# Problem Set 2

Applied Stats II

Due: February 28, 2022

## 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 class on Monday February 28, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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.**

   Before running the model, I will make the following assumptions which are necessary for logistic regression:

   1. We have independent observations.

   2. Covariates are linearly related to the logit of outcome variables.

   The global null hypothesis is: that all coefficients in the model are equal to zero: Beta = 0

   The alternative hypothesis is that the coefficients are not equal to zero.

   To fit an additive model the following code was used: I used the "." to add all the variables into the model:

```
reg <- glm(choice ~ ., data = climateSupport, family = "binomial")
summary(reg)
```

   This gave the following output:

```
Call:
glm(formula = choice ~ ., family = "binomial", data = climateSupport)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3976  -1.1490  -0.9679   1.1536   1.4025

Coefficients:
             Estimate Std. Error z value            Pr(>|z|)
(Intercept) -0.14458    0.04518  -3.200             0.00137 **
countries    0.32436    0.02689  12.062 < 0.0000000000000002 ***
sanctions   -0.12353    0.01964  -6.291        0.000000000315 ***
---
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: 11597  on 8497  degrees of freedom
AIC: 11603

Number of Fisher Scoring iterations: 4
```

The p-value for countries is 0.0000000000000002, which is significant, and for sanctions is 0.000000000315, which came out as also significant.

Describe the results and provide a conclusion:

The results show that both the explanatory variables ”countries” and ”sanctions” have significant beta values. This suggests that we reject the global null hypothesis that the coefficients in the model are equal to zero.

Firstly we can infer that for the variable ”countries”, that increasing the number of participating countries increases the log odds of supporting a policy.

Secondly, for the variables ”sanctions” we can infer that increasing the number of sanctions for missing emission reduction targets decreases the log odds of supporting a policy.

From this model, I can conclude that increasing the number of countries participating and decreasing the sanctions has a positive effect of the log odds that an individual will agree with a policy.

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)**

   To test this I fist wrote out the formula for the predictive model using the log odds:

```
yhat <- - 0.1445813 + 0.3243567(countries) - 0.1235335(sanctions) +
    error
```

   This is using the log odds, and will be transformed later.

   By matching the variables as factors to the meaning behind them I compared the original data to the transformed data, this showed:

   Sanctions 0 = None 1 = 5% 2 = 15%

   Countries 0 = 20 of 192 1 = 80 of 192 2 = 160 of 192

   For the first question I was interested in the 160 of 192 group: factor reference = 2

   And the changes from 5% to 15% , therefor using 1 and 2 in the sanctions group.

   This resulted in the following prediction equations used in R

```
yhat <- - 0.1445813 + 0.3243567(countries) - 0.1235335(sanctions) +
    error
yhat <- - 0.1445813 + 0.3243567*2 - 0.1235335(sanctions) + error
yhat_5<- - 0.1445813 + 0.3243567*2 - 0.1235335*1
yhat_15 <- - 0.1445813 + 0.3243567*2 - 0.1235335*2

#Using exp to transform from log odds to odds.
diff = exp(yhat_15) - exp(yhat_5)
diff
```

   The difference is: -0.1700309. This means that increasing sanctions from 5% to 15% in the 160 of 192 reference category, while holding all else constant at their standard means results in a decrease in the odds of whether an individual agreed with the policy by 0.1700309.

   (b) **For the policy in which very few countries participate [20 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)**

   To now look at the difference in a different country category I changed the country factor to 0. The factors for the change from 5% to 15% have stayed the same as the previous question.

```
yhat<- - 0.1445813 + 0.3243567(countries) - 0.1235335(sanctions) +
    error
yhat_l <- - 0.1445813 + 0.3243567*0 - 0.1235335*1
yhat_m <- - 0.1445813 + 0.3243567*0 - 0.1235335*2
```

```
diff_2 <- exp(yhat_m) - exp(yhat_l)
diff_2
```

The difference here is -0.08887817. This means that on average for those in the 20 to 192 category the difference of increasing sanctions from 5% to 15% resulted, while holding all else constant at their standard means, results in a decrease of -0.08887817 in the odds of whether an individual agreed with the policy.

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

To change the odds, which have been used in the previous two questions to the probability, I used the following formula:

$Probability = odds/(1 + odds)$

So I first calculated the odds for no sanctions if 80 of 192 countries were participating.

```
yhat_none <- - 0.1445813 + 0.3243567*1 - 0.1235335*0
```

This produced "0.1797754" as the logs odds, I calculated the exp of this to find the odds and was then able to insert it into the formula.

```
yhat_none <- exp(yhat_none)
```

The odds of an individual will support a policy if there are 80 of 192 countries participating with no sanctions is 1.196948.

From here I input this into the probability formula:

```
x <- yhat_none/(1 + yhat_none)
x
```

This gave the probability of 0.5448232. Meaning that there is a 54.48% chance that an individual will support a policy if there are 80 of 192 countries participating with no sanctions.

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

This could potentially change if the interaction term was significant in the model, as this would suggest that the interaction term could tell us something important about the model. To find out of this is significant I ran a interactive model.

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

```
reg2 <- glm(choice ~ countries + sanctions +countries:sanctions,
    data = climateSupport, family = "binomial")
summary(reg2)
```

This output the following:

```
Call:
glm(formula = choice ~ countries + sanctions + countries:
    sanctions,
```

```
    family = "binomial", data = climateSupport)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3993  -1.1490  -0.9693   1.1536   1.4009

Coefficients:
                      Estimate Std. Error z value          Pr(>|z|)
(Intercept)          -0.148144   0.057311  -2.585           0.00974 **
countries             0.328007   0.045036   7.283 0.000000000000326 ***
sanctions            -0.121111   0.030987  -3.908 0.000092917443707 ***
countries:sanctions  -0.002455   0.024288  -0.101           0.91950
---
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: 11597  on 8496  degrees of freedom
AIC: 11605

Number of Fisher Scoring iterations: 4
```

By running the interaction model it shows that the interaction term is non-significant with a p-value of 0.91950. This suggest it owuld not change our answers to 2a and 2b much. Below I have shown the difference between the two predictive equations when using either the additive or interactive model: additive (used earlier)

```
yhat <- - 0.1445813 + 0.3243567(countries) - 0.1235335(sanctions)
    + error
```

interactive(current)

```
yhat_int <- -0.148144162 + 0.328007182(countries) - 0.121110523(
    sanctions) - 0.002454648(countries*sanctions)
```

By comparing the two it shows that the interactive term will make very little difference as the coefficients are the same to the second decimal place.