

Due: February 19, 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 February 19, 2023. 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.csv` 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.

```r
# load data
load("/Users/samantanedzinskaite/Documents/GitHub/StatsII_Spring2023/problemSets/PS02/template/climateSupport.RData")
data <- climateSupport

#have to encode the 'choice' data to make binary/binomial variables

new_data <- as.data.frame(ifelse(data$choice == "Supported", 1, 0))


#the countries variable is a categorical variable with 3 levels, and the
#sanctions variable is a categorical variable with 4 levels.
#have to convert them into a dummy variable for the model. (1 if it matches the categorical variable, 0 if it does not.)
new_data$countries <- model.matrix(~ countries - 1, data = data)
new_data$sanctions <- model.matrix(~ sanctions - 1, data = data)

#rename columns for convenience
colnames(new_data)[1] <- "choice"
colnames(new_data)[2] <- "countries"
colnames(new_data)[3] <- "sanctions"

#now we can fit an additive model
model <- glm(choice ~ ., data = new_data, family = binomial(link = "logit"))
#summary output
summary(model)
stargazer(model, title = "Model Summary", out = "summary_output.html")
```

Model Summary

	Dependent variable:
	choice
countriescountries20 of 192	-0.648***
	(0.054)
countriescountries80 of 192	-0.312***
	(0.054)
countriescountries160 of 192	
sanctionssanctionsNone	0.304***
	(0.062)
sanctionssanctions5%	0.495***
	(0.062)
sanctionssanctions15%	0.170***
	(0.062)
sanctionssanctions20%	
Constant	0.072
	(0.054)
Observations	8,500
Log Likelihood	-5,784.130
Akaike Inf. Crit.	11,580.260
Note:	*p<0.1; **p<0.05; ***p<0.01

Global Hypothesis

Statistic	N	Mean	St. Dev.	Min	Max
Resid. Df	2	8,496.500	3.536	8,494	8,499
Resid. Dev	2	11,675.830	152.134	11,568.260	11,783.410
Df	1	5.000		5	5
Deviance	1	215.150		215.150	215.150
Pr(> Chi)	1	0.000		0	0

The p-value s is 0.000, which is less than the significance level of 0.05, providing good evidence against the null hypothesis, that the observed data fits the predicted distribution very well.

The countries variable has three dummy variables, with the reference category being 192 of 192. The coefficients for countries20 of 192 and countries80of192 are negative and statistically significant, which means that compared to 192of192 countries, the odds of supporting the policy are lower for individuals from countries that participated in the policy to a lesser extent (i.e., 20 or 80 out of 192 countries).

The sanctions variable has four dummy variables, with the reference category being sanctionsNone. The coefficients for sanctions5 and sanctions15 are positive and statistically significant, which means that compared to having no sanctions, the odds of supporting the policy increase when the sanctions are set to 5 percent and 15 percent of the monthly household costs (given 2 percent GDP growth) (as opposed to having no sanctions in place)

Drawing a conclusion from the results of the model analysis, it can be said that both the countries and sanctions variables have a statistically significant effect on an individual's likelihood of supporting the policy. Individuals from countries that participated in the policy to a lesser extent (i.e., 20 or 80 out of 192 countries) compared to those from countries that participated more (i.e., 160 out of 192) are less likely to support the policy. Additionally, the odds of supporting the policy increase when the sanctions are set to 5 percent and 15 percent of the monthly household costs (given 2 percent GDP growth).

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)

As can be seen from the summary output, the difference between the coefficients for sanctionssanctions15 (0.170) and sanctionssanctions5 (0.495) is 0.325. This means that when the sanction level increases from 5 percent to 15 percent, the log odds of supporting the policy increase by 0.325.

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

We can find the estimated probability using this formula: log(p / (1 - p)) = b0 + b1X1 + b2X2 + bk*Xk or the inverse **p = 1 / [1 + exp(- (b0 + b1X1 + b2X2 + ... + bk*Xk))]**

```
#probability = 1 / (1 + exp(-(intercept + b1*countries80of192 + b2*
    sanctionsNone)))
probability <- 1 / (1 + exp(-(-0.072 + (-0.312)*1 + 0.304*1)))
print(probability)
```

0.4800107

The estimated probability that an individual will support the policy if there are 80of192 countries participating with no sanctions is 0.48/48 percent.

(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.

```
#c) would an interaction be appropriate?
#a model with interaction

model2 <- glm(choice ~ countries * sanctions, data = new_data, family
    = binomial(link = "logit"))
#compare it to our original additive model
test_for_model_comparison <- anova(model, model2, test = "Chisq")
summary(test_for_model_comparison)
```

Model Comparison

Statistic	N	Mean	St. Dev.	Min	Max
Resid. Df	2	8,491.000	4.243	8,488	8,494
Resid. Dev	2	11,565.110	4.450	11,561.970	11,568.260
Df	1	6.000		6	6
Deviance	1	6.293		6.293	6.293
Pr(> Chi)	1	0.391		0.391	0.391

As we can see from the table above, the p-value is 0.3912, which is much greater than the significance level of 0.05. As a result, there is not great evidence to suggest that Model 2 with the interaction termis a better fit for the data than Model 1 with only the effects added together.