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.

   The following output was generated when fitting an added model with the following code:

```
reg1 <- glm(choice ~ countries + sanctions, data = climateSupport, family
    = binomial(link = "logit"), )
```

Table 1: Difflog and Presvote

	Dependent variable:
	choice
countries80 of 192	0.336^{***}
	(0.054)
countries160 of 192	0.648^{***}
	(0.054)
sanctions5%	0.192^{***}
	(0.062)
sanctions15%	-0.133^{**}
	(0.062)
sanctions20%	-0.304^{***}
	(0.062)
Constant	-0.273^{***}
	(0.054)
Observations	8,500
Log Likelihood	$-5,784.130$
Akaike Inf. Crit.	11,580.260
Note:	*p<0.1; **p<0.05; ***p<0.01

   The global null hypothesis in this case is:

$$H_0 = \beta_0 = \beta_1 = ... = \beta_p = 0$$

This means that the alternative hypothesis is that at least one of the β values is not equal to 0.
A likelihood ratio test was carried out to investigate this with the following code:

```
reg1_null <- glm(choice ~ 1, data = climateSupport, family = binomial(
    link = "logit"))
anova(reg1_null, reg1, 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
```

With the p-value being < 0.05, at least one predictor variable is a significant predictor in the logistic regression model. This means we can reject the global null hypothesis, as at least one predictor variable did not equal 0.

Interpreting the regression table:
The intercept in this table means that when the proposed policy included 20 countries and had no sanctions, the expected odds that an individual would expect the proposal would be $\exp^{-0.273} = 0.761$, meaning that an individual would be more likely to support the policy according to the model.
For the rest of the model, each of the predictor variables were deemed to have a non-zero effect on the log odds of whether an individual would support the policy. For example, holding all other variables constant, if the policy included 80 of the 120 countries it would increase the log odds of supporting the policy by 0.336.

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 interpret these coefficients, we first need to look at both logistic regression

equations (with the first one being the 5% sanction and the second being the 15% sanction)
:

$$ln(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}) = -0.273 + 0.648 + 0.192 = 0.567$$

$$ln(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}) = -0.273 + 0.648 - 0.133 = 0.242$$

As shown above, changing the sanctions from 5% to 15% decreases the log odds by 0.325. When converting the sanction coefficients these into their respective odds ratios:

$$e^{\beta_{5\%}} = e^{0.192} = 1.212$$

$$e^{\beta_{15\%}} = e^{-0.133} = 0.875$$

If the policy contains 5% sanctions, it increases the odds of support by $\approx 21\%$, while if the policy uses the 15% sanctions, it decreases the odds of support by $\approx 12\%$. As shown above, changing the sanctions from 5% to 15% decreases the log odds by 0.325. Using the formula to convert the log odds into estimated probability, the following results were obtained:

$$\hat{P} = \frac{\exp^{0.567}}{1 + \exp^{0.567}} = 0.638$$

$$\hat{P} = \frac{\exp^{0.242}}{1 + \exp^{0.242}} = 0.560$$

An estimated odds ratio can then be calculated by using the following formula:

$$\hat{OR} = \frac{\frac{0.638}{1-0.638}}{\frac{0.560}{1-0.560}} = 1.385$$

The estimated OR in this case is 1.385, meaning using the equation with 5% sanctions increases support by $\approx 38.5\%$ compared to using the equation with 15% sanctions.

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

The following code was used to create a dataframe with predicted probabilities and confidence intervals:

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

```

```
5 predicted_data <- cbind(predicted_data, predict(reg1,
6                                                 newdata = predicted_
    data,
7                                                 type = "response",
8                                                 se = TRUE))
9 predicted_data <- within(predicted_data,
10                           {PredictedProb <- plogis(fit)
11                           LL <- plogis(fit - (1.96 * se.fit))
12                           UL <- plogis(fit + (1.96 * se.fit))
13                           })
```

The following output was created: As can be seen in the fourth row, the estimated

Table 2: Predicted Data

	countries	sanctions	fit	se.fit	residual.scale	UL	LL	PredictedProb
1	80 of 192	15%	0.483	0.013	1	0.625	0.612	0.618
2	160 of 192	15%	0.560	0.013	1	0.642	0.631	0.637
3	20 of 192	15%	0.400	0.013	1	0.605	0.593	0.599
4	80 of 192	None	0.516	0.013	1	0.632	0.620	0.626
5	160 of 192	None	0.593	0.013	1	0.650	0.638	0.644
6	20 of 192	None	0.432	0.013	1	0.613	0.600	0.606
7	80 of 192	5%	0.564	0.013	1	0.643	0.631	0.637
8	160 of 192	5%	0.638	0.012	1	0.660	0.649	0.654
9	20 of 192	5%	0.480	0.013	1	0.624	0.612	0.618
10	80 of 192	20%	0.440	0.013	1	0.614	0.602	0.608
11	160 of 192	20%	0.518	0.013	1	0.633	0.620	0.627
12	20 of 192	20%	0.360	0.012	1	0.595	0.583	0.589

probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions is 0.516, with the lower and upper bounds of the 95% confidence interval being 0.542 and 0.490 respectively (alternate table provided in R project in GitHub). This is going off the `fit` column, as I believe there could be an error in the example code provided? If you take out the `type = "response"` from line 7 in the code, the `fit` column then contains the log odds of the equations, as the `type = "response"` changes it to the predicted probability value for you. Therefore, I think the `PredictedProb` column in the table above has taken the probability value and put it into the link function.

(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.

The answers to 2a and 2b could potentially change as when the new interaction variable is added, it affects the values of the previous coefficients, which would

in turn change the log odds when filling out the equations. To test whether an interaction variable is appropriate, a second regression was run with an interaction effect. From this, a likelihood ratio test was carried out to determine whether an interaction effect would be appropriate for the model:

```
Analysis of Deviance Table

Model 1: choice ~ countries * sanctions
Model 2: choice ~ countries + sanctions
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8488      11562
2      8494      11568 -6  -6.2928   0.3912
```

As shown above, the p-value from the likelihood ratio test is not significant ($<$ 0.5), meaning there is not sufficient evidence that including an interactive effect between the sanctions and number of countries would be a significant predictor in support for the policy.