独立行政法人
統計センター

信頼に応えてつくる確かな統計

National Statistics Center

統計センターにおける
欠測値補完に関する研究の一事例

2023年度統計関連学会連合大会
令和5年9月5日（火）

独立行政法人統計センター
村田 一郎

この資料はウェブサイトで公開しています

統計センター　学会発表

発表内容

1　統計センターで行っている研究

2　個人企業経済調査における
「0円」に着目した補完の検討

3　補完シミュレーションによる精度の分析

1　統計センターで行っている研究

統計センターの役割

- 国勢調査や消費者物価指数など、わが国の基本となる統計の作成
- 各府省や地方公共団体の統計整備の支援
- 政府統計の総合窓口（e-Stat）を初めとする公的統計の利用基盤の提供
- 公的統計のミクロデータ利用（統計データの二次的利用）の効率的かつ効果的な実施の支援
- 業務の高度化・効率化や統計ニーズの多様化への対応などに資するため、必要な研究開発を推進

など

統計センターにおける研究事例

- 分類自動格付技術の研究
  - 自然言語処理と機械学習を用いた教師付多クラス分類器の開発
  - 家計調査（総務省）への実務適用

- データエディティングに関する研究
  - データチェックや欠測値の補完方法の検討
  - 経済センサス-活動調査（総務省・経済産業省）、
    個人企業経済調査（総務省）への実務適用

個人企業経済調査についての研究

個人企業経済調査の概要

目的	個人企業の経営の実態を明らかにし、個人企業の基礎資料を得る	調査事項	・事業主及び従業員に関する事項 ・営業上の収支、棚卸及び設備投資に関する事項など
期日	毎年6月1日現在	調査方法	郵送またはオンライン
範囲	個人経営の事業所約37,000（抽出）	結果の公表	調査翌年3月までに

◆ 調査の見直し（令和元年度）に向けた研究

・対象産業・調査方法見直し後の結果精度確保のため、主要な項目の欠測についての補完を検討
・統計センターにおける実証研究を踏まえて、
個人企業経済統計研究会（～令和元年10月）において取りまとめ

売上金額	→	過去データを、時点調整したもので補完
仕入金額、経費計、給料賃金	→	他企業データで補完（最近隣法によるドナー補完、ドナー候補選定に当たっては外れ値処理を実施）
期首棚卸高、期末棚卸高	→	層化平均値で補完（平均値代入法）

→令和元年調査分から実務適用

・見直し後の複数年の調査データが蓄積した段階で、データの傾向などをさらに分析し、
補完方法の見直しを検討

※統計法（平成19年法律第53号）第33条第１項の規定に基づき調査票情報を利用

2　「0円」に着目した補完の検討

回答値における0円の割合

令和4年調査

母集団情報	産業	売上金額	仕入金額	棚卸高（期末）	棚卸高（期首）	経費計	給料賃金
売上高90%点未満	D 建設業	0.2%	11.0%	58.9%	59.0%	0.8%	71.6%
	E 製造業	0.1%	22.1%	50.6%	50.5%	1.1%	70.9%
	I 卸売業，小売業	0.2%	2.2%	15.1%	15.3%	0.9%	73.7%
	M 宿泊業，飲食サービス業	0.1%	0.8%	33.6%	34.2%	0.5%	53.9%
	N 生活関連サービス業，娯楽業	0.0%	14.1%	41.7%	42.3%	1.8%	84.0%
	R サービス業（上記産業を除く）	0.2%	69.4%	83.0%	83.3%	2.6%	78.2%
売上高90%点以上	D 建設業	0.0%	4.7%	49.7%	50.4%	0.1%	39.8%
	E 製造業	0.0%	9.7%	32.0%	32.5%	0.0%	23.0%
	I 卸売業，小売業	0.0%	0.2%	6.0%	6.1%	0.0%	12.6%
	M 宿泊業，飲食サービス業	0.0%	0.3%	11.7%	12.4%	0.0%	6.2%
	N 生活関連サービス業，娯楽業	0.0%	5.5%	17.1%	17.3%	0.0%	24.0%
	R サービス業（上記産業を除く）	0.0%	64.6%	73.6%	73.7%	0.0%	24.8%

※発表者による集計

0円に着目した補完について

- 調査項目のうち0円の割合が高い項目の補完について、
  0円を先に補完する方法を検討

1　ロジスティック回帰モデルにより0円か否かを推定 →（0円）→ 0円で補完

↓（0円でない）

2　現行と同じ最近隣法または平均値により補完

- 0円の割合が高い仕入金額、期首・期末棚卸高、給料賃金を対象

0円推定モデルの作成

- i番目のデータのある項目（仕入金額、期首・期末棚卸高、給料賃金のいずれか）が0円であるか否かの確率π_iを、ロジスティック回帰モデルにより推定

$$\log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$$

x_1: 産業分類（中分類、33区分）
x_2: 母集団情報の売上高階級（90%点以上／未満の2区分）
x_3: 従業者規模階級（3区分）
x_4: 都道府県
x_5: 売上に占める受託額の割合階級（5区分）

0円推定モデルの説明変数について

従業者数については、個人企業経済調査の結果表で用いられている
次の3区分の従業者規模階級を説明変数とした。

・事業主のみ	・事業主と家族で無給の人のみ	・雇用者あり

特に給料賃金の0円と0円以外を判別する上で寄与が大きい。

従業者規模階級別、給料賃金が0円と0円以外の企業数及び0円の割合

※従業者数不詳は省略

階級	令和元年			令和2年			令和3年		
	0円	0円以外	0円割合	0円	0円以外	0円割合	0円	0円以外	0円割合
事業主のみ	4766	396	92.3%	5322	278	95.0%	5843	386	93.8%
事業主と家族	2174	151	93.5%	2340	114	95.4%	1968	122	94.2%
雇用者あり	5069	15734	24.4%	5544	16332	25.3%	5192	14503	26.4%

※発表者による集計

0円推定モデルの説明変数について

受託の状況に関する質問のうち、
「受託額の売上げに占める割合」が特に棚卸高の0円と0円以外を判別する上で寄与が大きいため、説明変数とした。

受託額割合別、期末棚卸高が0円と0円以外の企業数及び0円の割合

※受託の状況不詳は省略

割合	令和元年			令和2年			令和3年		
	0円	0円以外	0円割合	0円	0円以外	0円割合	0円	0円以外	0円割合
50%未満	161	425	27.5%	189	418	31.1%	162	356	31.3%
50～100%未満	213	281	43.1%	254	305	45.4%	230	272	45.8%
100%	479	227	67.8%	829	231	78.2%	831	263	76.0%
受託がない	9331	16153	36.6%	9915	17037	36.8%	9157	15347	37.4%

※発表者による集計

3　補完シミュレーションによる精度の分析

補完シミュレーション

- 個人企業経済調査の調査データのうち、補完対象の項目について回答が得られているデータを使用
- 産業中分類（33区分）と、母集団情報の売上高階級（2区分）により補完クラスを構成

項目別補完シミュレーション方法

補完対象	補完方法	シミュレーション方法
共通 （0円の補完）	ロジスティック回帰モデルにより確率を推定し、その確率に基づいて0円を補完	確率の推定は10分割クロスバリデーションの要領で行う
仕入金額及び 給料賃金 （現行方法）	最近隣法によるドナー補完 最近隣法の距離計算は3種類 ・売上金額と経費計のマハラノビス距離 ・売上金額と仕入金額のマハラノビス距離 ・売上金額のみのユークリッド距離	Leave-One-Out法によりドナー選択 （補完クラス内の自身以外のすべてのデータをドナー候補とする）
期首棚卸高及び 期末棚卸高 （現行方法）	平均値代入法	自身も含めた補完クラス内平均値で補完

補完シミュレーション

- 0円を先に補完する場合としない場合の2とおりをシミュレーション
  （0円を先に補完してもその後のドナー選択や平均値への影響はない）

- 回答値と補完値の差を標準化二乗平均平方根誤差（NRMSE）で
  補完クラスごとに評価

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

x^{true}=回答値

x^{imp}=補完値

σ^{true}=回答値の標準偏差

※金額項目は裾の長い分布をしていることから、
  値に1を加えて対数変換したものを使用

- 補完値の分布についても分析

0円推定モデルの評価

・ROC曲線の下側面積（AUC）による評価値

※10分割のクロスバリデーションの平均値

	令和元年	令和2年	令和3年	令和4年
仕入金額	0.92	0.93	0.93	0.93
期首棚卸高	0.82	0.83	0.83	0.83
期末棚卸高	0.83	0.83	0.83	0.83
給料賃金	0.89	0.90	0.90	0.90

・シミュレーションで付与した各項目の確率に基づいて、
0円か否かの実現値を1回発生させたときの適合率・再現率

単位：％

	0円の推定	令和元年	令和2年	令和3年	令和4年
仕入金額	適合率（Precision）	58.15	61.23	61.68	62.19
	再現率（Recall）	58.52	61.86	62.30	62.54
期首棚卸高	適合率（Precision）	56.47	59.12	58.59	58.83
	再現率（Recall）	56.51	59.52	58.44	58.43
期末棚卸高	適合率（Precision）	56.63	58.12	58.44	58.79
	再現率（Recall）	55.84	58.45	57.85	59.17
給料賃金	適合率（Precision）	70.15	72.37	72.53	72.92
	再現率（Recall）	70.26	72.14	72.57	72.99

0円推定モデルの評価

- 最近隣法において0円が補完されたときの回答値との一致率と、0円推定モデルの正解率の比較

令和4年調査データによるシミュレーション結果

補完対象	0円の推定	最近隣法			0円推定モデル
		売上金額＋経費計	売上金額＋仕入金額	売上金額のみ	
仕入金額	適合率（Precision）	62.23	-	59.45	62.19
	再現率（Recall）	61.83	-	58.67	62.54
給料賃金	適合率（Precision）	73.44	70.24	69.67	72.92
	再現率（Recall）	77.81	71.54	67.52	72.99

単位：％

回答値と補完値の差の評価

仕入金額・給料賃金の対数変換後NRMSE（令和4年）

母集団情報	産業	仕入金額		給料賃金	
		最近隣法	0円+最近隣法	最近隣法	0円+最近隣法
売上高90%点未満	D 建設業	1.129	1.307	1.124	1.113
	E 製造業	1.115	1.144	1.081	1.086
	I 卸売業，小売業	0.834	0.913	1.075	1.022
	M 宿泊業，飲食サービス業	0.932	1.017	1.085	1.054
	N 生活関連サービス業，娯楽業	1.200	1.344	1.134	1.021
	R サービス業（上記産業を除く）	1.103	1.014	1.017	0.942
売上高90%点以上	D 建設業	1.093	1.349	1.083	1.176
	E 製造業	1.122	1.265	1.019	1.228
	I 卸売業，小売業	0.481	0.940	0.922	1.135
	M 宿泊業，飲食サービス業	1.132	1.232	1.084	1.493
	N 生活関連サービス業，娯楽業	1.184	1.426	0.991	1.227
	R サービス業（上記産業を除く）	0.865	0.837	0.899	1.004

売上金額と経費計のマハラノビス距離によるドナー選択

回答値と補完値の差の評価

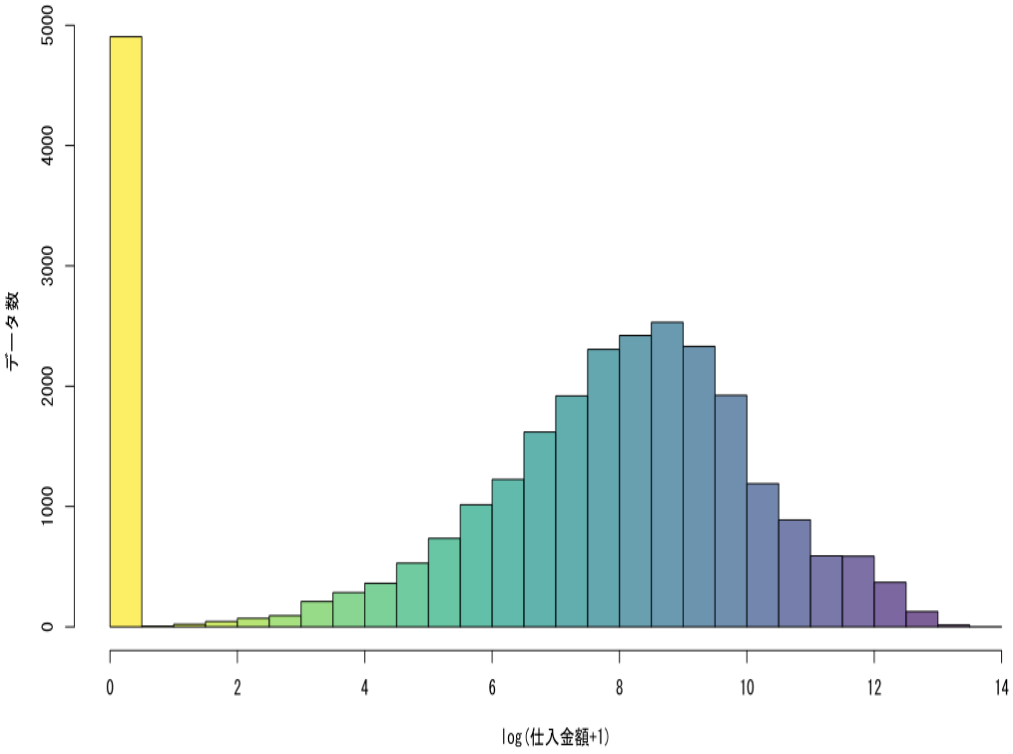
棚卸高の対数変換後NRMSE（令和4年）

母集団情報	産業	期末棚卸高		期首棚卸高	
		平均値補完	0円+平均値補完	平均値補完	0円+平均値補完
売上高90%点未満	D 建設業	1.537	1.344	1.532	1.312
	E 製造業	1.421	1.253	1.419	1.224
	I 卸売業，小売業	1.156	1.324	1.153	1.333
	M 宿泊業，飲食サービス業	1.257	1.322	1.263	1.322
	N 生活関連サービス業，娯楽業	1.340	1.353	1.344	1.352
	R サービス業（上記産業を除く）	2.043	1.183	2.070	1.219
売上高90%点以上	D 建設業	1.497	1.411	1.500	1.445
	E 製造業	1.265	1.265	1.265	1.280
	I 卸売業，小売業	1.101	1.270	1.104	1.309
	M 宿泊業，飲食サービス業	1.156	1.459	1.152	1.437
	N 生活関連サービス業，娯楽業	1.150	1.387	1.152	1.344
	R サービス業（上記産業を除く）	1.727	1.063	1.726	1.098

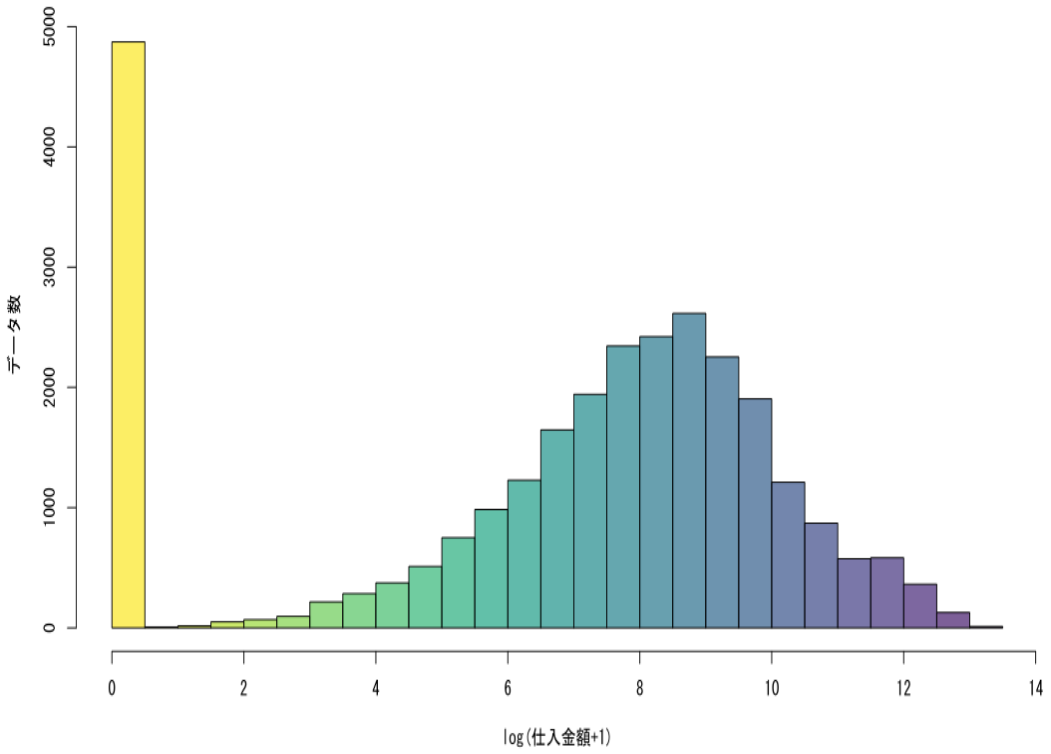
回答値と補完値の分布

令和4年の仕入金額（+1の対数）

回答値



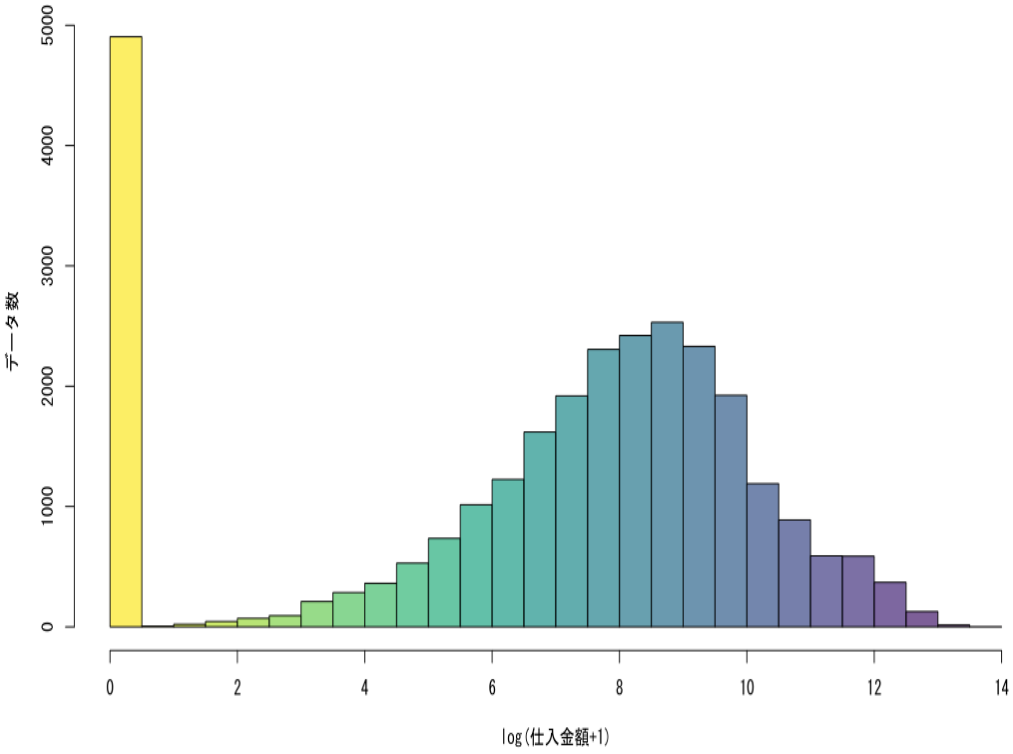
最近隣法の補完値



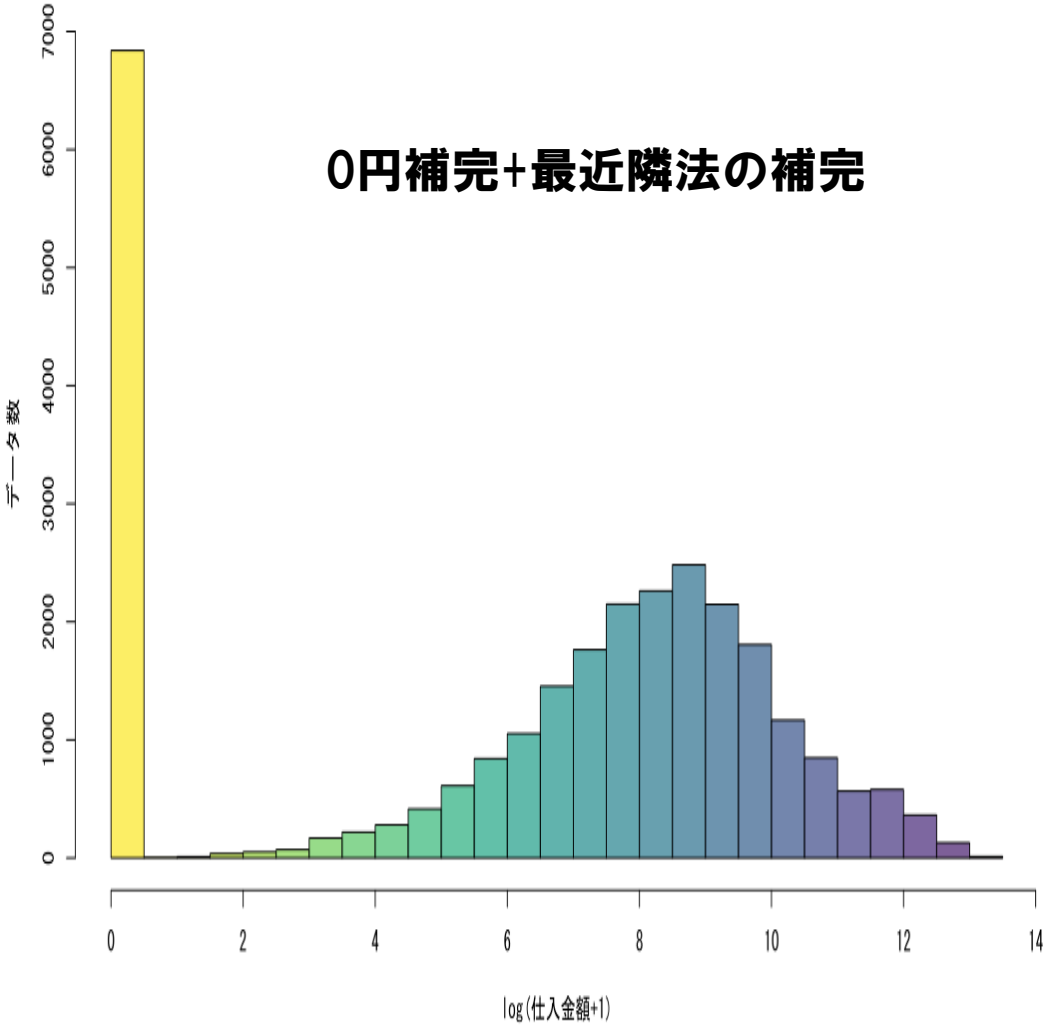
回答値と補完値の分布

令和4年の仕入金額（+1の対数）

回答値



0円補完+最近隣法の補完



補完シミュレーションの結果

- 0円を推定するモデルは、最近隣法のうちもっとも良い距離計算（売上金額＋経費計）を行った場合に近い精度となった

- 0円を推定して先に補完した場合、
  - 回答値と補完値の差のNRMSEは、
    0円の割合が高い区分において改善、低い区分においては悪化
  - 0円を先に補完した分、補完値における0円の割合が増える

まとめ

- 統計センターにおける技術的な研究事例を紹介

- 個人企業経済調査の欠測値補完について、既存の研究に加えて0円に着目した補完方法を検討した
  - 0円を推定するモデルは現行の補完方法と同程度の精度を示した
  - 補完値の分布の変化を見ることも重要
  - 現行の方法と組み合わせるには、0円以外の値の補完方法について更なる検討が必要