Assignment 3

GV903 Advanced Methods – 2020/2021
University of Essex, Department of Government

Instructions

This assignment is due on Wednesday, 17 February at 9:45 am. Please submit on FASER. Instructions can be found in the module handbook on Moodle.

Please create a new document with your answers. Use LaTeX in conjunction with `knitr` via RStudio for document creation, and submit both a PDF file (the compiled document) and the source files necessary for compilation (e. g., with the extensions `.Rnw`, `.bib` etc). Up to nine points are given for proper use of technology and proper formatting, as part of the maximum of 100 attainable points.

The number of points you can obtain for each question or task is given in square brackets; each task is worth a maximum of seven points. The points add up to 100 (including the points for use of technology as described above) and determine your final mark for this assignment. The final mark you earn for this assignment is given by the equation

$$m = t + \sum_{i=1}^{n} p_i, \tag{1}$$

where p_i denotes the number of points you earn for question or task i, n is the number of tasks in this assignment, and t is the number of points you earn for use of technology and formatting. All answers are evaluated on three criteria: how clearly the results are presented; how correct the results are; and how elegant, computationally efficient, or sophisticated the solution is (where applicable). Good luck!

Dataset: MAS party identification in Bolivia

In a brand new article hot off the press, Poertner (2021) examines how voters come to support new political parties. His article contends that "new types of locally organized, participant-based societal organizations—such as neighbourhood associations, informal sector unions, and indigenous movements—can play a crucial mediating role in securing electoral support for new parties" (Poertner, 2021).

The author bases his research in Bolivia and carries out an experiment and reports additional survey results to support his theory. Here in this assignment, we ignore the author's experimental evidence and focus on the observational, survey-based data for further analysis.

The paper reports two logit models, in which the dependent variable is whether a voter strongly identifies with the newly emerged MAS party (variable `MAS`). The most important explanatory variables are two dummies that indicate whether the individual is a member in local organisations that endorse the MAS party (variable `org_MAS_individual`) and whether somebody in the close social network of the individual is a member in such organisations (variable `org_MAS_network`), as reported in the first two rows in Model 2 in Table 1 in the article. Both effects are statistically significant and show a large magnitude.

Unfortunately, there are a few statistical problems in the analysis. They become evident if you inspect and re-run the replication script for the analyses reported in the paper. The replication dataset can be downloaded from the Harvard Dataverse at `https://doi.org/10.7910/DVN/4RQWXW`. The code for the analysis is contained in the script `Code_Original_Survey.R`, and the survey data used in that script can be found in the file `survey_cleaned.csv`. The analysis starts with the following preparatory steps, which you should repeat before completing the tasks below:

```
survey <- read.csv("survey_cleaned.csv")
PID_data <- subset(survey, MAS == 1 | (B8A == 14 & B8B == 1))
```

You can then work with the `PID_data` dataset. The gravest problem is that too many collinear variables are included in the model specification. This causes the number of observations to go down dramatically because many of these variables have `NA` values, which makes the respective observations drop out of the model. It also causes some variables to drive each other's effect sizes artificially up. To rectify this problem, you will re-estimate Model 2 as a reduced version with fewer variables in Task 1 below. Based on this reduced model, you will complete a few other tasks, before you will eventually move towards new model specifications and research questions based on the same `PID_data` dataset.

The original replication script contains frequency weights (variable `wt`) because some areas were undersampled while others were oversampled. The author uses the `glm` argument `weights = wt` to attach a weight to each observation (sometimes slightly above 1.0, sometimes slightly below 1.0). It does not make a major difference to the results. Therefore, please do *not* do this in your analyses below. You will get the opportunity to implement a logit model with frequency weights below in Task 10.

Based on the dataset, please complete the following tasks. Please include a brief description of the steps in your solutions for all tasks; the description will form part of the assessment.

1 Estimation and presentation of logit and probit models

Estimate a reduced version of Model 2 that contains the two main variables of interest, the left–right ideology score variable (`lr_score`), the indigenous ideology score variable (`ind_score`), as well as age and sex of the respondent. In addition to the logit model, also estimate the same model specification using a probit model. Report both in a `texreg` table with perfect formatting and coefficient labels as in the original table. [7 points]

2 Logit and probit coefficient interpretation

Interpret the findings from the table for the two main variables of interest: `Individual member` and `Social network member`. How do they compare to the original findings from Table 1 in the article? Also interpret the effect for the left–right scale. In addition to the interpretation of the logit coefficients, how would you interpret the probit coefficients? [7 points]

3 Predicted probabilities

Based on your first model (the reduced logit model), plot predicted probabilities for MAS identification at different levels of the left–right scale using `ggplot2`. Create two plots side by side. In the first one, plot predicted probabilities of an individual whose social network contacts are a

member of local organisations but who is not themselves a member, along with predicted probabilities for an individual who is neither a member nor has people in their close social network who are members in such organisations. Also plot the uncertainty around the prediction curve. In the second plot, again show two prediction curves, but this time individual members versus no membership. Keep all other variables constant at their means. Use the `predict` function in this task. Briefly interpret your diagrams. [7 points]

4 Manual prediction

For each observation in the `PID_data` dataset, create fitted values on the probability scale, based on your logit model. Do this manually, without using the `predict` function. [7 points]

5 Bootstrapping predicted probabilities

Create a new version of the first predicted probability plot from Task 3 using manual prediction (like in Task 4) and bootstrapped confidence intervals.

The `predict` function uses the delta method to generate predictions, i. e., the predictions are generated analytically. In Task 3, you created predicted probabilities along with confidence intervals using this technique. This time, we want to do it all manually.

Re-run your logit model on 1,000 bootstrap samples and generate bootstrapped predictions on the probability scale. Plot your manual prediction curve along with the bootstrapped confidence intervals. [7 points]

6 Bootstrapping the share of indigenous people

The `indigenous` variable indicates whether a respondent is an indigenous citizen. Compute the share of indigenous people in the sample. Unfortunately, this is only a single number and does not include any uncertainty due to the sampling process. It is the best estimator of the share of indigenous people in the population, but there is some underlying uncertainty around the population share. Use bootstrapping to test whether the share of indigenous people in the population is lower than 50 per cent. Also calculate the probability that the share is lower than 50 per cent. [7 points]

7 Prediction performance

Your reduced logit model from Task 1 contained a left–right score. The dataset also contains an alternative version of this: a self-reported ideology placement into seven categories (also included in the published Table 1). You want to know if the predictive performance of the model with the `lr_score` variable or an alternative version with the self-reported placement instead of `lr_score` is higher. Use receiver operating characteristic curves to make this comparison, and consider the area under the curve. [7 points]

8 AIC and BIC

In Task 1, you estimated a logit model. Compare the goodness of fit of this model with a reduced version of the same model in which the `age` variable is omitted. Use the log likelihod, AIC, and BIC as criteria, and interpret your findings. Show the equations, insert the values, and verify

your results by calculating the results manually in `R`. Interpret the results based on the three criteria. [7 points]

9 Likelihood ratio test

Show the equation for the likelihood ratio test. Insert the respective values for your logit model and the reduced version of the model computed in Task 8. Compute the test statistic, critical value, and p-value. Verify your results using `R`. Is the `age` variable necessary? Interpret also with reference to Task 8. [7 points]

10 Implementation of the weighted logit model

The original replication code used the `weights = wt` argument in the `glm` call. This weights observations in a way that is comparable to weighted least squares, just in a GLM context. It was necessary here because some units had a higher sampling probability than others to begin with.

The likelihood function for the weighted logit model is

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{w}) = \prod_{i=1}^{n} \left(\theta_i^{y_i}(1-\theta_i)^{1-y_i}\right)^{w_i} \tag{2}$$

with link function

$$\theta_i = \text{logit}^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) = \frac{1}{1+e^{-\mathbf{X}_i^\top \boldsymbol{\beta}}}. \tag{3}$$

The corresponding log likelihood function is

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{w}) = \sum_{i=1}^{n} w_i \left[y_i \log \theta_i + (1-y_i)\log(1-\theta_i)\right] \tag{4}$$

Re-estimate your logit model from Task 1 using this weighted logit model. For this purpose, write the log likelihood function in `R`, prepare starting values and input data, use the `optim` function, and retrieve and present your results. Optionally, you can verify your results using the `glm` function with the `weights` argument. Compare the results to the non-weighted version of the model; are there any differences? [7 points]

11 Ordered logit model

Use an ordered logit model to explain the respondents' household income quartile as a function of age, sex, and the left–right score. Interpret the results for age and ideology. [7 points]

12 Parallel regression assumption test

Test the parallel regression assumption in the model estimated in Task 11. Which model would you recommend for explaining the respondents' household income quartile as a function of age, sex, and the left–right score: a linear model, an ordered logit model, a multinomial logit model, or a conditional logit model? Justify your choice. [7 points]

13 Ordered logit prediction

For an age range between 18 and 75 years, hold all other variables constant at their means and plot predicted probabilities and cumulative probabilities based on the model in Task 11. What are your substantive findings? [7 points]

References

Poertner, M. (2021). The organizational voter: Support for new parties in young democracies. *American Journal of Political Science*. Forthcoming. URL: `https://doi.org/10.1111/ajps.12546`.