Problem Set 2

Applied Stats II

Zhuo Zhang/23346227

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"))
# check the data
head(climateSupport)
summary(climateSupport)
str(climateSupport)
# Fit an additive model
library(mgcv)
#  Forced conversion from character vector to logical vector
as.logical(ifelse(climateSupport$choice == "Supported", 1, 0))
# Convert counties and sanctions to unordered factors
climateSupport$countries <- factor(climateSupport$countries, ordered
    = FALSE)
climateSupport$sanctions <- factor(climateSupport$sanctions, ordered
    = FALSE)
# check the data
str(climateSupport)
# Fit a logistic regression model
model <- glm(choice ~ .,
             data = climateSupport,
             family = "binomial")
# Display summary output
summary(model)
```

Result:

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

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

Number of Fisher Scoring iterations: 4
```

Intercept indicates that when all independent variables are 0, the estimated logarithmic probability of the dependent variable is -0.27266.
The estimated value of countries80 of 192 is 0.33636, indicating that when a country is one of 80 out of 192 countries, the logarithmic probability of the dependent variable increases by 0.33636.
The estimated value of countries 160 of 192 is 0.64835, indicating that when a country is one of 160 out of 192 countries, the logarithmic probability of the dependent variable increases by 0.64835.
The estimated value of sanctions5% is 0.19186, indicating that when the severity of sanctions is 5%, the logarithmic probability of the dependent variable increases by 0.19186.
The estimated value of sanctions15% is -0.13325, indicating that when the severity of sanctions is 15%, the logarithmic probability of the dependent variable decreases by 0.13325.
The estimated value of sanctions20% is -0.30356, indicating that when the degree of sanctions is 20%, the logarithmic probability of the dependent variable decreases by 0.30356.
From the results, it can be seen that the P-values of all coefficients are less than the commonly used significance level of 0.05. This means that all coefficients are significant, and we have sufficient evidence to reject the assumption that the coefficients are zero, meaning that their impact on the dependent variable is significant.

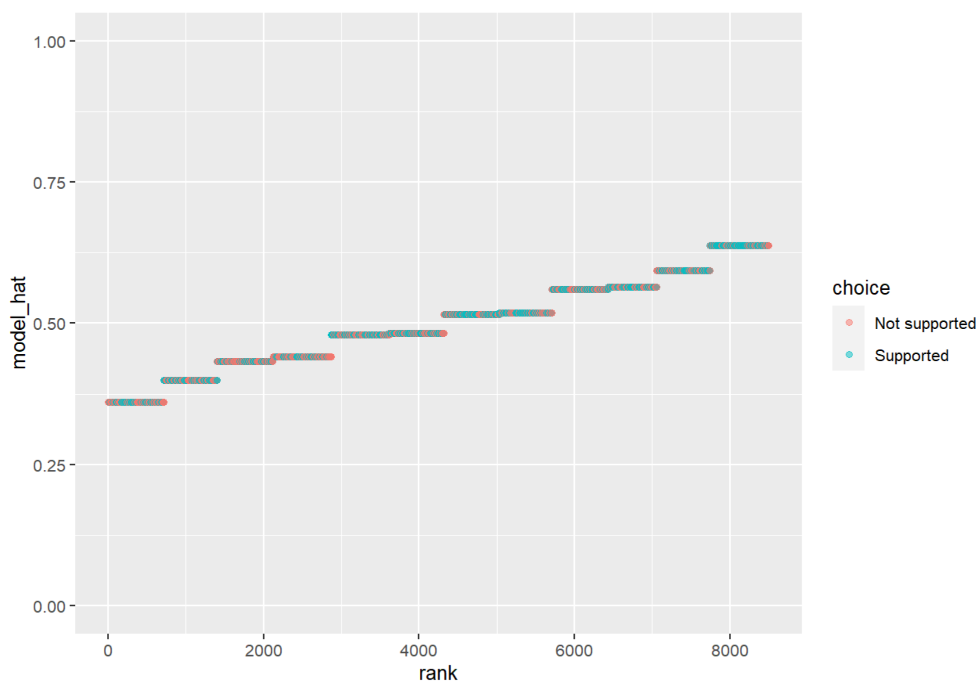
```r
#Testing the Global null Hypothesis and its p-value
#Testing the Global null Hypothesis
global_null_hypothesis <- summary(model)$null.deviance
cat("Global null hypothesis:", global_null_hypothesis, "\n")
# Obtain p-value
model_p_value <- summary(model)$coefficients[, "Pr(>|z|)"]
cat("Model p-value:", model_p_value, "\n")
# Obtain the number of independent variables in the model
df.null <- length(coef(model)) - 1
# Calculate the p-value of the global null hypothesis
global_null_p_value <- pchisq(summary(model)$null.deviance, df = df.
    null, lower.tail = FALSE)
cat("Global null hypothesis p-value:", global_null_p_value, "\n")
```

Result:
Global null hypothesis: 11783.41
Model p-value: 3.635432e-07 4.051815e-10 2.397037e-33 0.002025651 0.0318345 1.014753e-06
Global null hypothesis p-value: 0
In this model, the global null hypothesis is that all coefficients of the explanatory variables are equal to zero, indicating no influence of any explanatory variable on the outcome.
According to the provided results, the p-value for the global null hypothesis is 0,

indicating that we can reject the global null hypothesis.

```
#plot
# create dataframe
predicted_data <- data.frame(
  choice = climateSupport$choice,
  model_hat = model$fitted.values,
  model_interaction_hat = model_interaction$fitted.values
)

# Reorder and Draw
predicted_data %>%
  arrange(model_hat) %>%
  mutate(rank = row_number()) %>%
  ggplot(aes(rank, model_hat)) +
  geom_point(aes(colour = choice), alpha = 0.5) +
  scale_y_continuous(limits = c(0,1))
```



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)

```
#2a:
# Extract coefficients from the model
```

```
coefficients <- coef(model)
# Extracting coefficients related to interaction terms
interact_coefficients <- coefficients[grep("countries160 of 192:sanctions", names(coefficients))]
# Calculate the probability logarithm at 5% and 15% sanction levels
log_odds_5percent <- coefficients["(Intercept)"] + coefficients["countries160 of 192"] + coefficients["sanctions5%"]
log_odds_5percent
log_odds_15percent <- coefficients["(Intercept)"] + coefficients["countries160 of 192"] + coefficients["sanctions15%"]
log_odds_15percent
# Calculate the change in probability
odds_change <- exp(log_odds_15percent - log_odds_5percent)
# Print results
cat("As sanctions increase from 5% to 15%, the probability of individuals supporting policies changes as follows:", odds_change, "\n")
```

Result:
As sanctions increase from 5% to 15%, the probability of individuals supporting policies changes as follows: 0.7224531

(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:Calculate the estimated probability
log_odds_80_no_sanctions <- coef(model)['(Intercept)'] + coef(model)['countries80 of 192']
prob_80_no_sanctions <- exp(log_odds_80_no_sanctions) / (1 + exp(log_odds_80_no_sanctions))
cat("Estimated probability of 80 countries without sanctions:", prob_80_no_sanctions, "\n")
```

Result:
Estimated probability of 80 countries without sanctions: 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.

```
# 2c:Test interaction
model_interaction <- glm(choice ~ countries*sanctions,
                         data = climateSupport,
                         family = "binomial")
summary(model_interaction)
# Use anova() to compare models or view AIC/BIC
anova(model, model_interaction, test="Chisq")
```

Result:

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

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 ***
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
```

```
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
```

The chi-square statistic between the two models is 6.29, with 6 degrees of freedom and a p-value of 0.39. This p-value suggests that we cannot reject the possibility that the difference between the two models is due to randomness, and therefore it can be argued that the inclusion of an interaction term in the model is unnecessary. So in this case it is not appropriate to add an interaction term and therefore there will be no change in the answers for 2a and 2b.