Problem Set 2

Yajie Dong

Due: February 18, 2024

Question 1

Please answer the following questions:

Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.

Step 1:

Load and explore the Dataset: Use the provided dataset 'climateSupport.RData' to load the data into R. The dataset contains observations from 8500 individuals, with the response variable being 'choice' (1 if the individual supports the policy, 0 otherwise), and explanatory variables being 'countries' (number of participating countries) and 'sanctions' (level of sanctions for missing emission targets).

Let's do that by R:

```
#  set a seed for reproducibility when doing simulations
set.seed(123)
# Load necessary libraries
library(tidyverse)  # For data manipulation and visualization
library(broom)      # For tidying up model outputs
library(lme4)       # For fitting generalized linear mixed models (if needed)
library(ggplot2)    # For plotting
# Load the dataset
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/
    climateSupport.RData?raw=true"))
# Explore the dataset (Optional)
View(climateSupport)
summary(climateSupport)
# Use the summary function and capture the output
summary_output <- capture.output(summary(climateSupport))

```

Choice	Countries	Sanctions
Not Supported	20 of 192: 4264	None: 2119
	80 of 192: 2865	5%: 2133
	160 of 192: 2840	15%: 2111
Supported		20%: 2137

Table 1: Summary of climateSupport

From the data view R output we could know:

- The `climateSupport` dataset has 8500 observations, with 4264 not supporting and 4236 supporting the policy.
- The `countries` variable has three levels: "20 of 192", "80 of 192", and "160 of 192", indicating the number of countries participating in the agreement.
- The `sanctions` variable has four levels: "None", "5%", "15%", and "20%", representing the possible sanctions for non-compliance.

Step 2:

Fit an Additive Model: We'll need to **fit a logistic regression model since the response variable is binary.** The model will assess how 'countries' and 'sanctions' affect the likelihood of an individual supporting the policy.

Let' do that in R:

```r
# Ensure that 'countries' and 'sanctions' are factors with the correct levels
climateSupport$countries <- as.factor(climateSupport$countries)
climateSupport$sanctions <- as.factor(climateSupport$sanctions)
# Fit the logistic regression model
model <- glm(choice ~ countries + sanctions, data = climateSupport, family =
    binomial())
# Summary of the model to examine coefficients and significance
summary(model)

```

Output from R:

Table 2: Summary of the GLM Model

Term	Estimate	Std. Error	z value	Pr(¿—z—)
(Intercept)	-0.005665	0.021971	-0.258	0.7965
countries.L	0.458452	0.038101	12.033	< 2e-16 ***
countries.Q	-0.009950	0.038056	-0.261	0.7937
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

Step 3:Analysis of Logistic Regression Model

The logistic regression model aims to predict the likelihood of an individual supporting an environmental policy based on two main factors: the number of countries participating in an international agreement and the sanctions for not adhering to the agreement.
Based on the model summary provided, the logistic regression equation can be formulated with the given coefficients as:

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot X_{\text{countries.L}} + \beta_2 \cdot X_{\text{countries.Q}} + \beta_3 \cdot X_{\text{sanctions.L}} + \beta_4 \cdot X_{\text{sanctions.Q}} + \beta_5 \cdot X_{\text{sanctions.C}})}}$$

Where:

- $\beta_0 = -0.005665$ (Intercept)
- $\beta_1 = 0.458452$ (Coefficient for countries.L)
- $\beta_2 = -0.009950$ (Coefficient for countries.Q)
- $\beta_3 = -0.276332$ (Coefficient for sanctions.L)
- $\beta_4 = -0.181086$ (Coefficient for sanctions.Q)
- $\beta_5 = 0.150207$ (Coefficient for sanctions.C)

Each β coefficient represents the log odds change associated with a one-unit increase in the corresponding predictor variable, holding other variables constant. To interpret these coefficients in terms of odds ratios, which describe the factor by which the odds of support increase or decrease, exponentiate the coefficients:

- Odds Ratio for countries.L: $e^{0.458452}$

- Odds Ratio for countries.Q: $e^{-0.009950}$
- Odds Ratio for sanctions.L: $e^{-0.276332}$
- Odds Ratio for sanctions.Q: $e^{-0.181086}$
- Odds Ratio for sanctions.C: $e^{0.150207}$

1. Model Summary:

The logistic regression model, which includes 'choice' as the binary response variable (1 for support, 0 for no support), shows that the linear term for the number of participating countries ('countries.L') and all terms for sanctions ('sanctions.L', 'sanctions.Q', 'sanctions.C') have significant p-values, indicating a statistically significant effect on the probability of policy support. The 'countries.Q' term is not significant, suggesting a linear rather than a quadratic relationship between the number of participating countries and policy support. The negative coefficients for 'sanctions.L' and 'sanctions.Q' indicate that higher sanctions are associated with a lower likelihood of policy support. However, the positive coefficient for 'sanctions.C' introduces a non-linear component, suggesting a more complex relationship as sanctions increase.

2. Global Null Hypothesis and P-value:

- The null hypothesis ($H_0 : \beta_1 = \beta_2 = \ldots = 0$) suggests that none of the explanatory variables have an effect on the likelihood of policy support.
- The significant reduction in deviance from the null model (11783 to 11568) upon adding the explanatory variables indicates an improved fit of the model.
- The highly significant p-value for 'countries.L' ($p < 2 \times 10^{-16}$) leads us to reject the null hypothesis, affirming the influence of both the number of participating countries and the level of sanctions on policy support.

3. Conclusion:

- The number of countries participating in the agreement has a positive and significant influence on the likelihood of an individual supporting the policy, with a greater effect observed as the number of participating countries increases from 20 to 160.
- Sanctions for non-compliance exhibit a complex relationship with policy support, initially reducing the likelihood of support as they increase but showing a nuanced pattern due to the cubic term.
- The model elucidates the significance of international cooperation and enforcement mechanisms in shaping individual support for environmental policies.

4.Visualization:

Let's do that by R:

```r
#Visualization
# Assuming 'model' is our fitted logistic regression model
coefficients_df <- broom::tidy(model)
ggplot(coefficients_df, aes(x = term, y = estimate, fill = p.value < 0.05)) +
  geom_col() +
  geom_errorbar(aes(ymin = estimate - std.error, ymax = estimate + std.error),
    width = 0.2) +
  labs(title = "Effect Sizes of Predictors on Policy Support", y = "
    Coefficient Estimate", x = "") +
  scale_fill_manual(name = "Significance", values = c("TRUE" = "steelblue", "
    FALSE" = "grey"), labels = c("TRUE" = "Significant", "FALSE" = "Not
    Significant")) +
  theme_minimal()

# Generate a new dataset for predictions
new_data <- expand.grid(countries = levels(climateSupport$countries),
    sanctions = levels(climateSupport$sanctions))
new_data$predicted_prob <- predict(model, newdata = new_data, type = "response
    ")

# Plot
ggplot(new_data, aes(x = countries, y = predicted_prob, color = sanctions,
    group = sanctions)) + # Added 'group = sanctions' here
  geom_line() +
  geom_point() +
  labs(title = "Predicted Probability of Policy Support", y = "Predicted
    Probability", x = "Number of Participating Countries") +
  theme_minimal() +
  scale_color_brewer(palette = "Set1")

```

The plots from the R output are as follows:

The provided plots reinforce the conclusions drawn from the model. The 'Effect Sizes of Predictors on Policy Support' plot visualizes the impact of each predictor on the response variable, highlighting the significance of the number of countries and sanctions. The 'Predicted Probability of Policy Support' plot illustrates the predicted probabilities of support across different levels of participating countries and sanctions, offering a clear visual interpretation of the model's predictions.

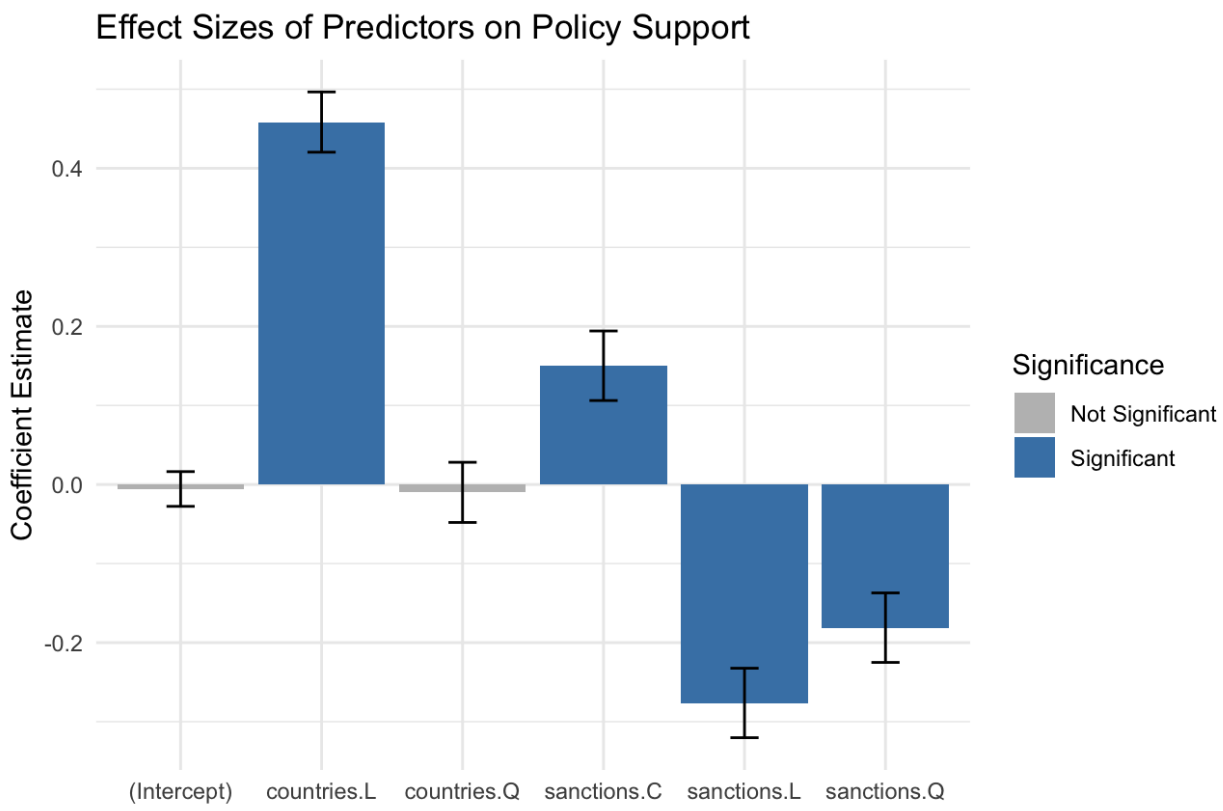

Figure 1: Effect Sizes of Predictors on Policy Support

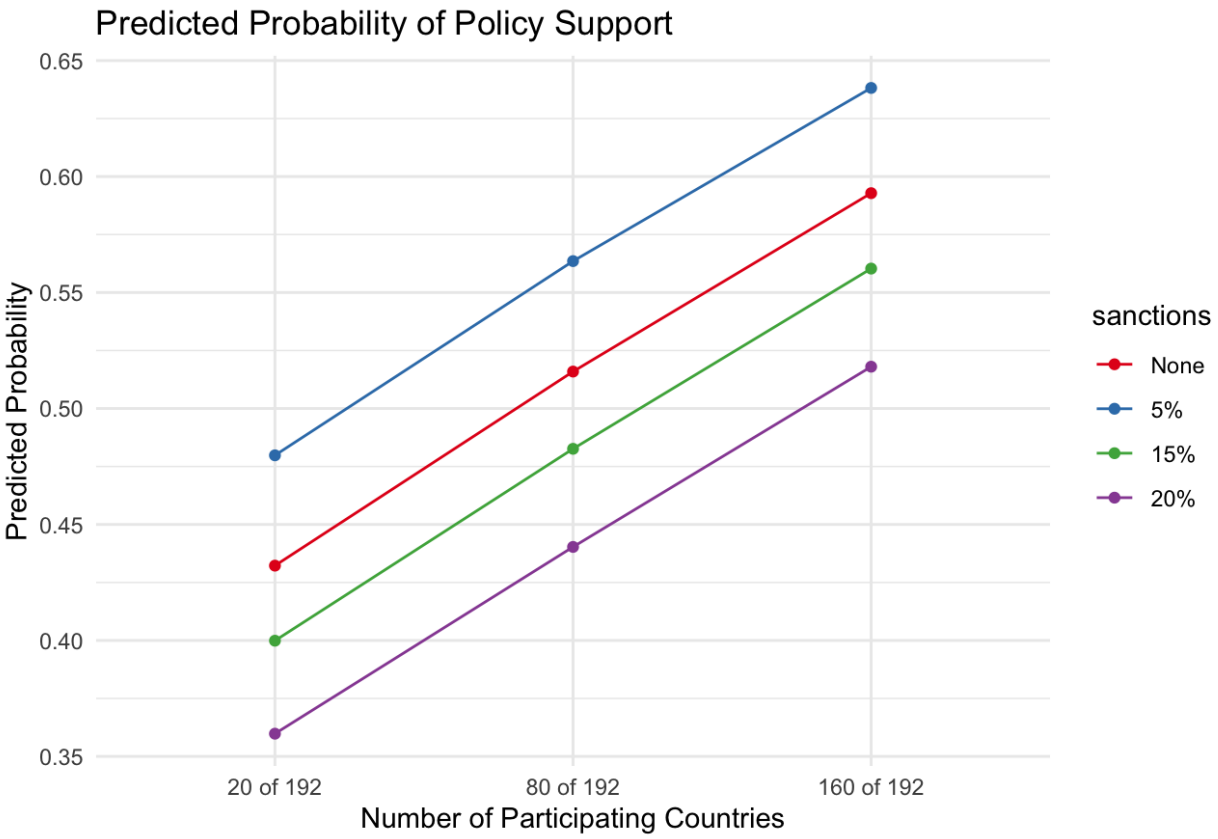

Figure 2: Predicted Probability of Policy Support

Question 2:If any of the explanatory variables are significant in this model, then:

1. **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)**

   The odds ratio can be calculated as follows:

$$\textbf{Odds Ratio} = \frac{\textbf{Odds of support at 15\% sanction}}{\textbf{Odds of support at 5\% sanction}}$$

   Let's calculate odds ratio for sanctions increase from 5% to 15% by R:

```
#2(a)
# Calculate odds ratio for sanctions increase from 5% to 15% with nearly
    full participation
# This needs to match the model's coefficient names.
```

```r
# First, we need to determine the coding used for the levels of sanctions
# For example, let's say that we have the following coding:
# None - 0, 5% - 1, 15% - 2, 20% - 3
# We need to calculate the predicted log odds for sanctions at 5% and 15%

# Data frame for prediction focusing on 160 of 192 countries participating
predict_data <- data.frame(
  countries = factor(rep("160 of 192", 2), levels = c("20 of 192", "80 of 192", "160 of 192")),
  sanctions = factor(c("5%", "15%"), levels = c("None", "5%", "15%", "20%"))
)

# Calculate predicted log odds for the sanctions levels
predicted_log_odds <- predict(model, newdata = predict_data, type = "link")

# The odds at 5% sanctions level
odds_5 <- exp(predicted_log_odds[1])

# The odds at 15% sanctions level
odds_15 <- exp(predicted_log_odds[2])

# Calculate the odds ratio
odds_ratio <- odds_15 / odds_5
print(paste("Odds Ratio for sanctions increase from 5% to 15%: ", odds_ratio))

```

The calculated odds ratio is shown below:

```
[1] "Odds Ratio for sanctions increase from 5% to 15%:  0.722453082248423"
```

The odds ratio of 0.7224531 indicates that for policies where nearly all countries participate (160 of 192), increasing sanctions from 5% to 15% is associated with a decrease in the odds of an individual supporting the policy by about 27.75% ($1 - 0.7224531$). This suggests that higher sanctions may reduce support for the policy among individuals when a large number of countries are participating.

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

Let's calculate it by R:

```r
# Create new data for prediction
new_data_80_none <- data.frame(countries = factor("80 of 192", levels = levels(climateSupport$countries)),
```

```
                                    sanctions = factor("None", levels = levels
    (climateSupport$sanctions)))

# Predict probability
prob_support_80_none <- predict(model, newdata = new_data_80_none, type =
    "response")
prob_support_80_none

```

The estimated probability result from R we get:0.5159191

The predicted probability of 0.5159191 suggests that when 80 of 192 countries are participating and there are no sanctions, an individual has about a 51.59% chance of supporting the policy. This indicates a **slightly higher** than even chance of support in this scenario.

3. **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.

The significance of the interaction term can be tested with an ANOVA between the model without the interaction term and one with it. This is performed to see if the model with the interaction term fits the data significantly better.
Now let's do that by R:

```
#2(c)
# Fit a model with interaction terms
interaction_model <- glm(choice ~ countries * sanctions, data = climateSupport
    , family = binomial())

# ANOVA to compare models
anova_test <- anova(model, interaction_model, test = "Chisq")
anova_test
```

The result from R output:

Table 3: Analysis of Deviance Table

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(¿χ^2)
1: choice ~ countries + sanctions	8494	11568			
2: choice ~ countries * sanctions	8488	11562	6	6.2928	0.3912

The ANOVA test comparing the model without the interaction term (Model 1) to the model with the interaction term (Model 2) shows a p-value of 0.3912 for the addition of the interaction terms. This p-value is not statistically significant at the common alpha level of 0.05, suggesting that the interaction between the number of participating countries and the level

of sanctions does not significantly improve the model's fit to the data. Therefore, based on this analysis, including the interaction term in the model may not be necessary, and the answers to parts 2(a) and 2(b) are unlikely to change significantly if the interaction term were included.

This comprehensive analysis provides valuable insights into how the number of participating countries and the level of sanctions influence individual support for the policy. It also highlights the importance of testing for interaction effects to understand if the influence of one predictor on the outcome depends on the level of another predictor. However, in this case, the interaction does not appear to be significant.