POLS/CSSS 510:
Maximum Likelihood Methods for the Social Sciences

Problem Set 5

Professor: Christopher Adolph, Political Science and CSSS

Fall Quarter 2017

Due in class on Tuesday 28 November 2017

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with R or other statistical packages, you should attach all code you have written and all (interesting) output. Materials should be stapled together in order by problem. The most readable and elegant format for homework answers incorporates student comments, code, output, and graphics into a seamless narrative, as one would see in a textbook.

Problem 1: Modeling vote choice with multinomial logit

In this problem you will again use data from the 1992 American National Election Study. The data are taken from a nationally representative sample of the US population. The survey was conducted in the fall of 1992 with a reinterview of the respondents following the election. In the data you will be using, only the vote report is taken from the post-election interview.

The dataset `nes92.csv` contains information on the vote for President and approval of then-President Bush along with several variables selected to capture the range of influences that current work on voting behavior expects to be important. Descriptions of the variables are at the end of this assignment.

Variable	Description
`vote92`	Vote 1992: 0 = Bush, 1 = Clinton, 2 = Perot
`bushapp`	Bush Approval, 1992: 0 = Strongly Disapprove, 1 = Disapprove, 2 = Approve, 3 = Strongly Approve
`rlibcon`	Ideological self-placement: 1 = very liberal, ... 7 = very conservative
`rbdist`	Ideological distance between respondent and Bush in absolute value
`econ`	Economy worse than year ago? 1 = much better, 2 = better, 3 = same, 4 = worse, 5 = much worse
`milforce`	Opposition to military force: 1 = would use force, ... 5 = would not use force
`gulfwar`	Was the Gulf War worth the cost? 0 = not worth the cost, 1 = worth cost
`partyid`	−3 = Strong Democrat, −2 = Democrat, −1 = Leans Democratic, 0 = Independent, ... 3 = Strong Republican
`yrsofed`	Years of education in actual years: 0–17
`nonwhite`	0 = white, 1 = nonwhite
`econ3`	Recoding of `econ` into three categories: 1 = better, 2 = same, 3 = worse

a. **[5 points]** Use a multinomial logit model to explain respondents' 1992 Presidential vote choice. Using `multinom()` from the `nnet` library, fit the model to the variable `vote92` with `rlibcon`, `rbdist`, `econ`, `gulfwar`, and `nonwhite` as the only covariates. Report the estimated parameters, their standard errors, and the value of the log likelihood at its maximum.

b. **[15 points]** Calculate the probability that a white respondent voted for Bush, Clinton, or Perot, given each self-reported position on the liberal-conservative continuum (`rlibcon` $= \{1, 2, 3, 4, 5, 6, 7\}$) and all other covariates held at their means. Next, calculate the probabilities for a nonwhite respondent.

c. **[15 points]** Present the expected probabilities of a vote for each presidential candidate calculated in part **b** in a way that would not require your reader to know anything about multinomial logit models. Include the uncertainty of your estimates in your presentation. Briefly discuss the results of the model.

d. **[5 points]** What bearing does the *independence of irrelevant alternatives* assumption have on your results?

Problem 2: Can a Model Be *Too* Good? Logit Revisited

a. [10 points] Consider two variables x and y where y is generated from x as a Bernoulli draw with probability of success $\text{logit}^{-1}(x\beta)$ where the true value of β is 2 and is unknown from our perspective. Suppose that we observe eight instances of these variables, $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0, 0)$ and $\mathbf{x} = (5, -2, -2, -2, 5, -2, -2, -2)$. Using the Bernoulli-logit likelihood, show the profile likelihood of β over the domain $[-10, 10]$. Then repeat the exercise with $\beta \in [-100, 100]$ and $\beta \in [-1000, 1000]$. Interpret the likelihood surfaces you have found.[1]

b. [10 points] Estimate a logit model of y as a function of x and report the estimated coefficients and their standard errors. Use both the logit MLE and GLM with a logit link to estimate the model. Do you trust these results? What do you think is happening?

c. [5 points] Now consider some real-world data, provided by Daniel Stegmüller. The file `India.Rdata` contains a dataframe called `govcollapse` which records incidents of government collapse in India (see table below for a description).[2] Fit a logistic regression to the variable `collapse` using all the covariates in `govcollapse`. Report the estimated parameters, their standard errors, and the log likelihood at its maximum using `glm()`. Do you have any concerns about these estimates?

d. [15 points] Examine the in-sample predicted probabilities of government collapse and plot them against each covariate. Do you notice any problems with these predictions? What might you do to improve them and would your proposed fix have any downside? Does this example raise general concerns about the proper use of logistic regression?

1 If you use the likelihood code provided in the course examples, you will need to make sure `x` is a matrix using `x <- as.matrix(x)` before passing it on to the likelihood function.

2 This R datafile can be loaded using the `load()` function.

Variable	Description
`collapse`	1 if government collapsed
`iou`	index of opposition unity
`ideology`	1 if coalition has chief minister from ideological party
`splinter`	1 if coalition included a splinter party
`opp`	Percentage of seats won by opposition
`extreme`	1 if coaltion includes both extreme left and right parties