mystyle backgroundcolor=, commentstyle=, keywordstyle=, numberstyle=, stringstyle=, basicstyle=, breakatwhitespace=false, breaklines=true, captionpos=b, keepspaces=true, numbers=left, numbersep=5pt, showspaces=false, showstringspaces=false, showtabs=false, tabsize=2 style=mystyle

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.

Answer:

Provided are both the R code used to do the logistic regression, and the table of the coefficients of the regression, in the R file an alternative method of fitting an additive model without dummy variables is also demonstrated (both give equal results).

[language=R] Creating dummy variables. climateSupport$choice <$
$-ifelse(climateSupport$choice $==$ "Supported", 1, 0) climate-
Support ¡- within(climateSupport, countries$_2$0 $< -ifelse(countries ==$
"20of192", 1, 0)$countries_8$0 $< -ifelse(countries ==$ "80of192", 1, 0)$countries_1$60 $<$
$-ifelse(countries ==$ "160of192", 1, 0)
sanctions$_N one < -ifelse(sanctions ==$ "$None$", 1, 0)$sanctions_5 <$
$-ifelse(sanctions ==$ "5$sanctions_1$5 $< -ifelse(sanctions ==$
"15$sanctions_2$0 $< -ifelse(sanctions ==$ "20)

Fitting the model using the glm function. Family = binomial(logit) tells us that the dependent variable is binary. Any two of the dummy variables can be omitted to act as the reference. (As long as one is countries, and one is sanctions). This avoids perfect multicollinearity, which would result in NAs. model below omits countries$_2$0$and sanctions_N one. I will use this model, as it makes sense to use the lowe$
$-glm(choice\ countries_8 0 + countries_1 60 + sanctions_5 + sanctions_1 5 + sanctions_2 0, data =$
$climateSupport, family = binomial(logit))$
summary(model)

Regression equation for model. Reference, countries$_2$0$and sanctions_N one$ $-0.27266 + 0.33$

Table 1:

	Dependent variable:
	choice
countries_80	0.336^{***}
	(0.054)
countries_160	0.648^{***}
	(0.054)
sanctions_5	0.192^{***}
	(0.062)
sanctions_15	-0.133^{**}
	(0.062)
sanctions_20	-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 constant term is the log odds that an individual agrees with the policy when the number of participating countries is 20 of 192, and the sanctions are None.

Each of the coefficient values represent the change in the log odds from the reference, when the corresponding predictor variable is coded as 1, meaning that predictor variable is true. When a predictor variable is coded as 0, it is canceled out resulting in no change from the reference for that variable.

In this model it is only ever possible for one of the "countries" and one of the "sanctions" predictor variables to be coded as true (1) at a time. This is because the scenarios they represent are mutually exclusive (e.g if the sanctions are 5%, they cannot be 20% at the same time).

The global null hypothesis:

H0: all slopes $(\beta_j) = 0$

The alternative hypothesis: HA: at least one slope $(\beta_j) \neq 0$

In order to test this in R, a likelihood ratio test would need to be used, which can be done in R with the following code:

[language=R] Create model$_n ullwithnoXdata, forcomparisonwith"model".model_n ull <$
$-glm(choice\,1, data = climateSupport, family = binomial(logit))summary(model_n ull$

Perform a likeihood ratio test. anova(model$_n ull, model, test = "LRT")

P-value = 2.2e-16 At least one predictor in the model is reliable.

As the p-value $= 2.2e-16 < \alpha = 0.01$ we can reject the global null hypothesis, and conclude that at least one predictor in the model is reliable.

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)

Answer:

To answer this I will use the regression equation, for logistic regression, in R. [language=R] -0.27266 + 0.33636*0 + 0.64835*1 + 0.19186*1 - 0.13325*0 - 0.30356*0 countries$_1$60 = 1, $sanctions_5$ = 1.0.56755$P(hat) = 1/(1 + exp(-(0.56755)))$0.6381977

-0.27266 + 0.33636*0 + 0.64835*1 + 0.19186*0 - 0.13325*1 - 0.30356*0 countries$_1$60 = 1, $sanctions_1$5 = 1.0.24244$P(hat) = 1/(1+exp(-(0.24244)))$0.5603149

0.6381977 - 0.5603149 0.0778828

Increasing the sanctions from 5% to 15% (when 160 out of 192 countries participate) equates to a decrease in the probability that an individual will support the policy **= 0.0778828, or 7.8%**.

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

Answer:

[language=R] -0.27266 + 0.33636*1 + 0.64835*0 + 0.19186*0 - 0.13325*0 - 0.30356*0 0.0637 P(hat)= 1/(1+exp(-(0.0637))) 0.5159196

The estimated probability that an individual will support a policy if there are 80 of 192 countries participating with no sanctions is **0.5159196, or 51.6%**

(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.

Answer:

It is unlikely there would be much change in the answers to 2a and 2b with the introduction of an interactive term into

the model. This is because the scenarios which the dummy variables represent are, as previously stated, mostly mutually exclusive. When one of the coutries variables is true (1), the others are false (0), and the same is true of sanctions, this would limit the potential for interactions.

To test if including an interaction is appropriate, a model including the interaction term can be written in R. This model can be tested against the model we have been using named "model" using a likelihood ratio test. The following R code does this.
[language=R] model$_interaction <- glm(choice\ countries_80 * countries_160 * sanctions_5 * sanctions_15 * sanctions_20, data = climateSupport, family = binomial(logit))$
anova(model$_interaction, model, test = "LRT")$
The p-value which results from this likelihood ratio test is $= \mathbf{0.3912}$ which far exceeds the level of significance $\alpha = \mathbf{0.01}$. This suggests that the model with the interaction term is not more useful at predicting estimates than the original model named "model". Including an interaction would be inappropriate in this case.
In other words, we have insufficient evidence to reject the null hypothesis that the model without the interaction provides an adequate fit for the data.

Bibliography:

- Logistic Regression in R, Clearly Explained!!!! (2018). Available at:
- StatQuest: Logistic Regression (2018). (Accessed: 15 February 2024).
- 6.2 Logistic Regression Models in R (2021). (Accessed: 17 February 2024).
- Charlie (2012) 'Answer to "How to handle multicollinearity in a linear regression with all dummy variables?"', Cross Validated. Available at: https://stats.stackexchange.com/a/30536 (Accessed: 17 February 2024).

- Logistic Regression in R Tutorial (no date). Available at: https://www.datacamp.com/tutorial/logistic-regression-R (Accessed: 15 February 2024).
- January (2014) 'Answer to "Recoding a variable with three levels into a dummy variable"', Cross Validated. Available at: https://stats.stackexchange.com/a/89439 (Accessed: 15 February 2024).
- Charlie (2012) 'Answer to "How to handle multicollinearity in a linear regression with all dummy variables?"', Cross Validated. Available at: https://stats.stackexchange.com/a/30536 (Accessed: 17 February 2024).
- Dr. Jeffrey Ziegler's lecture material.