Problem Set 2

Applied Stats II

Due: February 19, 2023

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 19, 2023. 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.csv` 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 recode "choice" as a binary outcome variable:
climateSupport$choice_binary <- ifelse(climateSupport$choice == "Supported", 1, 0)

I fit a generalised linear model with 'choice' as the outcome variable, and
'countries' and 'sanctions' as the predictor variables.
I use a binomial distribution due to the dependent variable being a binary outcome.

likelihood_model1 <-glm(choice_binary ~ countries + sanctions,
data = climateSupport, family = "binomial")
summary(likelihood_model1)

Here is the summary output:

Deviance Residuals:
Min       1Q   Median       3Q      Max
-0.6392  -0.4829  -0.3587   0.4844   0.6413
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.498607   0.005357  93.075  < 2e-16 ***
countries.L  0.112700   0.009248  12.186  < 2e-16 ***
countries.Q -0.002386   0.009319  -0.256 0.797937
sanctions.L -0.067615   0.010709  -6.314 2.86e-10 ***
sanctions.Q -0.044187   0.010720  -4.122 3.79e-05 ***
sanctions.C  0.036718   0.010723   3.424 0.000619 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2438998)
Null deviance: 2125.0  on 8499  degrees of freedom
Residual deviance: 2071.7  on 8494  degrees of freedom
AIC: 12136
Number of Fisher Scoring iterations: 2

Summary of results:
```

The intercept has a log odds value of 0.499. This suggests that individuals have a likelihood of about 62% to support an environmental agreement which is agreed by 20 of 192 countries, but which has no sanctions for missing emission reduction targets (the two baseline categories).

The coefficient estimate for 'countries.L' is positive, and highly significant ($p<0.001$). This suggests that having 80 participating countries is positively associated with support for an agreement, compare to the baseline category of 20 countries supporting.
However, the coefficient estimate for 'countries.Q' is not statistically significant, suggesting that an agreement having 160 countries participating does not have a significant effect on likelihood to support an agreement, compared to the baseline category of 20 countries supporting.

The three coefficient estimates for the three levels of sanctions are all highly statistically significant ($p<0.001$). Coefficients for 'sanctions.L' and 'sanctions.Q' are negative, whilst the coefficient for 'sanctions.C' is positive. This suggests that the level of sanctions for missing emission reduction targets are negatively associated with likelihood for individuals to support an agreement at 5% and 15% sanctions, relative to the baseline category of no sanctions, but that at 20% this effect is reversed.

In conclusion, the number of countries participating in the agreement matters for the likelihood of individuals to support an agreement, but only up to a point. After a certain level, additional countries' participation do not positively effect odds of supporting an agreement. On the other hand, higher levels of sanctions are negatively associated with support for an agreement at 5% and 15% sanctions, although the trend is reversed at the highest level of sanctions tested (20%).

Global NUll Hypothesis test:

```
null_model <- glm(choice_binary ~ 1, data = climateSupport)
null_test <- anova(null_model, likelihood_model, test = "Chisq")

Model 1: choice_binary ~ 1
Model 2: choice_binary ~ countries + sanctions
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499     2125.0
2      8494     2071.7  5   53.292 < 2.2e-16 ***
```

P < 0.05, therefore we can reject the global null hypothesis, that all of the coefficients of the independent variables in the model are equal to zero.

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)

   ```
   First I calculate the log odds for each of the two scenarios:

   log(odds) when sanctions are 5% = 0.498607 -0.067615 -0.002386 = 0.428606
   log(odds) when sanctions are 15% = 0.498607 -0.044187 -0.002386 = 0.452034

   Next I use the two sets of log odds to calculate the difference in odds:

   difference_in_odds <- exp(0.428606) - exp(0.452034)
   difference_in_odds
   Difference in odds = [1] [1] -0.0363893
   ```

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

   ```
   First I calculate the log odds:
   log(odds) = 0.498607 -0.002386 = 0.496221


   Then I use the log odds to calculate the probability that an individual will
   support a policy if there are 80 countries participating with no sanctions:
   probability <- 1/(1 + exp(1)^(-0.496221))
   probability = 0.6482389
   ```

   (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.

   ```
   It is possible that model fit could be improved by the inclusion of an
   interaction effect.

   In order to test if an interaction is appropriate in this case, I conduct a
   likelihood ratio test.
   ```

Firstly, I create a model with the interaction effect included:

```
likelihood_model_interaction <- glm(choice_binary ~ countries * sanctions,
data = climateSupport)
```

Next I calculate the deviance of both models:

```
deviance_original_model <- deviance(likelihood_model)
deviance_interaction_model <- deviance(likelihood_model_interaction)
```

I calculate the likelihood ratio statistic:

```
test_statistic <- abs(deviance_interaction_model - deviance_original_model)
test_statistic
# [1] 1.598523
```

I calculate the difference in the degrees of freedom:

```
df_diff <- df.residual(likelihood_model)
- df.residual(likelihood_model_interaction)
```

Finally, I calculate the p-value:

```
p_value <- pchisq(test_statistic, df_diff, lower.tail = FALSE)
p_value
# [1] 0.9526835
```

Because p>0.05, I fail to reject the null hypothesis, that the model without including interaction is sufficient, and that it is unnecessary to include an interaction effect.