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.

   Question 1:

```r
# load data
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/climateSupport.RData?raw=true"))
head(climateSupport)
# Converting the "choice" column into 1s and 0s
climateSupport$choice <- as.numeric(climateSupport$choice == "Supported")
head(climateSupport)
# checking for NAs
sum(is.na(climateSupport$sanctions))
sum(is.infinite(climateSupport$sanctions))
# transforming variables into numerical categories
climateSupport$countries_num <- as.numeric(sapply(strsplit(as.character(climateSupport$countries), "of"), "[", 1))
head(climateSupport$countries_num)
#
sanctions_mapping <- c("none" = 0, "5%" = 5, "15%" = 15, "20%" = 20)
climateSupport$sanctions_num <- sanctions_mapping[climateSupport$sanctions]
head(climateSupport$sanctions_num)
# and then running the glm()
formula <- choice ~ ns(countries_num) + ns(sanctions_num)
model_g <- glm(formula, data = climateSupport, family = binomial(link = "logit"))
summary(model_g)
#
Coefficients:
                   Estimate  Std. Error z value Pr(>|z|)
(Intercept)        -0.12687     0.04329  -2.931  0.00338 **
ns(countries_num)   0.80419     0.06688  12.025  < 2e-16 ***
ns(sanctions_num) -0.46328     0.06926  -6.689 2.25e-11 ***
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: 11593  on 8497  degrees of freedom
AIC: 11599Number of Fisher Scoring iterations: 4
```

Our null hypothesis is that none of the explanatory variables; number of participating countries and the sanctions imposed by it, have an effect on the support given to an international environmental agreement. For a 0.05

significance level, p-values smaller than 0.05 suggest that the explanatory variable is likely to have a significant effect on the outcome. The coefficients estimates indicate that there is a positive association between the number of countries participating in the agreement and it's support and a negative association between the sanctions and the support to an agreement. Or, a one unit increase(in this case the variables are categorical, not continuous) in the number of countries is associated with an average change of 0.08 in the log odds of the support taking on a value of 1. While a unit increase in the sanctions is associated with an average change of -0.046 in the log odds of the support taking on a value of 1. If X2 = null dev - res dev with p degrees of freedom. X2 = 11783 - 11593, with 2 degrees of freedom. Using a pvalue calculator (from the statology.org website) the p value is 0, which means that the model is useful

```
# Now I will fit a null model and use ANOVA to compare it to my model to check
# if what I said above holds
# fitting the model with no explanatory variables to get the null model
null_model <- glm(choice ~ 1, data = climateSupport, family = binomial(link = "logit"))
summary(null_model)
#
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.006588   0.021693  -0.304    0.761
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11783  on 8499  degrees of freedom
AIC: 11785Number of Fisher Scoring iterations: 3
#
anova_res <- anova(null_model, model_g, test = "Chisq")
print(anova_res)
#
 Analysis of Deviance Table
 Model 1: choice ~ 1
 Model 2: choice ~ ns(countries_num) + ns(sanctions_num)
      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
    1      8499      11783
    2      8497      11593  2   190.78 < 2.2e-16 ***
```

The pvalue from the ANOVA is very low, near zero, so I believe it supports the conclusion that my model is a better fit compared to the null 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)

Here scenario 1 is 160 out of 192 countries and 5 percent sanctions. While scenario 2 is 160 out of 192 with 15 percent sanctions.

```
# Defining scenario 1
scenario1 <- data.frame(countries_num = 160, sanctions_num = 5)
# Predicting the log odds of the scenario 1
logodds1 <- predict(model_g, newdata = scenario1, type = "link")
# Exponentiation the log odds to get odds ratio
odds1 <- exp(logodds1)
print(odds1)
#
1
1.529671
# scenario 2
scenario2 <- data.frame(countries_num = 160, sanctions_num = 15)
logodds2 <- predict(model_g, newdata = scenario2, type = "link")
odds2 <- exp(logodds2)
print(odds2)
#
1
1.270394
# Getting the coeffiecients from the model and extracting the
    coefficients for countries and sanctions
coef_summary <- summary(model_g)$coefficients
coef_countries <- coef_summary["ns(countries_num)", "Estimate"]
coef_sanctions <- coef_summary["ns(sanctions_num)", "Estimate"]
# calculating change in odds for each scenario
oddschange1 <- exp(coef_countries * (scenario1$countries_num -
    scenario2$countries_num) +                    coef_sanctions * (
    scenario1$sanctions_num - scenario2$sanctions_num))
print(oddschange1)
# 102.8021
#
oddschange2 <- exp(coef_countries * (scenario2$countries_num -
    scenario1$countries_num) +                        coef_sanctions * (
    scenario2$sanctions_num - scenario1$sanctions_num))
print(oddschange2)
# 0.00972743
```

The odds change from the 5 percent scenario to the 15 percent scenario is 102.802. Meaning that the odds of support for policies with 5 percent sanctions is 102.8 times higher than to policies with 15 percent sanctions, holding the number of countries constant. The first calculation were the original odds, the estimated odds for each scenario and the second the calculated change in odds, the multiplicative change in odds associated to the change in categories of sanctions, holding countries constant.

(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
# defining scenario 3
scenario3 <- data.frame(countries_num = 80, sanctions_num = 0)
# predicting the probability of support
prob_support <- predict(model_g, newdata = scenario3, type = "
    response")
print(prob_support)
# the probability is 0.537
```

The estimated probability is 0.537

(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.

Above in exercise 1, we already have a model without the interaction, we willl now fit a model with the interaction term

```r
#since we already have a model without the interaction we willl now
    fit a model with the interaction term
modelg_int <- glm(choice ~ ns(countries_num) * ns(sanctions_num),
    data = climateSupport, family = binomial(link = "logit"))
#performing  the test to compare models
anova_intest <- anova(modelg_int, model_g, test = "Chisq")
print(anova_intest)
# output
Analysis of Deviance Table
Model 1: choice ~ ns(countries_num) * ns(sanctions_num)
Model 2: choice ~ ns(countries_num) + ns(sanctions_num)
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       8496      11593
2       8497      11593 -1 -0.02172   0.8828
```

Sine the resulting pvalue is bigger than 0.05 it suggests that including the interaction term doesn't improve the fitness of the model.