Problem Set 2

Applied Stats II

Due: February 18, 2024///Wei Tang 23362496

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.

   Firstly, we should load the dataset and then check the data frame.

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

   Now we should clean the data frame, and then transform it into a proper format to proceed to the next step.

```r
# clean the data, transform the dummy variables
dummy_country_variables <- model.matrix(~ countries -1, data = dfm)
dummy_sanctions_variables <- model.matrix(~ sanctions -1, data = dfm)
dfm$choice<-ifelse(dfm$choice=="Supported",1,0)
dfm <- cbind(dfm, dummy_country_variables)
dfm <- cbind(dfm, dummy_sanctions_variables)
# delete some useless columns
dfm <- dfm[, -which(names(dfm) == "countries")]
dfm <- dfm[, -which(names(dfm) == "sanctions")]
dfm <- dfm[, -which(names(dfm) == "sanctionsNone")]
dfm <- dfm[, -which(names(dfm) == "countries20 of 192")]
# Replace the special symbols and whitespaces of the variable names
colnames(dfm) <- gsub(" ", "_", colnames(dfm))
colnames(dfm) <- gsub("%", "percent", colnames(dfm))
View(dfm)
```

   Now we get the summary of the model as below:

```
> summary(additive_model)

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 ***
sanctions5percent    0.19186    0.06216   3.086  0.00203 **
sanctions15percent  -0.13325    0.06208  -2.146  0.03183 *
sanctions20percent  -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 hypothesis` is:

`There’s no relationship between the predictors and the response variable.` $(\beta_1 = \beta_2 = ... = \beta_p = 0)$

Conclusion:
From the summary, we can notice that the p-value of all variables are less than 0.05. The p-value is less than the chosen significance level (0.05), so there is sufficient statistical evidence to reject the null hypothesis. Therefore, we reject the null hypothesis in favour of the alternative hypothesis.
`(Alternative Hypothesis: There is a relationship between at least one predictor and the response variable.)`

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)

   Interpretation: Holding the policy in which nearly all countries participate [160 of 192], increasing sanctions from 5% to 15% changes the `log odds` of an individual to support the policy by:$-0.13325 - 0.19186 = -0.32511$.

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

```r
# (b)
# create a new_data with 80 of 192 countries participating with no
    sanctions
new_data <-data.frame(countries80_of_192=1, countries160_of_192=0, ‘
    sanctions5percent‘ =0, ‘sanctions15percent‘=0, ‘
    sanctions20percent‘=0)
# predict individual support probability
```

```
predicted_probability <- predict(additive_model,newdata = new_data, type = "response")
# get the result of probability
print(predicted_probability)
```

Then we get the output as below:

```
> print(predicted_probability)
        1
0.5159191
```

So the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions is 0.5159191.

(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.

We can first build a model with interaction terms.

```
# train the additive model with interactive terms.
model_with_interaction <- glm(choice ~ `countries80_of_192` + `countries160_of_192` +  `sanctions5percent` + `sanctions15percent` + `sanctions20percent`+`countries80_of_192`:`sanctions5percent`+`countries80_of_192`:`sanctions15percent`+`countries80_of_192`:`sanctions20percent`+`countries160_of_192`:`sanctions5percent`+`countries160_of_192`:`sanctions15percent`+`countries160_of_192`:`sanctions20percent`,
                                  data = dfm,
                                  family = "binomial")
```

Then do a likelihood test between the additive_model and the model_with_interaction:

```
# do a likelihood test between the 2 models
likelihood_ratio_test <- anova(additive_model, model_with_interaction, test = "LRT")
print(likelihood_ratio_test)
```

We got the output as below:

```
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8494      11568
2      8488      11562  6   6.2928   0.3912
```

We noticed that the p-value of the likelihood test is much higher than 0.05, so we FAILED to reject the null hypothesis that there is no significant difference between the two models.

Then we can check the summary of the model with interactive terms to prove our conclusion.

```
# another way to check the necessity of interactive terms:
# check the summary to see the p-value of the corresponding coefficients
summary(model_with_interaction)
```

The output is as below:

```
Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -0.27469    0.07534  -3.646 0.000267 ***
countries80_of_192                     0.37562    0.10627   3.535 0.000408 ***
countries160_of_192                    0.61266    0.10801   5.672 1.41e-08 ***
sanctions5percent                      0.12179    0.10518   1.158 0.246909
sanctions15percent                    -0.09687    0.10822  -0.895 0.370723
sanctions20percent                    -0.25260    0.10806  -2.338 0.019412 *
countries80_of_192:sanctions5percent   0.09471    0.15232   0.622 0.534071
countries80_of_192:sanctions15percent -0.05229    0.15167  -0.345 0.730262
countries80_of_192:sanctions20percent -0.19721    0.15104  -1.306 0.191675
countries160_of_192:sanctions5percent  0.13009    0.15103   0.861 0.389063
countries160_of_192:sanctions15percent -0.05165    0.15267  -0.338 0.735136
countries160_of_192:sanctions20percent  0.05688    0.15367   0.370 0.711279
---
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
```

As we can see, the interactive terms do not exhibit significance, and they also put some influence on the significance of some original variables. In my opinion, this is caused by the collinearity between the original variables and interactive variables.

In light of this observation, we opt to keep the original model without interaction terms, aiming to exclude the influence of these non-significant variables in the modelling process.

In conclusion, the answers to 2a and 2b will NOT potentially change if we included the interaction term in this model.