# 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.

   First, I inspect the dataset to build familiarity. I note all variables are factors, with the two explanatory variables being ordinal (ie ordered) and levels are logically ordered (meaning I do not need to re-order them for analysis). I also confirm the variable names for future reference and check for any NAs which may need to be addressed (the dataset includes no NAs).

```
load(url(”https://github.com/ASDS–TCD/StatsII_Spring2024/blob/main/
    datasets/climateSupport.RData?raw=true”))

### Inspect dataset
head(climateSupport) # Check overall values of dataset
str(climateSupport) # Check shape, and names and classes of variables
summary(climateSupport) # Check summary stats ie frequencies of values
lapply(climateSupport, class) # Confirm explanatory variables are ordered
     factors
lapply(climateSupport, unique) # Check levels for each variable
levels(climateSupport$countries) # Check level ordering
levels(climateSupport$sanctions) # Check level ordering
```

   Next, I fit an additive model `CSmodel` using the ”binomial(link - logit)” parameter in order to build the model using log-odds:

   (a) `Constant` indicates the log-odds of a an individual's support for a policy given it has 20 of 192 countries supporting it AND no sanctions for noncompliance. However, this effect has a p-value above 0.05, indicating this result is not statistically significant.

   (b) `countries.L` indicates the average partial effect of a policy having 80 of 192 countries supporting it AND no sanctions for noncompliance upon the log-odds of an individual's preference (choice) for it. In other words, all other factors remaining constant, increasing the number of countries supporting a policy from 20 to 80 has an average increase of 0.458 upon the log-odds of an individual supporting said policy. This effect has a p-value below 0.05, indicating this result is statistically significant. Note this is the estimated linear function.

   (c) `countries.Q` indicates the average partial effect of a policy having 160 of 192 countries supporting it AND no sanctions for noncompliance upon an individual's preference (choice) for it. In other words, all other factors remaining constant, increasing the number of countries supporting a policy from 20 to 160 has an average decrease of 0.01 upon the log-odds of an individual supporting said policy. However, this effect has a p-value above 0.05, indicating this result is not statistically significant. Note this is the estimated quadratic function.

(d) `sanctions.L` indicates the average partial effect of a policy having 20 of 192 countries supporting it AND sanctions of 5% of the participating country's monthly household costs (given 2% GDP growth) upon an individual's preference (choice) for it. In other words, all other factors remaining constant, increasing sanctions for noncompliance from 0% to 5% has an average decrease of 0.276 upon the log-odds of an individual supporting said policy. This effect has a p-value below 0.05, indicating this result is statistically significant. Note this is the estimated linear function.

(e) `sanctions.Q` indicates the average partial effect of a policy having 20 of 192 countries supporting it AND sanctions of 15% of the participating country's monthly household costs (given 2% GDP growth) upon an individual's preference (choice) for it. In other words, all other factors remaining constant, increasing sanctions for noncompliance from 0% to 15% has an average decrease of 0.181 upon the log-odds of an individual supporting said policy. This effect has a p-value below 0.05, indicating this result is statistically significant. Note this is the estimated quadratic function.

(f) `sanctions.C` indicates the average partial effect of a policy having 20 of 192 countries supporting it AND sanctions of 20% of the participating country's monthly household costs (given 2% GDP growth) upon an individual's preference (choice) for it. In other words, all other factors remaining constant, increasing sanctions for noncompliance from 0% to 20% has an average increase of 0.150 upon the log-odds of an individual supporting said policy. This effect has a p-value below 0.05, indicating this result is statistically significant. Note this is the estimated cubic function.

```r
### Fit an additive model with response ("choice") and predictors ("countries", "sanctions")
CSmodel <- glm(choice ~ countries + sanctions,
               data = climateSupport,
               family = binomial(link = "logit"))
# Inspect new logistical model details
CSmodel_summary <- summary(CSmodel) # Review summary statistics
CSmodel_table <- stargazer(CSmodel,
          type = "latex",
          title = "Logistic Regression Model",
          align = TRUE)
```

Table 1: Logistic Regression Model

| | *Dependent variable:* |
|---|---|
| | choice |
| countries.L | $0.458^{***}$ |
| | (0.038) |
| countries.Q | $-0.010$ |
| | (0.038) |
| sanctions.L | $-0.276^{***}$ |
| | (0.044) |
| sanctions.Q | $-0.181^{***}$ |
| | (0.044) |
| sanctions.C | $0.150^{***}$ |
| | (0.044) |
| Constant | $-0.006$ |
| | (0.022) |
| Observations | 8,500 |
| Log Likelihood | -5,784.130 |
| Akaike Inf. Crit. | 11,580.260 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Next, I conduct the chi-square test to assess whether the partial effect of at least one explanatory variable is likely to be non-zero. Analysis shows the p-values for the coefficients ($\hat{\beta}$) for `countries` and `sanctions` are below 0.05 and therefore give evidence to reject the null hypothesis (no explanatory variables have any effect) and evidence to support the alternative hypothesis (at least one explanatory variable has a non-zero effect).

$$H_0 : \hat{\beta}_{\texttt{countries}} + \hat{\beta}_{\texttt{sanctions}} = 0$$
$$H_a : \hat{\beta}_{\texttt{countries}} + \hat{\beta}_{\texttt{sanctions}} \neq 0$$

Table 2: Analysis of Deviance: CSmodel and Chi-Squared Test

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Df | 2 | 2.500 | 0.707 | 2 | 3 |
| Deviance | 2 | 107.575 | 55.365 | 68.426 | 146.724 |
| Resid. Df | 3 | 8,496.667 | 2.517 | 8,494 | 8,499 |
| Resid. Dev | 3 | 11,662.780 | 109.924 | 11,568.260 | 11,783.410 |
| Pr(>Chi) | 2 | 0.000 | 0.000 | 0 | 0 |

```r
# First, using Chi-Squared test
global_null_chisq <- anova(CSmodel, test = "Chisq")
print(global_null_chisq)
```

However, we can explicitly use the `likelihood ratio test` (LRT) to compare the full model `CSmodel` to a nested model `CSmodel_null` to assess whether adding the given explanatory variables improves the fit of the model. The null hypothesis in this case is that the ”null” model is a sufficient ’fit’ for the observed data, and that there is no improvement in fit by adding at least one of the selected explanatory variables.

In this case, we again see the p-value is below 0.05, giving evidence to reject the null hypothesis and evidence to support the alternative hypothesis, that adding at least one of the given explanatory variables has a non-zero effect on the fit of the model. This approach is generally considered a more powerful approach to assessing the statistical significance of a logistic regression model as it directly compares the fits of the given model with its nested model(s).

```r
# Second, comparing with the nested null model
CSmodel_null <- glm(choice ~ 1,
                data = climateSupport,
                family = binomial(link = "logit"))
global_null_LRT <- anova(CSmodel_null, CSmodel, test = "LRT")
print(global_null_LRT)
```

Table 3: Analysis of Deviance: CSmodel and LRT Test

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Resid. Df | 2 | 8,496.500 | 3.536 | 8,494 | 8,499 |
| Resid. Dev | 2 | 11,675.830 | 152.134 | 11,568.260 | 11,783.410 |
| Df | 1 | 5.000 | | 5 | 5 |
| Deviance | 1 | 215.150 | | 215.150 | 215.150 |
| Pr(>Chi) | 1 | 0.000 | | 0 | 0 |

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)

   To answer this question, I consider the logistic regression model:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_{\texttt{countries.L}} \times \texttt{countries.L} + \beta_{\texttt{countries.Q}} \times \texttt{countries.Q} \\ &+ \beta_{\texttt{sanctions.L}} \times \texttt{sanctions.L} + \beta_{\texttt{sanctions.Q}} \times \texttt{sanctions.Q} \\ &+ \beta_{\texttt{sanctions.C}} \times \texttt{sanctions.C}\end{aligned}$$

   To calculate the change in odds with "countries.Q" when sanctions change from "sanctions.L" to "sanctions.Q," I compare the two scenarios:

$$\begin{aligned}\log(\texttt{sanctions.L}|\texttt{countries.C}) &= \beta_0 + \beta_{\texttt{countries.C}} \times \texttt{countries.C} \\ &+ \beta_{\texttt{sanctions.L}} \times \texttt{sanctions.L} \\ \log(\texttt{sanctions.Q}|\texttt{countries.C}) &= \beta_0 + \beta_{\texttt{countries.C}} \times \texttt{countries.C} \\ &+ \beta_{\texttt{sanctions.Q}} \times \texttt{sanctions.Q}\end{aligned}$$

   As this is an additive model, terms cancel out so that the change in log-odds is computed as the difference between the two "sanctions" coefficients, indicating an average increase of 0.09524625 in the log-odds for an individual supporting a policy when the sanctions increase from 5% to 15%, all other variables held constant. Further, computing this into its odds ratio indicates an average change in the odds by 1.09993, in other words, an average increase in the odds, holding all other variables constant:

$$\begin{aligned}\text{Change in Log-Odds} &= \beta_{\texttt{sanctions.Q}} - \beta_{\texttt{sanctions.L}}\\ &= (-0.181086) - (-0.276332)\\ &= 0.09524625\end{aligned}$$

```r
### Computing log-odds
log_odds_change <- as.vector(CSmodel$coefficients[5] - CSmodel$coefficients[4])
odds_ratio_change <- exp(log_odds_change)
print(odds_ratio_change)
```

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

I first calculate the log-odds for this event:

$$\begin{aligned}\log(\texttt{countries = 80}|\texttt{sanctions = 0\%}) &= \beta_0 + \beta_{\texttt{countries.Q}} \times \texttt{countries.Q}\\ &= -0.005665297 + (-0.009949894)\\ &= -0.01561519\end{aligned}$$

I can use the log-odds to calculate the estimated probability for an individual to support the policy in the given scenario using the logistic function. I find approximately 50% probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions:

$$\begin{aligned}P\left(Y = 1\right) &= \frac{1}{1 + e^{(-\texttt{log-odds})}}\\ P\left(\texttt{support}\right) &= \frac{1}{1 + e^{(0.01561519)}}\\ &= 0.4960963\end{aligned}$$

```r
### Estimate probability for 80 countries and no sanctions
log_odds_2b <- as.vector(CSmodel$coefficients[1] + CSmodel$coefficients[3])
probability_support_2b <- 1 / (1 + exp(-log_odds_2b))
print(probability_support_2b)
```

(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.

As the analysis of deviance shows in Table 4, the p-value is greater than 0.05, failing to give evidence to reject the null hypothesis (including an interaction between the explanatory variables has a non-zero effect on the fit of the model). In other words, adding the interaction term does not change the fit of the model. Therefore, the answers to 2a and 2b would likely not change.

Table 4: Analysis of Deviance: CSmodel and Interaction with LRT Test

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Resid. Df | 2 | 8,491.000 | 4.243 | 8,488 | 8,494 |
| Resid. Dev | 2 | 11,565.110 | 4.450 | 11,561.970 | 11,568.260 |
| Df | 1 | 6.000 | | 6 | 6 |
| Deviance | 1 | 6.293 | | 6.293 | 6.293 |
| Pr(>Chi) | 1 | 0.391 | | 0.391 | 0.391 |