Advanced Topics in Statistical Methods

PLSC 504

Exercise One

September 5, 2024

Part I: Separation, Simulated

Our in-class discussion on Wednesday discussed what happens when we encounter "separation" / "perfect prediction" in a logistic regression context. Here' we're going to explore that further using simulations, and in particular focus on the differences between standard and Firth / penalized logistic regression in the presence of total and near-separation.

1. Consider a large population of $\mathfrak{N}$ cases with two binary variables X and Y, where in the population there are equal numbers of individuals with $\{X = Y = 0\}$, $\{X = Y = 1\}$, and $\{X = 1, Y = 0\}$ but *no* individuals that have $\{X = 0, Y = 1\}$:

	$\underline{Y = 0}$	$\underline{Y = 1}$
$X = 0$	$\mathfrak{N}/3$	0
$X = 1$	$\mathfrak{N}/3$	$\mathfrak{N}/3$

   Begin with a small (say, $N = 20$) sample from this population. Fit standard logistic and Firth-corrected logit models of Y on X to the sample, and recover estimates of the estimated logit slope $(\hat{\beta}_X)$ and its associated standard error for each. Describe the distribution of the estimates of each,[1] and compare them across the two models.

2. Increase the sample size N by a factor of 10, and repeat part (1), discussing if and how a larger sample changes the resulting estimates from both types of models.

3. Increase the sample size N <u>further</u> by a factor of 20, and repeat part (2).

4. Repeat parts (1)-(3), but now sampling from a population where one in 100 individuals have $\{X = 0, Y = 1\}$:

	$\underline{Y = 0}$	$\underline{Y = 1}$
$X = 0$	$\mathfrak{N}/3.030303$	$\mathfrak{N}/100$
$X = 1$	$\mathfrak{N}/3.030303$	$\mathfrak{N}/3.030303$

   In particular, discuss how the standard and Firth logit models differ in this case, and how those differences change as the sizes of the samples increase.

[1] Note: A single simulation doesn't really tell you anything. What I'm asking for here is basically to {sample → fit the model to the sampled data → extract the slopes and their standard errors} some large number of times, and then summarize – via plots and/or tables – the values of those quantities across the many samples / simulations.

Part II: Data Analysis

With support from the National Endowment for the Arts, in June and July of 2018 a group of public opinion researchers conducted the *Self-Perceptions of Creativity & Arts Participation in the United States* survey.[2] The survey was designed "to measure the ways that American adults experience and exercise creativity in their daily lives," and comprised a national probability sample of 3447 adults in the U.S. The survey included a series of questions about respondents' participation in artistic and creative endeavors, including:

> "The following questions are about ways people make and do art. During the last 12 months (that is, between June 28, 2017 and June 28, 2018), did you do any painting, drawing, sculpture, or printmaking activities?"

Survey participants could respond with "yes" or "no." In this part of the exercise, we'll explore some personal / demographic correlates of that question.

The data from this survey are available on the PLSC 504 Github repository, in the folder creatively names "Exercises." The central variable of interest, `MakeArt`, is coded "1" if the respondent answered "yes" to the above question, and "0" if they answered "no." We'll use various forms of logistic regression to examine the association of that variable with a few others:

- `State` is the two-letter postal code for the state in which the respondent lived at the time of the survey.
- `Age` is a seven-category ordinal variable reflecting respondents' age categories (1 = age 18-24, 2 = age 25-34, 3 = age 35-44, 4 = age 45-54, 5 = age 55-64, 6 = age 65-74, and 7 = age 75+).
- `Female` is a dichotomous variable, coded 1 if the respondent self-identified as female and 0 if they self-identified as male.
- Four dichotomous variables reflect the racial/ethnic self-identification of the respondents, as measured by the survey designers: `White`, `Black`, `Asian`, and `Hispanic`.[3] For each, a value of 1 indicates that respondent identified as a member of that category, and a 0 indicates they did not.
- `Education` is a four-category ordinal indicator of the respondent's highest level of formal education (1 = no high school diploma, 2 = high school diploma or equivalent, 3 = some college, and 4 = BA degree or above).
- `Income` is an 18-category ordinal variable reflecting the respondent's annual household income (ranging in roughly $10,000 increments from "less than $5,000" to "more than $200,000").
- `HHUnder18` is the number of individuals under the age of 18 that the respondent indicated were currently living in their household; these are typically children.

[2]The full citation is: Novak-Leonard, Jennifer, Gwendolyn Rugg, Megan Robinson, and Norman Bradburn. 2018. "Self-Perceptions of Creativity & Arts Participation, United States." Inter-university Consortium for Political and Social Research [distributor], 2020-10-21.

[3]The omitted / reference category aggregates responses of "Other" and "Two or More Races."

Exercise

1. Begin by fitting a logistic regression model of predictors related to the `MakeArt` variable. In specifying your model, briefly discuss your specification decisions, including (a) why each variable was included in the regression, (b) the variable's functional form (including your decisions to dichotomize variables, include quadratic and other nonlinear terms, etc.) and (c) any multiplicative interactions you include.

2. Discuss your findings in substantive terms, using tables, plots, and other methods. At this point, Vincent Arel-Bundock's `modelsummary` and `marginaleffects` packages will likely be of substantial assistance.

3. Re-fit your model, this time using data only on respondents in Pennsylvania. Check for telltale signs of separation, and respond appropriately if that appears to be an issue.

This assignment is due *electronically*, as a *PDF file*, at 11:59 p.m. ET on Friday, September 13, 2024. You can submit your homework by emailing copies **both** to Dr. Zorn (`zorn@psu.edu`) and Ms. Herlihy (`mth5492@psu.edu`). This assignment is worth 50 possible points.