

経済センサス-活動調査のためのロバストな比率補定の方法について

独立行政法人統計センター 統計技術研究課 和田 かず美
独立行政法人統計センター 統計技術研究課 坂下 佳一郎

1. 比率補定とその一般化

比率補定 (ratio imputation) は、補定を行う目的変数 y が、ある単変量の説明変数 x との比がほぼ定数になる場合に使用される。 y と x の比 r を、 x と y がともに欠測のない観測値を用いて $\hat{r} = \sum_{k \in \text{obs}} y_k / \sum_{k \in \text{obs}} x_k$ により推定し、欠測値 y_i を推定値 $\hat{y}_i = r x_i$ により補定する (Cochran, 1977, p.31)。ここで、“obs”は欠測のない観測値のレコードを示す。一般的にこの補定のモデルは

$$y_i = r x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 x_i) \quad (1)$$

と表現され、その誤差項 ϵ_i は平均0で分散が x に比例する独立でランダムな変数である (Cochran, 1977, p.158)。これを、 x と関係を持たない誤差項 $\epsilon_i = \epsilon_i / \sqrt{x_i}$ により書き換えると、

$$y_i = r x_i + \epsilon_i \sqrt{x_i}, \quad \epsilon_i \sim N(0, x_i).$$

さらに、 β を任意の定数として、誤差項を $\epsilon_i = \epsilon_i x_i^\beta$ という形に拡張すると、そのモデルと推定量 \hat{r} は、

$$y_i = r x_i + \epsilon_i x_i^\beta, \quad \hat{r} = \frac{\sum y_i x_i^{1-2\beta}}{\sum x_i^{2(1-\beta)}}. \quad (2)$$

2. 提案するロバスト化手法

モデル(2) に M-推定量の考え方を導入したロバスト化比推定量 \hat{r}_{rob} 、そのモデル式と残差 $\hat{\epsilon}_i$ は、

$$\hat{r}_{rob} = \frac{\sum w_i y_i}{\sum w_i x_i}, \quad w_i y_i = r_{rob} w_i x_i + \epsilon_i \sqrt{w_i x_i}, \quad \hat{\epsilon}_i = \frac{y_i}{\sqrt{x_i}} - \hat{r}_{rob} \sqrt{x_i}. \quad (3)$$

この推定量の計算は、Bienias et al. (1997) による回帰のための繰返し加重最小二乗法 (IRLS: Iteratively Reweighted Least Squares) に準拠し、ウェイト関数は Tukey の biweight

$$w\left(\frac{\hat{\epsilon}_i}{\sigma_{AAD}}\right) = w(e) = \begin{cases} \left[1 - \left(\frac{e}{c}\right)^2\right]^2 & |e| \leq c \\ 0 & |e| > c. \end{cases}$$

残差の尺度を平均絶対偏差(AAD)として、その平均変化率が1%未満で収束とする。経験的に3から8の間でユーザーが任意に設定する調整定数 c は、ロバスト性が最も低い8を選択した。

3. モデル選択

売上高、給与総額及び費用総額の欠測値について、それぞれ費用総額、費用総額、売上高を説明変数として補定を行うために、 $\beta = 1$ と $1/2$ の場合について実データによるシミュレーションによるモデル選択を行い、 $\beta = 1/2$ を採用した。

参考文献

- Bienias, J. L., Lassman, D. M., Scheleur, S. A. and Hogan H. (1997) Improving Outlier Detection in Two Establishment Surveys. *Statistical Data Editing 2 - Methods and Techniques*. (UNSC and UNECE eds.), pp. 76-83.
- Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed., Wiley handbooks in survey methodology, John Wiley & Sons.
- 和田かず美 (2012) 多変量外れ値の検出~繰返し加重最小二乗(IRLS)法による欠測値の補定方法~. 統計研究彙報第 69 号, pp.23-52. 総務省統計研修所.