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.

```r

# load data
load(url("https://github.com/ASDS-TCD/StatsII_Spring2024/blob/main/datasets/climateSupport.RData?raw=true"))
model <- glm(choice ~ countries + sanctions, data=climateSupport, family=binomial(link="logit"))

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

# summary of the model
model_summary <- summary(model)

# conduct the ANOVA test
model_anova <- anova(model, test="Chisq")
```

   (a) model summary output:

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

   Deviance Residuals: Min 1Q Median 3Q Max -1.4259 -1.1480 -0.9444 1.1505 1.4298

   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

(b) ANOVA output 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

(c) Result and Conclusion;

glm() function in R is used to fit generalized linear models (GLMs), response variable (choice) is binary; Logistic regression is a type of GLM, the log-odds of the probability of the outcome is modeled as a linear combination of the predictor variables. The overall conclusion from the model summary output is both the number of participating countries and the structure of sanctions have a significant impact on the support for the policy. The likelihood of supporting policy tends to increase with the number of countries participating increase (countries. L: coefficient is 0.458, a statistically significant p-value:2e-16) The severity of sanctions increases with the likelihood of an individual supporting the policy decreases.(sanctions.L: coefficient is -0.276332. the p-value is statistically significant 3.15e-10)

The ANOVA function with test="Chisq" is used to perform a global null hypothesis test, by comparing two models to see if the full model with the predictors provides a significantly better fit to the data than the null model. The full model includes both the countries and sanctions variables as predictors of the choice outcome. The null model, which is a reduced model that includes only the intercept (no predictors)

Both country and sanctions variables significantly influence the binary choice outcomeWhen the country variable is included in the logistic regression model, it significantly improves the model's ability to predict the binary choice outcome. The reduction of residual deviance by 146.724 points towards a substantial improvement in the fit of the model, The p-value 2.2e-16 provides statistical evidence that this improvement is highly significant; The sanctions variable reduces the residual deviance by 68.426, with a p-value of 9.272e-15, also showing a significant effect but to a lesser degree compared to the country variable.

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)

The coefficient for sanctions.L is -0.276. This coefficient means the average change in the log odds of an individual supporting the policy for a one-unit change in sanctions if other variables are constant. Change from 5% to 15% sanctions, let's assume 10% increase corresponds to a one-unit increase in the sanctions variable. This increase would result in a decrease in the log odds of supporting the policy by the coefficient value multiplied by the change in units (-0.276 * 1 unit)

(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

intercept <- -0.005665
coef_L <- 0.458452
coef_Q <- -0.009950


# proportional to the number of countries
L_value <- 80 / 192
Q_value <- L_value^2

log_odds = intercept + (coef_L * L_value) + (coef_Q * Q_value)

# Convert log-odds to probability
probability = exp(log_odds) / (1 + exp(log_odds))
print(probability)
#0.5457787
```

Using the coefficients from the part 1 logistic regression model, sanctions are at 0; calculation base on L represents a direct scaling of the number of countries, Q represents the square of the linear term

(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.

```r

# Fit model with interaction
model_with_interaction <- glm(choice ~ countries * sanctions,
    family=binomial(link="logit"), data=climateSupport)

# Compare models
anova(model, model_with_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
>
```

There is no need to include the interaction term in the final model.
Model 1 without interaction term; Model 2 with interaction term. For Model 1, there are 8494 residual degrees of freedom, and for Model 2, there are 8488. The decrease in degrees of freedom by 6 from Model 1 to Model 2 indicates that the interaction term adds 6 parameters to the model. Model 1 has a residual deviance of 11568, and Model 2 has a lower residual deviance of 11562, the interaction term may provide a better fit for the data; The p-value is 0.39. it higher than the significance level of 0.05, which means no statistically significant evidence that the interaction term (between countries and sanctions) improves the model fit for explaining the support for the policy.(response variable: choice)