資料５

令和元年個人企業経済調査
～欠測値の補完について～

独立行政法人統計センター
技術研究開発課

１．欠測値補完の基本方針

1. 欠測値補完に使用するデータ

- ◆ 個人企業経済調査の当年データ
- ◆ 事業所母集団DB

2. 補完対象項目

売上金額	仕入金額
期首棚卸高	期末棚卸高
経費計	給料賃金

（計6項目）
（単位欠測は対象外）

3. 補完処理単位：補完クラス

２．補完処理の構成

補完対象項目	補完処理
売上金額	時点調整をした過去の同一企業データによる補完（LOCF法：Last Observation Carried Forward）
仕入金額	最近隣ホットデック補完
経費計	
給料賃金	
期首棚卸高	層化平均値補完
期末棚卸高	

2.1 時点調整済LOCF法

売上金額に対しては**時点調整済Last Observation Carried Forward(LOCF)法**によって補完を行う。

LOCF法
同一企業の前回の調査データを据え置きする方法。
（**時点調整**：過去からの変化率（比率）を加味した調整。）

調査データ

No	売上
001	5000
002	6325
003	NA
004	2500
005	985
006	6610
007	3908

過去値×比率※を代入

※＜比率算出式＞

$$\frac{\text{調査データの売上金額の合計}}{\text{欠測を除く事業所の事業所母集団DBの売上金額の合計}}$$

母集団名簿

No	売上
001	5300
002	6705
003	4131
004	2650
005	1044
006	7007
007	4142

2.2 最近隣ホットデック補完

- 仕入金額、経費計、給料賃金に対しては補完クラスを考慮した上で**最近隣ホットデック法**を用いる。

最近隣ホットデック法
性質の近いデータをドナーとして選び、そのドナーの値を代入する方法。

No.003企業に最も近い企業を探す

今回は欠測変数以外の多変数の線形関係も考慮し、**マハラノビス距離**を用いる。

$$D(\boldsymbol{x}_i) := \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{x}})^{\top} \boldsymbol{V}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{x}})} \quad (\mathrm{i} = 1, \ldots, \mathrm{n})$$

（補完対象項目のうち
欠測していない変数を用いる。）

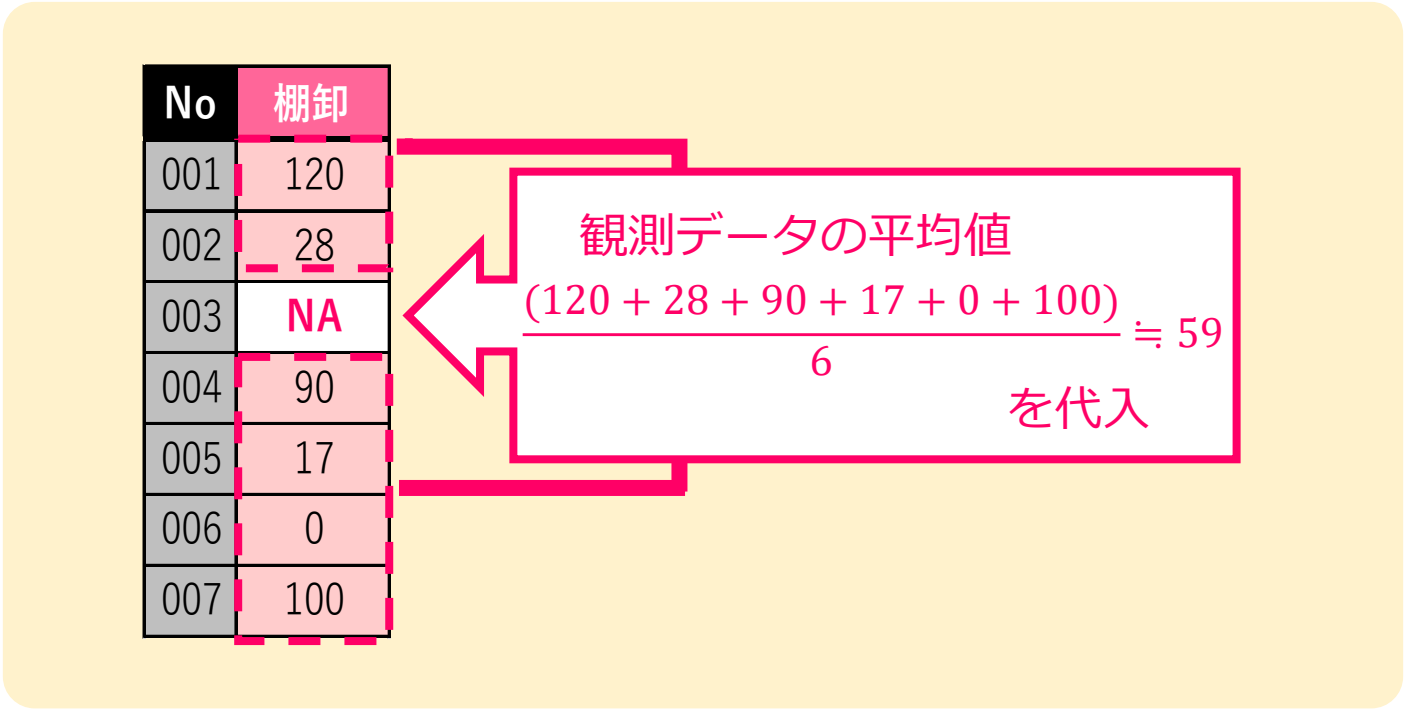
No	売上	仕入	経費	給料
001	5000	150	3050	0
002	6325	273	4797	1238
003	4173	225	2759	NA
004	2500	135	882	0
005	985	0	358	0
006	6610	970	3628	900
007	3908	175	1323	0

No.007の給料の値を代入

2.3 平均値補完

- 期首棚卸高、期末棚卸高に対しては補完クラスを考慮した上で**平均値代入法**を用いる。

平均値代入法
観測されているデータの平均値を代入する方法。



3．シミュレーションの概要

■使用データ

売上金額	仕入金額	経費計	給料賃金	期首棚卸高	期末棚卸高
H30個人企業経済調査（構造編）、H28経済センサス	H14～H29個人企業経済調査（構造編）				

■シミュレーション方法

1. 標本抽出時の層である産業大分類（4区分）、売上高90%点（２区分）別で補完クラスを設定する。
2. データの20%を（意図的にランダムに）欠測させ、各手法により補完することを10000回行う。最近隣ホットデック補完については欠測パターン別にシミュレーションを行う。

（参考：H14～H29年個人企業経済調査のデータサイズ）

産業大分類	対象企業		そのうち項目欠測20%	
	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

欠測のパターン

最近隣ホットデック補完に用いる欠測パターンは７通り

補完対象項目	売上金額	仕入金額	経費計	給料賃金	期首棚卸高	期末棚卸高
補完方法	時点調整済LOCF法	最近隣ホットデック法			平均値代入法	
	最近隣ホットデック補完に用いる欠測パターン					
1	○	×	×	×	○	○
2	○	×	○	×	○	○
3	○	○	×	×	○	○
4	○	×	×	○	○	○
5	○	×	○	○	○	○
6	○	○	×	○	○	○
7	○	○	○	×	○	○

○：観測（補完済含む）
×：欠測

※最近隣ホットデック補完に際し、売上金額は予め補完しておく。

シミュレーション結果 ⇒スライド１０へ

シミュレーション結果 ⇒スライド９へ

結果の評価方法

- **NRMSE (Normalized Root Mean Square Error)**

：標準平均平方誤差

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^{true} - x_i^{imp}}{\sigma}\right)^2}$$

RMSEを標準化した指標。
真値と補完値の乖離について、標準化することにより、データの元々の単位にかかわらず比較が可能。

- **RMSE (Root Means Square Error)**：二乗平均平方根誤差

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{true} - x_i^{imp}\right)^2}$$

- **MAE (Mean Absolute Error)**：平均絶対偏差

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{true} - x_i^{imp}\right|$$

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

- **誤差率**

（真値の平均値－補完後の平均値）／真値の平均値

売上金額（時点調整済LOCF法）

過去からの変化率（比率）による
シミュレーション結果と誤差率

比率	NRMSE	RMSE	MAE	誤差率
0.95	0.49	23437	3843.1	-0.39%

期首・期末棚卸高（平均値補完）

産業別の真値の平均値と補完後の平均値の誤差率

産業	期首棚卸高	期末棚卸高
E	1.90%	1.69%
I	0.12%	0.10%
M	-0.05%	-0.04%
Q	0.00%	-0.06%

仕入高・給料賃金（最近隣ホットデック）

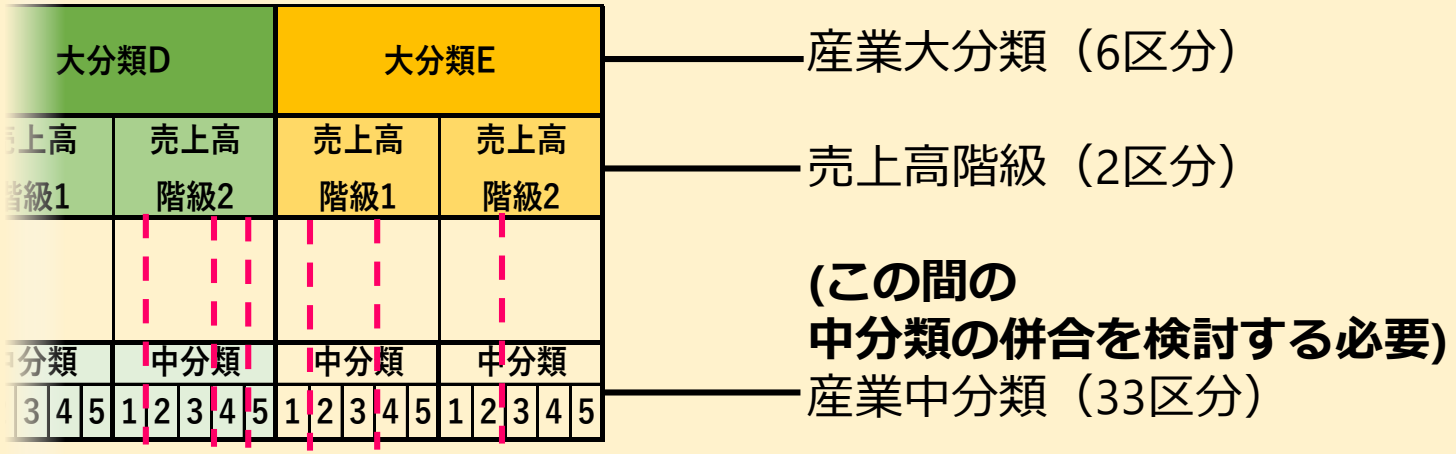
売上金額と経費計から最近隣法（マハラノビス距離）によりドナーを選択し、仕入高と給料賃金を補完した産業別結果

産業	NRMSE		RMSE		MAE		誤差率	
	仕入高	給料賃金	仕入高	給料賃金	仕入高	給料賃金	仕入高	給料賃金
E	0.90	0.87	3434	2306	2023	1185	-0.02%	0.05%
I	0.32	0.73	4600	2190	2742	1075	-0.18%	-0.08%
M	0.61	0.86	2018	1537	1323	934	-0.21%	-0.06%
Q	1.10	0.80	1416	1280	763	641	-0.32%	0.20%

（欠測のパターン２を用いている）

4．補完クラスについて

- **補完クラス**とは、欠測値の補完処理の際に参照するクラスである。できる限り欠測データと類似した補完クラスを決定する必要がある。
- 補完クラスを「売上高階級（2区分）」 ×「産業中分類（33区分）」としたいが、補完クラスのメンバー数が極端に少ない場合、特徴の似た中分類同士を併合する。



5．ドナー候補に対する外れ値処理

- 最近隣ホットデック法を適用する際のドナー候補から、ドナーとしての代表性が低い極端なデータを除外する必要がある。

- 相関関係を考慮した外れ値の検出にあたり、ロバストな推定量であるMSD推定量を用いる。

MSD(Modified Stahel-Donoho)推定量
平均や分散といった要約統計量のロバストな推定量。複数の補完対象項目とその相関を考慮するために用いる。

- 対象とするデータの特徴：単峰の楕円体分布
- 好ましい性質：分布の対象性が崩れても外れ値を適切に検出できる

5.1 外れ値処理のイメージ

補完クラスごとに要約統計量をMSD推定量によって算出し、外れ値の閾値を検討する