Problem Set 2

Applied Stats II

Due: February 19, 2023 - Marcus Ó Faoláin, 16327268

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.

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)

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

   (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.

Question 1

[] Fit an additive model.

In order to fit an additive model, we use the following code in R studio, with a + instead of a * in order to generate an additive rather than an interactive model.

```
logistic <- glm(choice ~ sanctions + countries, climateSupport, family = "binomial")
```

Provide the summary output, the global null hypothesis, and p-value.

Using the summary() function, we can get a summary output as follows:

```
summary(logistic)
```

```
Call:
glm(formula = choice ~ sanctions + countries, family = "binomial",
    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
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 ***
countries.L  0.458452   0.038101  12.033  < 2e-16 ***
countries.Q -0.009950   0.038056  -0.261 0.793741
---
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 global null hypothesis (H_0) states that none of the coefficients (β) of the logistic regression model have an influence on the dependent variable "choice". In more literal terms, this means that neither the level of sanctions for non enforcement nor the number of countries that have signed up for the policy has an effect on the binary dependent variable of whether people support the policy or not.

In mathematical terms, it looks as follows:

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_n = 0$

Where each β is a coefficient for each of the independent variables in relation to the outcome variable. This means that each β would have a slope equal to 0.

The alternative hypothesis (H_α) is that at least one of the β coefficients is not equal to and therefore has an influence on the outcome variable.

$H_\alpha : \beta_j \neq 0$

As we can see, four of the β coefficients have p-values less than 0.01. This means that we reject the global null hypothesis that all
Please describe the results and provide a conclusion.

The coefficient estimate β_1 in the case of "sanctions.:L" has a value of -0.276332. We can see from the three stars on the table and from the p-value of 3.15e-10 that the result is associated with a p-value of less than 0.05, which suggests that "sanction.L" does in fact have an influence on whether an individual supports an international environmental agreement.

Going from no sanctions up to 5% sanctions multiplies the odds of support for an international agreement by $e^-0.276332$, which is equivalent to 0.758561.

This means that a change from no sanctions to 5% sanctions decreases the odds of support for an international agreement by around 24.1%, when all other factors are controlled for.

We notice from the regression table that the coefficients of sanctions.Q (15%), sanctions.c (20%) and countries.L (80/192) are significant. We can interpret these coefficients similarly to the previous one.

Going from 5% sanctions up to 15% sanctions multiplies the odds of support for an international agreement by $e^-0.181086$, which is equivalent to 0.8343636.

This means that a change from 5% sanctions to 15% sanctions decreases the odds of support for an international agreement by around 16.6%, when all other factors are controlled for.

Going from 15% sanctions up to 20% sanctions multiplies the odds of support for an international agreement by $e^0.150207$, which is equivalent to 1.162075.

This means that a change from 15% sanctions to 20% sanctions increases the odds of support for an international agreement by around 16.2%, when all other factors are controlled for.

When the countries supporting it increases from 20/192 to 80/192, the odds in support of the international agreement are multiplied by $e^0.458452$, which is equivalent to 1.581624.

This means that when the number of countries supporting an international agreement increases from 20/192 to 80/192, it results, on average, in a 58.2% increase in the odds of support for the international agreement.

The p-value of the final coefficient for countries.Q (160/192 countries) is not below 0.05, and therefore we cannot reject the null hypothesis that it has an influence on the dependent variable, choice.

In conclusion, using the output table, we can create an equation for predicting the odds of support for an international environmental agreement, with the following formula

$$ln(\frac{P(Y_i=1)}{(1-P(Y_i=1))}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

By subbing in our values, we can get the following equation:

$$ln(\frac{P(Y_i=1)}{(1-P(Y_i=1))}) = -0.05 - 0.276332 X_1 + -0.181086 X_2 + 0.150207 X_3 + 0.458452 X_4$$

Which can then be transformed into:

$$P(Y_i = 1) = \frac{exp(-0.05-0.276332 X_1+-0.181086 X_2+0.150207 X_3+0.458452 X_4)}{1+exp(-0.05-0.276332 X_1+-0.181086 X_2+0.150207 X_3+0.458452 X_4)}$$

We can set up a function in R to calculate different odds for us as follows:

```
p <- function(x1, x2, x3, x4){
p <- exp(-0.05 - 0.276332*x1 -0.181086*x2 + 0.150207*x3 + 0.458452*x4)
q <- 1- exp(-0.05 - 0.276332*x1 -0.181086*x2 + 0.150207*x3 + 0.458452*x4)
```

```
return(p/q)
}
```

Question 2

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

We use the following code to get the predicted probabilities

```
predicted_data <- with(climateSupport, expand.grid(countries = unique(countries),
sanctions = unique(sanctions)))

predicted_data <- cbind(predicted_data, predict(logistic4,
newdata = predicted_data,
type = "response",
se = TRUE))
```

Which yields the following output:

	countries	sanctions	fit	se.fit	residual.scale
1	80 of 192	15%	0.4826196	0.01339475	1
2	160 of 192	15%	0.5603146	0.01320502	1
3	20 of 192	15%	0.3998931	0.01301632	1
4	80 of 192	None	0.5159191	0.01335758	1
5	160 of 192	None	0.5928323	0.01307316	1
6	20 of 192	None	0.4322534	0.01315368	1
7	80 of 192	5%	0.5635428	0.01347950	1
8	160 of 192	5%	0.6381958	0.01242082	1
9	20 of 192	5%	0.4798090	0.01326799	1
10	80 of 192	20%	0.4403193	0.01312026	1
11	160 of 192	20%	0.5180228	0.01349899	1
12	20 of 192	20%	0.3598012	0.01249970	1

2. (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)

Increasing sanctions from 5% to 15% will multiply the odds of supporting the agreement by $e^1 81086$, which is equivalent to 0.834636. This means that it will decrease the odds of supporting an international agreement by around 16.5%.

2. (b) For the policy in which very few countries participate [20 of 192], how does increasing sanctions from 5% to 15% change the odds that an individual will support the policy? (Interpretation of a coefficient)
Increasing sanctions from 5% to 15% changes the odds decreasing it 8% based on the predicted probabilities table.

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

We can see from the table that the predicted probability support for a policy at 160 countries with no sanctions is 0.5928323, which is around 59.2%

2. (d) Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?
The answers would change if we included an interaction term in the model, however, as shown in the following question, it is not appropriate to use an interaction term in the model. The probability would change by the exponent of the relevant interaction terms on page 13.

2. (e) Perform a test to see if including an interaction is appropriate.

One way we can check to see if an interaction term is by running another logistic regression model with an interaction term instead of an additive term. We can do this and examine the output as follows:

```
interaction_log <- glm(choice ~ sanctions*countries,
climateSupport,
family = "binomial")

summary(interaction_log)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4359  -1.1570  -0.9632   1.1349   1.4079

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.003809   0.022006  -0.173 0.862583
sanctions.L              -0.274221   0.043953  -6.239 4.41e-10 ***
sanctions.Q              -0.182289   0.044011  -4.142 3.45e-05 ***
sanctions.C               0.153245   0.044069   3.477 0.000506 ***
countries.L               0.457140   0.038078  12.005  < 2e-16 ***
countries.Q              -0.011167   0.038152  -0.293 0.769750
sanctions.L:countries.L  -0.001754   0.076700  -0.023 0.981755
sanctions.Q:countries.L  -0.007622   0.076156  -0.100 0.920278
sanctions.C:countries.L   0.095197   0.075608   1.259 0.208001
sanctions.L:countries.Q   0.133840   0.075554   1.771 0.076484 .
sanctions.Q:countries.Q   0.093425   0.076303   1.224 0.220806
sanctions.C:countries.Q   0.010449   0.077046   0.136 0.892123
---
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: 11562  on 8488  degrees of freedom
AIC: 11586

Number of Fisher Scoring iterations: 4
```

We can see that none of the interative terms have p-values below 0.05. This strongly suggests that the interaction is not a statistically reliable relationship, and we cannot reject the null hypothesis that the interaction between the variables has no effect.

Therefore we can conclude that an interaction is not appropriate in our model.