Problem Set 2

Student: Shekhar Kedia (23351315)
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.

   Solution:
   Firstly using `R`, I load the dataset. Then, I create levels for the explanatory variables (including a reference category) by running the following codes:

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

# Creating levels for factor variables and setting the reference (base)
    category
climateSupport$sanctions <- relevel(factor(climateSupport$sanctions,
    ordered=F), ref="None")
climateSupport$countries <- relevel(factor(climateSupport$countries,
    ordered=F), ref="20 of 192")
```

   Then, I run the logistic regression model (additive) and create the summary output using the following codes:

```
# Logit model (additive)
logit_add <- glm(choice ~ countries + sanctions , data = climateSupport,
                 family = binomial (link = "logit" ))

summary(logit_add) #Displaying output of the model
```

```
Call:
glm(formula = choice ~ countries + sanctions, family = binomial(link = "logit"),
data = climateSupport)

Coefficients:
                      Estimate  Std.Error  z-value     Pr(>|z|)
(Intercept)           -0.27266    0.05360  -5.087   3.64e-07 ***
countries80 of 192     0.33636    0.05380   6.252   4.05e-10 ***
countries160 of 192    0.64835    0.05388  12.033    < 2e-16 ***
sanctions5%            0.19186    0.06216   3.086    0.00203 **
sanctions15%          -0.13325    0.06208  -2.146    0.03183 *
sanctions20%          -0.30356    0.06209  -4.889   1.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11568  on 8494  degrees of freedom
AIC: 11580

Number of Fisher Scoring iterations: 4
```

The global null for the model is:
H_0 : All slopes $= 0$ i.e., $\beta_1 = \beta_2 = 0$
H_1 : At least one slope (β_j) is not equal to 0

A comparison of null deviance and residual deviance is used to test the global null hypothesis using a likelihood ratio test which follows a central χ^2 distribution under H_0 being true.
We first prepare the null model and then compare it against the full model using the following `R` codes:

```r
logit_null <- glm(choice ~ 1 , data = climateSupport,
                  family = binomial (link = "logit" ))
anova(logit_null, logit_add, test = "LRT")
```

We get the following output:

```
Analysis of Deviance Table

Model 1: choice ~ 1
Model 2: choice ~ countries + sanctions
     Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499      11783
2      8494      11568  5   215.15   < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the p-value is less than 0.05 (α), **we reject the null** hypothesis that all the slopes equal to 0. Therefore, we have found sufficient evidence to conclude that **at least one predictor is reliable** in the logistic 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)

Solution:
Firstly I change the reference (base) level for `sanctions` variable from "None" category to "5". Then, I run the same additive model and produce the summary output using the following codes:

```r
# Changing the reference level (base) to 5%
climateSupport$sanctions <- relevel(factor(climateSupport$sanctions, ordered=F), ref="5%")

# Logit model (additive)
logit_add_new <- glm(choice ~ countries + sanctions , data = climateSupport,
                   family = binomial (link = "logit" ))

summary(logit_add_new) #Displaying output of the model
```

```
Call:
glm(formula = choice ~ countries + sanctions, family = binomial(link = "logit")
data = climateSupport)

Coefficients:
                     Estimate  Std.Error  z-value  Pr(>|z|)
(Intercept)          -0.08081    0.05316  -1.520   0.12848
countries80 of 192    0.33636    0.05380   6.252  4.05e-10 ***
countries160 of 192   0.64835    0.05388  12.033   < 2e-16 ***
sanctionsNone        -0.19186    0.06216  -3.086   0.00203 **
sanctions15%         -0.32510    0.06224  -5.224  1.76e-07 ***
sanctions20%         -0.49542    0.06228  -7.955  1.79e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11568  on 8494  degrees of freedom
AIC: 11580

Number of Fisher Scoring iterations: 4
```

We interpret from the model output that for policy in which nearly all countries participate [160 of 192], increasing sanctions from 5% to 15% **reduces the log odds** that an individual will support the policy by **0.325 units** on average.
It is imperative to note that since it's an additive model, the number of participating countries (`countries`) variable is considered constant in determining the effect of (`sanctions`) variable in predicting the outcome and therefore we would make similar interpretation for any number of countries participating.

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

Solution:

$$\text{Estimated Pr}(\text{choice} = 1|\text{countries} = 80 \text{ of } 192, \text{sanctions} = \text{None}) = \hat{\pi}_i$$

$$or, \hat{\pi}_i = \frac{\exp^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}{1 + \exp^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}$$

Using the following `R` codes, we can predict the probability:

```
# Predicting probability
predict(logit_add, newdata = data.frame(countries="80 of 192",
    sanctions="None"), type="response")
```

We see that the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions is **0.516**.

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

Solution: Firstly we create another model with the interaction term and then look at the results using the following `R` codes:

```
# Creating interactive model
logit_mul <- glm(choice ~ countries * sanctions , data =
    climateSupport,
                     family = binomial (link = "logit" ))
summary(logit_mul) #Displaying the model output
```

```
Call:
glm(formula = choice ~ countries*sanctions, family = binomial(link = "logit"),
data = climateSupport)
Coefficients:
                                 Estimate  Std.error  z value Pr(>|z|)
(Intercept)                      -0.15291    0.07339  -2.083 0.037207 *
countries80 of 192                0.47033    0.10912   4.310 1.63e-05 ***
countries160 of 192               0.74275    0.10556   7.036 1.98e-12 ***
sanctionsNone                    -0.12179    0.10518  -1.158 0.246909
sanctions15%                     -0.21866    0.10687  -2.046 0.040751 *
sanctions20%                     -0.37439    0.10671  -3.508 0.000451 ***
countries80 of 192:sanctionsNone -0.09471    0.15232  -0.622 0.534071
countries160 of 192:sanctionsNone -0.13009   0.15103  -0.861 0.389063
countries80 of 192:sanctions15%  -0.14700    0.15368  -0.957 0.338798
countries160 of 192:sanctions15% -0.18173    0.15094  -1.204 0.228591
countries80 of 192:sanctions20%  -0.29192    0.15306  -1.907 0.056493 .
countries160 of 192:sanctions20% -0.07321    0.15196  -0.482 0.629984
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11562  on 8488  degrees of freedom
AIC: 11586

Number of Fisher Scoring iterations: 4
```

We see that, the responses to 2a and 2b would potentially change if we included the interaction term in the model as the **intercept and slope estimates would vary** depending on the categories of both explanatory variables (`countries` and `sanctions`).

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

  Solution:
  Next, I compare the model with the interaction term with the additive model to see if including an interaction is appropriate or not. We use the following `R` codes:

  ```r
  # Performing test to compare additive and interactive models
  anova(logit_add, logit_mul, test = "LRT")
  ```

  We get the following output:

  ```
  Analysis of Deviance Table

  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
  ```

  As the p-value is greater than 0.05 (α), **we fail to reject the null** hypothesis that having the interaction term is not a better model over additive model. Or in other words, there is not enough evidence to suggest that number of countries participating by sanction level has an effect in predicting likelihood of an individual supporting a policy.