The fmlogit Package: A Light Document

Xinde James Ji

Oct 10, 2016

This document provides documentations for the fmlogit package in R. Updates will be published at my github site. Any suggestions or concerns are welcomed.

What is the fractional multinomial logit model?

Fractional multinomial responses arises naturally in various occasions. For example, a municipality allocates its budgets to multiple departments, and we are interested in the proportion of the budgets that each department receives. Or, there are multiple candidates in a presendential election, and we are interested in the percentage of support for each candidate in each state.

The model is distinct in that 1) each of the response lies between 0 and 1, and 2) the share of all responses adds up to one. The fmlogit model utilizes the two distinct factors, and model it explicitly.

How to install fmlogit

Type the following code into your R console:

```
install(devtools)
library(devtools)
install_github("f1kidd/fmlogit")
library(fmlogit)
```

So, why do we even need fmlogit in R?

Don't we already have an fmlogit module in Stata? Yes, and you are very welcome to check that out.

However, this package offers several advantages over Stata's fmlogit module, namely, ## 1. Integration with the R Platform Implementating the model in R offers the opportunity to integrate the whole empirical process. With the help of numerous R packages, everything can be accomplished in R from data processing, estimation, post-estimation, to final manuscript writing. This is a huge advantage over stata.

2. Post-estimation Improvements

The marginal effect estimation in this package is much faster than Stata's fmlogit package. In this package user can specify which variable(s), and what effect to be calculated. This results in a huge gain in running time for the post-estimation commands.

Also, this package allows hypothesis testing for marginal and discrete effects while Stata does not. The standard error is calculated via Krinsky-Robb method, which allows empirical hypothesis testing without knowing the underlying distribution of the effects.

3. Estimation Flexibility

This package allows factor variable inputs, and automatically transform it into dummy variables. This is not (explicitly) allowed in Stata.

4. Extensions

This package also allows the user to calculate and infer from a vector of "willingness to pay" measures. The user can calculate the mean and the standard error of the effect on overall WTP increase, which is implicitly derived from the choice dynamics.

How the estimator works?

The estimator used here is an extension of Papke and Wooldridge(1996)'s paper, in which they proposed a quasi-maximum likelihood(QMLE) estimator for fractional response variables. As their approach applies to binary response variables, here we expand it to a multinomial response variables with fractional structure.

The basic step of the estimator is the following: ## Step 1. Construct the multinomial logit likelihood This step is straight forward. A simple multinomial logit transformation will do the job. For detailed derivations & formula, please see the technical document here where I explained the econometric steps in detail. ## Step 2. Maximize the sum of the log likelihood function Generally speaking, R is not the most efficient scientific computing machine that exists, and that is the tradeoff we have to face. Here the program offers several maximization methods provided in the *maxLik* package. The recommended algorithm is either conjugate gradients (CG), or Berndt-Hall-Hall-Hausman (BHHH). For a large dataset it may take a while (running for one hour is entirely possible, so don't terminate the program just yet). ## step 3. Calculate robust standard error Here the program follows Papke & Wooldridge (1996)'s paper, and construct the robust standard error estimator for the parameters. The program also offers a simple z-test for parameters based on the standard error.

How post-estimation commands works?

Interpreting marginal and discrete effects for limited dependent variables can be tricky, and this is especially true for multinomial logit models. The coefficients obtained in the regression model represents the logit-transformed odds ratio for that specific choice against the baseline choice. This is not intuitive at all in terms of what are the actual effects on that specific choice. The bottom line is, the coefficients and standard errors obtained in the original models are not the basis for evaluating hypotheses.

Marginal and Discrete Effects

Instead, researchers need to compute what's called the "marginal effect", as we usually do in linear models. However, the marginal effect in logit type models is tricky because the effects are heterogeneous across different observations. In other word, each unique set of Xb may have a unique marginal effect.

Typically, two types of summary measures are used to illustrate the global average marginal effects: the marginal effects at the mean (MEM), which is the effect at mean x values; and the average marginal effects (AME), which is the average effect across all observations. We allow both of the two options to be specified.

A more inclusive approach will be to plot the marginal effect of interest across all individuals. This is not provided in the function, but can certainly be implemented in future developments. Another possibility will be to calculate the so-called locally averaged treatment effect (LATE), where the effect of interest will be centered around a certain range of values.

Standard Errors

Here we adopt the simulation-based Krinsky-Robb method to compute standard errors for marginal and discrete effects as oppose to the empirical delta method used in Stata. These two methods should be asymptotically equivalent. However, using Krinsky-Robb allow us to perform hypothesis testing on the effects, while Delta method cannot accomplish that.

The hypothesis testing here is very simple: say we test $H_0 : D_j = 0$. We just need to compare 0 with our N draws, and see if it falls out of the 95% mass. This is a major advantage we provide here comparing with Stata's fmlogit module.

Practical Concerns

One of the concerns for the package is the computation speed of the estimation process. The maximization process can take somewhere from 20 seconds to 1 hour, depending on how large the dataset is. This is certainly a limit of this package. This is the inherent drawback of R's computation speed, and I can do nothing about that.

However, the loss in estimation will certainly be compensated in the post-estimation process. Stata's dfmlogit command is very slow (takes somewhere between 5-60 minutes), while here the effects calculation takes seconds to complete.

References

Papke, L. E. and Wooldridge, J. M. (1996), Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J. Appl. Econ., 11: 619-632.

Wulff, Jesper N. "Interpreting Results From the Multinomial Logit Model Demonstrated by Foreign Market Entry." Organizational Research Methods (2014): 1094428114560024.