公的統計調査における ロバストな比率の推定による 企業の経理項目の欠測の補定について

独立行政法人　統計センター
統計情報・技術部　統計技術研究課

坂下 佳一郎　床 裕佳子　和田 かず美

注：本報告の内容は、報告者個人の見解であり、必ずしも報告者所属機関の見解ではありません。

2015/09/09 @岡山大学
2015年度統計関連学会連合大会
公的統計(2)（午後・A会場） #6

本日のおしながき

1. 背景・研究目的
2. ロバストな比率補定(ratio imputation)
3. シミュレーションの概要
4. シミュレーション結果の評価方法
5. シミュレーション結果
6. まとめ
7. 今後の本研究の展開

1.背景・研究目的

背景

- 調査では欠測発生の懸念
- 仮にその欠測をそのままにすれば集計結果に影響
- 一方、欠測を補定(imputation)しても不適切な方法なら結果に別の影響

背景

- 経済センサス - 活動調査創設(2012)
  ⇒より詳細な経理項目を取得可能

- これまで企業の経理項目を大規模に補定を試みる実験は多くない

目的

- 経理項目間の比率を用いて補定を行う想定で、企業の経理項目の欠測を補定するより良い方法を検討
  - ロバストな方法で
  - 補定ドメイン（補定実施区分）も
- 公的統計調査データを活用したシミュレーションを実施し結果を分析

2.ロバストな比率補定(ratio imputation)

比率補定とは

- 公的統計で使われる基本的な方法
- 欠測に比率による偏りがない前提
- 補定対象２項目の一方のみ欠測の場合に有効
- ドメインごとに処理を実施

比率補定とは

- 補定対象2項目の比率y_i/x_iに着目し、各ドメインの補定比率$\bar{r}$を求める

$$\bar{r} = \frac{\sum_{i=1}^{n} \frac{y_i}{x_i}}{n}$$

（n: ドメイン内比率計算可能レコード数）

- 補定比率$\bar{r}$を欠測ではない値に掛けるもしくは割ることで欠測を補定

比率補定の長所

- ２つの項目を補定する際に必要な推定値が１つ
- 比率の分布の仮定を置く必要がない

比率補定適用の課題と対策

1. 補定比率のロバスト性
   …ロバスト比率補定（後述）を導入

2. ドメイン
   …補定に関するシミュレーションを
   実施し、良いドメインを見つける

ロバスト比率補定

- 回帰残差についてのIRLS（繰返し加重最小二乗法）のアプローチを、比率補定に適用
  - 比率$\bar{r}$の計算の際、各レコードの比率のドメイン内比率分布中心部からの乖離の大きさに応じてウエイトw_iを加重

$$\bar{r} = \frac{\sum_{i=1}^{n} \frac{y_i}{x_i} w_i}{\sum_{i=1}^{n} w_i}$$

【参考】IRLSについて

- M推定量を算出するためのアルゴリズム
- 計算は簡便で収束が早い
- 残差の大きさに応じてデータのウエイトを削り、外れ値の影響を緩和する

参考文献

Holland, P. W. and Welsch, R. E. (1977) Robust Regression Using Iteratively Reweighted Least-Squares, Communications in Statistics – Theory and methods, A6(9), pp.813-827

和田(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

ロバスト比率補定のアルゴリズム

1. 初期値としてドメイン内平均比率$\bar{r}^{(0)}$を算出

$$\bar{r}^{(0)} = \frac{\sum_{i=1}^{n} \frac{y_i}{x_i}}{n}$$

2. 各レコードの比率y_i/x_iと単純平均比率$\bar{r}^{(0)}$との差分$\varepsilon_i^{(0)}$を平均絶対偏差$s^{(0)}$により規格化した残差$e_i^{(0)} = \varepsilon_i^{(0)}/s^{(0)}$を求め、ウエイト関数に基づきウエイト$w_i^{(1)}$を算出

ロバスト比率補定のアルゴリズム（続き）

3. 2.の$w_i^{(1)}$を使い、加重平均比率$\bar{r}^{(1)}$を算出

$$\bar{r}^{(1)} = \frac{\sum_{i=1}^{n} \frac{y_i}{x_i} w_i^{(1)}}{\sum_{i=1}^{n} w_i^{(1)}}$$

4. 3.で$\bar{r}$を得る度にsを求め、前回のsとの変化率を比較。

1%以上：w_iを更新して$\bar{r}$を再計算。

1%未満：計算終了。

最後に得た比率$\bar{r}$を補定比率として採用

ロバスト比率補定のウエイト関数w_i

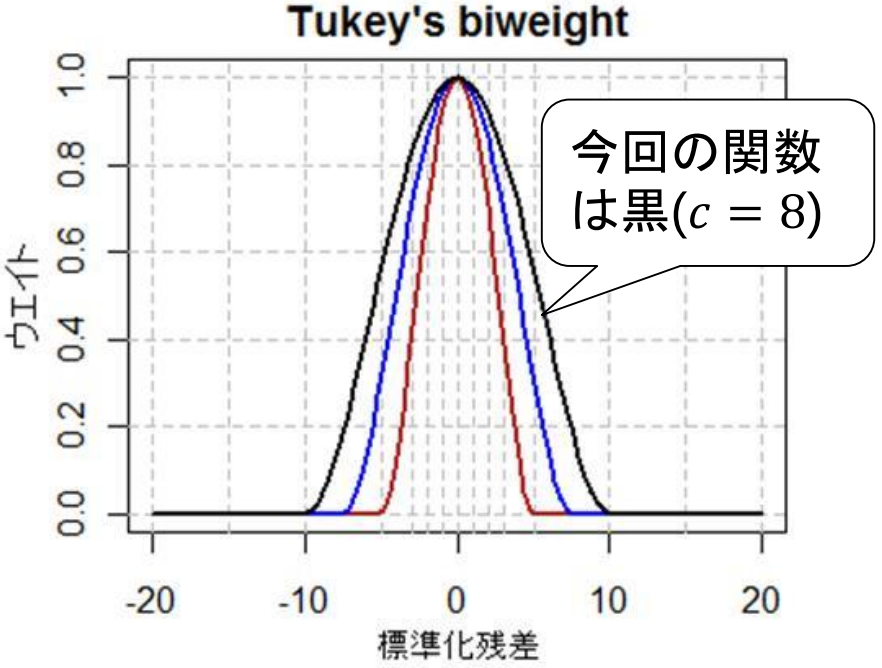
- Tukeyのbiweight関数を利用

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

（e：平均絶対偏差で規格化した残差／c：パラメータ）

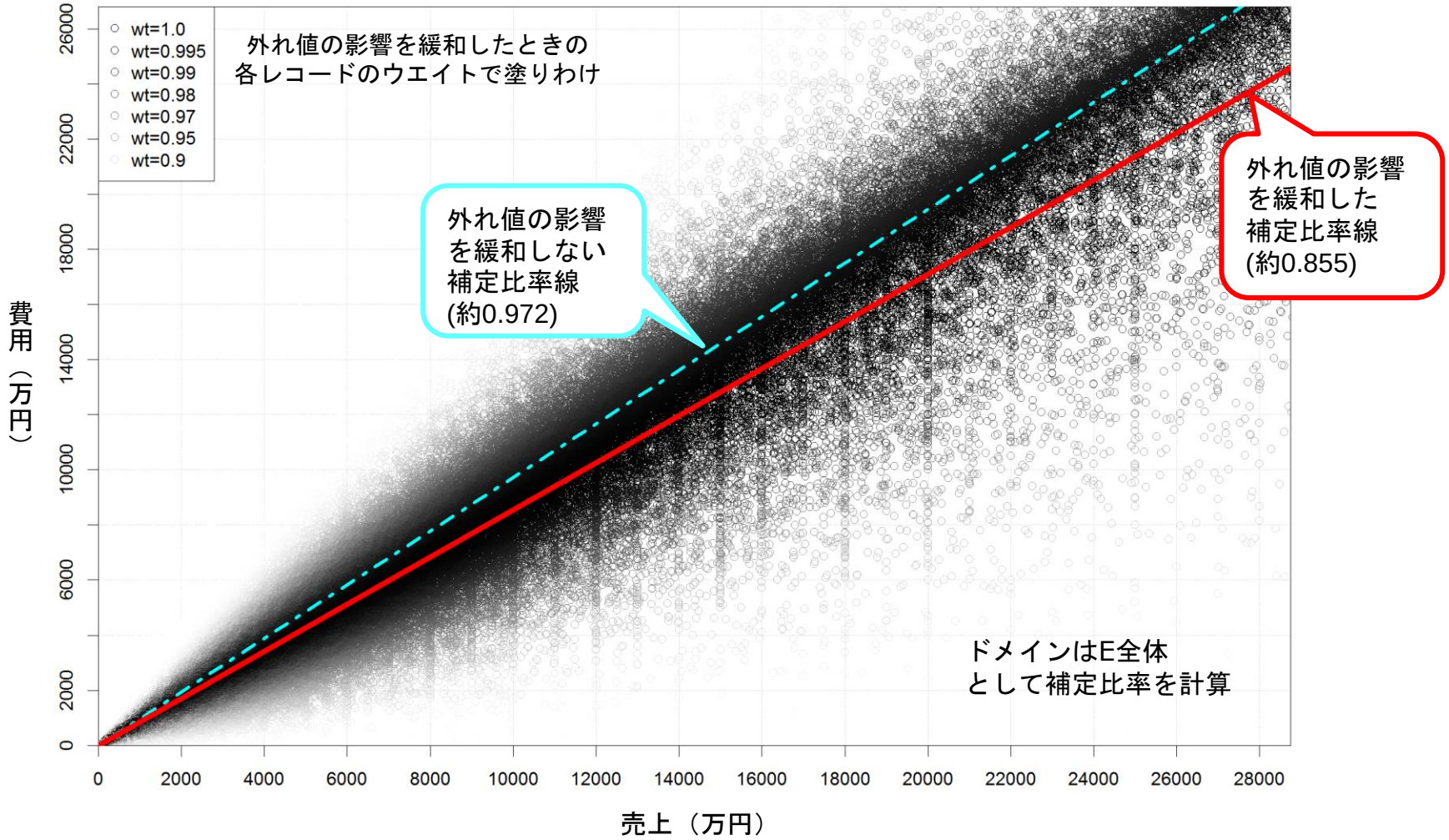
【参考】Tukeyのbiweight関数

- Huberのウエイト関数と同様によく使われる
- 極端な外れ値は影響排除（ウエイト0）
- 比率補定に適用する場合、推定パラメータが1つであるため無限ループの可能性はない



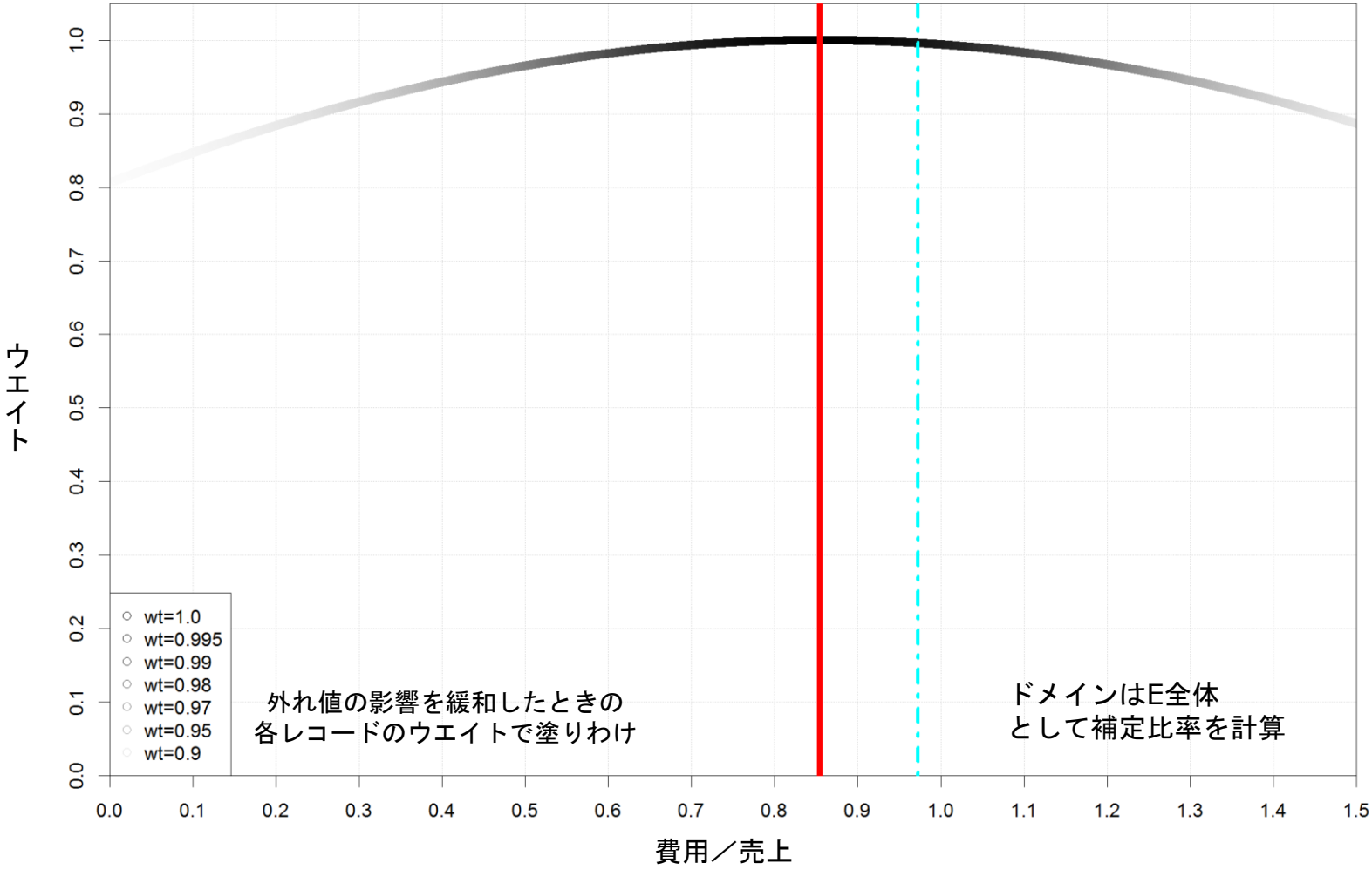
代表値のロバスト性（イメージ）

例）産業分類：E（製造業）　売上～費用の散布図



【参考】代表値のロバスト性（イメージ）

例）産業分類E（製造業）　比率～ウエイトの関係



3.シミュレーションの概要

シミュレーションの概要

- 使用データ

  平成24年経済センサス - 活動調査の企業レコード

- 補定対象：企業の経理項目
  - 売上総額
  - 費用総額

「費用／売上」比率を変数として活用

補定に活用する変数

- 費用総額／売上総額
  - 正の値をとる
    この値が1以上だと大雑把にいって赤字
    0に近いほど経営成績は良い
- 比率設定の際、分母項目は分子項目よりも値が一般的に大きいものを選択
  - 求まる補定比率が安定する

ドメイン設定に使う項目

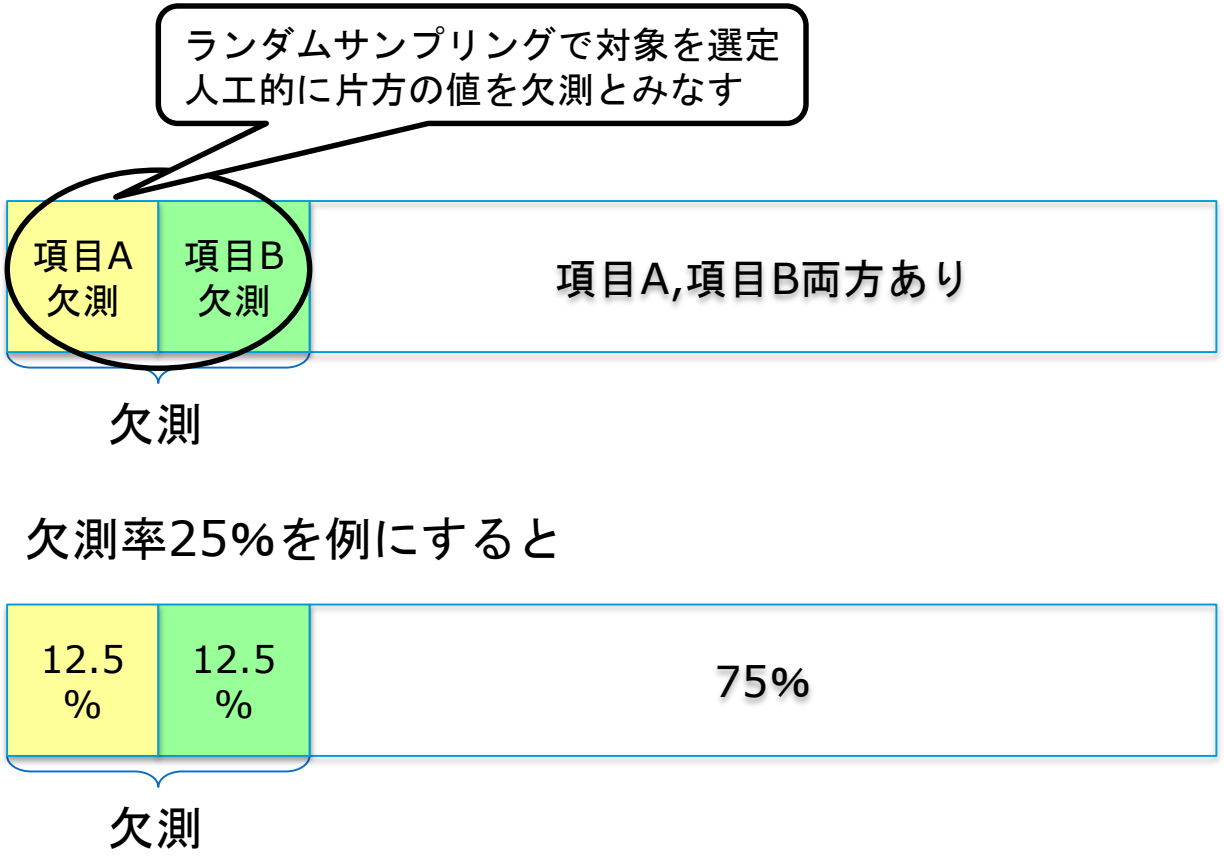
- 産業分類
  - 産業大分類(26)
  - 産業中分類(111)
  - 産業小分類(447)
  - 産業細分類(541)

※()内の数字は分類数

シミュレーションの概要

- データの一部に人工的に欠測を生成
  - 欠測レコードは毎回産業中分類毎にランダムサンプリング
  - 欠測率（比率が計算できない割合）は 5,10,15,20,25%
  - 欠測レコードの半分を一方の項目の、残りの半分を他方の項目の欠測レコードと扱う
- 欠測率毎に100回ずつ実施

欠測レコード生成のイメージ



4.シミュレーション結果の評価方法

シミュレーション結果の評価方法

- 産業大分類を評価層とする
- 評価層内全レコードの補定項目の
  真値合計 と 補定値を含む合計
  の差の絶対値を求める
  - 「補定値を含む合計」では生成欠測レコードの値として「比率補定値」を採用

シミュレーション結果の評価方法

- 各回の補定方法別の結果を比較し、差が最小の方法を「良い」と評価
- 異なるドメインによる補定結果を比較
- 100回のうち「良い」評価回数が一番多い条件を最終的に「良い」補定方法と評価

シミュレーション結果のイメージ

- シミュレーション結果のイメージ（数値は架空）

回	ドメイン1	ドメイン2	ドメイン3	ドメイン4	真値
1	70083	69964	70030	70030	70017
2	71200	71000	70362	70362	70017
3	70090	70078	70039	70039	70017
4	70077	70065	70027	70027	70017
5	70174	70148	69967	69967	70017

- 真値合計と補定値を含む合計の差（の絶対値）で評価

回	ドメイン1	ドメイン2	ドメイン3	ドメイン4	真値
1	66	53	13	13	****
2	1183	983	345	345	****
3	73	61	22	22	****
4	60	48	10	10	****
5	157	131	50	50	****

悪い

良い

5.シミュレーション結果

シミュレーション結果

- 売上ー欠測率　5%／c＝8
  評価層：産業大分類E（製造業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	0	4	23	73	100
【平均乖離率】	0.31%	0.20%	0.170%	0.168%	***
【RMSE】	1.029	0.698	0.579	0.575	$(\times 10^8)$

シミュレーション結果

- 売上ー欠測率15%／c＝8
  評価層：産業大分類E（製造業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	0	0	24	76	100
【平均乖離率】	0.91%	0.61%	0.510%	0.508%	***
【RMSE】	2.927	1.980	1.653	1.648	$(\times 10^8)$

シミュレーション結果

- 売上ー欠測率25%／c＝8
  評価層：産業大分類E（製造業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	0	0	29	71	100
【平均乖離率】	1.46%	0.99%	0.826%	0.822%	***
【RMSE】	4.657	3.177	2.640	2.625	$(\times 10^8)$

6.まとめ

まとめ

- 評価層を産業大分類とした補定シミュレーションで以下の結論を得た
  - 評価層毎に「良い」ドメインは異なる
    ⇒細かくドメインを設定しなくても良い評価層の場合、この補定処理の所要時間が短縮される期待

7.今後の本研究の展開

今後の本研究の展開

- 他に有効なドメインはあるか
  - 規模項目（常用雇用者数、支所数…）
  - 経営組織（個人経営、法人、非法人団体…）
  - 地域
- 評価層を変えたらどうなるか
  - 集計結果はもっと詳細に公表される
    （例：産業分類であれば細分類ごと）

今後の本研究の展開

- 補定比率を単純比率平均とした補定結果との比較・評価方法

参考文献

- 途中でご紹介したもののほかに…

Ito, T., Abe, Y., and Noro, T. (2013) The Best Stratification to Impute Missing Values of Turnover in Economic Surveys.
59th Session of the ISI World Statistics Congress Proceedings, 25-30, Aug. 2013, Hong Kong, China.

シミュレーション結果

- 売上ー欠測率　5%／c＝8
  評価層：産業大分類G2
  （情報サービス業，インターネット付随サービス業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	52	37	6	5	100
【平均乖離率】	0.173%	0.176%	0.1827%	0.1831%	***
【RMSE】	4.143	4.195	4.358	4.365	$(\times 10^6)$

シミュレーション結果

- 売上ー欠測率15%／c＝8
  評価層：産業大分類G2
  （情報サービス業，インターネット付随サービス業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	78	15	5	2	100
【平均乖離率】	0.44%	0.45%	0.46%	0.46%	***
【RMSE】	8.783	8.967	9.219	9.241	$(\times 10^6)$

シミュレーション結果

- 売上ー欠測率25%／c＝8
  評価層：産業大分類G2
  （情報サービス業，インターネット付随サービス業）

項目	ドメイン				合計
	大分類	中分類	小分類	細分類	
【誤差最小回数】	74	23	3	0	100
【平均乖離率】	0.69%	0.71%	0.72%	0.73%	***
【RMSE】	12.930	13.260	13.514	13.555	$(\times 10^6)$

補定比率のばらつき

- 売上／c＝8
  ドメイン：産業大分類E（製造業）

100回分の比率の標準偏差	欠測率		
	5%	15%	25%
【ロバスト比率】	0.0014	0.0020	0.0024
【単純平均比率】	0.0075	0.0115	0.0157