Economics Bulletin

Volume 40, Issue 4

Fractional logit estimation under varying spatial resolution

Jingyu Song
Nationwide Mutual Insurance Company

Paul V Preckel
Purdue University

Michael S Delgado
Purdue University

Abstract

We propose a method for estimating logit regression models in the case that the independent variables are measured at a finer-scale spatial resolution than the dependent variable. Whereas the traditional approach is to aggregate the fine-scale data to the resolution of the dependent variable prior to estimation, we propose integrating the aggregation directly into the regression so as to maximize the value of information contained at the fine-scale resolution. Monte Carlo simulations show reasonable finite sample performance and that the traditional approach is biased. Our estimator is applicable in many cases that use remotely sensed or GIS data, such as land use problems.

Citation: Jingyu Song and Paul V Preckel and Michael S Delgado, (2020) "Fractional logit estimation under varying spatial resolution", *Economics Bulletin*, Volume 40, Issue 4, pages 2959-2963
Contact: Jingyu Song - songjingyu686@gmail.com, Paul V Preckel - preckel@purdue.edu, Michael S Delgado - delgado2@purdue.edu.
Submitted: June 23, 2020.   **Published:** November 14, 2020.

1. Introduction

Databases used routinely by researchers in applied economics and social sciences are often organized according to a particular spatial structure. Traditional economic data (e.g., output) is frequently available only at the level of geopolitical regions (e.g., states, counties, or provinces); these geopolitical divisions can generally be referred to as "administrative units" (AUs). Data we wish to use as drivers for the economic data may be available at much finer spatial resolutions from sources such as satellite imagery or GIS databases. While the detail offered by these fine spatial scale data is tremendous, researchers are challenged to maximize the information value without sacrificing the data variation that is needed for econometrically identifying the parameters or mechanisms of interest. Often, researchers simply aggregate the data to a consistent (or similar) spatial unit. The simplest and perhaps most prominent example of the aggregation approach is to layer the pixelated independent variable data over the dependent variable AU measurements and average the pixelated data to the AU level. Yet, whether averaging the independent variables to the AU level eliminates useful information is unclear at best.

An example of the empirical problem we have in mind is the prediction of whether a particular land pixel will be dedicated to a particular use, such as being planted in a particular crop. The model we develop may be used to predict probabilities or shares across multiple uses (e.g. Papke and Wooldridge 1996; Mullahy 2015; Song et al. 2018), though for simplicity our example focuses on a singular use. A typical dataset might include fine-scale (pixel-level) data for the independent variables, such as soil quality or climate, and coarse-resolution (AU-level) data for the dependent variable, such as the share of land in the province planted in the crop. One may be interested in predicting either the probability that an entire pixel is put into a particular use (e.g., entirely planted in corn), or in predicting the share of land in a pixel that is put into the particular use (e.g., share of land planted in corn). The traditional approach would be to aggregate the fine-scale data to the province (e.g., AU) level, and estimate a standard logit regression to predict the probability (or share) of land use at the AU level. However, while all data are used, estimation and prediction are conducted at the lower spatial resolution, and the information regarding heterogeneity within the AU is lost.

We propose to re-orient the logit regression model such that the fine-scale resolution is more fully used, providing a framework for estimating pixel-level probabilities (or shares). Specifically, we integrate the data aggregation step directly into the econometric model, rather than *prior* to parameter estimation. Our approach exploits nonlinearity in the logit model, recognizing that logit regression using averaged data is not the same as an averaged logit regression using pixelated data. A Monte Carlo experiment demonstrates the excellent finite sample performance of our estimator and shows that the traditional averaged approach is, in general, biased.

2. Estimation Framework

The following notation facilitates the exposition of our estimator and is developed for the general, multiple use case:

j the index of AUs
i the index of pixels within an AU

k the index of the use categories (dependent variables)
$\boldsymbol{X}_{ij}$ the observations of the independent variables for pixel i in AU j
y_{jk} the observations of the dependent variables for AU j for variable k (fractions that sum to one across k)
$\boldsymbol{W}()$ a transformation function mapping $\boldsymbol{X}_{ij}$ to functions of the independent variables
$\boldsymbol{\beta}_k$ the vector of coefficients to be estimated for dependent variable k
A_{ij} area in pixel i within AU j

Using this notation, we define the logistic regression problem. Define the probability that pixel i in AU j is of type k by:

$$G_{ijk}\big(\boldsymbol{W}(\boldsymbol{X}_{ij}),\boldsymbol{\beta}_k\big) = \frac{\exp\,(\boldsymbol{W}(\boldsymbol{X}_{ij})\boldsymbol{\beta}_k)}{\sum_{k=1}^{K}\exp\,(\boldsymbol{W}(\boldsymbol{X}_{ij})\boldsymbol{\beta}_k)} \qquad (1)$$

where we make the usual normalization $\boldsymbol{\beta}_1 = 0$ for identification. Here, we use a generic form of $\boldsymbol{W}()$ to preserve flexibility. The area-weighted average of these probabilities is:

$$H_{jk} = \frac{\sum_{i\in I_j} G_{ijk}\big(\boldsymbol{W}(\boldsymbol{X}_{ij}),\boldsymbol{\beta}_k\big)A_{ij}}{\sum_{i\in I_j} A_{ij}}\,. \qquad (2)$$

Estimation proceeds following the quasi-likelihood approach, via the quasi-likelihood function:

$$\mathcal{L} = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk}\ln H_{jk}. \qquad (3)$$

3. Monte Carlo Experiment

Monte Carlo Design: We conduct a Monte Carlo experiment to demonstrate the properties and finite sample performance of our estimator. Following the nature of the empirical problem we envision, we artificially construct a land grid of “administrative units” that each consist of a number of smaller “pixels”. For simplicity, we restrict all pixels to have the same area; empirically, differences in area size across pixels and AUs is both permitted and expected. (The estimation framework shown previously allows for this heterogeneity of pixel size.) Given these AUs and pixels, we randomly generate two independent variables, from which we construct our dependent variable as a share of the total area of each pixel (so as to mimic a share of land devoted to a particular use, such as share of land planted with corn). To keep the model transparent and straightforward, we assume that each pixel is devoted to only two activities, so that the shares of “use” and “non-use” sum to one within each pixel.

The values of the independent variables are designed following Song et al. (2018), whereby the first independent variable we generate mimics temperature patterns across sub-global regions and the second independent variable mimics land slope. The key difference between these two variables is that temperature has moderate variation within an AU but more substantial variation across AUs. Specifically, the first independent variable is created by generating a normal random number with mean 15 and standard deviation 4, to which we add a uniform random variable for each pixel within the AU on [-2, 2]. The second independent variable, mimicking slope, is drawn from a uniform random variable on [0, 1] for all AUs and pixels within those AUs. To be clear, we construct these variables to mimic typical land use variables to create a more realistic experimental environment, but there is no theoretical requirement that would render our experimental results non-general.

For these exercises, we fix $\boldsymbol{W}()$ to be the identity, and choose $\boldsymbol{\beta}_2$ to be [–7, 0.6, –6], in which the components correspond to the intercept and coefficients for the first and second independent variables, respectively. Using this structure, we generate the pixel-level value of the fractions for the two land uses according to (1), and calculate the weighted average of these at the AU level according to (2). This "true value" of the use fraction is then multiplied by a randomly generated error that is drawn from a beta distribution with mean equal to one and with a support from zero to one over the true value of the use fraction variable. The two shape parameters of this beta distributed error are chosen so that their minimum is 4. Varying this value alters the signal to noise ratio, but does not change the qualitative results of our experiment.

Finally, the number of observations of our dependent variable is the number of AUs, for which the experimental design uses 50, 100, 500, 1,000, and 2,000 AUs as increasing sample sizes. Within each AU, the number of pixels is randomly generated as the square of the truncation of a uniform random variate on the interval [20, 51] giving a range of pixels per AU of 400 to 2500. We conduct 1,000 trials for each of these experiments.

Monte Carlo Results: The Monte Carlo results are displayed in the left panel of Table 1 titled "Pixel Data", indicating that the independent variables are defined at the pixel level as described above. For each set of experiments, we report the mean parameter estimate, as well as the standard deviation of the estimates and the root mean squared error (RMSE) as a measure of performance. We see that the coefficients are converging towards the true values, and both the standard deviation and RMSE values are uniformly decreasing, as the sample size (number of AUs) increases. These results demonstrate consistency.

How does this compare with the traditionally-used alternative of averaging the pixel-level independent variable observations up to the AU level? We maintain the notation from the Estimation Framework section, noting that $K = 2$ (i.e. $k = 1$ for non-use and $k = 2$ for use). To examine this question, we re-estimate our simulated model following an *a priori* averaging of the pixel level data to the AU level:

$$G_{jk}\big(\boldsymbol{W}(\bar{\boldsymbol{X}}_j),\boldsymbol{\beta}_k\big) = \frac{\exp\,(\boldsymbol{W}(\bar{\boldsymbol{X}}_j)\boldsymbol{\beta}_k)}{\sum_{k=1}^{K}\exp\,(\boldsymbol{W}(\bar{\boldsymbol{X}}_j)\boldsymbol{\beta}_k)} \tag{4}$$

where we again normalize $\boldsymbol{\beta}_1 = 0$, and $\boldsymbol{X}_{ij}$ is replaced in (1) with the area-weighted average:

$$\bar{\boldsymbol{X}}_j = \frac{\sum_{i\in I_j}\boldsymbol{X}_{ij}A_{ij}}{\sum_{i\in I_j}A_{ij}}. \tag{5}$$

The analogs to (2) and (3) are then given as:

$$H_{jk} = \frac{\sum_{i\in I_j} G_{jk}(\boldsymbol{W}(\bar{\boldsymbol{X}}_j),\boldsymbol{\beta}_k)A_{ij}}{\sum_{i\in I_j}A_{ij}} = \frac{G_{jk}(\boldsymbol{W}(\bar{\boldsymbol{X}}_j),\boldsymbol{\beta}_k)\sum_{i\in I_j}A_{ij}}{\sum_{i\in I_j}A_{ij}} = G_{jk}\big(\boldsymbol{W}(\bar{\boldsymbol{X}}_j),\boldsymbol{\beta}_k\big), \tag{6}$$

and

$$\mathcal{L} = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \ln H_{jk} = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \ln G_{jk}. \tag{7}$$

For comparison purposes, the same data from the pixel level simulations are used for these averaged data results, which are displayed in the columns labeled "Averaged Data" in Table 1. The mean parameter estimates are much further from the true values than with our proposed estimates. While the standard deviations and RMSE values are declining as the sample size increases, this estimator appears to be biased even at a relatively large sample size of 2,000 observations.

Table 1. Monte Carlo Results

	Pixel Data			Averaged Data		
No. of AUs	$\beta_{2,0} = -7$	$\beta_{2,1} = 0.6$	$\beta_{2,2} = -6$	$\beta_{2,0} = -7$	$\beta_{2,1} = 0.6$	$\beta_{2,2} = -6$
Mean						
50	-12.1152	1.1765	-14.985	-3.5357	0.3229	-3.6801
100	-8.7025	0.7732	-8.3620	-4.0540	0.3557	-3.7370
500	-7.2654	0.6304	-6.4875	-4.4253	0.3831	-3.9178
1,000	-7.1830	0.6367	-6.8603	-4.3881	0.3816	-3.9447
2,000	-7.0741	0.6184	-6.4651	-4.3844	0.3847	-4.0523
Std. Dev.						
50	7.2139	0.6104	13.4019	4.3441	0.0154	8.6483
100	2.9552	0.1835	6.0554	3.7126	0.0115	7.4008
500	0.9113	0.0570	1.9835	1.6278	0.0044	3.2567
1,000	0.3261	0.0425	0.7645	1.0927	0.0031	2.1836
2,000	0.2168	0.0290	0.5362	0.8078	0.0023	1.6119
RMSE						
50	8.8405	0.8394	16.1295	5.5546	0.2775	8.9498
100	3.4092	0.2523	6.4970	4.7380	0.2446	7.7355
500	0.9487	0.0646	2.0415	3.0457	0.2170	3.8641
1,000	0.3739	0.0561	1.1506	2.8311	0.2184	2.9980
2,000	0.2290	0.0343	0.7096	2.7374	0.2153	2.5277

4. Conclusion

We propose a method of integrating independent variable data organized at a different spatial resolution from the independent variables into a logit regression so as to maximize the value of the information contained in the fine-scale spatial data while maintaining econometric tractability. Monte Carlo experiments show that our estimator performs well, and that a traditional pre-averaging approach is generally biased. Our estimator is applicable in many empirical contexts in which independent variables are available at a less aggregated level than the dependent variable. Finally, our approach provides the wherewithal to predict at the less aggregated level.

References

Mullahy, J. (2015) “Multivariate fractional regression estimation of econometric share models” *Journal of Econometric Methods* **4(1)**, 71–100.

Papke, L. E. and Wooldridge, J. M. (1996) “Econometric methods for fractional response variables with an application to 401(k) plan participation rates” *Journal of Applied Econometrics* **11**, 619–632.

Song, J., Delgado, M. S., Preckel, P. V. and Villoria, N. B. (2018) “Downscaling of national crop area statistics using drivers of cropland productivity measured at fine resolutions” *PLOS ONE* **13(10)**: e0205152. https://doi.org/10.1371/journal.pone.0205152