Topics in Political Methodology

PLSC 504

Exercise Two

September 12, 2024

Part I: MNL, IIA, ETC.

We'll start by investigating the IIA assumption using simulations. Because we know that logistic and normal distributions are quite similar, we'll generate some simulated data from a multivariate Normal distribution; this will allow us to easily vary the latent associations between the outcomes.[1]

More specifically, consider a simple example with $J = 3$ possible outcomes on Y (that is, $Y_i \in \{1, 2, 3\}$) as a function of a single, exogenous, binary predictor X. So:

$$Y_i = j \text{ iff } U_{ij} > U_{ik} \, \forall \, j \neq k \, \in J = \{1, 2, 3\}$$

where:

$$\begin{aligned} U_{i1} &= 0 - X_i + u_{i1} \\ U_{i2} &= 0 + X_i + u_{i2}, \end{aligned}$$

with $U_{i3} = 0$, $X \in \{0, 1\}$ with $\Pr(X_i = 1) = 0.5$, and

$$\{u_{i1}, u_{i2}\} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}\right).$$

In words: An observation i's underlying utility for choices 1 and 2 depend linearly on an exogenous binary variable X. The utility for choice 1 is on-average lower when $X = 1$, while that for choice 2 is on-average higher when $X = 1$ (more specifically, $\beta_{11} = -1$ and $\beta_{12} = 1$); the utility for choice 3 is normalized to zero. Those utilities also depends on two choice-specific random error terms, which are distributed as multivariate normal with means of zero, variances of one, and covariance (here, equivalent to the Pearson's correlation) equal to σ_{12}. Each observation's observed outcome is the one with the highest level of utility.

Starting with this framework:

1. Begin by generating a relatively large (maybe $N = 1200$ or so) data set according to the framework above, with $\sigma_{12} = 0$ (that is, where the latent utilities' stochastic components are uncorrelated with each other, corresponding to the IIA assumption). Fit a MNL model of Y on X for those data, then use one or more of the approaches (tests, etc.) talked about in class to assess whether the IIA assumption is violated.

2. Repeat / simulate step (1) many times, keeping the results each time. Discuss the distribution of the values of what you find.

3. Repeat step (2), this time with data where $\sigma_{12} = 0.4$.

4. Repeat step (2) again, this time with data where $\sigma_{12} = -0.8$.

Hints: (a) You can easily generate multivariate normal data (here, the u_is) using the `mvtnorm` routine in the `MASS` package; (b) you can select the choice with the highest value of U using `apply` and `which.max`, e.g. `foo <- apply(U,1,which.max)`; (c) note that the `mlogit` package (and command) wants the data to be in "long form;" see the code from September 4 (around line 370) and/or the help for `mlogit.data` for tips.

[1] Note that it is not hard to generate multivariate logistic data; there's even an R package for it. But we'll stick with the multivariate normal here, just because.

Part II: "A covert ethnic-pride celebration for red-state whites of Northern European descent."

In a 2019 publicity stunt, a Fox News commentator began her show by attempting to drink a light-bulb-bedazzled steak through a straw. In her commentary on the bit, she noted that "it (the steak) has everything the Democrats hate. If I could have put an SUV on this, I would have."

What does this have to do with political science? The answer perhaps ought to be obvious.

Back in 2005, ABC News and the Washington Post commissioned a poll about public opinion on traffic. Among other things, pollsters asked 1204 lucky, randomly-selected Americans:

> "What kind of vehicle do you usually drive – a car, an SUV, a pickup truck, or what?"

In this part, we'll explore the political dynamics of car ownership, using the data from the 2005 ABC/WP poll. The main variable of interest is `cartype`, coded one for cars, two for SUVs, and three for pickup trucks. Covariates include dummy variables for `urban` residence, being `married`, having `kids`, and being `black` and/or `female`, as well as a naturally coded variable for `age` and an ordinal variable for level of `education`. Best of all, we also have two dichotomous variables for political party (`democrat` and `GOP`, with independents as our baseline) and a four-point ordinal scale indicating each respondent's approval or disapproval for then-President Bush. All data are available on the PLSC 504 Github repository, in the folder creatively names "Exercises."

Your instructions for this part of the exercise are:

1. After examining summary statistics, estimate a multinomial logit (MNL) model of vehicle type ownership. Report your estimation results, and interpret these findings, in statistical and substantive terms. Are the results in the "expected directions"? Discuss their statistical significance.
2. Using the MNL results, generate and examine the predicted probabilities of each type of vehicle, across the range of values for *one* of your more important independent variables, using tables or graphs of the probabilities. Interpret these results in substantive terms.
3. Examine / test for whether the data/model conform to the MNL model's IIA assumption. Use whatever methods you can / are aware of, and discuss your findings on this point in both statistical and substantive terms.
4. Irrespective of the results of the IIA test(s), reestimate the same specification using HEV and multinomial probit models. Compare those findings to the MNL results, and briefly discuss similarities and differences. You might find the vignettes for the `mlogit` package useful here.
5. Finally, briefly discuss *in substantive terms* what your statistical conclusions suggest for the relationship between political ideology/preferences and automobile ownership.

This assignment is due *electronically*, as a *PDF file*, on or before 11:59 p.m. ET on Friday, September 20, 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.