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.

   Set assumptions
   Null hypothesis (H0) : The coefficient of the explanatory variable in the model is zero, that is, the number of participating countries and the degree of sanctions have no influence on the supporting policy.
   Alternative hypothesis (H1) : The coefficient of at least one explanatory variable in the model is non-zero, that is, the number of participating countries and the degree of sanctions have at least one influence on supporting policies.

```r
# Load the required libraries
library(mgcv)

# View data structure
str(climateSupport)
# 'data.frame': 8500 obs. of  3 variables:
#  $ choice    : Factor w/ 2 levels "Not supported",..: 1 1 1 1 1 1 2 1 2 1 ...
#  $ countries: Ord.factor w/ 3 levels "20 of 192"<"80 of 192"<..: 2 3 3 2 3 1 3 2 3 3 ...
#  $ sanctions: Ord.factor w/ 4 levels "None"<"5%"<"15%"<..: 3 3 1 3 2 3 2 4 2 4 ...


# Look at the first few lines of data
head(climateSupport)
# choice  countries sanctions
# 1  Not supported  80 of 192       15%
# 9  Not supported 160 of 192       15%
# 17 Not supported 160 of 192      None
# 25 Not supported  80 of 192       15%
# 33 Not supported 160 of 192        5%
# 41 Not supported  20 of 192       15%


# Convert ordered categorical variables to ordered factor variables
climateSupport$countries <- factor(climateSupport$countries, ordered = TRUE)
climateSupport$sanctions <- factor(climateSupport$sanctions, ordered = TRUE)

```

```r
# Fit the addition model, then easily provide summary output, and test the global null hypothesis
model <- glm(choice ~ countries + sanctions, data = climateSupport, family = binomial)

# Carry out global null hypothesis test and output the result
global_test <- anova(model, test = "Chisq")
global_test
# Analysis of Deviance Table

# Model: binomial, link: logit

# Response: choice

# Terms added sequentially (first to last)


# 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 ***
# ---
# Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1



# Extract the "choice" variable
choice <- climateSupport$choice

# Merge the "choice" variable with the converted dataframe
model_data_df <- cbind(choice, model_data_df)

# View model summary
summary(model)
# Call:
#   glm(formula = choice ~ countries + sanctions, family = binomial,
#       data = climateSupport)

# Coefficients:
#  Estimate Std. Error z value Pr(>|z|)
# (Intercept) -0.005665   0.021971  -0.258 0.796517
# countries.L  0.458452   0.038101  12.033  < 2e-16 ***
#   countries.Q -0.009950   0.038056  -0.261 0.793741
# sanctions.L -0.276332   0.043925  -6.291 3.15e-10 ***
#   sanctions.Q -0.181086   0.043963  -4.119 3.80e-05 ***
#   sanctions.C  0.150207   0.043992   3.414 0.000639 ***
#   ---
#   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



# Explain the coefficient of the significant variable
coef_summary <- coef(summary(model))
significant_vars <- coef_summary[coef_summary[, "Pr(>|z|)"] < 0.05, ]
significant_vars
# Estimate Std. Error   z value     Pr(>|z|)
# countries.L  0.4584525 0.03810109 12.032529 2.397037e-33
# sanctions.L -0.2763322 0.04392471 -6.291041 3.153443e-10
# sanctions.Q -0.1810859 0.04396287 -4.119065 3.804131e-05
# sanctions.C  0.1502066 0.04399173  3.414428 6.391603e-04



# Extract the p-value for each coefficient
p_values <- summary(model)$coefficients[, 4]
p_values
# (Intercept)  countries.L  countries.Q  sanctions.L  sanctions.Q
    sanctions.C
# 7.965174e-01 2.397037e-33 7.937408e-01 3.153443e-10 3.804131e-05
    6.391603e-04
```

Conclusion:

The global null hypothesis test results show that:
We reject the null hypothesis (p lt; 0.05), that is, the impact of these variables on supporting policies is significant.

Model summary results interpretation:
The coefficient estimates and standard errors in the model show the extent and uncertainty of each explanatory variables influence on the supporting policy.

The P-values of ”countries.L” and ”sanctions.L” are very small, much less than 0.05, indicating that the linear part of the number of participating countries and the degree of sanctions has a significant impact on the support policy.

”Sanctions.q” and ”sanctions.c” also have smaller p-values, indicating that the effects of the secondary and tertiary components of sanctions degree on support policies are also significant.

In summary, our model shows that the number of participating countries and the degree of sanctions have a significant impact on supportive policies, which provides important clues for us to understand and predict the tendency of supportive policies.

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)

```r
coef_summary
# Estimate Std. Error     z value      Pr(>|z|)
# (Intercept)              0.3144475 0.03816294    8.239605 1.727804e-16
# `countries20 of 192` -0.6483497 0.05388308 -12.032529 2.397037e-33
# `countries80 of 192` -0.3119888 0.05386949   -5.791568 6.973252e-09
# sanctions.L             -0.2763322 0.04392471   -6.291041 3.153443e-10
# sanctions.Q             -0.1810859 0.04396287   -4.119065 3.804131e-05
# sanctions.C              0.1502066 0.04399173    3.414428 6.391603e-04

# Extract the coefficient of sanctions
sanctions_coef <- coef(model)["sanctions.L"]
log_odds_change <- exp(0.1 * sanctions_coef) - 1
log_odds_change
# sanctions.L
# -0.02725491
```

   When the sanction level increases from 5 to 15 percent, the odds (ratio of support to non-support) of individual support for the policy is expected to increase log odds change to -0.02725491 In other words, the probability that individuals will support the policy will increase at this rate.

   (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
# Create a new data frame to predict
new_data <- data.frame(countries = "80 of 192", sanctions = "None")

# Predict the probability of supporting the policy
predicted_prob <- predict(model, newdata = new_data, type = "response")
predicted_prob
# 0.5159191
```

   Print results:
   When 80 countries participate and there are no sanctions, the probability of individual support for the policy is predicted to be 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.

Establish the hypothesis:
H0: There is no significant difference between models that include interaction terms and models that do not.
H1: There are significant differences between models that include interaction terms and those that do not.
The results show that according to the p-value (0.3912) of the hypothesis test, we cannot reject the null hypothesis, that is, the difference between the two models is not significant.

```
# Perform hypothesis testing of the interaction item
interaction_model <- glm(choice ~ countries * sanctions, data = climateSupport, family = binomial)

# Fit the first logistic regression model (excluding interaction terms)
model1 <- glm(choice ~ countries + sanctions, data = climateSupport, family = binomial)

# Fit the second logistic regression model (including interaction terms)
model2 <- glm(choice ~ countries * sanctions, data = climateSupport, family = binomial)

# Compare two models using anova function
anova_result <- anova(model1, model2, test = "Chisq")

# Print comparison results
print(anova_result)

# 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
# p-value: 0.3912
```

Result Explanation

Based on the results of hypothesis testing, we find that the difference between the Model that includes interaction terms (Model 2) and the model that does not include interaction terms (Model 1) is not significant (p gt; 0.05).
This means that, in this case, the model that includes the interaction terms does

not significantly account for more variation. Therefore, we can continue to answer questions 2a and 2b using a model that does not include interaction terms.

Conclusion:
Combining our hypothesis testing results, we conclude that the results obtained in answers 2a and 2b do not change significantly even if we include the interaction terms. Therefore, we can still trust the results of 2a and 2b without considering the interaction terms.
Since the p-value is 0.3912, we cannot reject the null hypothesis. This means that at the current level of significance, we do not have enough evidence to support the hypothesis that there is a significant difference between models that include interaction terms and models that do not.
Therefore, we can conclude that including interaction terms in this model does not significantly improve the ability to explain policy support.