Gov 2001: Problem Set 5

Due Wednesday, March 9th at 6pm

Instructions

You should submit your answers and R code to the problems below using the Quizzes section on Canvas.

Problem 1 - Replicating Fearon and Laitin (2003)

For this problem you will replicate and explore the results of the article "Ethnicity, Insurgency and Civil War," by James Fearon and David Laitin (*American Political Science Review*. 97 (1) February 2003.). This article uses a number of institutional, geographic and social variables which have been hypothesized to predict the onset of civil conflict.

The complete dataset is available on the website in the .dta format. You will use the following variables in your analysis:

- onset: civil war onset that year
- warl: a 'distinct' civil war ongoing in previous year
- gdpenl: lagged per capita income, thousands of US dollars
- lpopl1: lagged natural logarithm of population
- lmtnest: log % of country mountainous terrain
- ncontig: country is non-contiguous geographically
- Oil: country is an oil exporter
- nwstate: country achieved independence in past two years
- instab: significant change in polity score in last 3 years, lagged
- polity2l: lagged level of democratization
- ethfrac: a measure of ethnic fragmentation
- relfrac: a measure of religious fragmentation

Once you have loaded the data, you will need to subset it to include only the above variables. Then delete all rows containing missing data using `na.omit()`. Finally, fix the value of onset that was incorrectly coded as a 4. You may assume for the time being that it was supposed to be a 1.

You may not use canned R functions for this problem except to check answers.

1.A) Provide the appropriate stochastic component, systematic component, and independence assumptions for the logit model used in the paper.

1.B) Derive the log-likelihood for the model parameters, $\ell(\beta|Y_i, X_i, n)$.

1.C) Critically evaluate one modelling assumption made by the authors and explain your reasoning.

1.D) Write a function in R to implement the log-likelihood you found in 1.B.

1.E) Replicate Model 1 from Table 1, using your function from 1.D and `optim()` with `method = "BFGS"`. Be careful in selecting the starting values for your optimization (a vector of zeroes should be adequate). Report selected coefficient estimates and their standard errors.

1.F) Using your estimates from 1.E what is the predicted probability of civil war for oil exporters, with all other covariates are held at their median?

1.G) Conduct a likelihood ratio test which compares the unrestricted model (i.e. the full specification from Model 1) with a restricted model which excludes the ethnic and religious fractionalization variables. Treating the restricted model as your null hypothesis, what is your test statistic and what do you conclude (at the $\alpha = .05$ rejection level)?

1.H You want to get a rough sense of how well the model fits the data. Using your estimates from 1.E, calculate your predicted probabilities $\hat{\pi}_i$ for each observation. Compute the Brier (1950) score, which is a metric for assessing model fit, defined as

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - Y_i)^2$$

Problem 2: Different Link Functions

In this problem, we are going to investigate what happens to our model when we choose another link function to connect our covariates to the Bernoulli probability. Still assume that we observe n observations with outcome Y_i and covariates X_i. Each $Y_i \sim$ Bernoulli(π_i), but instead of the logit link, we assume that

$$\pi_i = 1 - \exp\left[-\exp(X_i\beta)\right]$$

2.A) Derive the log-likelihood for the model parameters, $\ell(\beta|Y_i, X_i, n)$

2.B) Write a function in R to implement the log-likelihood you found in 2.A.

2.C) Re-estimate Model 1 from Table 1 of Fearon and Laitin (2003) using your new function from 2.B and `optim()` with `method = "BFGS"`. Note that you will likely need to use different starting values to achieve algorithm convergence – we suggest using 0.15 for all of the β parameters as good starting values. Try different starting values if you obtain an error. Report selected coefficient estimates and their standard errors.

2.D) Using your estimates from 2.C what is the predicted probability of civil war for oil exporters, with all other covariates are held at their median?

2.E) You want to compare the fit using your new link function to the fit from your logit model. Using your estimates from 2.C, calculate your predicted probabilities $\hat{\pi}_i$ for each observation. Compute the Brier (1950) score, which is a metric for assessing model fit, defined as

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - Y_i)^2$$

2.F) Based on your answers to 1.G/1.H and 2.D/2.E, do the two models give qualitatively different results? Which model should a researcher choose (if there is a clear winner) and why?

R Code

Please submit all your code for this assignment as a .R file. Your code should be clean, commented, and executable without error.