Problem Set 2

Ana Sofia Jesus 19327602

18/02/2024

Question 1

After importing the data and before creating an additive model, I converted the explanatory variables (countries and sanctions) into non-ordered factors, to ensure they are appropriately formatted for analysis.

```
climateSupport$countries <- factor(climateSupport$countries, ordered = FALSE)
climateSupport$sanctions <- factor(climateSupport$sanctions, ordered = FALSE)
```

After recoding the variables, the logistic regression model is fitted using the glm() function to fit an additive model to predict the likelihood of an individual supporting a policy based on the number of countries participating and the possible sanctions for non-compliance.

```
model1 <- glm(choice ~ countries + sanctions, data = climateSupport, family = binomial
(link = "logit"))
```

I have obtained the following output:

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

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.08081    0.05316  -1.520  0.12848
countries80 of 192   0.33636    0.05380   6.252 4.05e-10 ***
countries160 of 192  0.64835    0.05388  12.033  < 2e-16 ***
sanctionsNone       -0.19186    0.06216  -3.086  0.00203 **
sanctions15%        -0.32510    0.06224  -5.224 1.76e-07 ***
sanctions20%        -0.49542    0.06228  -7.955 1.79e-15 ***
---
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
```

The intercept term (-0.08081) represents the estimated log odds of the response variable when all predictor variables are at their reference levels. The coefficients for the predictor variables represent the difference in log odds of the response variable between each category and the reference category, holding all other predictors constant.
The overall output suggests that the number of participating countries and the severity of sanctions both

significantly influence individuals' likelihood to support the environmental policy. Specifically, as the number of participating countries increases and the severity of sanctions increases, the odds of supporting the policy also increase.
In the following step, a Global Null Hypothesis Test is performed to evaluate whether the model with all the predictors provides a better fit to the data compared to a reduced model with no predictors.
The null hypothesis for this test predicts that all coefficients of the explanatory variables (countries and sanctions) in the logistic regression model are equal to zero, implying that these variables do not significantly improve the model's fit nor do they contribute to explaining the variation in the response variable (choice).

In the following code, a null model is fitted using the glm() function with only an intercept term.

```
null_mod <- glm(choice ~ 1, data = climateSupport, family = binomial())
```

Finally, an ANOVA test is conducted to compare the fitted model (model1) with the null model. The ANOVA test examines whether adding the explanatory variables (countries and sanctions) significantly improves the model fit compared to just having an intercept term. Both chi-squared and likelihood ratio tests are performed (test = "Chisq" and test = "LRT"), which are equivalent and provide consistent results.

```
anova(null_mod, model1, test = "Chisq")
anova(null_mod, model1, test = "LRT")

Analysis of Deviance Table

Model 1: choice ~ 1
Model 2: choice ~ countries + sanctions
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499      11783
2      8494      11568  5   215.15 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test produced a p-value very close to zero and below the critical value of 0.05, suggesting that we can reject the null hypothesis that the variables (countries and sanctions) do not increase the model fit. In other words, the inclusion of these two predictors in this model contribute significantly to explaining the variation in the response variable.

Question 2

0.1 a.

In order to answer this question I first identified the relavant coefficient: -0.32510
Based on this, increasing sanctions from 5% to 15%, on average, decreases the log odds that an individual will support the policy by 0.33, for all number of participating countries.

0.2 b.

In order to calculate the estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions, I began by defining the conditions for the probability of interest:

```
countries_condition <- "80 of 192"
sanctions_condition <- "None"
```

In the following step, I have created a data frame to store the specified conditions while using the factor() function to convert the conditions to factor variables, ensuring that they match the levels of the corresponding variables in the original dataset. Then, I have predicted the probability using the predict() function and applying it to the logistic regression (model 1).

```
conditions_df <- data.frame(countries = factor(countries_condition,
    levels = levels(climateSupport$countries)),
    sanctions = factor(sanctions_condition,
    levels = levels(climateSupport$sanctions)))

estimated_probability <- predict(model1, newdata = conditions_df, type = "response")
```

I have obtained the following estimated probability:

```
> estimated_probability
        1
0.5159191
```

The above output indicates that the estimated probability of supporting the policy under the specified conditions (80 of 192 countries participating and no sanctions) is approximately 0.5159, or 51.59%.

0.3 c.

In order to assess if a model including an interaction term would have a better fit than model 1, I have applied the same method as the one used for Question 1.
I have fitted a logistic regression model (model_interaction) that includes an interaction term between the variables countries and sanctions. The interaction term allows to examine whether the effect of one predictor on the response variable (choice) varies depending on the level of another predictor.
I began by fitting a model with the interaction term:

```
model_interaction <- glm(choice ~ countries * sanctions, data = climateSupport,
family = binomial(link = "logit"))
```

Next, I performed a Likelihood Ratio Test to compare the goodness of fit between the model without the interaction term (model1) and the model with the interaction term (model_interaction). I have obtained the following output.

```
> anova(model1, model_interaction, test = "Chisq")

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 p-value for the LRT (0.3912) is higher than the critical significance level of 0.05. We are then unable to reject the null hypothesis according to which including the interaction term does not significantly improve the model's fit compared to the model without it. This suggests that there is no significant improvement in model fit when including the interaction term compared to the model without it.
Therefore, based on this test result, including the interaction term may not be useful as it does not significantly change the model's explanatory power in predicting support for the policy. To conclude, the answers to questions 2a and 2b are unlikely to change based on the inclusion of the interaction term.