Problem Set 2

Kaley Burg

February 17, 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.

   My answer:

   - First I imported the dataset, changed the outcome variable to binary (0, 1). I used a table to ensure that this worked properly. Then I made the countries and sanctions variables into unordered variables to be able to use them as dummy variables in the regression. I also removed the spaces in the countries variable labels
   - Then I ran the logistic regression using glm() and family = 'binomial', shown in the code below:

```
#first we change the outcome variable to 0 and 1
climateSupport$choice

climateSupport$choice_bin <- ifelse(climateSupport$choice == "
    Supported", 1, 0)

#run a table to make sure it worked

table(climateSupport$choice, climateSupport$choice_bin)

#it worked

climateSupport$countries <- factor(climateSupport$countries, levels =
     levels(climateSupport$countries), ordered = F)
climateSupport$sanctions <- factor(climateSupport$sanctions, levels =
     levels(climateSupport$sanctions), ordered = F)

class(climateSupport$sanctions)
class(climateSupport$countries)


#change the levels to unordered because this has the regression treat
#the variable as dummy variables
levels(climateSupport$countries) <- gsub(" ", "", levels(
    climateSupport$countries))
levels(climateSupport$sanctions) <- gsub(" ", "", levels(
    climateSupport$sanctions))
```

```r
mod <- glm(choice_bin ~ sanctions + countries,
           data = climateSupport,
           family = "binomial")


summary(mod)
```

- **Here is the summary output:**

Table 1:

	Dependent variable:
	choice_bin
sanctions5%	0.192***
	(0.062)
sanctions15%	−0.133**
	(0.062)
sanctions20%	−0.304***
	(0.062)
countries80of192	0.336***
	(0.054)
countries160of192	0.648***
	(0.054)
Constant	−0.273***
	(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

- The odd ratios are also as follows:
- these are calculated from this code:

```r
odds_ratios <- exp(mod$coefficients)
```

Table 2:

(Intercept)	sanctions5%	sanctions15%	sanctions20%	countries80of192	countries160of192
0.761	1.211	0.875	0.738	1.400	1.912

- **The global null and alternative hypotheses are as follows:**

$$H_0: \text{ All slopes are equal to 0}$$
$$H_a: \text{ At least one slope is not equal to 0}$$

- We can test this hypothesis using an ANOVA, shown here:

```
modnull <- glm(choice_bin ~ 1,
               family = binomial(link = "logit"),
               data = climateSupport)

globalnulltest <- anova(modnull, mod, test = "LRT")
```

Table 3: ANOVA Results

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	8,499	11,783.410			
2	8,494	11,568.260	5	215.150	$< 2.2e-16^{***}$

Note: *p<0.1; **p<0.05; ***p<0.01

- **Interpretation of results and conclusions:**
- Here is the results written out as a logistic regression equation

$$\begin{aligned}\text{logit}[P(Y_i = 1|X_i)] = & 0.192 \times \text{sanctions5\%} \\ & - 0.133 \times \text{sanctions15\%} \\ & - 0.304 \times \text{sanctions20\%} \\ & + 0.336 \times \text{countries80of192} \\ & + 0.648 \times \text{countries160of192} \\ & - 0.273 \times \text{Constant}\end{aligned}$$

- **Interpretation of coefficients**

- When 20 countries participate in the international agreements and there are no sanctions for missing emission reduction targets, then on average the log odds of an individual agreeing with the policy is **0.273** This is a statistically significant relationship.
  - Furthermore, the baseline odds ratio can be represented as $e^{\hat{\beta}_0} = e^{-0.273} = 0.7613211$ meaning that 0.76 is the estimated baseline odds that an individual supports the policy
  - Interpretation of this results means that on average, people do not support the policy
  - 0.761 indicates that the odds of an individual supporting the policy are approximately 0.761 times lower than the odds of an individual not supporting it.
- A change from no sanctions to a 5% sanction is associated, on average, with a **0.192** increase in the log odds of an individual agreeing with the policy, holding the number of countries participating constant. This is a statistically significant relationship.
  - A change from no sanctions to a 5% sanction increases the odds of supporting the policy by a multiplicative factor of 1.211; it increases the odds by $\approx$ 21.1%
- A change from a 5% sanction to a 15% sanction is associated, on average, with a **0.133** increase in the log odds of an individual agreeing with the policy, holding the number of countries participating constant. This is a statistically significant relationship.
  - A change from a 5% sanction to a 15% sanction decreases the odds of supporting the policy by a multiplicative factor of 0.875; it decreases the odds by $\approx$ 12.5%
- A change from a 15% sanction to a 20% sanction is associated, on average, with a **0.304** increase in the log odds of an individual agreeing with the policy, holding the number of countries participating constant. This is a statistically significant relationship.
  - A change from a 15% sanction to a 20% sanction decreases the odds of supporting the policy by a multiplicative factor of 0.738; it decreases the odds by $\approx$ 26.2%
- A change from 20 of 192 to 80 of 192 countries participating in the international agreement is associated, on average, with a **0.336** increase in the log odds of an individual agreeing with the policy, holding sanctions constant. This is a statistically significant relationship.
  - A change from 20 of 192 to 80 of 192 countries participating in the international agreement increases the odds of supporting the policy by a multiplicative factor of 1.400; it increases the odds by $\approx$ 40.0%

- A change from 80 of 192 to 160 of 192 countries participating in the international agreement is associated, on average, with a **0.648** increase in the log odds of an individual agreeing with the policy, holding sanctions constant. This is a statistically significant relationship.
  - A change from 80 of 192 to 160 of 192 countries participating in the international agreement increases the odds of supporting the policy by a multiplicative factor of 1.912; it increases the odds by $\approx 91.2\%$
- **Interpretation of ANOVA**
- Since the p value is below a significance level of $\alpha = 0.05$, this indicates that there is sufficient evidence to reject the global null hypothesis which states that all of the β values are equal to 0.

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)

   - There a few different ways to solve this problem. First, the reference category in the regression can simply be set to the 5% category. The code is shown here and the results are on the next page:

```
climateSupport$sanctionstest <- relevel(climateSupport$sanctions,
    ref = "5%")

mod2 <- glm(choice_bin ~ sanctionstest + countries,
            data = climateSupport,
            family = "binomial")


summary(mod2)
```

Table 4:

	Dependent variable:
	choice_bin
sanctionstestNone	-0.192^{***}
	(0.062)
sanctionstest15%	-0.325^{***}
	(0.062)
sanctionstest20%	-0.495^{***}
	(0.062)
countries80of192	0.336^{***}
	(0.054)
countries160of192	0.648^{***}
	(0.054)
Constant	-0.081
	(0.053)
Observations	8,500
Log Likelihood	$-5{,}784.130$
Akaike Inf. Crit.	11,580.260
Note:	*p<0.1; **p<0.05; ***p<0.01

- We can see that the change in sanctions from 5% to 15% is associated with a 0.32510 decrease in the log odds of supporting the policy
  - Therefore, a change from 5 to 15% sanctions in which nearly all countries are participating in the international agreement decreases the odds of supporting the policy by a multiplicative factor of 0.722; it decreases the odds by $\approx$ 27.8%
- It is also important to note that since this is an additive model, the change between 5 and 15% sanctions is the same across all levels of country participation, as the slopes by country participation level are made equivalent.
- This can also be solved by hand when the reference category is set to "None" as follows:

$$\text{SanctionsNone} + \beta_1 = 0.19186$$
$$\text{SanctionsNone} + \beta_2 = 0.13325$$

$$\begin{aligned} &(\text{SanctionsNone} + \beta_1 = 0.19186) \\ -&(\text{SanctionsNone} + \beta_2 = 0.13325) \\ &\qquad\qquad\qquad\quad = \beta_1 + \beta_2 = 0.19186 + 0.13325 \end{aligned}$$

$$\beta_1 - \beta_2 = 0.32511$$
$$-\beta_2 - \beta_1 = -0.32511$$

- As shown above, we get approximately the same answer as if we change the reference category

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

- To find a predicted probability, we can use this equation:

$$P(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

- Using the β coefficient associated with 80 of 192 countries participating and sanctions held at their reference level of "None," then we can plug in the values to get this equation:

$$\frac{1}{1 + \exp\left(-(-0.27266 + 0.33636)\right)}$$

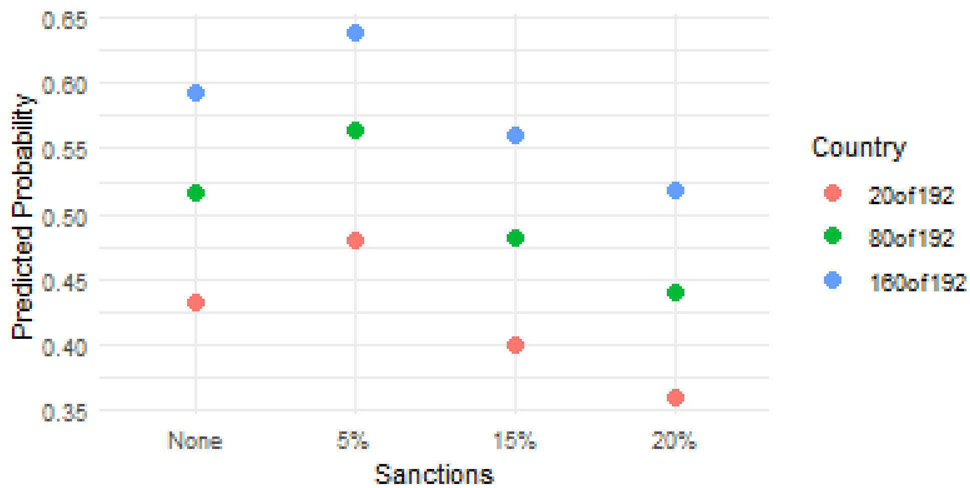
- This simplifies to 0.5159191
- Therefore, the estimated probability of an individual supporting the policy when 80 of 192 countries participate and there are no sanctions is **0.5159191**
- This could also be solved in R using the plogis() function
-
```
plogis(0.33636 - 0.27266)
```

- Additionally, we can use this function along with code adapted from the lecture slides to plot the predicted probabilities. The code for this is:

```r
predicted_data <- with(climateSupport, data.frame(countries,
    sanctions))
predicted_data$PredictedProb <- plogis(predict(mod, newdata =
    predicted_data, type = "link"))

#print result
predicted_data

# Plotting predicted probabilities
png("mod_predprobs.png", width = 400, height = 200)
ggplot(predicted_data, aes(x = sanctions, y = PredictedProb,
    color = countries)) +
  geom_point(size = 3) +
  labs(x = "Sanctions", y = "Predicted Probability", color = "
   Country") +
  theme_minimal()
```



- Then, using this list of predicted values we can subset by sanctions set to None and countries set to 80 of 192. The code is as follows:

```r
predval <- predicted_data[predicted_data$sanctions == 'None' &
    predicted_data$countries == '80of192', ]

#and check it

unique(predval)
unique(predicted_data)
```

- This returns these values on the next page:

Table 5: Your table caption here

Column 1	**Column 2**	**Column 3**
80of192	15%	0.4826196
160of192	15%	0.5603146
160of192	None	0.5928323
160of192	5%	0.6381958
20of192	15%	0.3998931
80of192	20%	0.4403193
160of192	20%	0.5180228
20of192	5%	0.4798090
80of192	5%	0.5635428
80of192	None	0.5159191
20of192	20%	0.3598012
20of192	None	0.4322534

- We can see here that we get the same predicted value of 0.52

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

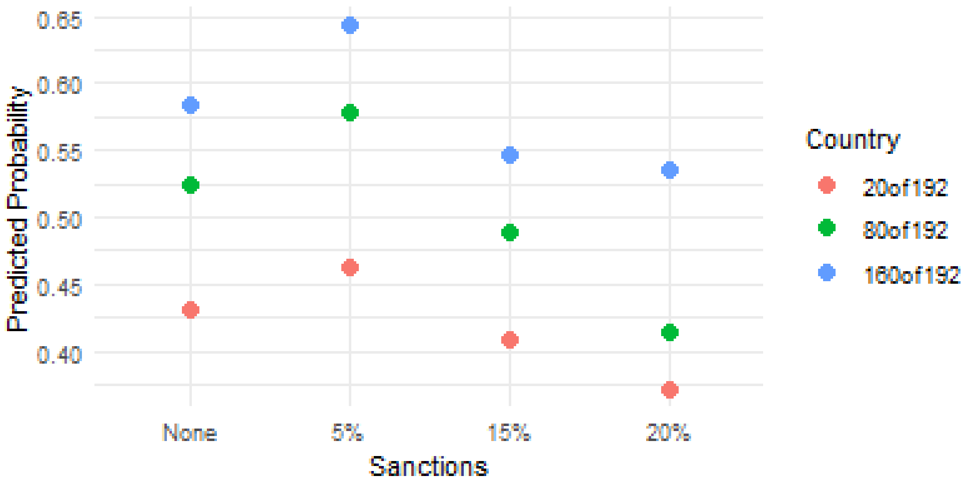
- Yes, the answers to 2a and 2b would potentially change if we included an interaction term. This is because with an interaction term, the slopes are not necessarily equal across each of the variables.
- For example, in 2a, the change from 5 to 15% sanctions is the same across all levels of country participation as we are modeling it to not vary along with country participation. Therefore, including an interaction term would potentially change this.
- We can see this when we run the interactive model:
-
```
mod_interaction <- glm(choice_bin ~ sanctions * countries,
                       data = climateSupport,
                       family = "binomial")
summary(mod_interaction)
```

Table 6:

	Dependent variable:
	choice_bin
sanctions5%	0.122
	(0.105)
sanctions15%	−0.097
	(0.108)
sanctions20%	−0.253**
	(0.108)
countries80of192	0.376***
	(0.106)
countries160of192	0.613***
	(0.108)
sanctions5%:countries80of192	0.095
	(0.152)
sanctions15%:countries80of192	−0.052
	(0.152)
sanctions20%:countries80of192	−0.197
	(0.151)
sanctions5%:countries160of192	0.130
	(0.151)
sanctions15%:countries160of192	−0.052
	(0.153)
sanctions20%:countries160of192	0.057
	(0.154)
Constant	−0.275***
	(0.075)
Observations	8,500
Log Likelihood	−5,780.983
Akaike Inf. Crit.	11,585.970
Note:	*p<0.1; **p<0.05; ***p<0.01

- Given the change in β values, we can see that the associated change between 5 and 15% sanctions for 160 of 192 countries participating would be associated with a -0.182 change in log odds of support for the policy (a decrease in odds by a multiplicative factor of 0.8336013)
- This is in comparison to without the additive model (change in log odds of -0.32510)

- For 2c, the predicted values also are not guaranteed to have the same slopes. This is best demonstrated visually:



- We can see that the connections between points of the same country level have different slopes than that of the earlier plot. Therefore, the predicted values would change if we use an interactive model.
- This can also be demonstrated with the same code as above

-
```
predval2 <- predicted_data2[predicted_data2$sanctions == 'None' &
    predicted_data2$countries == '80of192', ]

#and check it

unique(predval2)
unique(predicted_data2)
```

- The results are included in the next page

Table 7: Your table caption here

Column 1	**Column 2**	**Column 3**
80of192	15%	0.4879433
160of192	15%	0.5472222
160of192	None	0.5836972
160of192	5%	0.6433289
20of192	15%	0.4081633
80of192	20%	0.4136546
160of192	20%	0.5355030
20of192	5%	0.4618474
80of192	5%	0.5786963
80of192	None	0.5252101
20of192	20%	0.3711485
20of192	None	0.4317549

- We can see that the predicted value at 80 of 192 countries and no sanctions changes slightly from the previous table
- Perform a test to see if including an interaction is appropriate.
  - This can be tested using an ANOVA with the additive model and interactive model
  - ```
    int_anova <- anova(mod, mod_interaction , test = "LRT")
    ```
  - These are the ANOVA results:

Table 8:

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

  - We can see that the p value is 0.3912, so there is **not** sufficient evidence to able to reject the null hypothesis, meaning that the interaction is not appropriate.
    ```