Problem Set 2

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.

```r
# load data
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/climateSupport.RData?raw=true"))
####
pset2_data<-climateSupport
head(pset2_data)
############################
# In the 'choice' column/variable contains logical values Not supported and Supported
# I Convert 'Supported' to 1 and 'Not supported' to 0, also I convert the countries and sanctions variables too.

pset2_data$choice <- ifelse(pset2_data$choice == "Supported", 1, 0)
pset2_data$countries <- factor(pset2_data$countries, ordered=FALSE)
pset2_data$sanctions <- factor(pset2_data$sanctions, ordered=FALSE)
#####
head(pset2_data)
View(pset2_data)
dim(pset2_data) # Rows/Observations= 8500; Columns/Variables=3
names(pset2_data) # "choice" ;   "countries" ; "sanctions"
##
with(pset2_data, table(pset2_data$choice)) # This line of the code provide the frequency of Not supported and Supported
#  0          1
# 4264     4236
```

   The given dataset has three variables such as: choice that need to be changed to binary where I replaced Not supported by zero and Supported by 1. After running R code using with() function to check the frequency table, I have found that 4264 or 50.16% don't support and 4236 or 49.84% support international environment agreement or policies people support (Bechtel and Scheve 2013).

```r
## Fit logistic regression model
fit_myFull_logit_supporting_policy <- glm (choice ~ countries + sanctions, family=binomial(link="logit"),
                                           data=pset2_data)
# Summary of the logistic regression model
summary(fit_myFull_logit_supporting_policy)
####
###Answer of Summary Output
###
#Call:
```

```
#  glm(formula = choice ~ countries + sanctions, family = binomial(
    link = "logit"),
#      data = pset2_data)
#
#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
#
```

Table 1: Coefficients

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.27266	0.05360	-5.087	3.64e-07 ***
countries 80 of 192	0.33636	0.05380	6.252	4.05e-10 ***
countries 160 of 192	0.64835	0.05388	12.033	¡ 2e-16 ***
sanctions 5%	0.19186	0.06216	3.086	0.00203 **
sanctions 15%	-0.13325	0.06208	-2.146	0.03183 *
sanctions 20%	-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

(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

Each predicted regression coefficient is statistically significant since their p-values in the last columns is less than 0.05 , (i.e., $p < 0.05$). Let recall the regression by using the R code:

```
coef(fit_myfull_logit_supporting_policy)
```

then it will produce the table bellow:

Table 2: Coefficients

Variable	**Coefficient**
(Intercept)	-0.2726631
countries80 of 192	0.3363609
countries160 of 192	0.6483497
sanctions5%	0.1918553
sanctions15%	-0.1332475
sanctions20%	-0.3035641

In a logistic regression, the response being modeled is the log(odds) that Y=1. The regression coefficients gives the change in log(odds) in the response for a unit change in the predictor variables holding all other predictor variables constant.Since log(odds) are difficult to interpret, I have decided to exponentiate them (i.e., estimated parameters) to put the results on odd scale, first I need to write this R code given by:

```
exp(coef(fit_myFull_logit_supporting_policy))
```

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

Table 3: Coefficients

I would like to explore the dataset in more detail and I am interested in finding the 95

An option for making a data.frame of confidence intervals and coefficients using R code:

confMod1 = data.frame (cbind (lower = exp (confint (fit_myFull_logit_supporting_policy) [, 1]) , co

Table 4: Coefficient Confidence Intervals

	Lower	**Coefs**	**Upper**
(Intercept)	0.6853331	0.7613492	0.8455914
countries80 of 192	1.2598455	1.3998442	1.5556530
countries160 of 192	1.7209643	1.9123823	2.1257297
sanctions5%	1.0725857	1.2114952	1.3685566
sanctions15%	0.7749409	0.8752484	0.9884554
sanctions20%	0.6535281	0.7381826	0.8336468

The table above provide the confidence interval regarding the estimated parameters. Now I am interested in Calculating the Global Null Hypothesis and p-value. The table above show the global null hypothesis: In summary, the rejection of

Table 5: Analysis of Deviance Table

Model	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(¿Chi)**
NULL	8499	11783	-	-	-
countries	2	146.724	8497	11637	< 2.2e-16 ***
sanctions	3	68.426	8494	11568	9.272e-15 ***

the global null hypothesis signifies that the predictors (countries and sanctions) collectively have a significant impact on the response variable (choice), as determined by the likelihood ration test. This implies that the model including these predictors provides better explanation of the response variable.

Extract p-value for the global null hypothesis by writing this r code:

```
my_global_null_p_value <- my_lr_test$"Pr(>Chi)"
print(my_global_null_p_value)
```

The R code above provide the following p-values:

- $p_{\text{value\_sanctions}} = 1.378383 \times 10^{-32}$
- $p_{\text{value\_countries}} = 9.271817 \times 10^{-15}$

Both p-values are less than 0.05 that indicate they are statistically significant.

(a) If any of the explanatory variables are significant in this model, then:

i. 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)

Question 2.
(2a) The coefficient for "sanctions" 15Calculate the odds ratio for sanctions 15

The general mathematical expression for the logistic regression model is given by:

$$\begin{aligned}\text{logit}\left(\frac{p}{1-p}\right) = &-0.2726631 \\ &+ 0.3363609 \times \text{countries}_{\text{80 of 192}} \\ &+ 0.6483497 \times \text{countries}_{\text{160 of 192}} \\ &+ 0.1918553 \times \text{sanctions}_{5\%} \\ &- 0.1332475 \times \text{sanctions}_{15\%} \\ &- 0.3035641 \times \text{sanctions}_{20\%}\end{aligned}$$

To answer question (2a) we need to extract from general equation the following:

$$\text{logit}\left(\frac{p_{5\%}}{1-p_{5\%}}\right) = -0.2726631 + 0.6483497 \times \text{countries}_{\text{160 of 192}} + 0.1918553 \times \text{sanctions}_{5\%}$$

$$\text{my\_5} = -0.2726631 + 0.6483497 \times 1 + 0.1918553 \times 1 = 0.5675419$$

$$\text{my\_15} = -0.2726631 + 0.6483497 - 0.1332475 = 0.2424391$$

Now I am going to calculate the difference between my_15 and my_5:

$$\text{diff\_my\_15\_and\_my\_5} = \text{my\_15} - \text{my\_5}$$

$$\text{diff\_my\_15\_and\_my\_5} = -0.3251028$$

$$\exp(\text{diff\_my\_15\_and\_my\_5}) = 0.7224531$$

That mean that 72.22

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

(2b)

For equation (2b), we have:

$$\text{logit}(p_{80}/(1-p_{80})) = -0.2726631 + 0.3363609 \cdot \text{countries}_{\text{80 of 192}}$$

Now, computing the value of my_80:

$$\text{my\_80} = -0.2726631 + 0.3363609 \times 1 = 0.0636978$$

Then, the probability prob_80 is:

$$\text{prob\_80} = \frac{1}{1 + \exp(-\text{my\_80})} = 0.5159191 \text{ or } 51.59191\%$$

That means that 51.59

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

$$\text{fit\_interaction\_logit} \leftarrow \text{glm}\,(\text{choice} \sim \text{countries} \times \text{sanctions}, \text{family} = \text{binomial}(\text{link} = \text{"logit"}), \text{data} = \text{pset2\_data})$$
$$\text{summary}(\text{fit\_interaction\_logit}) \quad (1)$$

The interaction logistic regression output table is given by:

Table 6: Coefficients

Variable	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 ***
sanctions5%	0.12179	0.10518	1.158	0.246909
sanctions15%	-0.09687	0.10822	-0.895	0.370723
sanctions20%	-0.25260	0.10806	-2.338	0.019412 *
countries80 of 192:sanctions5%	0.09471	0.15232	0.622	0.534071
countries160 of 192:sanctions5%	0.13009	0.15103	0.861	0.389063
countries80 of 192:sanctions15%	-0.05229	0.15167	-0.345	0.730262
countries160 of 192:sanctions15%	-0.05165	0.15267	-0.338	0.735136
countries80 of 192:sanctions20%	-0.19721	0.15104	-1.306	0.191675
countries160 of 192:sanctions20%	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

Reduced Model for logistics Regression Model:

$$\text{logit.fit.reduced} \leftarrow \text{step}(fit\_interaction\_logit) \quad (2)$$

The table for the output of the reduced logistic regression:

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

Table 7: Model Summary

Start	AIC	choice ~ countries * sanctions	
Df	Deviance	AIC	
- countries:sanctions	6	11568	11580
<none>		11562	11586
Step	AIC	choice ~ countries + sanctions	
Df	Deviance	AIC	
<none>		11568	11580
- sanctions	3	11637	11643
- countries	2	11715	11723

```
summary(logit.fit.reduced)
```

The Output Table is given by:

```
#Coefficients:
#                     Estimate S td. 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
```

From the see above I can see by doing the interactive logistic model regression did not improve the estimated parameter since we have some p values that are not statistically significant. Therefore by doing the reduced logistic regression I can find that all the predicted parameters are statistically significant. Conclusion: there is no need to perform the interactive logistic regression.

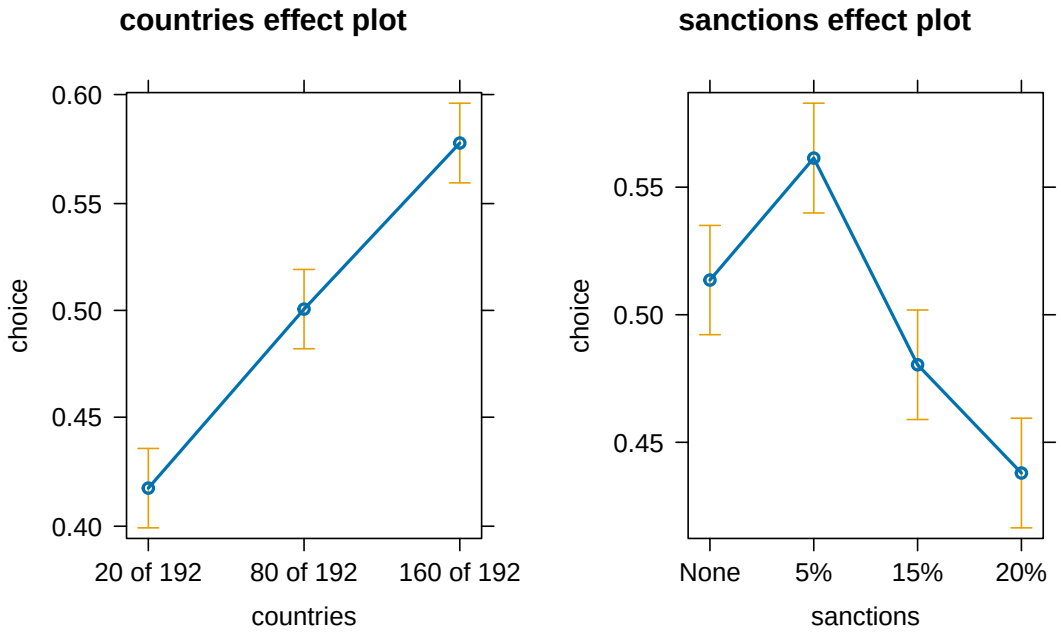

Figure 1: Extra work: Data Visualization - Plotting an interactive graphic using packages effects and plotly