Imputation of Suppressed CBP Employment Records

D. Lane Register[1], D.M. Lambert[1], B.C. English[1], K. L.Jensen[1], R.J. Menard[1], and J.P. Brown[2]

[1]Department of Agricultural and Resource Economics, University of Tennessee, Knoxville, TN.

[2]USDA Economic Research Service, Washington, D.C.

Introduction

- Suppression of employment records in the US County Business Patterns (CBP) data sets constrains the detail of new methods and recent advances in the analysis of the geographic distribution of firms and employment.
- Data sets created by imputation procedures can be purchased, but cost often puts them beyond the reach of many research budgets.
- Fortunately, methods exist whereby researchers can impute suppressed employment records.
- A comparison of these procedures is necessary to assess the accuracy and flexibility of each.

Objectives

This research compares an existing goal programming approach created by Zhang and Guldmann (2009) (the ZG approach) and a new method, an Iteratively Constrained Rebalanced Matrix (ICRM) procedure with known, randomly suppressed data sets, constructed similar to the CBP.

Exploring the CBP

FIPS	NAICS	empflag	emp	est	n1_4	n5_9	n10_19
4001	22----	B	0	9	7	1	1

- Data is arranged in a hierarchy of geographical (FIPS) and industry (NAICS) classification.
- Suppressed record noted by letter in "empflag" column.
- Each employment flag corresponds with a range of employment in which the suppressed "emp" record falls.
- Establishment size ranges provide additional information about employment. In the example above, there are 9 establishments, 7 of which have between 1 and 4 employees, 1 has between 5 and 9 employees, and 1 has between 10 and 19. These intervals extend out to a range of 5000+ employees.
- From the flag and site interval information, the uncertainty about the suppressed employment value is narrowed from 20-99 employees, to 22-56 employees (57% decrease in uncertainty).

Methods

The ZG Approach (Zhang and Guldmann, 2009)

The goal programming approach proposed by ZG minimizes the sum of deviations of imputed estimates from constraints based on employment flag intervals, establishment size ranges, and known county and sector employment totals.

The ICRM Procedure

This approach uses nonlinear programming and a Gauss-Seidel minimization routine, in a process similar to the standard RAS algorithm. The procedure is completed iteratively, using the following steps:

1. The objective function minimizes the cross-entropy of row and column scaling factors,

$$\min_{y_i, z_j} \sum_i y_i \ln(y_i) + \sum_j z_j \ln(z_j)$$

subject to

$$\sum_j a^t_{0(ij)} \cdot y_i \cdot z_j = c_i \quad \forall\, j$$

$$\sum_i a^t_{0(ij)} \cdot y_i \cdot z_j = s_j \quad \forall\, i$$

2. The values for are imputed such that

$$a^{t*}_{ij} = \begin{cases} pmin_{ij} \le a^{t-1}_{0(ij)} \cdot y^*_i \cdot z^*_j \le pmax_{ij} & if\ a_{0(ij)}\ is\ suppresssed \\ a_{0(ij)} & otherwise \end{cases}$$

where

$$pmin_{ij} = MAX\left(fmin_{ij}, \sum_k tmin_k * t_{ijk}\right)$$

$$pmax_{ij} = MIN\left(fmax_{ij}, \sum_k tmax_k * t_{ijk}\right)$$

3. Convergence of the vectors is checked such that the difference in $\boldsymbol{A}^t$ and $\boldsymbol{A}^{t-1}$after t iterations is a very small number.
4. If convergence is not reached, return to step one.

Identifier	Description
i	county index
j	sector index (1→21 for NAICS 11→99)
k	establishment size interval index (1→12)
	number of establishments in county *i,* sector *j*, and
t_{ijk}	establishment interval *k*
s_j	total employment for sector *j* across spatial units
c_i	total employment for county *i* across all sectors
$tmax_k$	employment for upper bound for interval *k*
$tmin_k$	employment for lower bound for interval k
$fmax_{ij}$	maximum employment of sector *j* in county *i,* if flagged
$fmin_{ij}$	minimum employment of sector *j* in county *i,* if flagged

Monte Carlo Simulation

- Simulated employment data sets that replicate the structure of a CBP data set.
- Increasing percentages of data were randomly suppressed in the simulated data sets.
- Goodness of fit was assessed for each method as the sum of the squared deviations between the true and imputed values to compare precision of the imputation procedures.

Results

- The ZG imputation procedure was feasible in 100% of the Monte Carlo trials.
- The unconstrained and constrained ICRM procedures had difficulty converging at lower levels of data suppression (<50%).
- All approaches were superior to the mid-point imputation approach.
- Compared to the "true" simulated employment patterns, mean absolute percent deviations of the constrained ICRM were lower then the ZG approach as the percent of data suppressed increased.

Conclusions and Further research

- Proposed method appears promising, computationally and in terms of precision.
- Method appears to work well when about half of the data are suppressed.
- Determine frequency of unconstrained ICRM estimated missing employment data outside relevant bounds.
- Incorporate ICRM bound constraints directly into nonlinear program.
- Investigate adaptive constraints, updating bounds at each iteration for ICRM approach.

Citations

- Zhang, S., and J.M. Guldmann. 2009. Estimating Suppressed Data in Regional Economic Databases: A Goal-Programming Approach. *European Journal of Operational Research* 192(2009): 521-537.

Poster presented at the Agricultural and Applied Economics Association 2012 Annual Meeting, August 12-14, 2012, Seattle, Washington. **Corresponding Author:** Dayton Lambert, Department of Agricultural Resource Economics, The University of Tennessee, 321C Morgan Hall, 2621 Morgan Circle, Knoxville, TN 37996. email: dlamber1@utk.edu. The views expressed here are those of the authors and do not reflect their respective institutions.