資料３

令和元年個人企業経済調査
～欠測値の補完について～

（独）統計センター
技術研究開発課

目　次

I. 補完についての研究の前提

最適な補完方法を検討するという統計局の方針に基づき本研究を行った

1) 補完の一般的な手順
2) 補完対象項目
3) 補完に使用するデータ
4) 補完処理の構成
5) 補完クラスについて

I.1 補完の一般的な手順

1) 同一客体のデータ内で、欠測値が一意に決まるような場合、その一意に決まる値を補記する

   例：内訳を全て合計すれば総計と一致するような調査項目において、内訳が全て記入され、総計が欠測していれば、内訳の合計を総計欄に補記する

2) 欠測のある客体について、過去データが存在し、それが経年で安定的なものであれば利用する

   例：欠測のある特定の調査対象企業について、前回調査あるいは欠測項目を調査している他調査データがあれば、その情報を利用して補記する

3) 上記の1)及び2)に該当しない場合に、他の客体のデータを使用して何らかの方法で欠測値を推定し使用する

I.2　補完対象項目

項目番号	経理項目
05	**売上金額**
06	費用総額※
07	**期首棚卸高**
08	**仕入高**
09	**期末棚卸高**
10	**経費計**
11	**給料賃金**
12	**地代家賃**
13	**減価償却費**
14	**租税公課**
15	**損害保険料**
16	**福利厚生費**
17	**外注工賃**
18	**利子割引料**

項目番号	従業者項目
041	**事業主の家族で無給**
042	**常用雇用者**
043	**臨時雇用者**

※印の「06 費用総額」は調査項目ではないが、補完作業のために中間データとして他の項目から作成する

- 紫色の項目番号は、H28経済センサス-活動調査に同じ項目が存在する
- 太字の経理項目名は、H30個人以前の旧調査に同じ項目が存在することを示すが、新調査で追加される拡大産業についてデータがない

I.3 補完に使用するデータ

補完には、新調査データの他に、時点の異なる二種類のデータを使用する

① 令和元年個人企業経済調査［**R1新調査**］

② H28経済センサス-活動調査［**H28センサス**］

③ H29事業所母集団DBから作成する抽出用母集団名簿 (H28センサス以降の新設企業情報を含む)［**H29DB**］

I.4 補完処理の構成

1) 過去の同一企業データからの補完

[H28センサス及びH29DB]

a. 経理項目（売上高、費用総額、地代家賃、給与賃金、減価償却費、租税公課）　※ 時点調整あり

b. 従業者数合計　※ 時点調整なし

2) 比率ホットデックによる補完(ドナー選択は最近隣法)

[R1新調査]

a. 経理項目（売上高、費用総額、地代家賃、給与賃金、減価償却費、租税公課）

※ ドナー候補のクリーニング・時点調整あり

b. 従業者数合計

※ ドナー候補のクリーニングなし

Ⅰ.5 補完クラスについて

- 無回答により起こる結果数値のバイアスを減らし、推定精度を上げるために設定する
- クラスは全ての調査単位について欠測のない変数により、クラス内の補完対象ができるだけ均質で、クラス間の差異は大きくなるように設定する
- クラスを構築する変数は、回答・無回答の傾向に関係するものを使用する

例) 大企業と中小企業で、回答する企業の割合が異なる場合

補完クラスを規模に関する変数、例えば従業者数等を用いて設定し、クラス内の回答割合をできるだけ均一にする

R1調査にあたり、現在対象産業を網羅する分析データが存在しないため、標本抽出時の層（都道府県×6区分の産業分類×売上高90%点）をベースとし、クラス内のデータ量を確保するために都道府県別を外した（6区分の産業分類×売上高90%点）。本報告の試算はこれに基づいている。

Kalton, G. and Kasprzy, D. (1986) The treatment of missing survey data, *Survey Methodology*, Vol. 12, pp. 1-16.

Ⅱ. 経理項目の補完

ここでは、前回の研究会での方向性を受けて行った、経理項目の補完に関する試算結果について報告する

1) 比率ホットデックの仕組み
2) シミュレーションの概要
3) 主要経理項目について
4) 棚卸高について
5) 経費計内訳について

Ⅱ.1 比率ホットデックの仕組み

再掲

例: **負の値をとらない変数 $y_1, \dots, y_m$ について、**
$y_1 + \cdots + y_m = y_{tot}$ という制約がある場合を考える

i 番目のレコード$\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,m})$について、頭から t 個の変数が観測され、その後ろの m-t 個の変数が欠測しているとき、次のような手順で補完する

① 欠測値の合計 $r_i = y_{i,tot} - y_{i,1} - \cdots - y_{i,t}$ を計算
② 任意の方法で、欠測のないレコードの中からドナーを選択し、それが d 番目のレコードとする
③ 補完すべき変数のドナー値の合計 $r_d = y_{d,\mathrm{t+1}} + \cdots + y_{d,m}$ を計算
④ 下式により補完値 $\tilde{y}_{i.j}$ を計算し、補完

$$\tilde{y}_{i.j} = \frac{r_i}{r_d} y_{d,j}, \qquad (j = t+1, \dots, m)$$

ドナーは、制約を満たすチェック済データなので、同時に補完する値は自動的に制約を満たすことになる

比率ホットデックによる補完例

制約条件: 内訳1 + 内訳2 + 内訳3 = 合計

r_d =1500-1000=500

r_i =1200-800=400

$$\tilde{y}_{i,\text{内訳}2} = \frac{r_i}{r_d} y_{d,\text{内訳}2} = \frac{400}{500} \times 300 = 240$$

$$\tilde{y}_{i,\text{内訳}3} = \frac{r_i}{r_d} y_{d,\text{内訳}3} = \frac{400}{500} \times 200 = 160$$

800 + 240 + 160 = 1200

実際の補完例

[06費用総額][07期首棚卸高][08仕入高]が欠測のパターン

補完した費用総額を使って
期首棚卸高、仕入高を補完する

ドナー選択: 費用総額(補完値)
期末棚卸高
経費計

比率の計算:
費用総額＋期末棚卸高－経費計

最近隣法: マハラノビス距離*で最も近いドナーを選択

	売上金額	費用総額	期首棚卸高	仕入高	期末棚卸高	経費計
欠測レコード	49,246	35,680			1,200	16,600
ドナーレコード	39,188	34,332	1,032	17,819	946	16,427
補完済レコード	49,246	**35,680**	**1,110**	**19,170**	1,200	16,600
真値	49,246	36,161	1,559	19,202	1,200	16,600

比率
1.0758

Ⅱ.2 シミュレーションの概要

■シミュレーション方法*

個人企業経済調査構造編データを産業大分類、売上高90%点別で補完クラスを設定し、ランダムに一定割合（20%）を欠測とみなし、欠測パターン別に補完する

■データサイズ

産業大分類	集計対象企業のデータサイズ		20%欠測とした場合	
	売上高階級		売上高階級	
	90%以上	90%未満	90%以上	90%未満
E 製造業	1079	9944	215	1988
I 卸売業、小売業	1922	17631	384	3526
M 宿泊業、飲食サービス業	1092	9815	218	1963
Q サービス業	1402	12400	280	2480

結果の評価方法

➢標準平均平方誤差 NRMSE*

(Normalized Root Mean Square Error)

$$\mathrm{NRMSE} = \sqrt{\frac{\sum\left(\frac{x^{true} - x^{imp}}{\sigma}\right)^2}{n}}$$

x^{true}：真値
x^{imp}：補完値
σ：x^{true}の標準偏差

欠測値補完の精度の評価指標の一つで、値が1に近づけば、補完値と真値の差異の分散が真値の分散に近く、0に近づくほど補完値の誤差が小さい

主要経理項目(07～10)の欠測パターン

組み合わせは以下の26通り

欠測パターン	05 売上金額	06 費用総額	07 期首棚卸高	08 仕入高	09 期末棚卸高	10 経費計
a	○	×	×	○	○	○
b	○	×	○	×	○	○
c	○	×	○	○	×	○
d	○	×	○	○	○	×
e	○	○	×	×	○	○
f	○	○	×	○	×	○
g	○	○	○	×	×	○
h	○	○	×	○	○	×
i	○	○	○	×	○	×
j	○	○	○	○	×	×
k	○	×	×	×	○	○
l	○	×	×	○	×	○
m	○	×	○	×	×	○

欠測パターン	05 売上金額	06 費用総額	07 期首棚卸高	08 仕入高	09 期末棚卸高	10 経費計
n	○	×	×	○	○	×
o	○	×	○	×	○	×
p	○	×	○	○	×	×
q	○	○	×	×	×	○
r	○	○	×	×	○	×
s	○	○	×	○	×	×
t	○	○	○	×	×	×
u	○	×	×	×	×	○
v	○	×	×	×	○	×
w	○	×	×	○	×	×
x	○	×	○	×	×	×
y	○	○	×	×	×	×
z	○	×	×	×	×	×

Ⅱ.3 主要経理項目の結果

15枚目スライドに示す26パターン全てについて、比率ホットデックの手順と、ドナー選択方法について試算を行い、複数の選択肢があるものについては最もNRMSEの小さい方法を選択した。

選択された全ての方法を別紙1に示し、対応するNRMSEの数値を別紙2に添付する。

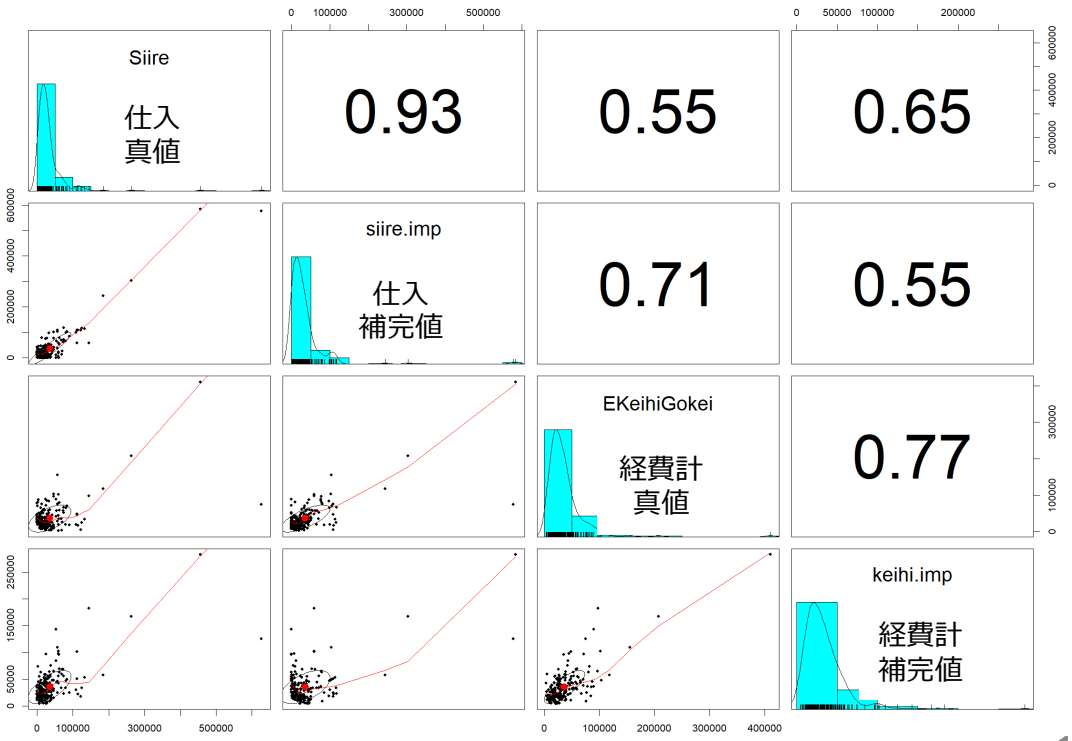
さらに、いくつかのパターンについて、以下のスライドに例示する。

仕入高と経費計が欠測の場合

費用総額、期首棚卸高、期末棚卸高からマハラノビス距離*でドナーを選択し、比率（07-09）を乗じて仕入高と経費計を補完

補完結果《NRMSE*》

産業	売上高階級	仕入高	経費計
E	90%以上	0.40	0.65
I		0.41	1.25
M		0.70	0.58
Q		0.84	0.87
E	90%未満	0.85	0.60
I		0.38	0.89
M		0.64	0.54
Q		1.04	0.44



経理項目すべてが欠測の場合（単位欠測）

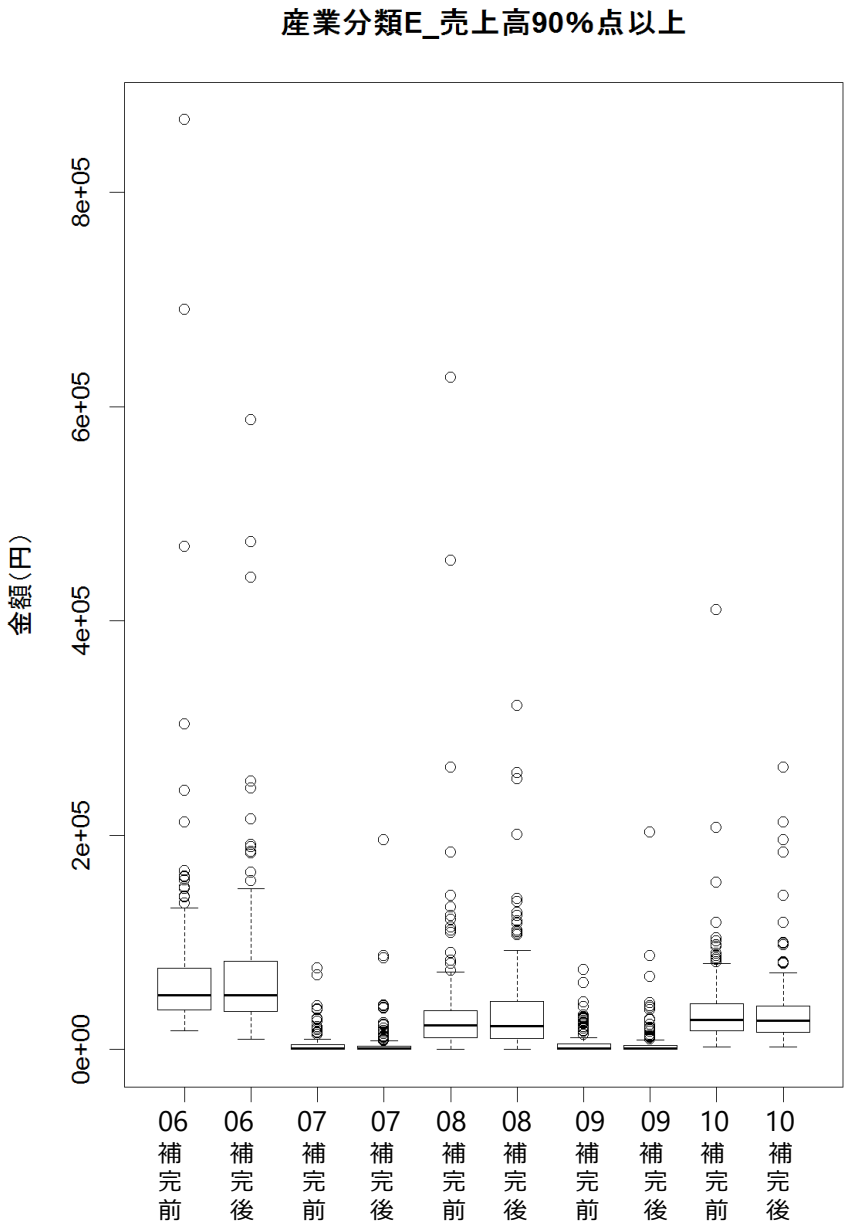
　売上金額のみが補完されているため、売上金額が最も近いレコードをドナーとして選択し、ドナーレコードの売上金額との比率を欠測項目にかけて補完する

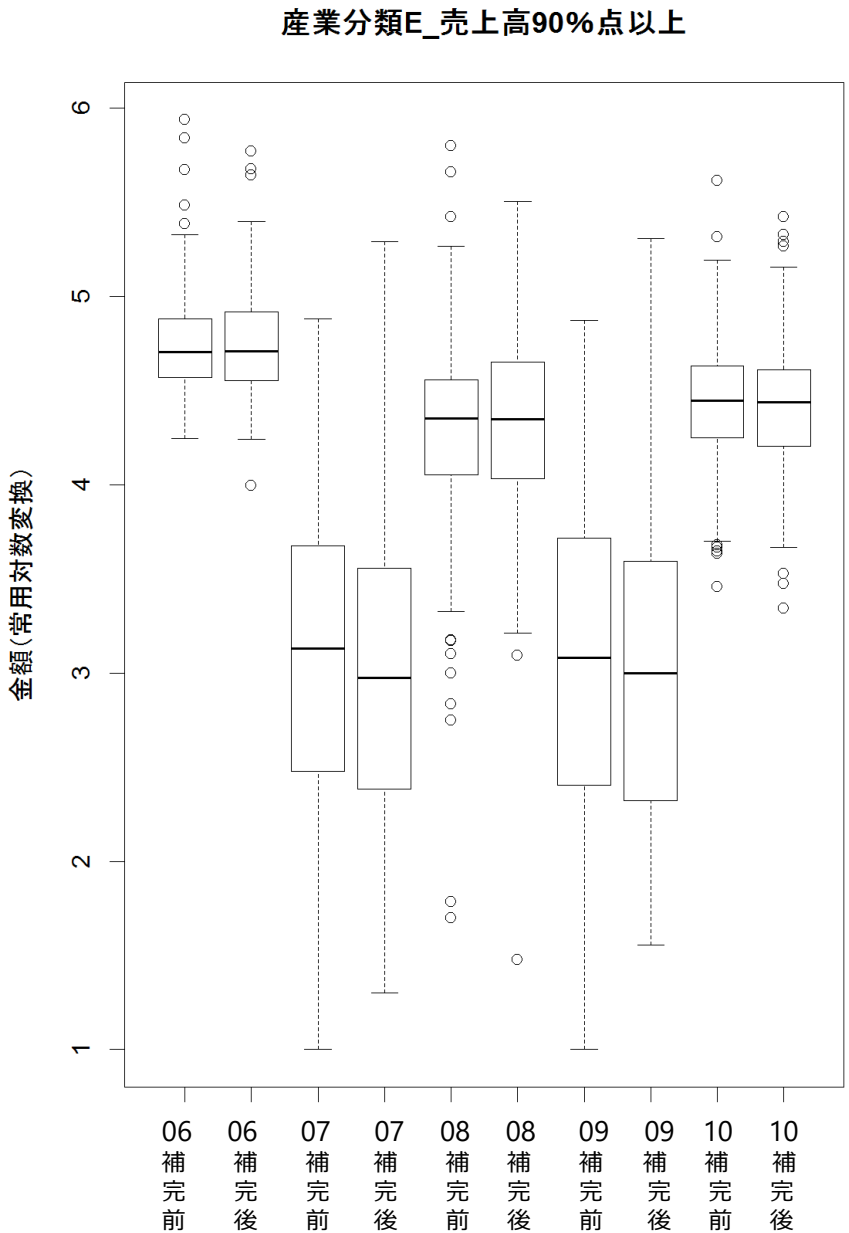
⇒売上金額との相関が高い項目の補完値は誤差が小さくなるが、売上金額との相関が低い期首・期末棚卸高は補完値との誤差が大きくなる

補完結果《NRMSE*》

産業	売上高	費用総額	期首棚卸高	仕入高	期末棚卸高	経費計
E	90%以上	0.37	1.85	0.60	1.85	0.76
I		0.38	1.96	0.45	1.78	1.25
M		0.44	1.21	0.75	1.20	0.80
Q		0.45	1.54	0.96	1.56	1.05
E	90%未満	0.58	1.26	1.03	1.26	0.78
I		0.30	1.35	0.48	1.36	1.00
M		0.53	1.29	0.67	1.31	0.84
Q		0.64	1.24	1.16	1.24	0.74

補完前後のデータ分布の比較





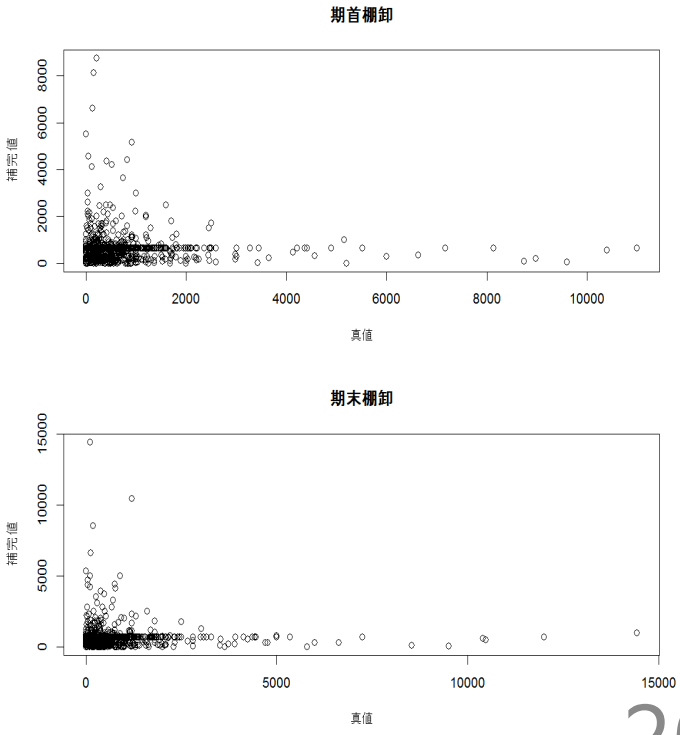
Ⅱ.4 棚卸高について

[07期首棚卸高]と[09期末棚卸高]が欠測の場合

期末棚卸高は期首棚卸高との相関が高いが、他の項目との相関関係が低いため、両者が欠測すると誤差が大きくなる傾向がみられ、おおむね平均値補完の方が良い推定となる

補完結果《NRMSE*》

産業	売上高	比率ホットデック		平均値	
		期首棚卸高	期末棚卸高	期首棚卸高	期末棚卸高
E	90%以上	1.48	1.61	1.00	1.00
I		1.19	1.27	1.00	1.00
M		1.06	1.12	1.00	1.00
Q		1.31	1.39	1.00	1.00
E	90%未満	1.09	1.12	1.00	1.00
I		0.92	0.98	1.00	1.00
M		0.95	1.04	1.01	1.01
Q		1.22	1.31	1.00	1.00



棚卸高を平均値補完した場合の試算例

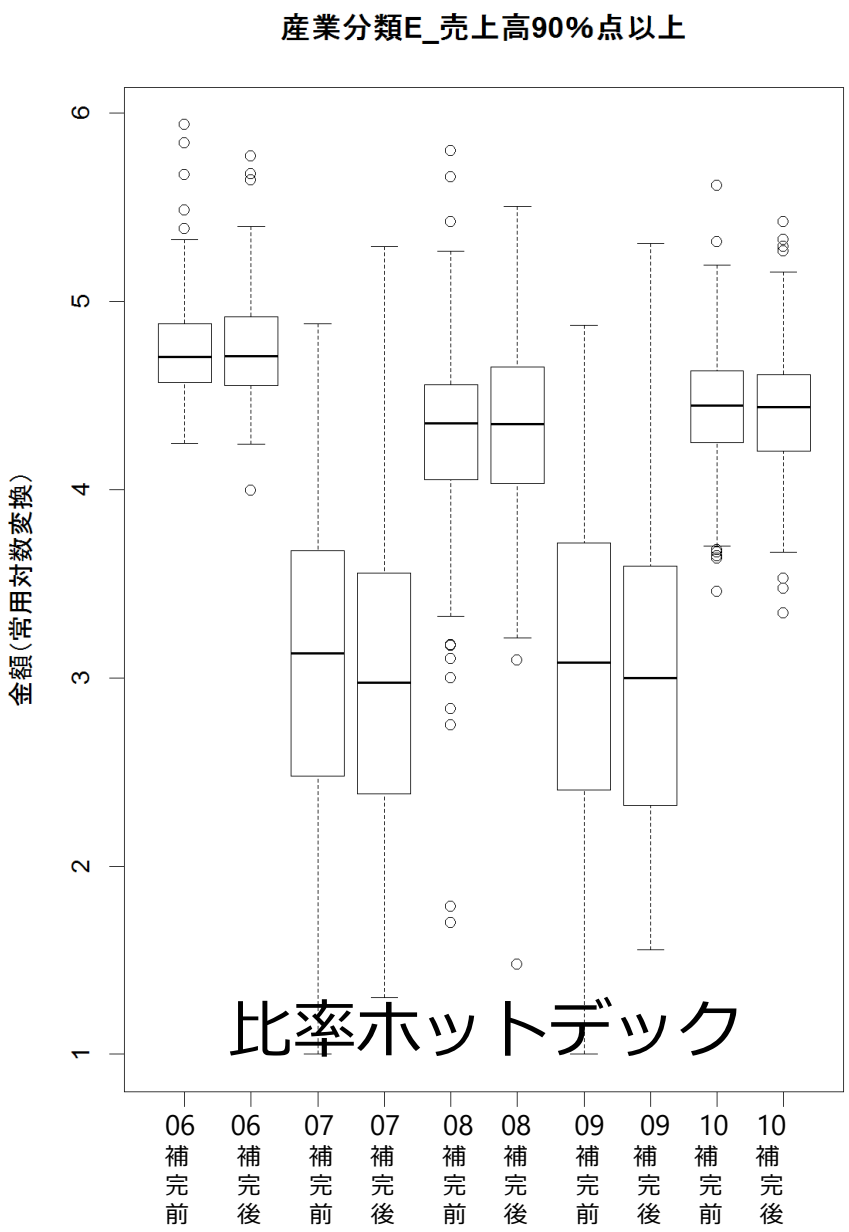
① [07期首棚卸高]と[09期末棚卸高]を平均値補完

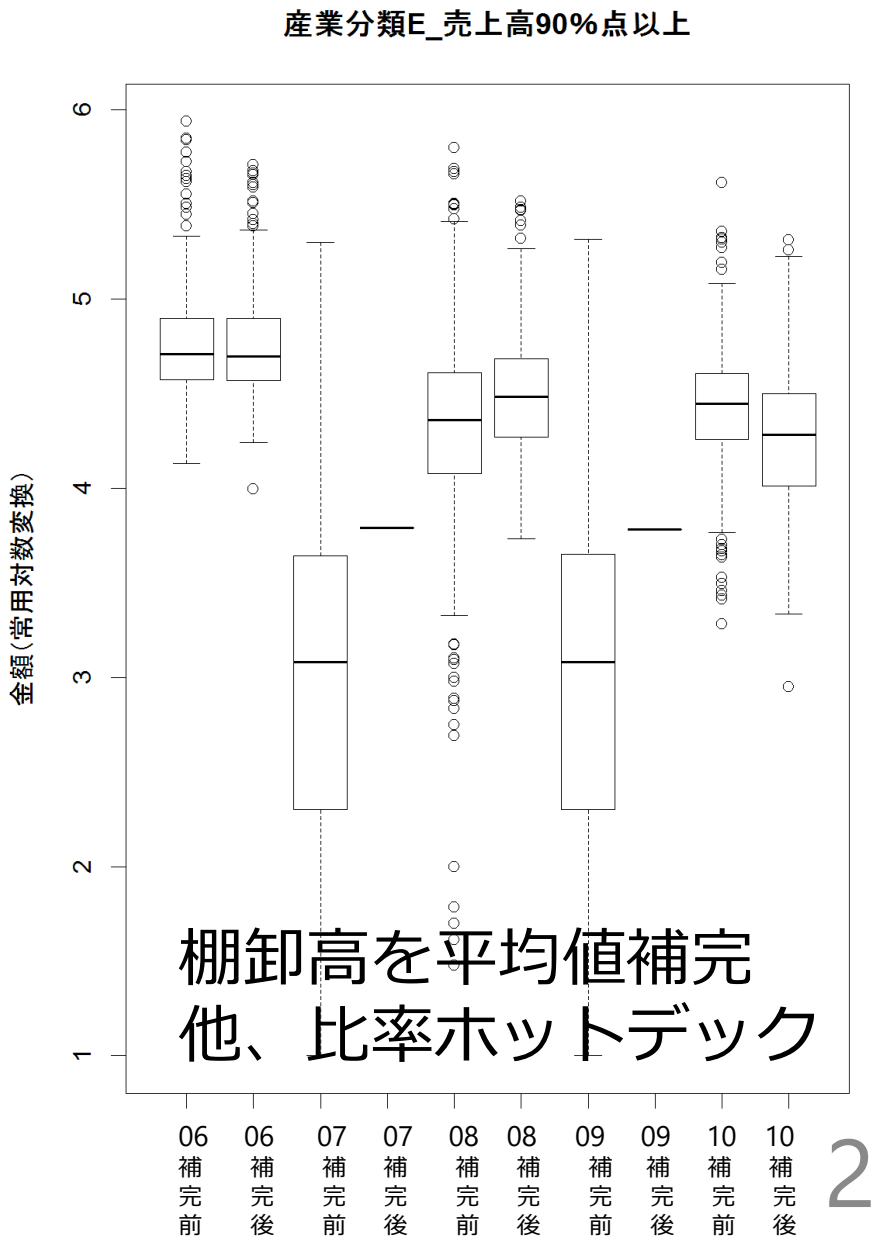
② 他の項目を比率ホットデックで補完(パターンo)

補完結果《NRMSE*》

産業	売上高	比率ホットデック					棚卸高、平均値補完 他、比率ホットデック				
		費用総額	期首棚卸高	仕入高	期末棚卸高	経費計	費用総額	期首棚卸高	仕入金額	期末棚卸高	経費計
E	90%以上	0.37	1.85	0.6	1.85	0.76	0.52	1	0.64	1	0.84
I		0.38	1.96	0.45	1.78	1.25	0.59	1	1.05	1	0.81
M		0.44	1.21	0.75	1.2	0.8	0.26	1	0.4	1	1.16
Q		0.45	1.54	0.96	1.56	1.05	0.28	1	0.44	1	0.89
E	90%未満	0.58	1.26	1.03	1.26	0.78	0.46	1	0.79	1	0.69
I		0.3	1.35	0.48	1.36	1	0.53	1	0.82	1	0.75
M		0.53	1.29	0.67	1.31	0.84	0.5	1.01	1.22	1.01	0.9
Q		0.64	1.24	1.16	1.24	0.74	0.67	1	1.19	1	0.76

棚卸高の補完方法の違いによるデータ分布の変化





Ⅱ.5 経費計内訳(11~18)について

項目番号	経理項目
05	売上金額
06	費用総額
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

A) 10経費計が欠測の場合

経費計補完時のドナーレコードの経費計内訳に比率をかけて補完する。

B) 10経費計が観測の場合

経費計が最も近いレコードをドナーとして選択*し、経費計で比率を決めて11給与賃金~18利子割引料を補完する。

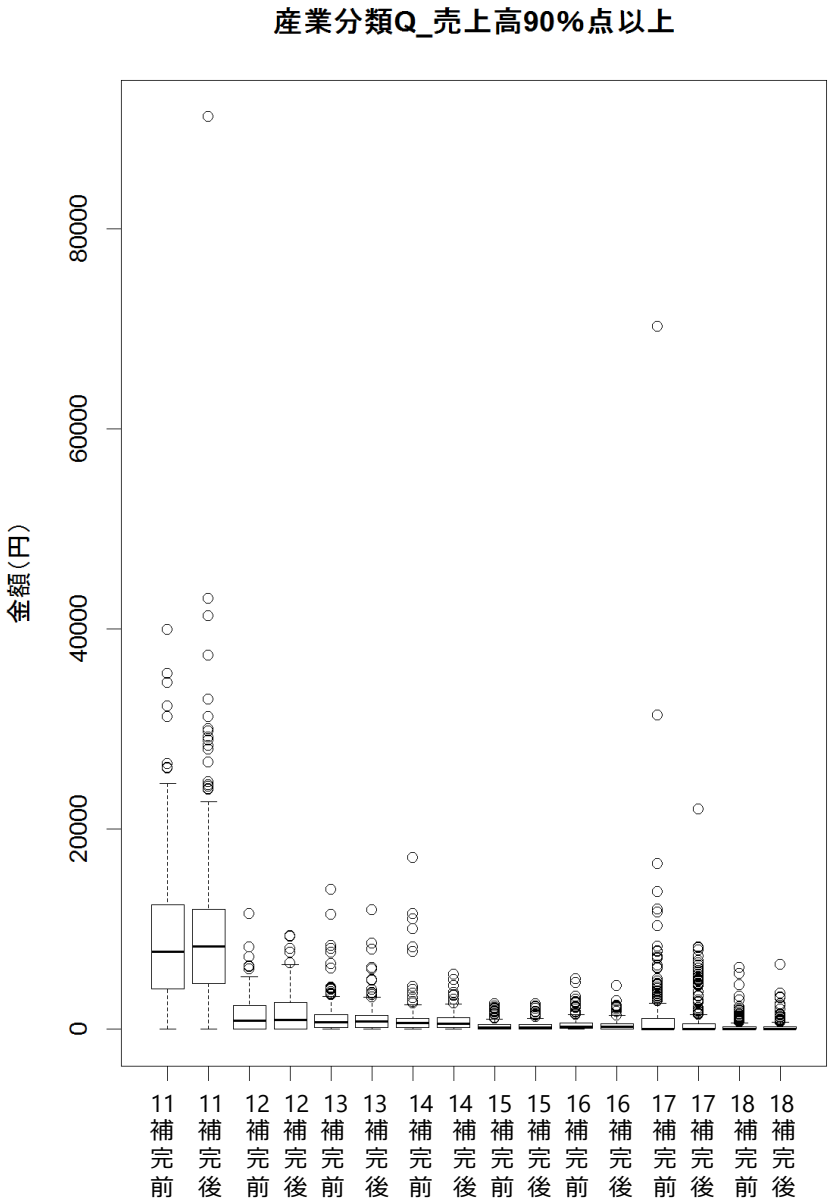
シミュレーション結果

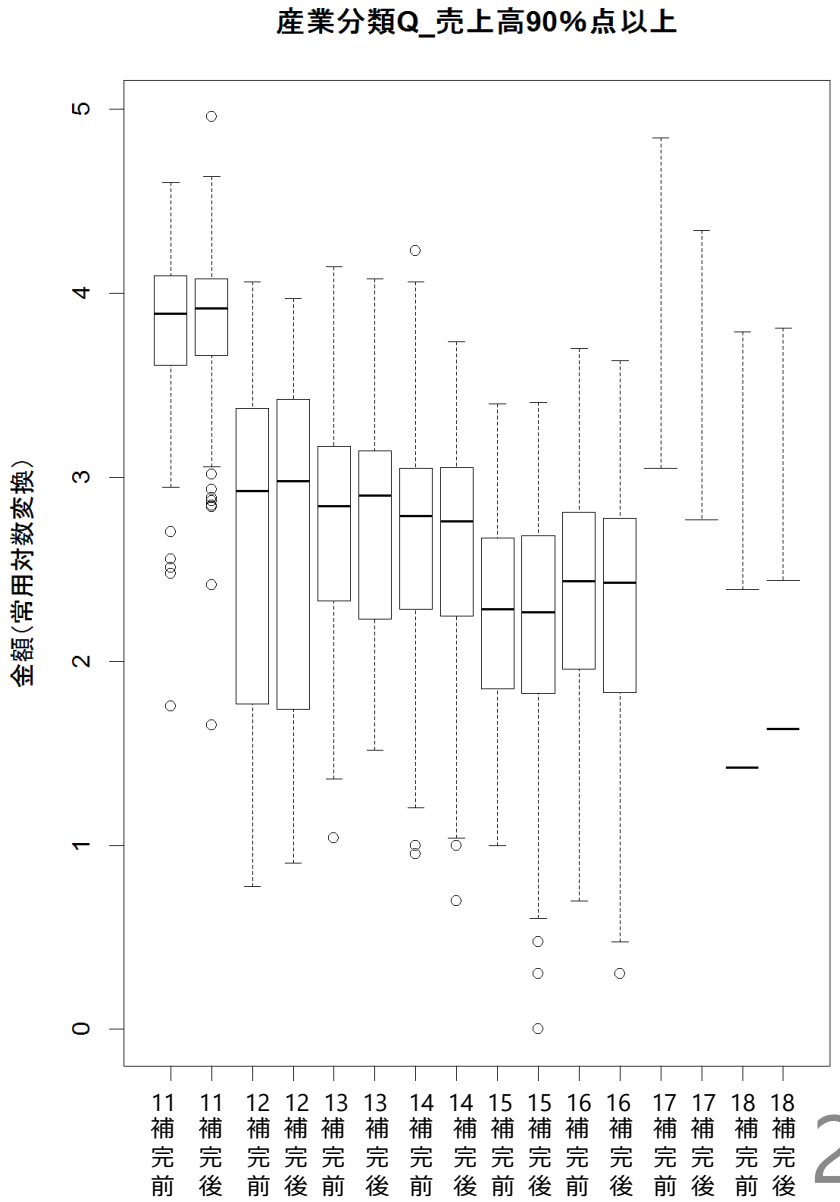
補完結果《NRMSE*》Normalized Root Mean Square Error

産業	売上高	給与賃金	地代家賃	減価償却費	租税公課	損害保険料	福利厚生費	外注工賃	利子割引料
E	90%以上	0.74	1.35	1.38	1.16	1.48	1.10	0.96	1.55
I		1.28	1.65	1.62	1.25	2.06	1.17	1.03	2.33
M		0.94	2.04	1.33	0.98	1.22	1.35	0.99	1.24
Q		0.84	1.44	1.26	1.42	1.23	1.25	1.08	1.26
E	90%未満	0.90	1.28	1.14	1.30	1.40	1.27	1.17	1.67
I		0.79	1.37	1.30	1.41	1.21	1.28	1.14	1.21
M		0.89	1.31	1.30	1.42	1.19	1.31	1.79	1.39
Q		0.86	1.21	1.26	1.27	1.37	1.27	1.48	1.33

給与賃金は経費計との相関が高いため、補完値との誤差は小さいが、その他の項目は相関が低いため、誤差も大きくなる。

補完前後のデータ分布の比較





経費計内訳の0値の状況

✓経費計内訳の各項目の0値の割合では、
多くの項目の値が0となっている。

産業	売上高	経費計	給料賃金	地代家賃	減価償却費	租税公課	損害保険料	福利厚生費	外注工賃	利子割引料
E	90%以上	0.0	6.1	38.4	12.3	5.1	7.4	15.8	37.2	25.6
I		0.0	1.7	52.0	34.7	22.2	19.4	25.3	89.2	50.7
M		0.0	0.9	24.9	12.5	6.6	11.6	15.0	89.6	35.8
Q		0.0	6.7	25.8	15.8	7.5	10.8	15.0	60.1	39.2
E	90%未満	0.2	57.0	54.5	36.3	13.8	22.2	64.0	56.0	65.1
I		0.2	56.1	47.1	33.8	14.4	22.3	66.8	82.7	67.1
M		0.0	41.6	36.3	37.9	25.7	29.2	65.1	96.1	71.2
Q		0.3	65.5	47.0	40.0	24.5	35.2	71.5	84.3	77.5

経費計及び経費計内訳の金額が0円となる個人企業の割合(%)
平成14～29年個人企業経済調査（構造編）のうち、除外なしのデータを対象

Ⅲ. 従業者数関連項目の補完

ここでは、前回の研究会での方向性を受けて行った、従業者数関連項目の補完に関する試算結果の概要と、「離散値を連続値のための方法で推定しているのでは」というご指摘に対応するために行った代案についての試算結果について報告する。

1) 比率ホットデックによる補完
2) シミュレーションの結果
3) 離散値のための補完方法の検討

Ⅲ.1 比率ホットデックによる補完

項目041～043の従業者内訳がすべて欠測の場合、

比率ホットデック補完する。

① 従業者合計と給料賃金を補助変数とし、最近隣法によりドナー選択を行う。従業者合計により比率を算出し、従業者内訳に比率をかけて補完する。

② 使用する距離関数は、従業者合計と給料賃金をロバスト標準化した値を用いて、ユークリッド距離*で測る。

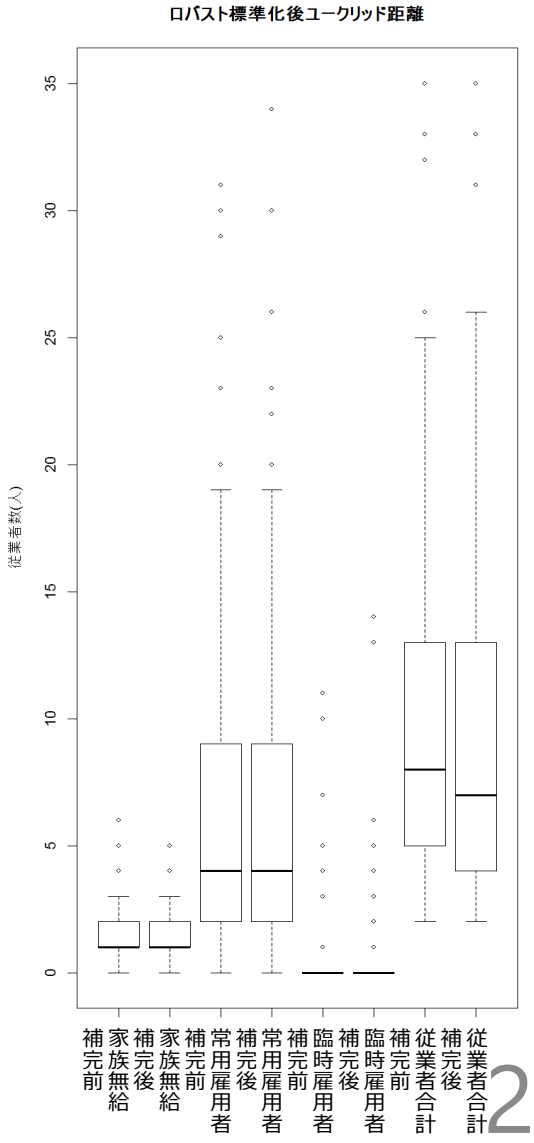
04	**従業者数**
041	**事業主の家族で無給**
042	**常用雇用者**
043	**臨時雇用者**

Ⅲ.2 シミュレーション結果

- 結果

NRMSE*

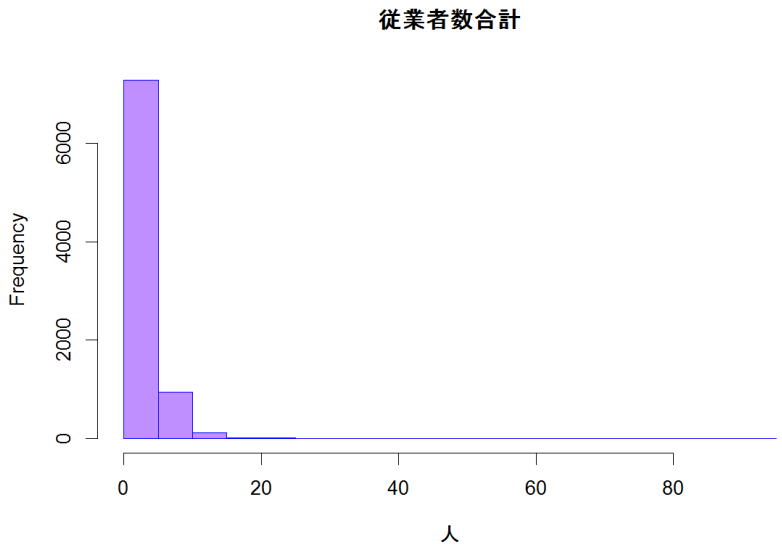
産業	売上高階級	家族無給	常用雇用者	臨時雇用者
E	90%以上	1.28	0.36	0.95
I		1.26	0.36	1.78
M		1.51	0.51	0.96
Q		1.18	0.61	3.48
E	90%未満	1.38	0.94	1.80
I		1.02	0.42	1.38
M		1.07	0.61	0.96
Q		1.71	1.10	1.05

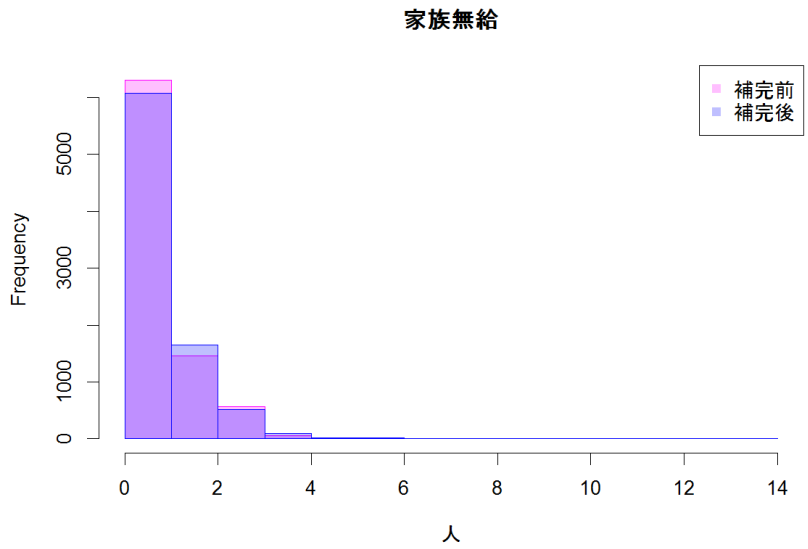


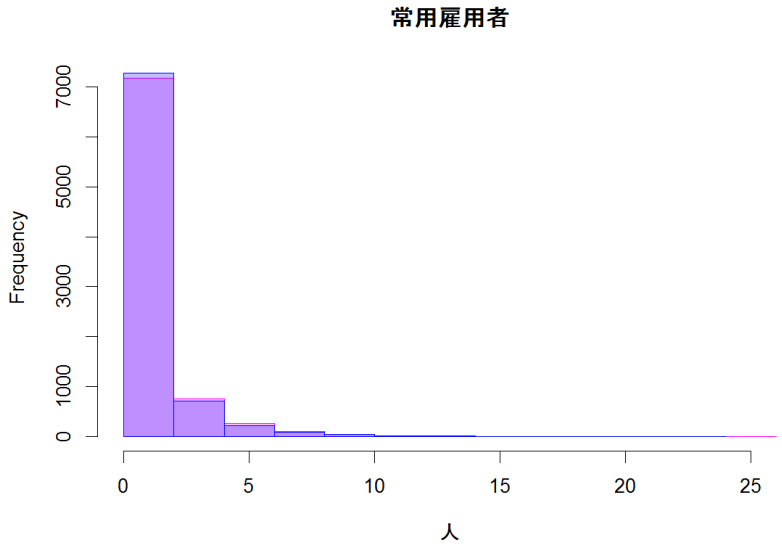
- 従業者内訳が合計と一致しない

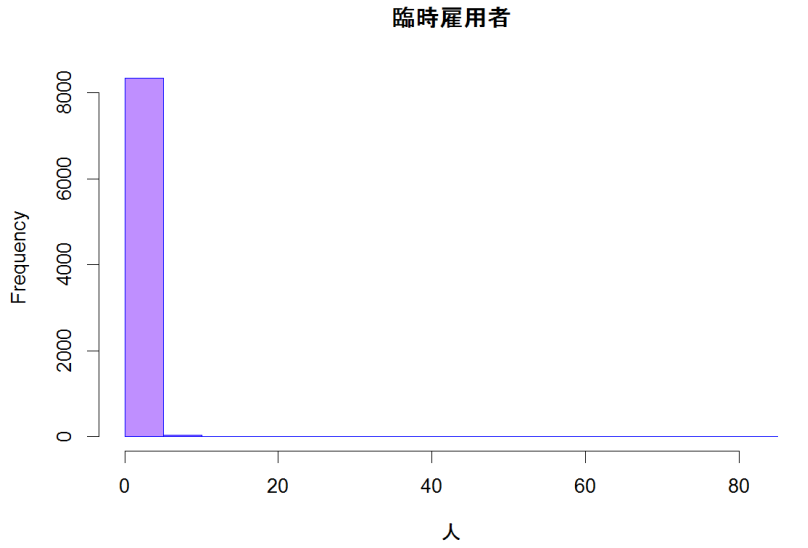
  補完した従業者内訳と従業者合計が一致しない補完値がある場合は、内訳の合計を合わせるように比率をかけた値で調整する

補完前後のデータ分布


従業者数合計
Frequency
0
2000
4000
6000
0
20
40
60
80
人


家族無給
補完前
補完後
Frequency
0
1000
3000
5000
0
2
4
6
8
10
12
14
人


常用雇用者
Frequency
0
1000
3000
5000
7000
0
5
10
15
20
25
人


臨時雇用者
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
人

Ⅲ.3 離散値のための補完方法の検討

- ポアソン分布を仮定した一般化線形モデルに基づく従業者数内訳の補完方法

一般化線形モデル

確率分布：ポアソン分布

リンク関数：対数関数

線形予測子：従業者数合計＋給与総額

予測に使用する変数は比率ホットデックの場合と同じ

シミュレーションの結果

補完結果《NRMSE*》Normalized Root Mean Square Error

産業	売上高階級	①比率ホットデック			②poisson分布			③poisson分布の結果から常用雇用者数を計算した場合		
		家族無給	常用雇用者	臨時雇用者	家族無給	常用雇用者	臨時雇用者	家族無給	常用雇用者	臨時雇用者
E	90%以上	1.28	0.36	0.95	1.00	0.67	5.98	0.99	0.37	0.88
I		1.26	0.36	1.78	1.06	1.58	1.01	1.06	0.27	1.01
M		1.51	0.51	0.96	0.98	0.95	0.72	0.98	0.39	0.72
Q		1.18	0.61	3.48	1.03	1.15	0.99	1.03	0.46	0.99
E	90%未満	1.38	0.94	1.80	**23.82**	**135.31**	**722.53**	1.04	0.57	0.89
I		1.02	0.42	1.38	1.02	1.06	1.02	1.02	0.45	1.02
M		1.07	0.61	0.96	1.00	0.80	0.97	1.00	0.73	0.97
Q		1.71	1.10	1.05	1.10	0.67	0.86	1.10	0.77	0.86

① 比率ホットデック補完：従業者数合計と給与総額をロバスト標準化後にユークリッド距離*で近いデータをドナーとし、比率をかけて内訳項目を補完する方法
② 一般化線形モデルのポアソン分布で推定した予測モデルを当てはめた方法
③ ②の結果から常用雇用者数を従業者数合計と算出した方法
計算式⇒従業者合計 －（１＋予測家族無給＋予測臨時雇用者）（※条件：＞０）

②については、NRMSEが非常に高くなるクラスが発生し、内訳と合計が必ずしも一致しない
③は合計が内訳と一致し、精度も高いが、ピンク色部分でマイナス値が発生する

IV. 過去の同一企業データの時点調整

前回の研究会での方向性を受け、時点調整のための比率計算について、外れ値の影響緩和のできる比推定モデルのロバストな推定量も検討した

1) **外れ値の影響を受けにくい比率の求め方について**
2) **シミュレーションの方法**
3) **試算結果**

IV.1 外れ値の影響を受けにくい比率の求め方

H28 経済センサス-活動調査の経理項目の補完に用いた、ロバストな一般化比推定量についても検討

モデル　$$\frac{y_i}{x_i^{\gamma}} = \beta x_i^{(1-\gamma)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

ロバスト化推定量　$$\hat{\beta} = \frac{\sum w_i y_i (w_i x_i)^{1-2\gamma}}{\sum (w_i x_i)^{2(1-\gamma)}}$$

残差が大きい観測値ほど小さな値になるウェイトw_iを乗じて、外れ値の影響を自動的に緩和するもの。γの値により性質の異なる推定量となり、γ=1/2で通常の比推定モデルとなる。

参考資料: H28.03 経済センサス-活動調査研究会（第4回）
資料2及び参考
https://www.stat.go.jp/info/kenkyu/e-census/katsuken/sidai04.html

IV.2 シミュレーションの方法

γ=1 の場合の推定量Aと、γ=1/2 の推定量Bについて、H30年個人企業経済調査とH28経済センサスデータでマッチングできた企業のみを対象として、二時点間の各項目の比率を算出した

条件	γ の値	モデル	推定量	疑似誤差
A	$\gamma = 1$	$y_i = \beta x_i + \varepsilon_i x_i$	$\hat{\beta}_{\mathrm{robA}} = \frac{1}{\sum w_i} \sum \mathrm{w_i} \frac{y_i}{x_i}$	$\tilde{r_i} = \frac{y_i}{x_i} - \beta \sim \mathrm{N}(0, \sigma^2)$
B	$\gamma = 1/2$	$y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$	$\hat{\beta}_{\mathrm{robB}} = \frac{\sum \mathrm{w_i} y_i}{\sum \mathrm{w_i} x_i}$	$\tilde{r_i} = \frac{y_i}{\sqrt{x_i}} - \beta \sqrt{x_i} \sim \mathrm{N}(0, \sigma^2)$
通常の比率		$y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$	$\hat{\beta} = \frac{\sum y_i}{\sum x_i}$	$\tilde{r_i} = \frac{y_i}{\sqrt{x_i}} - \beta \sqrt{x_i} \sim \mathrm{N}(0, \sigma^2)$

通常の比率のモデルはBの場合と同じ

時点調整の対象

05	**売上金額**
06	費用総額
07	**期首棚卸高**
08	**仕入高**
09	**期末棚卸高**
10	**経費計**
11	**給料賃金**
12	**地代家賃**
13	**減価償却費**
14	**租税公課**
15	**損害保険料**
16	**福利厚生費**
17	**外注工賃**
18	**利子割引料**

調査データに欠測があり、他の二種類のデータに同一企業の情報が含まれている場合は、まずそのデータを用いて補完する (従業者数データ以外は時点の違いを調整する)

時点調整を行う項目

- H28センサスに同一項目が存在する**番号が紫色**の項目
- H28センサスとR1新調査データ間で、補完項目の時点間の比率を算出
- 比率は、補完クラス・補完対象項目毎に算出する
- 比率計算上のゼロ値問題を回避するため、事前に全データに1を加える

IV.3 試算結果

■シミュレーション方法
H28センサスデータとH30年個人企業構造編の同一企業データをキー項目でマッチングし、全産業について項目別に通常の比率と推定量A及びBを算出する

■データサイズ
(マッチング後)

産業	データサイズ
E製造業	409
I卸売業、小売業	991
M宿泊業、飲食サービス業	637
Qサービス業	733

	売上金額	費用総額	給料賃金	地代家賃	減価償却費	租税公課
比率	0.93	1.09	0.82	0.95	0.80	0.88
推定量A	0.94	1.25	582.5	2140.82	3968.61	6796.75
推定量B	0.95	1.04	0.81	0.88	0.72	0.78

結果の評価

地代家賃、減価償却費及び租税公課において0値が多く、このために結果数値が特異な値になる推定量Aの使用は適切ではない　=>　**推定量B?**

- 推定量Aは、個々の企業の二時点データの比率をとり平均するため、二つの時点のどちらか片方だけゼロ値の場合に、極端に大きい値や小さい値が発生しやすい
- 推定量Bは、各時点のデータを全て足してから比率をとるため、ゼロ値があってもその影響は小さい

時点調整は、時間の経過に伴う景気変動の反映が目的であるが、項目別の比率の数値にかなり差異がある

この比率が、単に二時点でマッチングされた企業の個々の項目の変動状況を反映しているのであれば、時点調整処理の妥当性に疑問が残る

V. 経理項目の補完ドナー候補に対する多変量外れ値の検出

ホットデック補完の対象となる経理項目は、規模が大きいほど分散が大きいため、規模の大きいところほど、最近隣法によるドナー選択でも傾向の異なるデータが選択されてしまう可能性がある。あまり代表性が高くないと思われる極端なデータを、事前にドナー候補からは除外したい、というご依頼に基づき行った検討結果について報告する

1) **外れ値検出処理の目的**
2) **使用する多変量外れ値検出法の選択**
3) **MSD法とそのR関数**
4) **データ変換について**
5) **閾値の調整**

V.1 外れ値検出処理の目的

他の大部分のデータとは大きく傾向が異なる可能性のあるドナー候補を除外するため、多変量外れ値検出法を用いてクリーニングを行う

使用する外れ値検出法は、単峰の楕円体分布を対象とし、多変量正規分布データであれば、99.9%のデータが正常値と判定される範囲を標準の閾値としている。

ここでは、外れ値検出法の適用にあたり、以下の二点について検討した。

- 使用するデータを対称な楕円体分布に近づけるためのデータ変換
- データ変換の後のデータは多変量正規分布よりも裾が長いため、これに合わせた閾値の調整

V.2 使用する多変量外れ値検出法の選択

- 楕円分布に近いデータを適用対象とする手法
- 絶対基準でベストなものがあるわけではなく、データの状況により最適な手法は変わりうる
- 適用対象となるデータの分布は、変換を施したとしても必ずしも対称ではなく、分布の裾が長い場合が多い
- 分布の裾が長く歪みがあり、外れ値を非対称に加えた乱数データを用いて評価

比較結果 : Modified Stahel-Donoho (MSD) 法

- 詳細は、uRos (International Conference on the **u**se of **R** in **O**fficial **S**tatistics) 2018 において発表 [スライド: https://www.nstac.go.jp/services/society.html]
- さらに、"Comparison of multivariate outlier detection methods for nearly elliptical distributions" という題目で、Austrian Journal of Statistics の uRos2018特集号に投稿・採択され、今後掲載される

V.3 MSD法とそのR関数

MSD (Modified Stahel-Donoho) estimators

- カナダ統計局が実用化 （**カナダ版**）
  Franklin, S., & Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method. In *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 186-191).
- 改良版（**Euredit版**）
  Béguin, C., & Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. *Euroedit Deliverable* D4/5.2, 1(2).
- カナダ版及びEuredit版の実装と公開
  和田かず美. (2010). 多変量外れ値の検出~ MSD 法とその改良手法について~. 統計研究彙報, (67), 89-157. [https://github.com/kazwd2008/MSD]
- Euredit版の並列化実装・公開
  Wada, K., & Tsubaki, H. (2013, December). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection. In *2013 International Conference on Cloud Computing and Big Data* (pp. 304-311). IEEE.

Béguin and Hulliger (2003) に基づく ***msd*** 関数を使用

V.4 データ変換について

調査データは歪みがあり、分布の右裾が長いため、楕円分布を対象とする外れ値検出法の適用には何らかの**データ変換**が必要

例) 製造業データ

	最小値	Q1	中央値	平均値	Q3	最大値
05 売上高	185	5,262	11,364	20,353	22,467	761,461
06 費用	67	3,452	7,930	17,428	18,886	760,180
07 期首棚卸高	1	100	305	1,875	1,022	134,000
08 仕入高	5	958	2,911	8,185	6,041	498,602
09 期末棚卸高	1	100	326	1,875	1,032	140,100
10 経費計	50	1,930	4,623	9,243	11,261	261,578

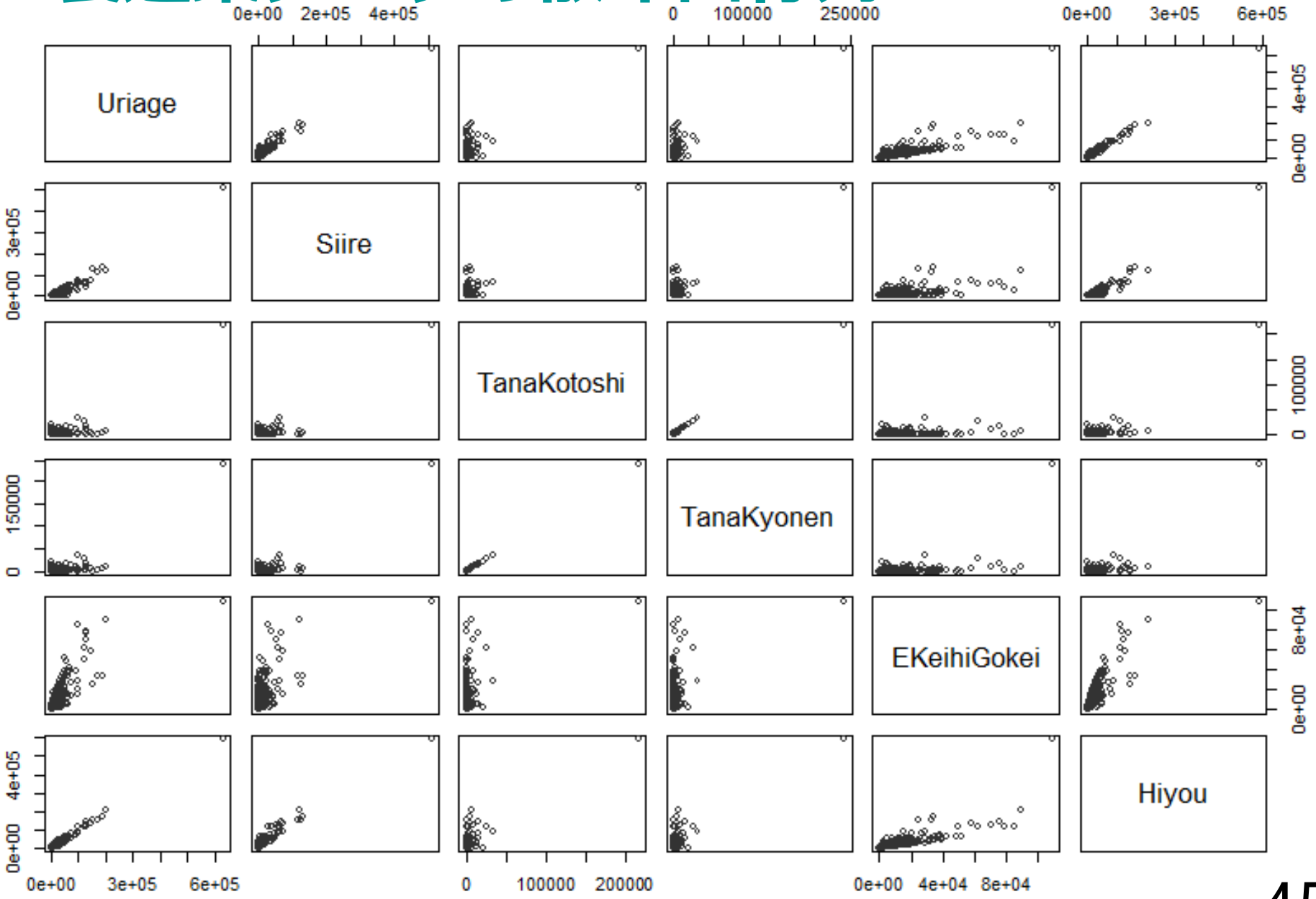
製造業データ

- 相関係数

	ピアソンの相関係数						スピアマンの順位相関係数					
	05	**06**	**07**	**08**	**09**	**10**	**05**	**06**	**07**	**08**	**09**	**10**
05	1.00	0.99	0.79	0.98	0.78	0.94	1.00	0.96	0.44	0.83	0.43	0.90
06	0.99	1.00	0.80	0.98	0.78	0.95	0.83	0.85	0.49	1.00	0.48	0.68
07	0.79	0.80	1.00	0.81	1.00	0.75	0.44	0.47	1.00	0.49	0.95	0.41
08	0.98	0.98	0.81	1.00	0.79	0.88	0.43	0.45	0.95	0.48	1.00	0.39
09	0.78	0.78	1.00	0.79	1.00	0.73	0.90	0.95	0.41	0.68	0.39	1.00
10	0.94	0.95	0.75	0.88	0.73	1.00	0.96	1.00	0.47	0.85	0.45	0.95

変数間の相関は比較的高く、ピアソンの相関がスピアマンよりもかなり高いところが散見される。これは、前者の相関を上げるような極端に大きな値を持つ外れ値の存在を示唆する。（次スライド参照）

製造業データの散布図行列

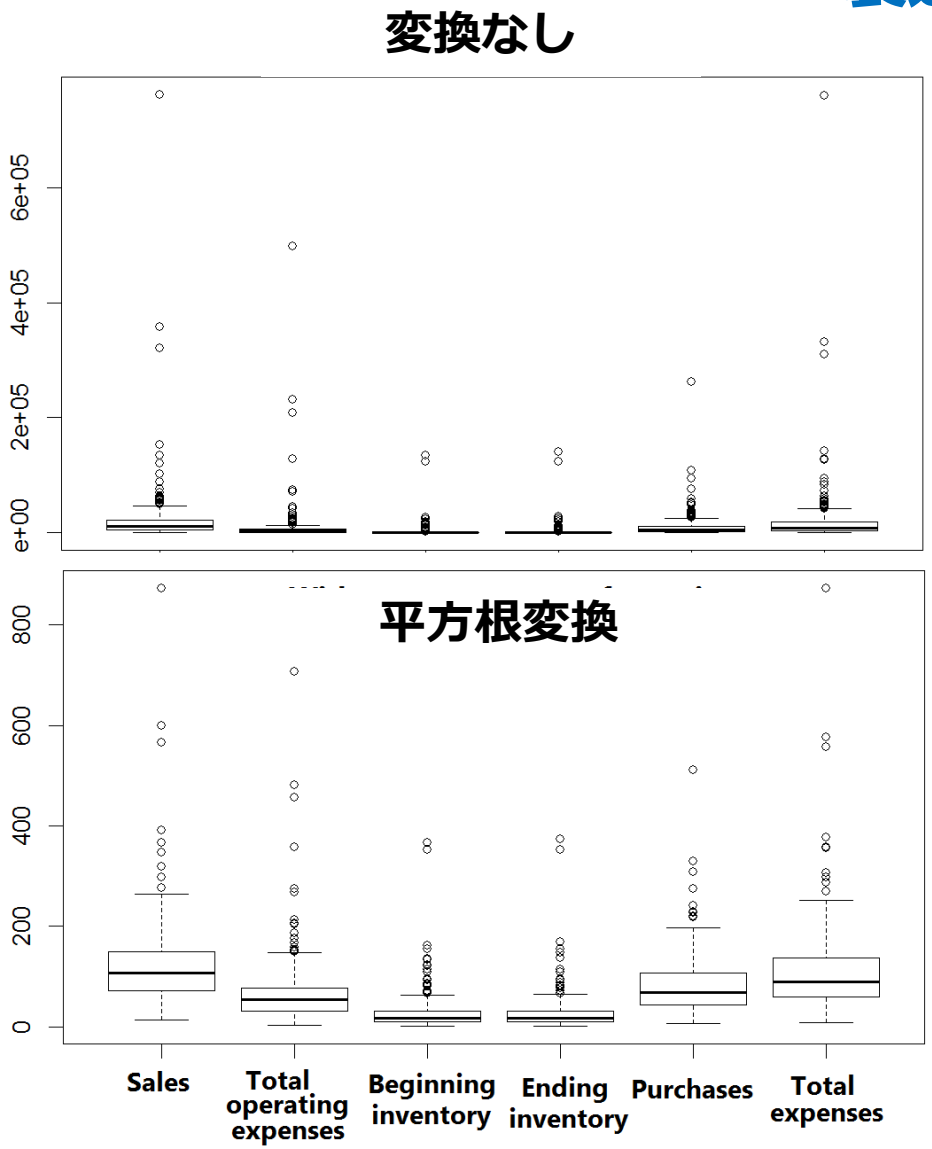


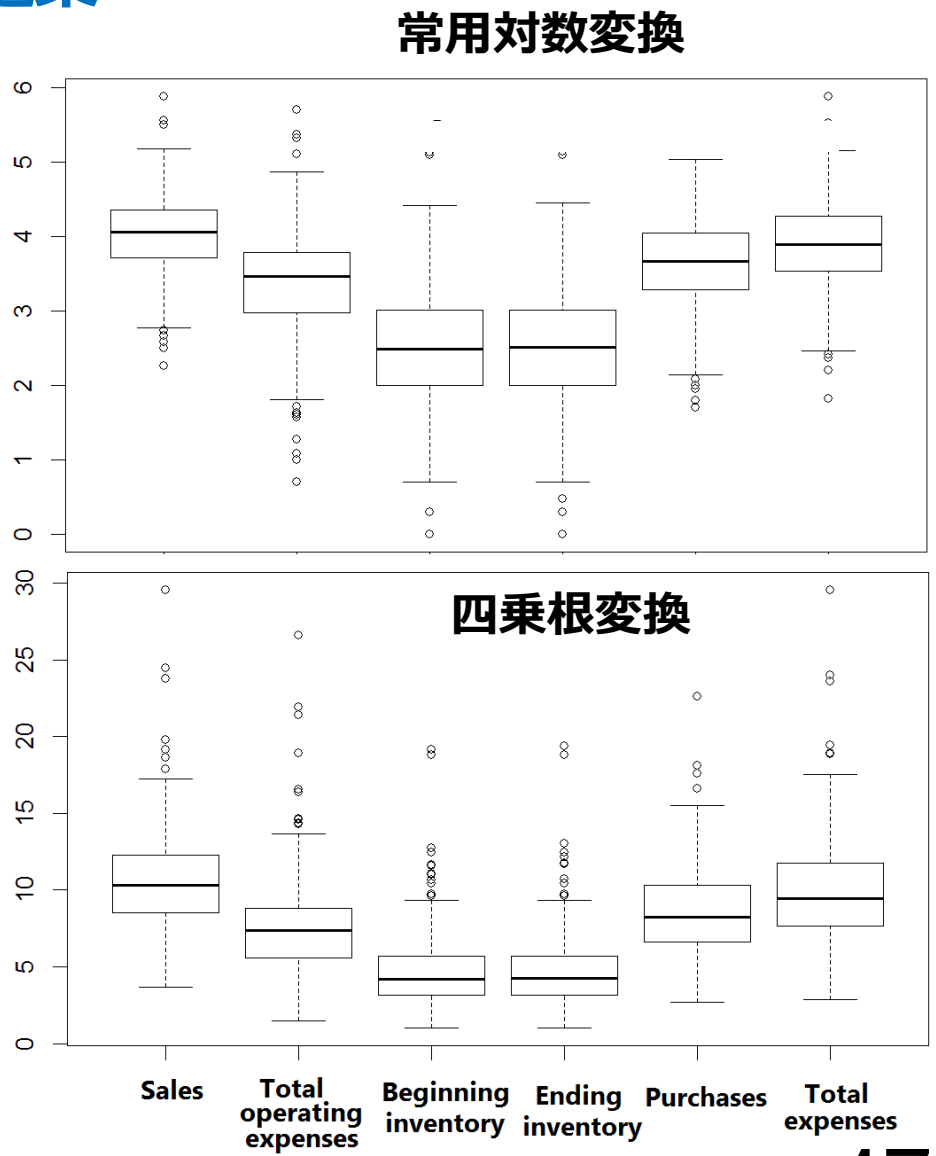
データ変換についての考え方

- 変数間の関係性を変えないように、全ての変数に同じ変換を施す
- 多変量外れ値検出の前処理としてのデータ変換により、検出される外れ値が変わる
- どのようなデータ変換を採用するかを決めるために、一般的にBox-Cox変換が使用されるが、この方法はロバストではないため、結果の利用には留意が必要

異なるデータ変換後の箱ひげ図

製造業





Box-Cox変換のラムダの推定

変数	05	06	07	08	09	10
ラムダの推定値	0.321	0.336	0.152	0.151	0.246	0.317

これは、λ=0.5 は平方根変換、0.25で四乗根変換、0で対数変換を示す

データに大きな外れ値が含まれると、ラムダの値は本来適切な値よりも小さくなるため、このλは適切なものよりも強い変換を示唆している可能性が高い

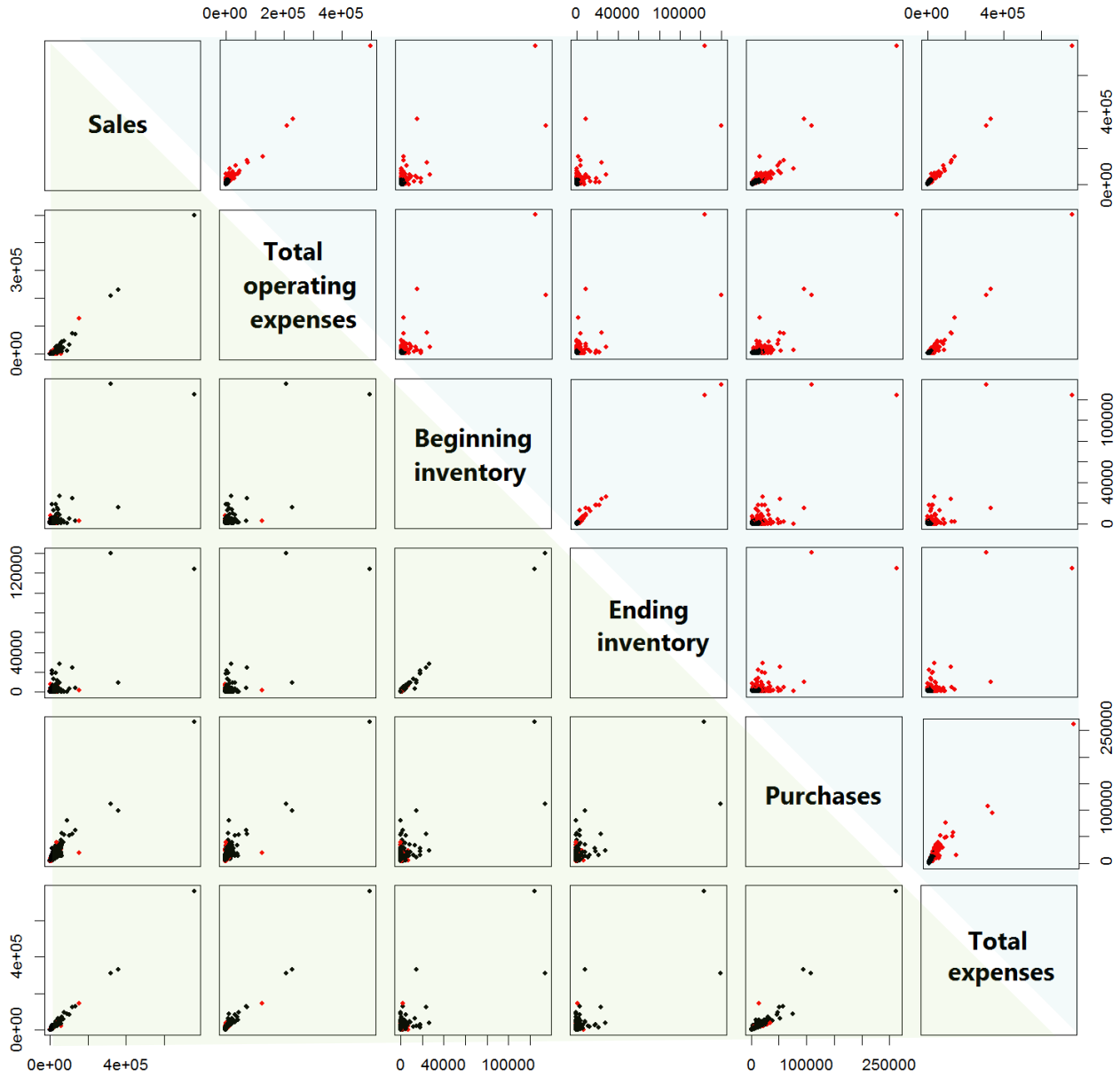
検出される外れ値数

産業	前処理のデータ変換	MSD No.	MSD %
製造業	平方根変換	47	12.05%
	四乗根変換	28	7.18%
	常用対数変換	41	10.51%

データの汚れ方にもよるが、変換により楕円分布のモデルにフィットすれば、検出される外れ値の数は減る

検出される外れ値の違い(1)

製造業



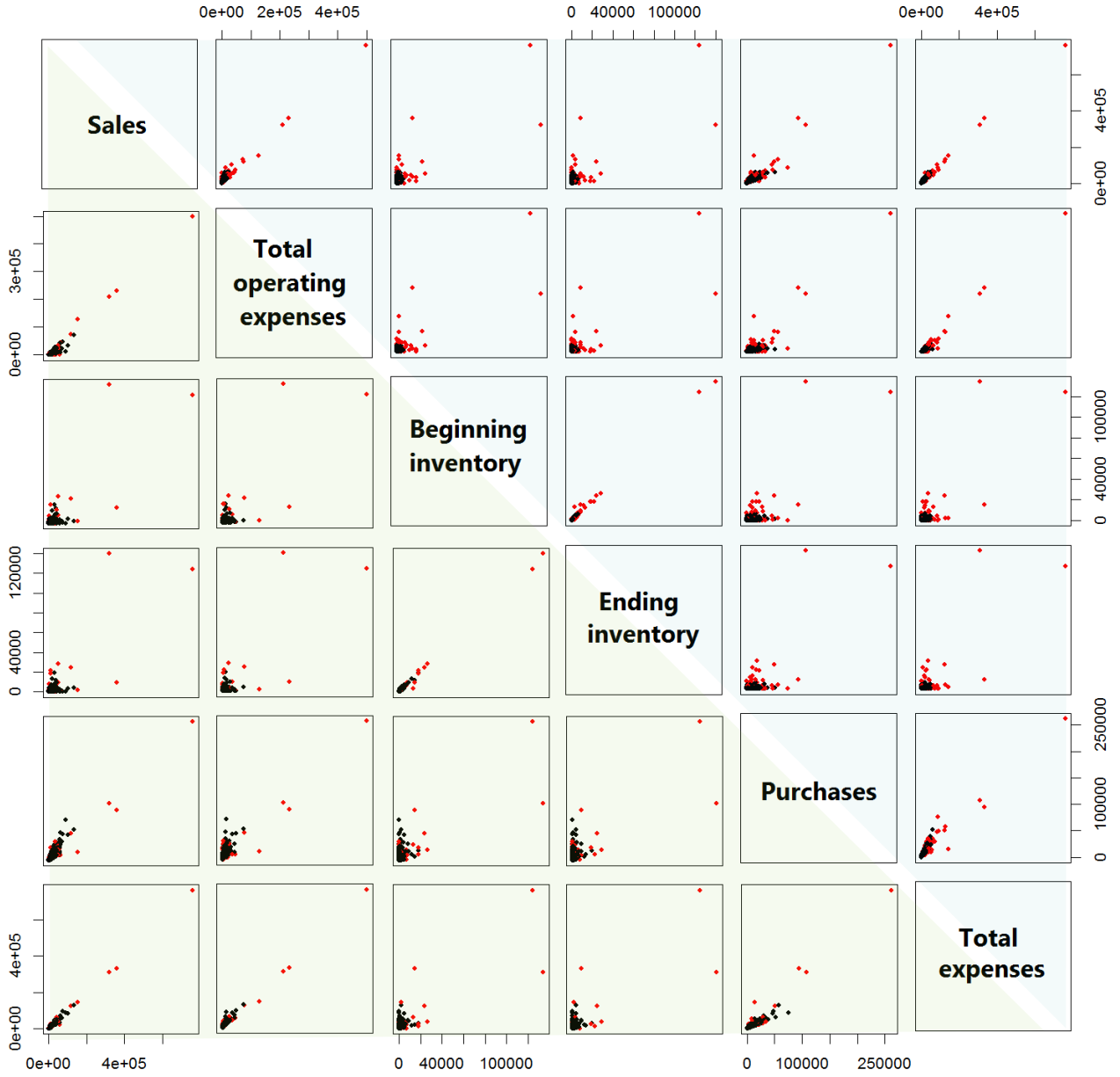
プロットは変換のない実軸表示で、それぞれの変換を施して検出された外れ値を赤で表示

上三角行列: 無変換

下三角行列:
常用対数変換

検出される外れ値の違い(2)

製造業



プロットは変換のない実軸表示で、それぞれの変換を施して検出された外れ値を赤で表示

上三角行列: 平方根変換

下三角行列: 四乗根変換

強い変換では検出対象は相関の外れ値にシフトし、弱い変換では特定の値が大きいものがより検出されやすくなる

ご提案は四乗根変換

V.5 外れ値判定の閾値の検討

平成30年個人企業経済調査の構造編データと、平成28年経済センサス- 活動調査データによる、時点調整をした過去データによる補完と、ドナー候補のクリーニング処理を含めた補完の総合テストを行い、その詳細は別紙３に示す。

別紙3の表4に、補完クラス別・閾値の乗数別の外れ値検出数をまとめたが、その内容に基づき、おおむね売上高90%未満の補完クラスで閾値乗数1.3、売上高90%以上の補完クラスで閾値乗数1.5を試算用の設定に使用している。

ただし、データ変換と同様に、閾値の乗数もデータ分布に依存し、ここでの試算は特定産業のみで、データ量も新調査とは異なっている。今後改めて一年目調査データを用いて、データ変換及び閾値の検討を必要とする。

VI. R1新調査データを用いた今後の分析予定

R1新調査前は、必要な分析データが網羅できず、ここまででご報告した試算結果は、特に拡大産業部分を中心に、必ずしも新調査で同様の結果が得られることは保証されない。

年度内にR1新調査データを用いて必要な分析・検証を行う。

また、今後さらなる改善の余地あると思われる資料内に*印のある部分を含め、より良い補完を目指す研究を行う。

VI. R1新調査データを用いた今後の分析
これまでの経理項目関係の試算の限界

番号	経理項目
05	**売上金額**
06	費用総額*
07	**期首棚卸高**
08	**仕入高**
09	**期末棚卸高**
10	**経費計**
11	**給料賃金**
12	**地代家賃**
13	**減価償却費**
14	**租税公課**
15	**損害保険料**
16	**福利厚生費**
17	**外注工賃**
18	**利子割引料**

- 番号が紫の項目は、H28センサスに存在する
- **項目名が太字**の項目は、H30個人に存在する

- 新調査の対象産業全てについてデータが存在するのは、番号が紫の部分のみ
- 旧調査データは全ての項目についてデータが存在するが、新調査の対象産業全ては網羅していない

前回及び今回の報告は全て既存データによる試算に基づいており、新調査で新たに追加される対象産業について試算どおりの結果になることは保証されない

年度内の作業予定

一年目の調査データを使用して、既存データで困難であった以下のような分析・検討を進める

- 拡大産業部分について、経理項目の補完方法の検証
- 補完クラスの決定
- 棚卸高の補完についての追加検証
- ドナー候補の外れ値検出について、データ変換と閾値の検証