The fmlogit Package: A Document

Xinde James Ji

Thursday, May 19, 2016

This document provides documentations for the fmlogit package in R. Updates will be published at my github site: https://github.com/f1kidd/fmlogit. Any suggestions or concerns are welcomed[1].

Motivation

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.

However, fractional multinomial logit model is *underrepresented* in the booming era of statistical softwares. The model itself is a coupling of *fractional response models*, which deals with responses which are proportional or fractional, and *multinomial response models*, which deals with binary responses of multiple options. As there are multiple softwares that deals with fractional logits and multinomial logits, the only software package that deals with fractional multinomial logits is Stata's fmlogit package by Maarten Buis. This package will contribute to the sea of software packages by providing an implementation of the model in R.

Note here that fractional multinomial logit model is a consistent estimator of fractional multinomial responses. Other estimation strategies are certainly useful in estimating fractional multinomials. Dirichlet regression models and two-limit tobit can also deal with responses that are ranged between 0 and 1. However, those models do not have the additional restriction that the proportions sums up to one, and thus will not be preferrable in this case.

Econometric Model

The basis of this function is 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.

We start by writing:[2]

$$E(y_{ij}|x_i) = G(x_i\beta_j)$$

for the j^{th} choice of the i^{th} obsevation , where G(.) is a know function satisfying 0<G(z)<1 for all $z \in \mathbb{R}$. Note that here we only allow for common covariates of x_i, and not for choice-specific attributes. Following the logit convention, G(.) is chosen to be the multinomial logit function, with the form:

$$G(z_j) = \frac{exp(z_j)}{\sum_{k=1}^{J} exp(z_k)}$$

And the multinomial likelihood function, is thus given by

$$ln(L_i) = \sum_{j=1}^{J} y_{ij} ln(G(x_i\beta_j))$$

[1] email: xji@vt.edu

[2] The demonstration below is in individual specific notation, but matrix notation is not hard to obtain from the individual specific notations. The actual function uses matrix calculation, which increases algorithm speed.

And Papke and Wooldridge(1996) showed that the QMLE estimator of β, obtained by the the maximazation problem

$$argmax_\beta \sum_{i=1}^{N} ln(L_i)$$

is a consistent estimator for β if G(z) is the correct functional form for E(y|x).

To estimate the standard error for the QMLE estimator, define $g(z_j) \equiv \partial G(z_j)/\partial z_j$, the partial derivative of the multinomial logit function with respect to choice j. Specifically, $g(z_j)$ has the following functional form:

$$g(z_j) = \frac{\hat{E}\hat{S} - \hat{E}^2}{\hat{S}^2}$$

where $\hat{E} = exp(x_i\beta_j)$, and $\hat{S} = \sum_{k=1}^{J} exp(x_i\beta_k)$.

A robust asymptotic standard error , $\hat{\beta_j}$, is given by the diagonal element of the following matrix:

$$\hat{A_j}^{-1}\hat{B}_j\hat{A_j}^{-1}$$

where

$$\hat{A} = \sum_{i=1}^{N} \frac{\hat{g}_{ij}^2 \mathbf{x}_i' \mathbf{x}_i}{\hat{G}_{ij}(1-\hat{G}_{ij})}$$

$$\hat{B} = \sum_{i=1}^{N} \frac{\hat{u}_{ij}^2 \hat{g}_{ij}^2 \mathbf{x}_i' \mathbf{x}_i}{[\hat{G}_{ij}](1-\hat{G}_{ij})]^2}$$

in which $\hat{u}_{ij}$ is the residual for the j^{th} choice of the i^{th} observation, given by $\hat{u}_{ij} = y_{ij} - G(x_i\beta_j)$. Specifically, $\hat{A}$ is the information matrix, which is not a consistent estimator itself, and $\hat{B}$ is a weight correction for A.

In most binary / multinomial response models, the convention is to treat one of the choices as a baseline. Here we apply the same logic, and treat j=1 as the baseline scenario. This implicitly generates a restriction that beta_1=0, and all other betas are the marginal difference to the baseline case.

Practical Concerns

Optimization Method

This function calls *optim()* to maximize the quasi-likelihood function. The *optim()* function provides several different maximization methods, including quasi-Newton, conjugate gradients, simulated annealing, etc., and the choice of optimization method can create vastly different parameter estimates. Here it is recommended that the conjugate gradients(CG) method being used in QMLE, as CG brings comparable results to Stata's fmlogit package. Quasi-Newton method is known to create a different set of parameter estimates, and it can be as large as 40% off from CG results for some parameters.

Robust Standard Error

It is worth noting that the robust standard error created in this function is consistently lower than that created in Stata's fmlogit package, typically by about 20%. However, the robust SE here is a consistent estimator following Pakpe and Wooldridge(1996)'s $\hat{A_j}^{-1}\hat{B_j}\hat{A_j}^{-1}$ estimator, so it is recommended that the number should be used with causion.

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.