Problem Set 2

Applied Stats II

Maiia Skrypnyk 23371609

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.

   Preparing the data:

```r
#Loading data
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/climateSupport.RData?raw=true"))

#Recoding 'Choice' to binary
climateSupport$choice <- as.numeric(climateSupport$choice == "Supported")
#Recoding 'Countries' and 'Sanctions' to unordered factors, setting reference levels
climateSupport$countries <- relevel(factor(climateSupport$countries, ordered = FALSE), ref = "20 of 192")
climateSupport$sanctions <- relevel(factor(climateSupport$sanctions, ordered = FALSE), ref = "None")
```

   Variables `countries '20 of 192'` and `sanctions 'None'` were set as reference categories.

   Logistic regression formula:

$$\text{logit}(P(\text{choice} = 1)) = \beta_0 + \beta_1 X_{\text{countries80}} + \beta_2 X_{\text{countries160}} + \beta_3 X_{\text{sanctions5\%}} + \beta_4 X_{\text{sanctions15\%}} + \beta_5 X_{\text{sanctions20\%}}$$

   Summary output:

```r
#Additive model
glm_full <- glm (choice ~ countries + sanctions,
                 data = climateSupport,
                 family = binomial (link = "logit"))
summary(glm_full)
```

   (Please see the summary table on the next page)

Table 1:

	Dependent variable:
	choice
countries80 of 192	0.33636***
	(0.054)
countries160 of 192	0.64835***
	(0.054)
sanctions5%	0.19186***
	(0.062)
sanctions15%	−0.13325**
	(0.062)
sanctions20%	−0.30356***
	(0.062)
Constant	−0.27266***
	(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

```r
#Calculating odds ratio of the coefficients
odds_ratios <- exp(glm_full$coefficients)
print(odds_ratios)
```

```
(Intercept)        0.7613492
countries80 of 192 1.3998442
countries160 of 192 1.9123823
sanctions5%        1.2114952
sanctions15%       0.8752484
sanctions20%       0.7381826
```

```r
#Calculating percentages
percentages <- (odds_ratios-1)*100
print(percentages)
```

```
(Intercept)        -23.86508
countries80 of 192 39.98442
countries160 of 192 91.23823
sanctions5%       21.14952
sanctions15%      -12.47516
sanctions20%      -26.18174
```

Interpretation of the coefficients:

- β_0 **(intercept, -0.27266):** On average, when 0 countries participate in the agreement and there are no sanctions, the log(odds) of an individual supporting a policy equals -0.27266. The estimated odds of this are ≈ 0.76 ($\approx 24\%$)(baseline odds ratio).
- β_1 **(countries80 of 192, 0.33636):** On average, holding all the other variables constant, having 80 of 192 countries participating in the agreement is associated with an increase in the log(odds) of an individual supporting a policy by ≈ 0.336 and an increase in the odds by a factor of ≈ 1.4 ($\approx 40\%$), compared to the reference category.
- β_2 **(countries160 of 192, 0.64835):** On average, holding all the other variables constant, having 160 of 192 countries participating in the agreement is associated with an increase in the log(odds) of an individual supporting a policy by ≈ 0.65 and an increase in the odds by a factor of ≈ 1.9 ($\approx 91\%$), compared to the reference category.
- β_3 **(sanctions5%, 0.19186):** On average, holding all the other variables constant, setting a level of sanctions for missing emission reduction targets at 5% is associated with an increase in the log(odds) of an individual supporting a policy by ≈ 0.19 and an increase in the odds by a factor of ≈ 1.22 ($\approx 21\%$), compared to the reference category.
- β_4 **(sanctions15%, -0.13325):** On average, holding all the other variables constant, setting a level of sanctions for missing emission reduction targets at 15% is associated with a decrease in the log(odds) of an individual supporting a policy by ≈ 0.133 and a decrease in the odds by a factor of ≈ 0.87 ($\approx 12.5\%$), compared to the reference category.
- β_5 **(sanctions20%, -1.13325):** On average, holding all the other variables constant, setting a level of sanctions for missing emission reduction targets at 20% is associated with a decrease in the log(odds) of an individual supporting a policy by ≈ 1.33 and a decrease in the odds by a factor of ≈ 0.74 (($\approx 26\%$)), compared to the reference category.

Overall, having more countries participating in the agreement is associated with higher support for the policy, while increasing sanctions generally decreases such support, with milder sanctions (5%) having a smaller impact compared to harsher ones (15% and 20%).

Global null hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
Or, in more general form, $H_0 : \beta_j = 0$
$H_a : \beta_j \neq 0$ (at least one of the coefficients is not equal to zero)

P-value: To establish whether at least one predictor is a significant predictor in our logistic regression model, we should compare the full model to the reduced model. We can use `anova()` function to perform a likelihood ratio test.

```
#Reduced model
glm_reduced <- glm(choice ~ 1, data = climateSupport, family =
    binomial(link = "logit"))

#Likelihood ratio test to determine p-value
anova(glm_full, glm_reduced, test="LRT")
```

```
Model 1: choice ~ countries + sanctions
Model 2: choice ~ 1
Resid.  Df Resid. Dev Df Deviance  Pr(>Chi)
1      8494      11568
2      8499      11783 -5  -215.15 < 2.2e-16 ***
```

P-value $< 2.2e^{-16}$ is extremely small and close to zero, therefore, we have found strong evidence to reject the null hypothesis (the additional parameters within the full model do not significantly improve the fit as compared to the reduced model) at the α level $= 0.05$.

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 our model is additive (not interactive), any β change in `sanctions` will have the same effect on the odds of an individual supporting the policy (`choice` $= 1$), regardless of the number of `countries` participating in the international agreement, be it 20, 80, or 160.

   So, we could calculate the effect of increasing sanctions from 5% to 15% as a difference between β_4 and β_3 as follows:

```
change <- glm_full$coefficients["sanctions15%"] - glm_full$
    coefficients["sanctions5%"]
print(change)
```

```
-0.3251028
```

Interpretation: On average, holding all the other variables constant, increasing sanctions from 5% to 15% is associated with a 0.3251028 decrease in the log odds of an individual supporting the policy.

```
#Calculating the odds ratio
odds_ratio1 <- exp(change)
print(odds_ratio1)
```

```
0.7224531
```

```
#Calculating the percentage
percentage <- (odds_ratio1 -1)*100
print(percentage)
```

```
27.75469
```

Therefore, on average, holding all the other variables constant, increasing sanctions from 5% to 15% is associated with $e^{-0.3251028} \approx 0.722 \approx 27.76\%$ decrease in the odds of an individual supporting the policy.

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

The formula for such estimation could be given by:

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_{\text{countries80}}+\beta_2 x_{\text{sanctionsnone}}}}{1+e^{\beta_0+\beta_1 x_{\text{countries80}}+\beta_2 x_{\text{sanctionsnone}}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{\text{countries80}}+\beta_2 x_{\text{sanctionsnone}})}}$$

Since we chose the "None" category for `sanctions` as a reference category, it is not included in the model, and we can assume that its coefficient equals to zero. Given that $X_{\text{countries80}} = 1$ and $\beta_2 = 0$,

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1\cdot 1+0\cdot X_{\text{sanctionsnone}}}}{1+e^{\beta_0+\beta_1\cdot 1+0\cdot X_{\text{sanctionsnone}}}} = \frac{1}{1+e^{-(\beta_0+\beta_1)}}$$

```
#Calculating the probability
plus <- glm_full$coefficients["(Intercept)"]+glm_full$coefficients["
    countries80 of 192"] #0.06369783
probability <- 1 / (1 + exp(-plus))
print(probability)
```

```
0.5159191
```

```
#Checking with the predict() function
probability2 <- predict(glm_full, newdata = data.frame(countries = "
    80 of 192", sanctions = "None"), type = "response")
print(probability2)
```

```
0.5159191
```

There is an estimated probability of approximately $0.52 = 52\%$ 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.

First, let's run an interactive logistic regression model:

```r
#Interactive model
interactive_glm <- glm(choice ~ countries * sanctions,
                       data = climateSupport,
                       family = binomial(link = "logit"))
summary(interactive_glm)
```

(Oops, it is a really long summary table – please see the next page)

Table 2:

	Dependent variable:
	choice
countries80 of 192	0.376^{***}
	(0.106)
countries160 of 192	0.613^{***}
	(0.108)
sanctions5%	0.122
	(0.105)
sanctions15%	-0.097
	(0.108)
sanctions20%	-0.253^{**}
	(0.108)
countries80 of 192:sanctions5%	0.095
	(0.152)
countries160 of 192:sanctions5%	0.130
	(0.151)
countries80 of 192:sanctions15%	-0.052
	(0.152)
countries160 of 192:sanctions15%	-0.052
	(0.153)
countries80 of 192:sanctions20%	-0.197
	(0.151)
countries160 of 192:sanctions20%	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

Even at a glance, we can see that the coefficients for all the interactive terms are not statistically significant, which can give us a notion that adding interaction to the model does not vastly contribute to explaining the variability in the response variable `choice`. Let's perform a likelihood ratio test to support or reject this hypothesis:

```
#Comparing additive and interactive models
lrt <- anova(glm_full, interactive_glm, test = "LRT")
print(lrt)
```

```
Model 1: choice ~ countries + sanctions
Model 2: choice ~ countries * sanctions
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8494      11568
2      8488      11562  6   6.2928   0.3912
```

Given the **p-value** of 0.3912, we have not found evidence to reject the null hypothesis (adding an interaction to the model does not significantly improve its fit) at the α level = 0.05. This suggests that the association between the response variable and the predictor variables can be explained by the main effects of the predictors alone, without considering their interactions.

Furthermore, including the interaction terms would potentially change the answers both to 2(a) and 2(b):

- **2(a)**: A given β change in `sanctions` may not have the same effect on the odds of an individual supporting the policy (`choice` = 1), as it may now be dependent on the specific number of `countries` participating in the international agreement ('levels' of the other explanatory variable). However, the coefficient for the interactive term is not statistically significant.
- **2(b)**: Let's estimate the same probability (that an individual will support a policy if there are 80 of 192 countries participating with no sanctions) using the interactive model now, and then compare:

```
#Calculating the probability (interactive model)
probability3 <- predict(interactive_glm, newdata = data.frame(
    countries = "80 of 192", sanctions = "None"), type = "response
    ")
print(probability3)
```

```
0.5252101
```

Calculating the difference of two probabilities:

```
probability3-probability2
```

```
0.009291008
```

Overall, the difference of two probabilities is quite minor, but it is still present.