参　考

独立行政法人
統計センター
統 計 技 術 研 究 課

ロバストな比率補定法について

■　比率補定（ratio imputation）

統計実務において、比率補定は、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。r を y と x の比とすると、欠測値 y_i は次のような推定値により補定される。

$$\tilde{y_i} = rx_i \tag{1}$$

一般に、r は未知なので、x と y がともに欠測のない観測値により以下のように推定する[De Waal et al. (2011)]。

$$\hat{r} = \frac{\sum_{k\in obs} y_i}{\sum_{k\in obs} x_i} \tag{2}$$

ここで、”obs”は欠測のない観測値のレコードを示す。この補定のモデルを次のように表すと、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 [Rao (1996)]。

$$y_i = rx_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 x_i) \tag{3}$$

x と y の相関が高いほどこの比推定量の推定効率は高いが、回帰と同様に外れ値に弱いことが知られている [e.g. Farrell and Barrera (2007)]。また、それぞれの変数の値を合算して比をとるという性質上、この推定量は特に数値の大きい観測値の影響を受けやすい。

■　比推定量のロバスト化

M-推定量の考え方を用いて、上述の比推定量をロバスト化する。M-推定量は、y の推定値からの乖離を示す誤差の大きさに応じて観測値に重み付けをすることにより、外れ値の影響を緩和することができる。

モデル(3)の誤差項 ϵ_i の分散には、x に比例して大きくなる不等分散性があり、このまま加重に用いれば外れ値と判定されるデータは x が大きいものに偏る。このため、分散が x と関係性を持たない新たな誤差項 ε_i を使用してモデル(3)を再表現したい。

$\varepsilon_i \sim N(0, \sigma^2)$ とすると、$\varepsilon_i = \epsilon_i / \sqrt{x_i}$ という関係があるため、(3)式は(4)式のように表

現することができる。

$$\frac{y_i}{\sqrt{x_i}} = r\sqrt{x_i} + \varepsilon_i$$

$$y_i = rx_i + \varepsilon_i\sqrt{x_i} \tag{4}$$

これに対応する残差 $\check{\varepsilon}_i$ は、式(5)のようになる。

$$\check{\varepsilon}_i = \frac{y_i}{\sqrt{x_i}} - \hat{r}\sqrt{x_i} \tag{5}$$

これにより、ロバスト化比推定量 $\hat{r}_{rob}$ とそのモデル式はそれぞれ式(6)及び(7)、対応する残差$\check{\varepsilon}_i$ は式(8)により得ることができる。

$$\hat{r}_{rob} = \frac{\sum w_i y_i}{\sum w_i x_i} \tag{6}$$

$$w_i y_i = r_{rob} w_i x_i + \varepsilon_i\sqrt{w_i x_i} \tag{7}$$

$$\check{\varepsilon}_i = \frac{y_i - \hat{r}_{rob} x_i}{\sqrt{x_i}} = \frac{y_i}{\sqrt{x_i}} - \hat{r}_{rob}\sqrt{x_i} \tag{8}$$

回帰 M-推定量の計算には、繰返し加重最小二乗法（IRLS: Iteratively Reweighted Least Squares）と呼ばれる計算アルゴリズムが、収束が早く計算が簡便なためによく使用される。Bienias et al.（1997）において紹介されている回帰のための IRLS アルゴリズムを、以下のように比推定に適用する。

1) 通常の比率補定と同様に、(2)式により $\hat{r}$ を算出し、これを初期値とする。あわせて、(5)式により残差$\check{\varepsilon}_i$を算出し、$\check{\varepsilon}_i$ とその尺度パラメータとなる平均絶対偏差(AAD: Average Absolute Deviation)から、後述のウェイト関数に基づいてw_iを得る。
2) 1)で得られたw_iを用いて、(6)式による $\hat{r}_{rob}$ と(8)式による残差 $\check{\varepsilon}_i$ 、$\check{\varepsilon}_i$ の AAD を算出し、再びウェイト関数に基づいて新たなw_iを算出する。
3) 新たなw_iを用いて再び 2)を繰り返す。このとき、直近とその一つ前の残差 $\check{\varepsilon}_i$ の AAD の変化率が 1%未満であれば、収束とみなして計算を終了し、最新の$\hat{r}_{rob}$ の値を補定に用いる比率とする。

ここで、欠測のない観測値のサイズを *n* とすると、AAD は下式により得ることができる。

$$\sigma_{\mathrm{AAD}} = \frac{1}{n}\sum_{i=1}^{n}|\check{\varepsilon}_i|$$

w_iを算出するためのウェイト関数には様々な選択肢があるが、ここでは Bienias et al.（1997）に準拠して次のような Tukey の biweight 関数を使用する。

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

この式は、残差 $\check{\varepsilon}$ を尺度パラメータσ_{AAD}で標準化した標準化誤差 e の絶対値が、定数 c より大きい場合のウェイト w は 0、c よりも小さい場合はその数字からウェイトを計算することを表している。c は調整定数と呼ばれ、経験的に 3 から 8 の間でユーザーが任意に設定することによりロバスト性を調整する。c が小さいほどウェイトが削られることになり、結果として推定のロバスト性が強くなる。

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.

De Waal, T., Pannekoek, J., Scholtus, S. (2011) *Handbook on Statistical Data Editing and Imputation,* Wiley handbooks in survey methodology. Hoboken, New Jersey: John Wiley & Sons.

Farrell, P. J. and Saliban-Berrera, M. (2006) A Comparison of Several Robust Estimators for a Finite Population Mean, *Journal of Statistical Studies*, Vol.26, pp.29-43.

Rao, J. N. K. (1996) On Variance Estimation With Imputed Survey Data, *Journal of the American Statistical Association*, Vol.91, pp.499−506.

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.

CART による補定ドメインの設定方法ついて

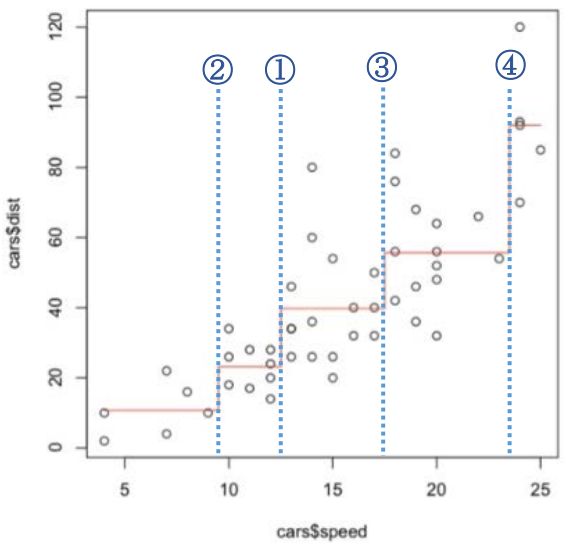
■ CART (Classification And Regression Tree: 分類木/回帰木)

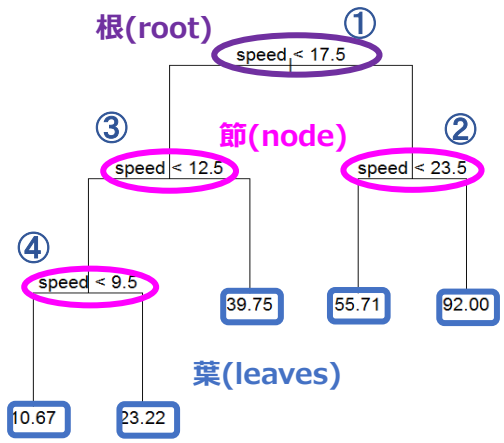
CART は、データマイニングや機械学習等の分野において、分類や予測のために広く使われている決定木（decision tree）と呼ばれる手法の一種である。目的変数がカテゴリ変数の場合は分類木、目的変数が連続変数の場合は回帰木と呼ばれる。

補定のためのドメイン設定において目的変数になるのは、カテゴリではなく連続値をとる補定のための比率 y/x であるため、回帰木について取り上げる。

■ 回帰木の例

統計ソフト R に組み込まれている cars データは、車のスピードと制動距離についての二変量データである。目的変数を制動距離、説明変数をスピードとしてこのデータに CART を適用すると、下の右側の図のような決定木が得られる。この決定木の最上部は根（root）、最下部が葉（leaves）、その間の分岐は節（node）と呼ばれる。





まず、根に注目すると、変数 speed が 17.5 より大きいかどうかで最初にデータが分割される。これは、左の散布図上で、①の線により示されている。CART での分岐の基準は、不純度を表すジニ係数で、分岐により不純度（目的変数のばらつき）が最も減少するよう

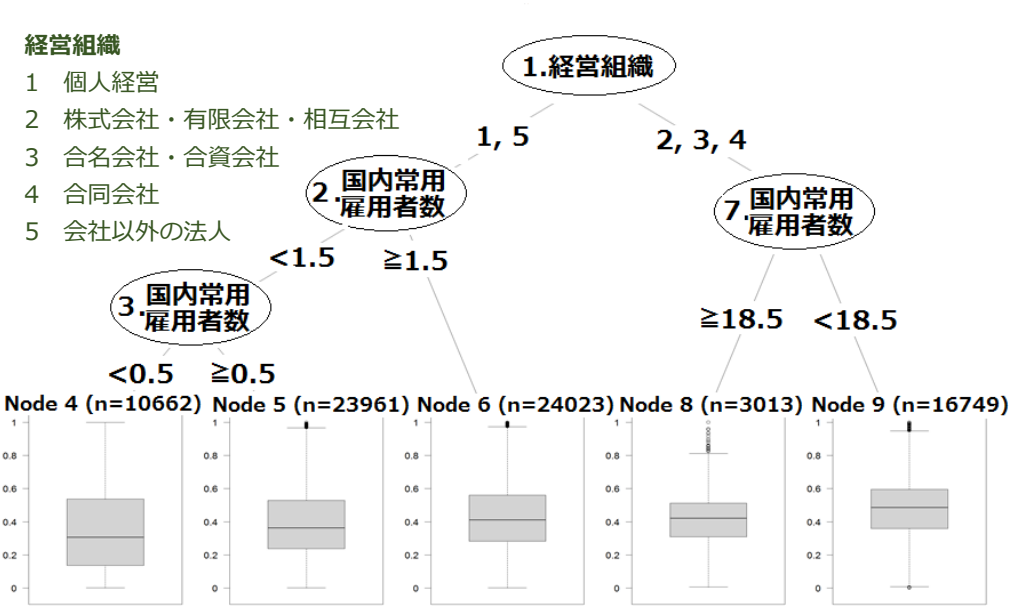
に分岐点が決定される。

最初の①の分岐で領域を二つに分割した後、②及び③の節でさらに各領域が二分され、そ下でさらに④により二分される領域もあるため、最終的に分割された領域（葉）の数は 5 つになる。この結果は、左側の散布図の中の階段状の赤線で表現されており、縦線部分は領域の境界を、横線部分は領域毎の目的変数の代表値を示している。

■ CART によるドメインの設定方法

補定のためのドメイン設定では、目的変数は補定のための比率 y/x、説明変数は質的変数の経営組織と人単位で整数値をとる国内常用雇用者数として、産業分類ドメインをさらに分割する可能性を検討する。結果のイメージは下図のとおりで、この場合、産業により設定された補定ドメインは、次のような条件によりさらに 5 分割できることがわかる。

[Node 4] 経営組織が 1 及び 5 で、国内常用雇用者数が 0 人
[Node 5] 経営組織が 1 及び 5 で、国内常用雇用者数が 1 人
[Node 6] 経営組織が 1 及び 5 で、国内常用雇用者数が 2 人以上
[Node 8] 経営組織が 2,3 及び 4 で、国内常用雇用者数が 18 人以下
[Node 9] 経営組織が 2,3 及び 4 で、国内常用雇用者数が 19 人以上



参考文献

Hastie, T., Tibshirani, R. Friedman, J. (2001) The Elements of Statistical Learning, Springer-Verlag, p.267-270.

ベリー, M. J. A., リノフ, G. (1999)「データマイニング手法」、海文堂.

ドメイン設定への U 検定の利用方法について

観測値のグループ（ここでは補定のためのドメイン）間の差異を見る方法の一つとして、下表のような統計的仮説検定法がある。ここでパラメトリックな方法とは、母集団に正規分布の仮定を置いたものを、ノンパラメトリックな方法は母集団に分布の仮定は置かないものを指す。

	パラメトリックな方法	**ノンパラメトリックな方法**
多群（対応なし） 等分散が前提	一元配置分散分析	クラスカル・ウォリス検定
二群（対応なし） 等分散が前提 不等分散対応	スチューデントの t 検定 ウェルチの t 検定	マン・ホイットニーの U 検定 ブルンナー・ムンツェル検定

例えば、スチューデントの t 検定は、母集団が正規分布、標本の分散がカイ二乗分布に従うという前提を置いている。今回使用する経理項目の比率は、正規分布の仮定を置くことができないため、ノンパラメトリックな方法を使用する必要がある。

三つ以上のグループについての検定にはクラスカル･ウォリス検定を用いる。ただしこの検定は、グループ内の差異の有無はわかるが、具体的にどのグループが違うのかはわからない。どのグループが違うのかを知るためには、U 検定など二群の差異を知るための検定を用いる必要があるが、同じグループ内で複数回検定を行う多重比較になる場合、グループ数が多くなるほど有意差は出にくい。

一方で、ここで必要なのは、同一母集団かどうかの仮説検定自体ではない。n が足りないドメインの統合先候補として、最も近いドメインがどれかを判別することが目的である。そこで、多重比較問題などは考慮せず、単に検定統計量の p 値が最も大きいドメインを候補として選び、データの分布や分散を確認した上で実際にドメインを統合するかどうかを決定する。

■　マン・ホイットニーの U 検定（Mann-Whitney's U-test）

U 検定は、ウィルコクソン・マン・ホイットニー検定やウィルコクソンの順位和検定とも呼ばれ、正規分布の仮定を必要としない二つのグループの差異についての検定方法である。この検定は、二つのグループが同じ母集団から標本抽出されたという仮定を置くため、実質的には等分散を仮定していることになる。

次のような仮説について、検定統計量 U を算出し、その p 値が有意水準 0.05 よりも小さければ帰無仮説が棄却される。

[帰無仮説 H_0]　両群がまったく同じ分布をしている

[対立仮説 H_1]　両群の分布が異なる

この方法は、実質的に順位を t 検定することとほぼ同じであるため、二つのグループの分散が異なる場合は、数値を順位に直して等分散の仮定を置かない t 検定（バートレットの方法）か、あるいはブリュンナー・ムンツェル検定を用いることもできる。

2 組の数，$x_1, x_2, \ldots, x_n$ と、$y_1, y_2, \ldots, y_m$ があるとき、$x_i > y_j$ を満たす (i, j) の組の数に、$x_i = y_j$ を満たす組の数の半分を足したものを U とすると，もしこれらの $n + m$ 個の数の並び順がランダムであれば，U の確率分布はある漸化式から計算できる。さらに、n, m が大きければ、U の分布はほぼ正規分布 $N(nm/2, nm(n+m+1)/12)$ になることを利用する。この U の値は、$n + m$ 個の中での $x_1, x_2, \ldots, x_n$ の順位の和から $n(n + 1)/2$ を引いた値に等しいので、U を使った検定は順位和を使った検定とみなすこともできる。

[出典:奥村(2005)]

<参考文献・資料>

奥村晴彦（2015）「Wilcoxon-Mann-Whitney 検定（WMW 検定）」、
[https://oku.edu.mie-u.ac.jp/~okumura/stat/wmw.html]

竹内啓編（1989）統計学辞典、東洋経済新報社.

Henkel, R. E. (1982)「統計的検定―統計学の基礎―」、松原・野上訳、朝倉書店.

Kasuya, E.（2001） Mann-Whitney U test when variances are unequal. Animal Behaviour,61:1247-1249.