＜資料編＞

資料 2-3　主催講演

「調査における『欠損値補完』について」

（立正大学データサイエンス学部 教授　高部勲）

日本マーケティングリサーチ協会講演会
（2024年3月8日 於：日本マーケティングリサーチ協会）

公的統計の欠損値補完について

高部　勲
立正大学データサイエンス学部

1

Contents

2

☞1．自己紹介

2．欠測値補完の概要

3．参考文献の紹介

4．公的統計における欠測値補完の事例

3

自己紹介

https://www.ris.ac.jp/ds/staff/009_isao-takabe.html

立正大学データサイエンス学部
Faculty of DATA SCIENCE

立正大学 150th RISSHO

Isao Takabe

教授

博士(統計科学)

高部 勲

たかべ いさお

専攻（担当分野）
公的統計、統計科学

略歴

2001年	早稲田大学理工学部卒業
2008年	政策研究大学院大学政策研究科政策専攻修士課程修了、修士（公共政策）
2019年	総合研究大学院大学複合科学研究科統計科学専攻博士課程修了、博士（統計科学）
2002～2021年	総務省（統計局、大臣官房、統計研究研修所等）、独立行政法人統計センター
2020～2021年	統計数理研究所客員准教授
2020～2021年	滋賀大学特任准教授
2021年	現職

研究テーマ

(1)公的統計データ・企業データのデータリンケージ・統計的マッチング、(2)状態空間モデル・時系列回帰モデル等に基づく消費関連データ（公的統計データ、POSデータ）の予測・経済指標の開発、(3)公的統計ミクロデータの高度な利活用方法などについて研究しています。

担当科目、ゼミ

ＡＩ入門Ⅰ、ＡＩ入門Ⅱ、統計調査法、社会調査の設計と実査、社会調査実習Ⅰ、社会調査実習Ⅱ、ゼミナールⅠ、ゼミナールⅡ、ゼミナールⅢ、ゼミナールⅣ、卒業研究・卒業論文

4

１．自己紹介

☞**２．欠測値補完の概要**

３．参考文献の紹介

４．公的統計における欠測値補完の事例

5

調査票の点検・補正と無回答・欠測

調査票の点検･補正

- 回収された調査票:
  ･無回答
  （回答されてはいるものの）
  ･部分的に無回答
  ･回答内容に不備　など
  ⇒こうした回答の不備を可能な限り発見･修正するのが「調査票の点検」
- 特に記入のミスなどは、（記入の）現場に近いところで直した方が効率的
  ⇒調査における誤りは、その発生場所に近いほど、
  調査対象への確認もしやすく、事実に即した確認･訂正が可能
- 原則は、「回答者に確認して補正」
  ⇒前後関係などから推測できる場合には「訂正」
  ⇒最終的に確認･訂正ができない場合には「無回答」

6

調査票の点検・補正と無回答・欠測

●統計調査においては，無回答や無記入により，
調査対象や調査項目の一部についての情報を
得られない場合がある
⇒このような欠落した値を欠測値（Missing data）
という

●欠測値を，何らかの補助情報などを用いて
埋めることを，欠測値の補完や補定（Imputation）
という

※「欠測値」…統計調査で用いられることが多い

※「欠損値」…財務諸表で用いられることが多い

7

調査票の点検・補正と無回答・欠測

●欠測値の補定を行う際に，以下の２種類の方法がある
・ホットデック法
・コールドデック法

●ホットデック法（Hot-deck imputation）：
実施している統計調査における調査票の情報を用いて
補定を行う方法
⇒類似の他の世帯の情報で埋めるなど

●コールドデック法（Cold-deck imputation）：
当該統計調査以外の他の情報（前回の調査結果，
外部の補助情報など）を用いて補定を行う方法
⇒行政記録情報によって統計調査の欠測値を
埋めるなど

8

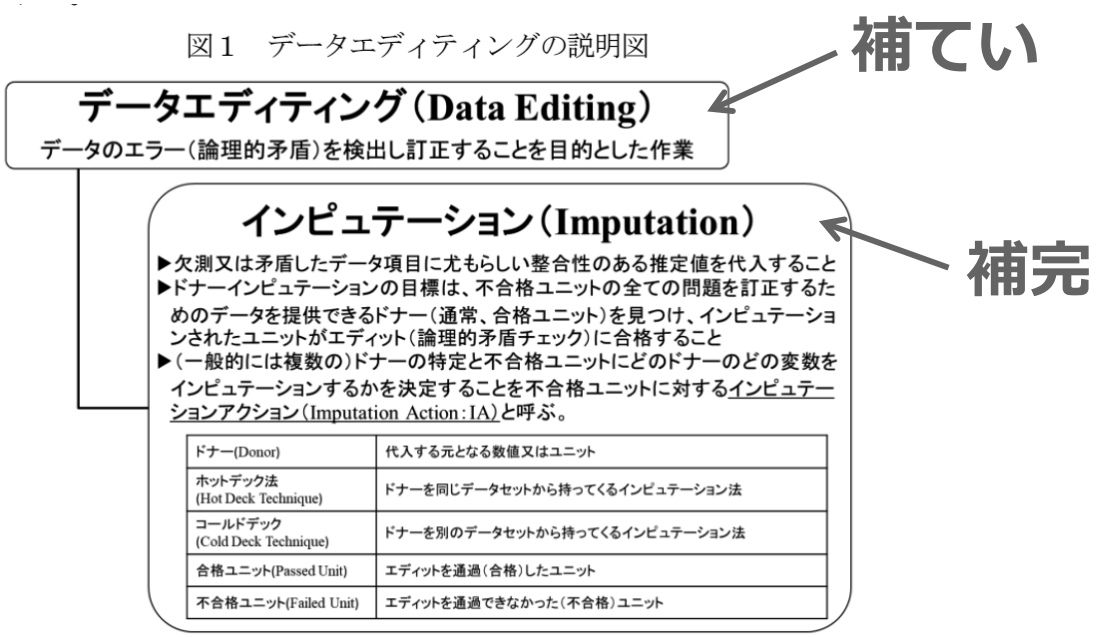
調査票の点検・補正と無回答・欠測

- 事業所を対象とした統計調査では,
  - ・産業・地域の平均値
  - ・前回の調査結果
  - ・行政記録情報

  などにより，欠測値の補定が行われる場合がある
- 世帯や個人を対象とした統計調査では,
  属性が類似している他の世帯または個人を選定し,
  それらの結果を用いて補定を行う場合がある

欠測値補完の種類

図１　データエディティングの説明図



北原、寺垣内（2023）諸外国の国勢調査におけるインピュテーション方法
(統計研究研修所 統計研究彙報　第80号)
https://www.stat.go.jp/training/2kenkyu/ihou/80/pdf/2-2-808.pdf

調査における欠測の分類と対応

欠測が生じる状況

- 脱落(Drop out)：
  - ・長期の調査である時点のデータが欠落
- 無回答：
  - ・調査全体に対する拒否（ユニット非回答）
  - ・一部質問に対する回答忘れ・拒否（項目非回答）
- 打切り(Censoring)：
  - ・ある変数が決められた範囲外の値をとることでデータが欠測（生存時間分析など）
- 切断(Truncation)：
  - ・欠測となったデータの件数も欠落

（統計委員会「統計データの補完に関する調査」報告書から引用）

11

調査における欠測の分類と対応

欠測の種類

- ユニット非回答（Unit Non Response）
  - ・一部質問に対する回答忘れ・拒否（項目非回答）
- 項目非回答（Item Non Response）
  - ・一部の変数に対するデータが欠測

欠測のパターン

- 単調な欠測

(1) 単調な欠測パターン

単調な欠測パターンとは、欠測を含むデータの変数と回答者（ケース）を以下のように入れ替えられるものである。

図表 単調な欠測パターンのイメージ（○：回答、×：欠測）

回答者＼変数	①	②	③	④	⑤	⑥	…
A	○	○	○	×	○	○	
B	×	○	×	×	○	×	
C	○	○	○	○	○	○	
D	×	○	×	×	○	×	
E	○	○	○	○	○	○	
F	×	×	×	×	○	×	
⋮							

変数の並び順を入れ替え
回答者の並び順を入れ替え

回答者＼変数	⑤	②	①	⑥	③	④	…
C	○	○	○	○	○	○	
E	○	○	○	○	○	○	
A	○	○	○	○	○	×	
B	○	○	×	×	×	×	
D	○	○	×	×	×	×	
F	○	×	×	×	×	×	
⋮							

- 単調でない欠測…複雑な反復計算が必要

（統計委員会「統計データの保管に関する調査」報告書から引用）

12

調査における欠測の分類と対応

欠測のメカニズム

- **完全にランダムな欠測（ＭＣＡＲ）：**
  - **・欠測の発生する確率が当該変数の値及び他の観測されている変数の値に依存しない場合**
- **ランダムな欠測（ＭＡＲ）：**
  - **・欠測の発生する確率が当該変数の観測された値及び他の観測されている変数の値には依存するが当該変数 の欠測 となった値には依存しない場合**
- **ランダムでない欠測（ＭＮＡＲ）：**
  - **・欠測の発生する確率が当該変数自体の値に依存**
  - **⇒欠測メカニズムをモデル化する必要**

（統計委員会「統計データの保管に関する調査」報告書から引用）

13

調査における欠測の分類と対応

欠測データの取扱い

- **完全ケースに基づく分析：**
  - **・一部の変数でも欠測を含む場合は、そのようなレコードをすべて取り除いて分析**
  - **・最も簡単であるが、欠測メカニズムがMCAR以外の場合にデータの偏り、結果への影響が生じる**
- **利用可能なケースに基づく分析：**
  - **・その変数が得られてるケースを全て対象として分析（相関係数の計算）**
  - **・完全ケースと同様の問題**

（統計委員会「統計データの保管に関する調査」報告書から引用）

14

欠測値補完の考え方

欠測による影響

- 調査項目の一部について未記入・無回答などの欠測が発生した場合、そのまま集計すると母集団としての代表性が損なわれ、平均値などの結果に偏りが発生するおそれ

補完に関する視点

- 欠測に対しては、基本的にはデータ収集段階においてできる限り発生しないように対処することが必要
- 最終的に発生した欠測値に対しては、統計的な処理として 、可能な限り 補完を 行うことにより統計表及び平均値などの結果の有用性を確保
- 欠測データメカニズムも考慮した適切な対応が必要
- 単純に欠測のない標本のみを用いると、欠測のメカニズムが完全にランダムな場合以外は結果に偏りが生じる

（統計委員会第９回評価分科会資料から引用）

15

欠測値補完の方法

２ 補完を行うための主な方法

欠測の種類	主な補完方法	補助変数の利用
項目の欠測	層化平均値代入	標本の層化
	回帰代入	説明変数
	比率補完	比率の計算
	ホットデック法	標本間の距離の計算、傾向スコアの算出等
	LOCF	伸び率の計算
ユニット単位での欠測	ウェイト調整	回答確率を計算するクラス、傾向スコアの算出

（統計委員会第９回評価分科会資料から引用）

15 16

欠測値補完の方法

2 補完を行うための主な方法

＜項目の欠測（item nonresponse）への対応＞

（1）（層化）平均値代入（Mean Imputation）

○手法の概要

・欠測値に対し、観測された標本の値の平均値を代入する。層化平均値代入は、すべての標本を適切に層化した上で、その層内の平均値を代入に用いる。

○手順

・層化平均値代入では、すべての標本について観測されている適切な項目（補助変数）に基づき、欠測のある標本を含めて標本を層化する。

・各層内で、欠測値に対し、観測されている標本の平均値を代入する。

○利用上の注意点等

・平均値代入は簡易な方法であるが、欠測が完全にランダムに発生している場合以外は、母平均の推定値には偏りが発生する。その改善として、適切な項目により標本を層化した上で代入を行うことにより、偏りを緩和することができる。

・補助変数として利用する項目には、欠測している項目と関連を有し、欠測のし易さとも関連する項目を使うのがよい。

・なお、平均値の補完に伴い、標本分散については過小に評価される。

（統計委員会第９回評価分科会資料から引用）

17

欠測値補完の方法

2 補完を行うための主な方法

（2）回帰代入（Regression Imputation）

○手法の概要

・欠測値に対し、回帰モデルに基づく推定値を代入する。

○手順

・欠測が生じていない標本を用いて、欠測している項目を従属変数とし、観測されている項目を説明変数とする回帰モデルを推定する。

・当該回帰モデルにより推定した値（回帰直線上の理論値）を代入値とする。

○利用上の注意点等

・回帰モデルは、欠測値に対しよい予測値を与える可能性があるが、そのためには適切なモデリングが必要となる。

・説明変数に用いる変数には、連続値のほか、カテゴリカル変数などがある。
なお、説明変数を一定の層への所属を表わすダミー変数とした場合には、層化平均値代入と同じものを表わす。

・線形回帰モデルによる理論値の代入に伴い、標本分散については過小に評価される。欠測値のばらつきを考慮して、予測値に誤差項（乱数）を加える方法は確率的回帰代入と呼ばれる。

（統計委員会第９回評価分科会資料から引用）

18

欠測値補完の方法

2 補完を行うための主な方法

（3）比率補完（Ratio Imputation）

○手法の概要

・欠測が発生している項目と他の項目との比率を利用して、代入値を算出する。

○手順

・欠測が発生していない標本を用いて、補完の対象とする項目（y）と他の項目(x)との比率(r)を計算する。

・欠測が生じている標本において観測されている項目(x)に当該比率(r)を乗じることで得られた値を欠測値への代入値とする。

・比率の算出は、観測されている項目を利用して適切な層区分を設定し、それら層区分ごとに行う。

○利用上の注意点等

・比率を算出する際に利用する項目としては、欠測が生じている項目に対して相関が高い項目を利用するのがよい。

（統計委員会第９回評価分科会資料から引用）

欠測値補完の方法

2 補完を行うための主な方法

（4）ホットデック法（Hot Deck Methods）

○手法の概要

・欠測値に対し、同じデータセットの中で、欠測が生じている標本と類似した標本（ドナー）を探し出し、ドナーの観測値を欠測値の代わりとして代入する。

・標本間の距離を定義し、欠測がある標本に近い標本をドナーとする。

○手順

・欠測が生じている標本と欠測が生じていない標本について、共通して観測されている項目（補助変数）の値を基に一定の距離を計算し、最も距離の近い標本の観測値を欠測値に代入する。

○利用上の注意点等

・回帰代入のようなモデルの仮定を要しないが、類似した標本を探し出すための作業が必要となる。

・用いる距離としては、標本に関する補助変数のベクトルに関するユークリッド距離や、マハラノビス距離などがある。マハラノビス距離は、変数間の相関を考慮した距離である。

（統計委員会第９回評価分科会資料から引用）

欠測値補完の方法

2 補完を行うための主な方法

・また、標本のすべてについて傾向スコア(注)を推定し、傾向スコアを距離としてその値が最も近い(差の絶対値が最小となる)標本の観測値を代入値とする方法もある。

(注)傾向スコア:標本ごとの補助変数の値に応じて標本が回答する確率を表わし、標本全体を用いてロジットモデルなどにより推定されたモデルを基に推定される。傾向スコアは、欠測の発生がランダムの下、モデルが正確であることが必要となる。

・距離に基づく以外の方法としては、観測されている項目に基づきすべての標本をセルに分類し、欠測のある標本と同じセル内に存在する欠測のない標本からランダムに選んで、その観測値を代入値とする方法などがある。

・なお、ドナーを同一のデータセットではなく、過去の調査結果など別のデータセットから探す場合はコールドデックと呼ばれる。過去の調査結果を利用する場合、利用するデータが経年で安定的なものであることなどが必要と考えられる。

(統計委員会第９回評価分科会資料から引用)

欠測値補完の方法

2 補完を行うための主な方法

(５)LOCF(Last Observation Carried Forward)

○手法の概要

・同一の客体を複数時点にわたって調査する場合(パネルデータ)において、欠測が発生した以降の各時点の値として、直近の観測値を代入値とする。

・欠測の発生以降、長期に適用するなどの場合は、経時による変化等を反映させるため、何らかの調整を行うことが考えられる。

○手順

・欠測が発生している標本について、直近の観測値を欠測値に代入する。

・経時による調整としては、欠測が生じている項目について、直近の観測値からの伸び率を欠測のない標本を用いて算出し、欠測が発生している標本の直近の観測値に乗じた値を代入値とする。

○利用上の注意点等

・欠測が発生した以降、当該項目の値は変化しないとみなすものであるが、補完の対象とする項目によっては長期に固定して用いた場合、妥当な推計とならない可能性がある。

(統計委員会第９回評価分科会資料から引用)

欠測値補完の方法

2 補完を行うための主な方法

（6）その他

○演繹的補完（Deductive Imputation）

・欠測が生じている標本において、観測されている項目間の関係から、欠測している項目の値を論理的に定めることができる場合、その値により補完する。

（例）費用合計の回答があり、内訳の一つにのみ欠測が生じていた場合、引き算で算出した欠測値を補完するなど

・補完に際して、一番初めに取り組むべき方法と考えられる。

○他の統計調査の結果、公開情報、行政記録情報等の活用

・欠測が生じている標本について、他の情報（他の統計調査、公開情報、行政記録情報、事業所母集団データベースの情報等）を用いて補完する。

ただし、情報の把握時点の違いや、統計上用いている定義との違いなどに注意する必要がある。

（統計委員会第９回評価分科会資料から引用）

欠測値補完の手順

3 補完の処理の主な手順

①欠測の発生状況の確認

○欠測が発生しており、補完の対象となる項目の確認

○欠測値が生じている標本について、欠測の発生状況や、他の項目・特定の属性等との関係が欠測の発生のし易さに影響していないかなどの特徴を把握

※分布を確認することや、全ての標本において観察されている適当な項目で標本を層化し、層ごとの回収率を確認するなど

⇩

②補完に利用可能な補助変数等の検討

○①の確認結果を踏まえ、欠測している変数と関係の強い変数や欠測のし易さに関連していると見られる項目（変数）などの利用可能性を検討

○その他、欠測の内容に応じて他の情報（当該調査の前回等の結果や外部の関連情報等）の適切な利用可能性についても検討

⇩

（統計委員会第９回評価分科会資料から引用）

欠測値補完の手順

3　補完の処理の主な手順

③適切な補完方法の検討

○②により利用可能な補助変数等も考慮し、適切な補完の方法について検討

・まず、演繹的な補完や、過去の結果から経年で安定的なものであれば利用を検討

・項目の欠測に対し、補助変数を基に欠測値を適切に予測できそうな場合は回帰補完や比率補完、ホットデック法等の検討。他には、層化平均値代入、LOCF（時点調整を含む）等の検討

・ユニット単位での欠測の場合はウェイト調整法を検討

○補完を行う上で層化を行う場合、適切な層区分の方法についても検討（欠測している項目と関連し、欠測のし易さにも関連する項目（変数）で層化を行うのがよい）

○調査項目ごとに異なる複数の補完方法を用いる場合は、補完の手順等を検討

○適用する補完方法間の比較を行うには、以下の様な方法がある

・観察されている項目の一部を欠測させるなどのシミュレーションを行い、推定値の真値からの乖離を表わす指標（平均平方誤差（RMSE）など）を利用する方法

・推定結果をセンサス調査等他の情報源と比較する方法

⇩

④補完方法の選択

○実務上の実行性等も勘案し、適切な補完方法を決定

（統計委員会第９回評価分科会資料から引用）

17

25

1．自己紹介

2．欠測値補完の概要

☞**3．参考文献の紹介**

4．公的統計における欠測値補完の事例

26

参考文献の紹介

資料1

評価分科会の設置について

平成30年11月 28 日
総務省統計委員会担当室

1．分科会設置経緯

〇統計改革推進会議最終取りまとめ（平成29年５月）（抄）

４．報告者負担の軽減と統計業務・統計行政体制の見直し・業務効率化、基盤強化
（２）統計業務の見直し・業務効率化及び各種統計の改善
③ 「評価チーム」による統計の有用性・信頼性の向上

個別統計について、正確性やユーザーのニーズへの適合性、公表の適時性、統計データの解釈可能性などの品質を確保し、その有用性・信頼性の向上に資するため、統計委員会の通常の取組とは独立して個別統計の品質の評価を行う評価チーム（仮称。以下同じ。）を、統計委員会の必置機関として設置する。

評価チームは、個別統計の品質の評価を、諮問を受けることなく、自らの把握した情報等に基づき、自ら課題を設定して調査審議を行い、評価結果を統計委員会・各府省に報告する。このため、評価チームは、ユーザーのニーズ、調査環境の実情、現場の課題等を積極的に把握することとする。

また、評価チームによる評価結果及びそれを受けた統計委員会・各府省における対応と考え方については、それぞれ公表する。

さらに、評価チームについては、評価組織にふさわしい自律性・中立性を確保することとし、そのための組織・運営の基本的考え方は以下のとおりとする。

- 評価チームは、統計委員会を通じることなく、評価結果を述べることができるようにすること
- 評価チームによる評価の際に委員等の意見の一致をみなかった場合、評価結果報告書には、その旨を明記すること
- 評価チームの委員等のうち、統計委員会内の他部会等に属する委員等は、その半数を超えないものとすること。その際、評価チームと統計委員会の他部会等を兼ねる委員等は、同一の統計について双方で議決権を行使することのないよう、当該他部会等で自ら関与した統計については、評価チームでは、議決権を行使しないものとすること
- 評価チームの委員等のうち、統計委員会内の他部会等に属しない委員等も、形式的には統計委員会（本委員会）の委員等であることから、同一の統計について双方で議決権を行使することのないよう、統計委員会（本委員会）では議決権を行使しないものとすること

1

27

参考文献の紹介

28

参考文献の紹介

第5回評価分科会

日時

令和2年1月27日(月)10:00～12:00

場所

総務省第2庁舎6階特別会議室

配布資料

議事次第
資料1 評価分科会に所属する委員、臨時委員及び専門委員の内閣総理大臣指名について
資料2 当面の評価分科会の検討の進め方(案)
資料3 精度検査報告書において、平成30年度までに実施すべきとされた事項についての関係府省の取組の現状(財務省、国土交通省関係)
資料4 法人企業統計調査の欠測値補完について
資料5 民間給与実態統計調査における欠測値補完等について
資料6 造船造機統計調査における調査対象事業所の整理について
資料7 自動車輸送統計調査における欠測値補完に関する取組等について
資料8 建築着工統計調査補正調査の見直しについて
資料9 諸外国における欠測値補完について
資料10 各府省の統計作成支援のための業務相談窓口について
資料11 「欠測値補完に関する調査研究」について(内閣府)
参考1 統計委員会令
参考2 統計委員会運営規則
参考3 欠測値への対応に関する各府省研究成果
参考4 欠測値への対応に関する各府省等職員研究発表の紹介
参考5 平成28年度統計法施行状況に関する審議結果報告書(統計精度検査関連分)
参考6 令和元年度 統計委員会評価分科会審議結果報告書(第1回～第4回審議分)
参考7 第4回評価分科会議事概要

議事次第

(1)分科会長の互選、分科会長代理の指名について
(2)当面の検討の進め方について
(3)精度検査報告書※提言に対応した取組について(法人企業統計調査、民間給与実態統計調査、造船造機統計調査、自動車輸送統計調査及び建築着工統計調査の補正調査)
(4)諸外国における欠測値補完及び総務省による各府省の統計作成支援について
(5)欠測値への対応に関する内閣府の研究成果について
※(平成28年度統計法施行状況に関する審議結果報告書(統計精度検査関連分)
(平成30年3月統計委員会)

29

参考文献の紹介

第6回評価分科会

日時

令和2年2月19日(水)16:00～18:00

場所

総務省第2庁舎 6階 特別会議室

配布資料

議事次第
資料1 精度検査報告書において、平成30年度までに実施すべきとされた事項についての関係府省の取組の現状(厚生労働省関係)
資料2 薬事工業生産動態統計調査回収率の管理について
資料3 賃金構造基本統計調査の欠測値補完について
資料4 個人企業経済調査欠測値の補完について
資料5 令和元年度個人企業経済調査～欠測値の補完について～
資料6 「特定サービス産業実態調査等における推計手法の確立に関する調査研究」について

参考1 「特定サービス産業実態調査等における推計手法の確立に関する調査研究」調査報告書
参考2 第5回評価分科会議事概要

議事次第

(1) 精度検査報告書※提言に対応した取組について(薬事工業生産動態統計調査及び賃金構造基本統計調査
(2) 欠測値への対応に関する総務省・統計センターの研究成果について
(3) 欠測値への対応に関する経済産業省の研究成果について

※平成28年度統計法施行状況に関する審議結果報告書(統計精度検査関連分)
(平成30年3月統計委員会)

30

参考文献の紹介

第9回評価分科会

日時

令和3年3月26日（金）10:00～11:30

場所

総務省第2庁舎6階特別会議室

議事次第

（1）欠測値の補完に係る主な方法等について
（2）その他

配布資料

議事次第
資料　欠測値の補完に係る主な方法等について（素案）
参考1 欠測値補完に係る主な方法等参考資料
参考2 第8回評価分科会議事概要

31

参考文献①

平成 24 年度内閣府大臣官房統計委員会担当室請負調査

統計データの補完推計に関する調査

報告書

平成 25 年 3 月

株式会社 野村総合研究所

第1章　調査研究と結果の概要

1.　調査研究の目的

企業の情報管理意識や個人の情報保護意識の高まり等により、近年の統計調査を取り巻く環境が厳しさを増している中で、統計データの補完推計は、統計精度の維持・向上を図る上で重要となっている。

また、東日本大震災などの大規模災害の際に、特定の地域で調査の実施が困難になるという事態も発生しており、「平成 22 年度統計法施行状況に関する審議結果報告書」（※）においても、全国を対象とする基幹統計調査で調査対象地域の一部除外等の取扱をした場合、被災地の状況を踏まえて可能な限り補完的、補足的な調査、推計等の措置を講ずる必要があるとされている。

そこで本調査研究では、今後の統計委員会における審議等に資することを目的として、欠測値の補完方法の種類、国内における補完推計の研究事例、海外における欠測値補完の適用事例等について、文献調査及びヒアリング調査を通じて整理するとともに、各種の統計調査における欠測値補完方法の適用の方向性について考察を行った。

（※）「平成 22 年度統計法施行状況に関する審議結果報告書」（平成 23 年 9 月 22 日 統計委員会）
http://www5.cao.go.jp/statistics/report/report.html#11

本調査研究において調査対象とした調査（国内）及び調査機関（海外）

＜国内＞		
	総務省	経済センサス－活動調査
	厚生労働省	国民生活基礎調査
	経済産業省	特定サービス産業実態調査
	日本銀行	全国企業短期経済観測調査
	統計数理研究所	日本人の国民性調査
＜海外＞		
	アメリカ	米国統計局（US Census Bureau） 米国労働統計局（BLS: US Bureau of Labor Statistics） ミシガン大学（University of Michigan） westat 社
	カナダ	カナダ統計局（Statistics Canada）

アメリカにおいては、統計調査を受託する調査会社である westat 社に加えて、アカデミック分野における取組や見解を整理するためにミシガン大学を調査対象とした。

2)　有識者による研究会の設置

本調査研究では、文献調査、国内における研究事例、海外事例調査等の情報を基に、有識者の知識・視点等を反映するため、学識経験者を中心とする 4 名から構成される研究会を設置し、適宜、助言を受けながら研究を進めた。

「統計データの補完推計に関する調査」
研究会委員名簿

座長	岩崎　学	成蹊大学理工学部教授
	川崎　茂	日本大学経済学部教授
	保田　時男	関西大学社会学部准教授
講師参加		
	鳩貝　淳一郎	日本銀行調査統計局経済統計課企業統計グループ企画役補佐

なお、海外現地調査には、吉森雅代　大阪大学大学院基礎工学研究科博士課程 3 年/学術振興会特別研究員（DC2）が同行することで、学術面からのアドバイス・支援を受けた。

32

参考文献①

目 次

－ i －

－ ii －

33

参考文献②

欠測値補完に関する調査研究報告書

平成 29 年３月

内閣府経済社会総合研究所
景　気　統　計　部

本報告書は、平成 28 年９月～11 月にかけて内閣府経済社会総合研究所景気統計部において開催した「欠測値補完に関する調査研究」研究会における議論に基づき、作成したものである。
本報告書の作成に当たっては、「欠測値補完に関する調査研究」研究会座長の星野 崇宏 慶應義塾大学経済学部・大学院経済学研究科教授、同研究会委員の土屋 隆裕 情報・システム研究機構統計数理研究所データ科学研究系教授及び元山 斉 青山学院大学経済学部准教授から貴重な御意見やコメントを頂いた。

目次

34

参考文献③

リサーチペーパー第 48 号　　Research Paper No.48

統計調査の欠測値補完方法に関する研究動向について（主に米国とオランダ）

Current Trends of Research in Imputation Methods for Statistical Surveys
in Foreign Countries (the United States and the Netherlands)

坂下　信之
統計研究研修所統計研修研究官

SAKASHITA Nobuyuki
SRTI Senior Researcher for Statistical Training

令和 2 年 9 月
September 2020

総務省統計研究研修所
Statistical Research and Training Institute (SRTI)
Ministry of Internal Affairs and Communications

統計調査の欠測値補完方法に関する研究動向について（主に米国とオランダ）

坂下　信之

概要

政府統計の精度維持・向上が喫緊の課題となる中で、欠測値や外れ値への対応はその重要な要素である。世界的にも 1980 年代半ばから今日でも参照される文献が現れ、今世紀に入ってからは、国連などの場で盛んに議論されるようになっている。

本リサーチペーパーでは、各国で継続して研究されてきたテーマに焦点を当て、テーマを特定した継続的な研究の見られるアメリカ合衆国及びオランダについて重点的に論文を収集した。また、一般用ミクロデータを用いて、諸外国では頻繁に用いられているが我が国では適用例の少ないホット・デック法の数値シミュレーションを試みた。

その結果、少なくとも今回対象にした２カ国については、インピュテーションシステム、ベイズモデル、制約条件下のインピュテーションなどの継続的な課題を設定し、豊富な過去の蓄積の上で新たな検討を行っていることが分かった。また、我が国の統計データにおける欠測値への対処にホット・デック法を適用することは可能であると考えられるが、ドナーの選び方などの具体的な方法は、適用する調査に応じて子細かつ実務的に検討する必要がある。

キーワード：データ・エディティング、欠測値補完、インピュテーション、ホット・デック法

35

参考文献④

Working Paper Series

「全国企業短期経済観測調査」における
欠測値補完の検討

宇都宮 浄人*・園田 桂子**

Working Paper 01-11

日本銀行調査統計局

〒100-8630　東京中央郵便局私書箱 203 号
＊　e-mail:kiyohito@ier.hit-u.ac.jp
＊＊ e-mail:katsurako.sonoda@boj.or.jp

日本銀行ワーキングペーパーシリーズ

ビジネスサーベイにおける欠測値補完の検討
——全国企業短期経済観測調査（短観）のケース——

平川貴大*
takahiro.hirakawa@boj.or.jp

鳩目淳一郎*
junichirou.hatogai@boj.or.jp

No.12-J-8
2012 年 8 月

日本銀行
〒103-8660　郵便事業（株）日本橋支店私書箱第 30 号

＊ 調査統計局

36

参考文献⑤

統計研究彙報　第80号　2023年３月　(137-162)

諸外国の国勢調査におけるインピュテーション方法

北原　昌嗣†
寺垣内　雅子†

The Imputation Method on the Population Census in Foreign Countries

KITAHARA Masatsugu
TERAGAUCHI Masako

国勢調査のような大規模調査データに欠測値や項目間における矛盾が生じた場合、それら全てのデータに対して調査回答者に確認を取ることは、現実的には極めて困難であるため、その対処方法として、むもらしい整合性のあるデータを当該項目に代入して訂正（補正）をする、いわゆるインピュテーションを実行することで、各国統計局は結果を公表している。現在、このインピュテーション方法には、いくつかの種類があり、各国統計局が自国のデータに適合したインピュテーションを実施しているところである。多くの国では、同じ調査データの中から、欠測（又は矛盾）した箇所（値）に極めて類似していると推定される者を検索し、それを欠測箇所へ代入するホットデック法が利用されているが、ホットデック法も細分化すると、様々なやり方、手順が存在し、各国統計局はより良いインピュテーションを実行できるように創意工夫をしている。公表日が定まっている統計調査では、インピュテーション処理に多くの時間を費やすことはできず、手順を変更するだけでも処理速度の向上が見込めるインピュテーション処理には、試行錯誤をしていく価値が十分にあると考えられる。

本稿では、国勢調査で使われるインピュテーション方法について、国連統計部が発行しているガイドラインにおける記述を紹介した上で、各国統計局へのヒアリング調査結果を比較し考察する。

キーワード：国勢調査、補定処理、国際比較

In case of clear inconsistencies, or when the item is blank in the large-scale survey data such as the Population Census, it is practically extremely difficult to confirm to respondents on all such data. As the solution, the National Statistical Organizations (NSOs) publish results corrected by substituting similar data into their items, so-called "Imputation." Presently, there are several types of this imputation method, and NSOs use imputation method that is appropriate for own country's data. Many countries use the Hot-Deck Technique, which looks for value that is estimated to be similar to the missing (or inconsistent) value and substitutes it into the missing value. Various methods and procedures exist for the Hot- Deck Technique, and the NSOs are devising it so that better imputation can be implemented. It is not possible to spend much time on imputation work for statistical surveys with a fixed publication date. Therefore, the imputation processing, which can be expected to improve processing speed simply by changing the procedure, is thought to be well worth studying.

This paper introduces the description of imputation methods used in the Population Census from the guidelinesissued by the United Nations Statistics Division, and then compares and considers the results of interviews with NSOs.

Key words: Population Census, Imputation, International Comparison

† 総務省統計局統計調査部国勢統計課

参考文献⑤

「欠測値補完に関する調査研究報告書」手順書

平成29年3月

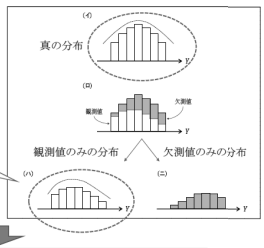
内閣府経済社会総合研究所
景気統計部

欠測データに伴う問題

- 統計調査において、無回答や無記入により調査客体又は調査項目の一部の情報が得られない場合、**欠測**（本来観測されるべきだが観測されない値）が生じる
- 欠測を含むデータについて、観測された値のみを用いた推定を行うと、以下の問題が生じる

①**欠測バイアスの可能性**：
欠測によって標本が母集団の縮図としての性格を失うことで推定に生じるバイアス

②推計精度の低下：
失われた情報の分だけ推計精度が低下



真の分布と比べて偏り
⇒母集団の縮図でなくなることでバイアスが生じる

特に欠測バイアスを軽減するような統計的処理が必要

1

欠測データの処理手順（全体像）

Step1：統計調査の推定目標を確認
↓
Step2：欠測の発生状況を確認
↓
Step3：欠測データメカニズム、補助変数の利用可能性を検討
↓
Step4：適切な欠測データ処理方法の候補を検討
↓
Step5：適切な処理方法を選択

【参考】
Step5 上級編：シミュレーション実施により適切な処理方法を選択
Step6：主な単一代入法の実施手順

2

欠測データの処理手順（Step1,2）

Step1：統計調査の推定目標を確認

- 推定対象が平均・総計等の推定か、又は分散・推定値の標準誤差等の推定か

※公的統計の大部分は平均・総計等の推定にとどまる。

Step2：欠測の発生状況を確認

- 調査客体単位、調査項目単位の欠測率はどの程度か

※欠測率が十分低い場合、異なる処理方法を用いた際のパフォーマンスの差異が小さいため、複雑な処理方法を用いる必要がない。（ただし、欠測のある変数が裾野の広い分布を持つ場合には留意が必要（裾野の広さが変数加工の重要性に影響）。）

- 時系列でみた場合、調査客体ごとの欠測パターンに特徴はないか

＜例1＞過去20年の客体企業ごとの欠測状況（回答状況）を見てみる

○観測　×欠測

調査年	1997	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	欠測回数	欠測率
客体企業A	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0	0%
客体企業B	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0	0%
客体企業C	○	○	○	×	○	○	○	○	×	○	○	○	○	×	○	○	○	○	×	○	4	20%
客体企業D	○	○	×	○	○	×	×	○	○	×	○	○	○	○	○	○	○	○	○	○	4	20%
客体企業E	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	×	×	5	25%

過去に欠測となった調査客体が再び欠測しやすい傾向（C,D,E）
何らかの理由（業績悪化等）により、一度欠測すると欠測が継続する可能性（E）

3

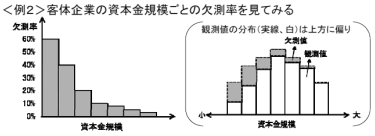
参考文献⑤

欠測データの処理手順（Step2）

● 欠測となりやすい調査客体に特徴はあるか（調査客体の企業規模、売上高、所得水準、資産保有額、就業状態等が欠測しやすさに影響していないか）

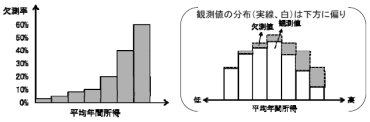
⇒ 欠測と相関の強い変数、欠測しやすさを説明する説明変数になり得る変数である「**補助変数**」が観測されていないか

＜例2＞客体企業の資本金規模ごとの欠測率を見てみる



中小企業ほど欠測しやすい傾向：資本金規模が補助変数の候補となり得る

＜例3＞調査対象世帯の年間所得水準ごとの欠測率を見てみる



高所得者ほど欠測しやすい傾向：平均年間所得が補助変数の候補となり得る

欠測データの処理手順（Step3）

Step3：欠測デ―タメカニズム、補助変数の利用可能性を検討

● 処理方法の適性を決める条件：
欠測データメカニズム、統計調査の推定目標、欠測率・欠測パターン　等

● 欠測データメカニズム（＝欠測が生じるしくみ）の種類：

種類	定義	例
①完全にランダムな欠測（MCAR）	変数の欠測確率が、当該変数及び他の観測されている変数の値に依存しない	コインの表裏によって、調査に協力するかどうか決める場合 ⇒**欠測バイアスは生じない**
②ランダムな欠測（MAR）	変数の欠測確率が、当該変数の観測値及び他の観測されている変数の値には依存するが、当該変数の欠測となった値には依存しない	回答者の大半が学生や無職者、無回答者の大半が就業者である調査で、金融資産保有額の欠測確率が就業状態の値に依存する場合 ⇒母集団の金融資産保有額の推定には、学生・無職者側への**下方バイアスあり**
③ランダムでない欠測（MNAR）	変数の欠測確率が、その変数自体の値に依存する	金融資産保有額平均を推定する調査で、上位資産階級ほど当該情報を秘匿する傾向が強い場合 ⇒標本が**低中位資産階級に偏る**

②→「補助変数」を利用し欠測バイアスの緩和が可能 ※6頁参照

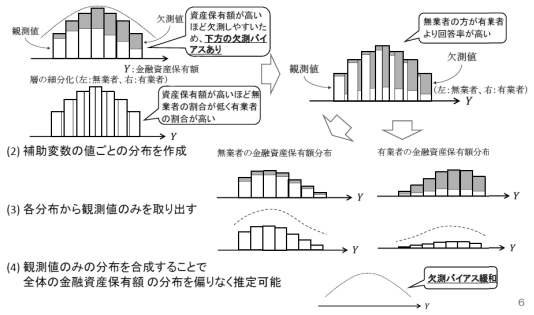
③→観測情報では欠測バイアスの緩和が不可能：モデル化が必要

欠測データの処理手順（Step3）

ランダムな欠測（**MAR**）の下での欠測バイアスの緩和の例

＜例えば、金融資産保有額の欠測確率が就業状態の値に依存する場合＞

(1) 各資産階級ごとに観測値と欠測値、補助変数の値（無業者か有業者か）に応じて標本を分割



(2) 補助変数の値ごとの分布を作成

(3) 各分布から観測値のみを取り出す

(4) 観測値のみの分布を合成することで全体の金融資産保有額 の分布を偏りなく推定可能

欠測データの処理手順（Step4）

Step4：適切な欠測データ処理方法の候補を検討

●欠測データメカニズムと（欠測バイアス緩和のための）適切な処理方法

①完全にランダムな欠測（MCAR） ―― **観測値のみ利用**（欠測値の処理を行わなくとも欠測バイアスは生じない）

②ランダムな欠測（MAR）
- 平均・総計等の推定
  - (1) **各種単一代入法**：欠測値を規則に基づいて決められた値に置換え
    - a. 補助変数（資本金規模、就業状態、資産保有高等）が利用可能な場合
      - ・**層化平均値代入**：層ごとの観測値の平均値を代入
      - ・**回帰代入**：欠測が生じた変数を被説明変数、補助変数を説明変数とする回帰モデルを推定し、得られた理論値を代入
      - ・**確率的回帰代入**：回帰代入の代入値に誤差項を加えた値を代入
      - ・**マッチング代入**：似た者同士を対応付け、似た者の観測値を代入
    - b. 当期と前期の観測値の間に高い正の相関がある場合
      - ・**横置き代入（LOCF）**：欠測が生じた標本の直近の観測値を代入
  - (2) **ウエイト調整法**：回答標本におけるウエイト（各標本が母集団の要素何単位分を代表しているか）を調整することで回答標本の偏りを補正
- 分散・推定値の標準誤差等の推定
  - (1) **多重代入法**：確率的回帰代入の考え方に基づき、疑似的な完全データ（欠測を含まないデータ）を複数作成。単一代入法と異なり、欠測値の背後にあるデータ生成過程に関する不確実性に対応した方法。
  - (2) **IPW法**：ウエイト調整法の一つ。「母集団の各要素が標本に含まれ、かつ回答する確率」の逆数を調査客体ごとのウエイトとする

③ランダムでない欠測（MNAR）
- MCAR,MARと異なり、完全データのデータ生成過程のみならず、欠測データメカニズムをモデル化した上で推定を行う必要（**尤度法**）
- 欠測データメカニズムに関し、あらゆる想定可能な前提条件に対して分析を実行し、結果を比較すること（**感度分析**）が望ましい

参考文献⑤

欠測データの処理手順（Step5）

Step5：適切な処理方法を選択

（1）対象となる統計調査の欠測データに対し、Step4で候補となった各処理方法及び現行方法を用いて処理を実施

（2）各方法を用いた場合の処理後のデータを比較し、現行方法の妥当性を検証
- ① 各方法間に大きな違いがない場合：
  現行方法を選択して問題ないとみられる
- ② 特異な結果を出す少数の方法と、同様な結果を出す多数の方法に分かれ、現行方法が後者（多数派）に含まれる場合：
  現行方法に問題があるという強い推論は得られない
- ③ 特異な結果を出す少数の方法と、同様な結果を出す多数の方法に分かれ、現行方法が前者（少数派）に含まれる場合：
  現行方法より**他の方法を選択した方がよい可能性**

※ケース②・③の場合、一部の方法で特異な結果を出す原因について、個票レベルで チェックを行う

【参考】 主な単一代入法の実施手順（Step6）

●ランダムな欠測（MAR）であり、平均・総計等の推定の場合に有効な欠測データ処理方法の例として、**層化平均値代入法**、**回帰代入法**、**傾向スコアマッチング代入法**及び**横置き代入法（LOCF）**の実施手順を紹介

※いずれの方法も適切な補助変数の利用により欠測バイアスを緩和

●次頁以降の具体的な数値例については、以下のケースを想定
- ・個人20人を調査客体とした統計調査
- ・「今月末の対前月末体重変化分（kg）」（変数yとする）の調査項目に欠測あり
- ・変数yと相関の高い変数（補助変数）として2つの変数が利用可能
  - 変数x_1：前月末の対前々月末体重変化分（kg）
    ⇒欠測なし、変数yとの間に正の相関
  - 変数x_2：今月の対前月1日当たり運動量変化分（時間/日）
    ⇒欠測なし、変数yとの間に負の相関

【参考】
Step5 上級編：シミュレーション実施により適切な処理方法を選択

● Step4で候補となった処理方法及び現行方法についてシミュレーションを実施

（1）対象となる統計調査の観測データに対し、2種類の欠測データメカニズム（ランダムな欠測（**MAR**）、ランダムでない欠測（**MNAR**））を仮定し(※)、一定の確率で機械的に欠測を生じさせる

（※）観測可能な情報からはMAR、MNARのどちらが成立しているか見分けがつかないため

（2）欠測を生じさせたデータに対し、
- ・ 複数の補助変数の組合せ(※)
- ・ 複数の変数の加工方法（水準、差分、対数等）

を設定し、各処理方法及び現行方法で欠測データ処理を実施

（※）「本年の所得」項目に欠測があり、調査対象者の「就業状態」及び「前年の所得」が補助変数として利用可能な場合：①「就業状態」、②「前年の所得」、③「就業状態」及び「前年の所得」の3種類の組合せを用いる

（3）各処理方法及び現行方法についてRRMSE(※)で評価
現行方法より優れた方法があれば当該方法の選択を検討

(※) $\mathrm{RRMSE} = \sqrt{E[(\text{推定値} - \text{真値})^2]}/\text{真値}$

【参考】 主な単一代入法の実施手順（Step6）

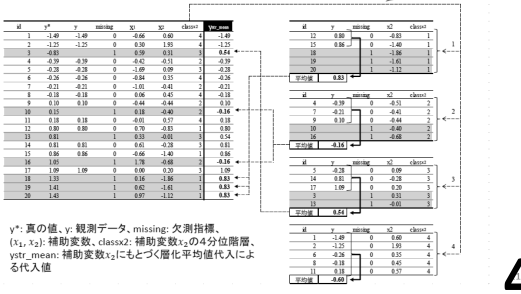
● 層化平均値代入法

（1）標本を補助変数（x_1）の値を用いて層化（グループ分け）

（2）層（グループ）ごとに目標変数（y）の観測値の平均値を算出

（3）欠測に対し、層ごとの観測値の平均値を代入

x_2の値の4分位により標本を4分割（層化）
※x_1のみ、x_1とx_2の両方を用いて分割してもよい



y*：真の値、y：観測データ、missing：欠測指標、
(x_1, x_2)：補助変数、classx2：補助変数x_2の4分位階層、
ystr_mean：補助変数x_2にもとづく層化平均値代入による代入値

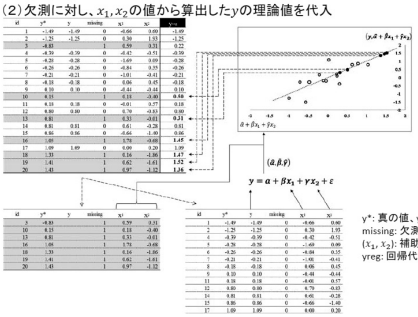
参考文献⑤

【参考】 主な単一代入法の実施手順（Step6）

● 回帰代入法

（1）目標変数（y）を被説明変数、補助変数（x_1, x_2）を説明変数とする回帰分析を実施

（2）欠測に対し、x_1, x_2の値から算出したyの理論値を代入



id	y*	y	missing	x1	x2	yreg
1	-1.49	-1.49	0	-0.66	0.60	-1.49
2	-1.25	-1.25	0	0.30	1.93	-1.25
3	-0.83		1	0.59	0.31	0.22
4	-0.39	-0.39	0	-0.42	-0.51	-0.39
5	-0.28	-0.28	0	-1.69	0.09	-0.28
6	-0.26	-0.26	0	-0.84	0.35	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	-0.21
8	-0.18	-0.18	0	0.06	0.45	-0.18
9	0.10	0.10	0	-0.44	-0.44	0.10
10	0.15		1	0.18	-0.40	0.18
11	0.18	0.18	0	-0.01	0.57	0.18
12	0.80	0.80	0	0.70	-0.83	0.18
13	0.81		1	0.33	-0.01	0.31
14	0.81	0.81	0	0.61	-0.28	0.81
15	0.86	0.86	0	-0.66	-1.40	0.86
16	1.05		1	1.78	-0.68	1.45
17	1.09	1.09	0	0.00	0.20	1.09
18	1.33		1	0.16	-1.86	1.47
19	1.41		1	0.62	-1.61	1.52
20	1.43		1	0.97	-1.12	1.26

id	y*	y	missing	x1	x2
3	-0.83		1	0.59	0.31
10	0.15		1	0.18	-0.40
13	0.81		1	0.33	-0.01
16	1.05		1	1.78	-0.68
18	1.33		1	0.16	-1.86
19	1.41		1	0.62	-1.61
20	1.43		1	0.97	-1.12

id	y*	y	missing	x1	x2
1	-1.49	-1.49	0	-0.66	0.60
2	-1.25	-1.25	0	0.30	1.93
4	-0.39	-0.39	0	-0.42	-0.51
5	-0.28	-0.28	0	-1.69	0.09
6	-0.26	-0.26	0	-0.84	0.35
7	-0.21	-0.21	0	-1.01	-0.41
8	-0.18	-0.18	0	0.06	0.45
9	0.10	0.10	0	-0.44	-0.44
11	0.18	0.18	0	-0.01	0.57
12	0.80	0.80	0	0.70	-0.83
14	0.81	0.81	0	0.61	-0.28
15	0.86	0.86	0	-0.66	-1.40
17	1.09	1.09	0	0.00	0.20

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、yreg: 回帰代入による代入値

【参考】 主な単一代入法の実施手順（Step6）

● 傾向スコアマッチング代入法（続き）

id	y*	y	missing	x1	x2	pscore	idpsmatch	distance	ypsmatch
1	-1.49	-1.49	0	-0.66	0.60	0.99			-1.49
2	-1.25	-1.25	0	0.30	1.93	0.94			-1.25
3	-0.83		1	0.59	0.31	0.52	14	0.18	0.81
4	-0.39	-0.39	0	-0.42	-0.51	0.91			-0.39
5	-0.28	-0.28	0	-1.69	0.09	1.00			-0.28
6	-0.26	-0.26	0	-0.84	0.35	0.99			-0.26
7	-0.21	-0.21	0	-1.01	-0.41	0.99			-0.21
8	-0.18	-0.18	0	0.06	0.45	0.97			-0.18
9	0.10	0.10	0	-0.44	-0.44	0.92			0.10
10	0.15		1	0.18	-0.40	0.64	17	0.22	1.09
11	0.18	0.18	0	-0.01	0.57	0.90			0.18
12	0.80	0.80	0	0.70	-0.83	0.18			0.80
13	0.81		1	0.33	-0.01	0.63	17	0.22	1.09
14	0.81	0.81	0	0.61	-0.28	0.35			0.81
15	0.86	0.86	0	-0.66	-1.40	0.88			0.86
16	1.05		1	1.78	-0.68	0.01	12	0.17	0.80
17	1.09	1.09	0	0.00	0.20	0.86			1.09
18	1.33		1	0.16	-1.86	0.27	14	0.08	0.81
19	1.41		1	0.62	-1.61	0.10	12	0.07	0.80
20	1.43		1	0.97	-1.12	0.06	12	0.11	0.80

id	y	missing	pscore
14	0.81	0	0.35

id	y	missing	pscore
17	1.09	0	0.86

id	y	missing	pscore
12	0.80	0	0.18

id	y	missing	pscore
14	0.81	0	0.35
12	0.80	0	0.18
12	0.80	0	0.18

$$pscore_i = \frac{exp(\hat{\alpha} + \hat{\beta}x_{1i} + \hat{\gamma}x_{2i})}{1 + exp(\hat{\alpha} + \hat{\beta}x_{1i} + \hat{\gamma}x_{2i})}$$

$$(missing_i, x_{1i}, x_{2i}) \qquad (\hat{\alpha}, \hat{\beta}, \hat{\gamma})$$

$$\max_{(\alpha,\beta,\gamma)} \sum_{i=1}^{20} \left(missing_i \ln \frac{1}{1 + exp(\alpha + \beta x_{1i} + \gamma x_{2i})} + (1 - missing_i) \ln \frac{exp(\alpha + \beta x_{1i} + \gamma x_{2i})}{1 + exp(\alpha + \beta x_{1i} + \gamma x_{2i})} \right)$$

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、pscore: 傾向スコア、idpsmatch: 傾向スコアマッチングによって合された相手レコードのid、distance: マッチングの相手との間の距離（傾向スコアの差）、ypsmatch: 傾向スコアマッチング代入による代入値（数式はロジットモデル）

【参考】 主な単一代入法の実施手順（Step6）

● 傾向スコアマッチング代入法

（1）全標本を用いて「**傾向スコア**」（補助変数の値によって条件付けた観測確率）を推定、各標本の「傾向スコア」を得る

※全標本（20人分）のデータを用い、観測確率（観測＝1, 欠測＝0）を2つの補助変数（x_1, x_2）で説明する2項回帰モデルを推定

※傾向スコア 0.99は「補助変数（x_1, x_2）から推定した結果、約99%の確率で回答する」意味

（2）無回答者と回答者の間で傾向スコアの差の絶対値を総当たりで算出

（3）各無回答者に対し、（2）の値が最も小さい回答者の値を代入

【参考】 主な単一代入法の実施手順（Step6）

● 横置き代入法（LOCF） ※パネルデータが利用できる場合のみ適用可能

（1）同一標本について、当期の値（y）と前期の値（x_1）の間に正の相関があることを確認

（2）欠測に対し、同一標本の前期の値を代入

id	y*	y	missing	x1	x2	ylocf
1	-1.49	-1.49	0	-0.66	0.60	-1.49
2	-1.25	-1.25	0	0.30	1.93	-1.25
3	-0.83		1	0.59	0.31	0.59
4	-0.39	-0.39	0	-0.42	-0.51	-0.39
5	-0.28	-0.28	0	-1.69	0.09	-0.28
6	-0.26	-0.26	0	-0.84	0.35	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	-0.21
8	-0.18	-0.18	0	0.06	0.45	-0.18
9	0.10	0.10	0	-0.44	-0.44	0.10
10	0.15		1	0.18	-0.40	0.18
11	0.18	0.18	0	-0.01	0.57	0.18
12	0.80	0.80	0	0.70	-0.83	0.80
13	0.81		1	0.33	-0.01	0.33
14	0.81	0.81	0	0.61	-0.28	0.81
15	0.86	0.86	0	-0.66	-1.40	0.86
16	1.05		1	1.78	-0.68	1.78
17	1.09	1.09	0	0.00	0.20	1.09
18	1.33		1	0.16	-1.86	0.16
19	1.41		1	0.62	-1.61	0.62
20	1.43		1	0.97	-1.12	0.97

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、ylocf: LOCFによる代入値

欠測値補完のメカニズムを考慮した推定

【諸外国の国勢調査におけるインピュテーション方法一覧表】

	アメリカ	カナダ	フランス	イギリス	ドイツ	イタリア	オランダ	オーストラリア	ニュージーランド	中国
調査方法	伝統的センサス	伝統的センサス	ローリングセンサス	伝統的センサス	複合型センサス	複合型センサス	レジスターベースセンサス	伝統的センサス	伝統的センサス	伝統的センサス
調査周期	10年	５年	毎年	10年	10年	毎年	10年	５年	５年	10年
インピュテーション方法（システム）	ホットデック法（最近隣法）	CANCEIS	シーケンシャルホットデック法	CANCEIS	CANCEIS	DIESIS	必要に応じて適切なインピュテーション方法を採用	ホットデック法（最近隣法）	CANCEIS	必要に応じて適切なインピュテーション方法を採用
詳細内容	▸欠測の内容によって、インピュテーション処理を計数インピュテーション（人数）と特性インピュテーション（内容）に分けている。 ▸可能な限りのデータを収集した後、わずかな欠測値、無効データ、不整合データに対してエディットや特性インピュテーションを実施する。 ▸エディットでは無効データや不整合データを発見し、特性インピュテーションで欠測値をインピュテーションする。 ▸総人口数が確定した後にエディットや特性インピュテーションを実施するため、総人口数には影響しない。 ▸エディットルールを適用する際に最小変化量は意識していない。 ▸未回答住戸には調査員を派遣し、在宅の場合には、携帯電話を使って聞き取り調査を実施	▸CANCEISは大規模データのエディットインピュテーションを効率的に処理可能にしたシステム ▸CANCEISは、まずドナー候補を探索し、そのドナー候補に対しインピュテーションアクションを検討することで優れたデータ駆動を実行することができ計算上の大きな利得を得る。 <今後の展望> ▸国勢調査のため、引き続きCANCEISを利用していく。 ▸CANCEISは今後もエラーが見つかればその都度改善することで性能を向上させ、新機能も導入していくだろう。	▸住所変数でソートされたファイルから順に地理的により類似の特徴を持ったドナーを探索する。 <メリット> ▸未回答率が低い場合やデータファイルが住所変数によって正確にソートできる場合には効率的 <デメリット> ▸データファイルが正確にソートできない場合は、地理的にはるか遠くのドナーを選択する可能性があり、インピュテーションの質に影響を与える。 <今後の展望> インピュテーション方法の変更は、現状では計画にない。	<CANCEIS採用理由> ▸ライセンスは必要だが大規模データに特化して設計されたインピュテーションシステムを無料で利用可能 ▸CANCEISの機能及び基礎となるインピュテーション方法が、2001年以前の国勢調査においてイギリス国家統計局が設計・開発したシステムよりも優れていた。 ▸国連発行の国勢調査エディティングハンドブックでも従来のトップダウン法より優れた多変量エディットシステムとしてCANCEISが例示されている。 <今後の展望> ▸国勢調査でのCANCEISの利用は将来の調査計画や評価段階における意思決定過程の一部である。	<CANCEIS採用理由> ▸CANCEISを採用した一番の理由は、エディットとインピュテーションの同時実行が可能、処理速度も許容範囲で効率的なアルゴリズムであるため。 <今後の展望> ▸今後も様々なインピュテーション方法を利用していく。 ▸インピュテーション方法の選定は調査データに依存する。 ▸現状のインピュテーション方法やシステムを維持した上で、適切なツールが開発されれば、それを取り入れていくだろう。	▸DIESISでは、世帯レベルと個人レベルにおいて、質的変数と量的変数を同時に処理可能 ▸DIESISには、"First donors then fields"と"First fields then donors"の２つのアルゴリズムによる「データ駆動」と「（理論上の）最小変化量」の2つのエディット方法が実装されている。 ▸これら二つのアルゴリズムが、データ保持とインピュテーションアクションの妥当性とのバランスを取るために、併用されている。	▸現在では多くの変数が行政記録情報由来であるため、大規模調査をしていた25年前よりもインピュテーションの必要性がかなり低い。 ▸行政記録情報から情報をほとんど得ることが可能であるが、唯一の例外として、「学歴」に関する行政記録情報は未完成である。 <今後の展望> ▸「学歴」については、多項ロジスティック回帰モデルを使ってインピュテーションをする計画がある。	▸最近隣法を備えた独自のオラクルベースのインピュテーションシステムを利用している。これまでの国勢調査でも利用されおり、メソドロジーも経時的に改良されている（例えば、インピュテーション対象とドナーの適合基準など）。 ▸このインピュテーションシステムはオーストラリア統計局の他のデータ処理システムと統合されている。 <今後の展望> ▸2026年国勢調査のインピュテーション計画については、まだ何も決まっていない。	<CANCEIS採用理由> ▸CANCEISのインピュテーション処理はフェレギ-ホルト法よりもはるかにデータ駆動性が優れている。 ▸CANCEISはカナダ統計局による公式なサポートがあり、継続的な改善がある。 ▸汎用性とカスタマイズ性が優れている。 <今後の展望> ▸CANCEISを標準的なツールの一つとして今後も使用し続ける可能性が高いが、多重インピュテーションに関する研究もしているところである。	▸第７回国家人口センサス（2020年）においては、デジタル手法を用いてデータを収集 ▸データ収集プログラムにデータチェック機能を組み込んだことで、データ欠測やエラーの発生が大幅に削減された。 <今後の展望> ▸例えば、住宅面積の欠測値については、近隣の回答世帯の平均値をインピュテーションに利用するだろう。

（注）調査方法の種類

1．伝統的センサス：調査員調査のことであり、紙媒体及び/又はインターネット調査票を使用した実地調査に基づく調査方法である。登記情報や行政記録情報を活用することもあるが、あくまで補助的な使用であり、直接的に調査項目を把握するために使用するものではない。
2．レジスターベースセンサス：登記情報や行政記録情報のみに基づき、踏査を実施しない調査方法。ただし、国勢調査を目的としていない既存の調査結果を活用することは可能である。（例：労働力調査の結果を国勢調査に利用）
3．複合型センサス：登記情報や行政記録情報から直接的に調査項目を把握するほか、悉皆又はサンプルによる踏査を実施し、調査項目を把握する調査方法である。
4．ローリングセンサス：累積的な連続したサンプル調査のことであり、長期間に渡って全国全てを網羅する調査方法である。

北原、寺垣内（2023）諸外国の国勢調査におけるインピュテーション方法(統計研究研修所 統計研究彙報　第80号)
https://www.stat.go.jp/training/2kenkyu/ihou/80/pdf/2-2-808.pdf

43

欠測値補完の事例

平均値補完の事例

活用事例

統計調査名	府省	全数・標本調査の別	調査周期	活用事例
個人企業経済調査	総務省	標本調査	年次	期首・期末棚卸高について、層化平均値により補完
特定サービス産業実態調査	経済産業省	標本調査	年次	「主たる業務」の年間売上高の業務種類別割合、契約先産業別割合について、前回個票をもとに、規模別等のグループに分け（例：主業事業従事者数、売上高など）し、グループ毎の内訳項目の構成比を算出し、補完を行う個票の該当する規模別等のグループの構成比により補完 ※前回値が有る個票の場合は、前回値により算出。
経済産業省生産動態統計調査	経済産業省	一定規模以上全数調査	月次	生産金額の回答があり、数量が無回答の場合、双方の回答がある企業の平均単価を計算し、回答金額を当該平均単価で割り戻して補完

（統計委員会第５～第９回評価分科会資料から引用）

44

平均値補完の事例

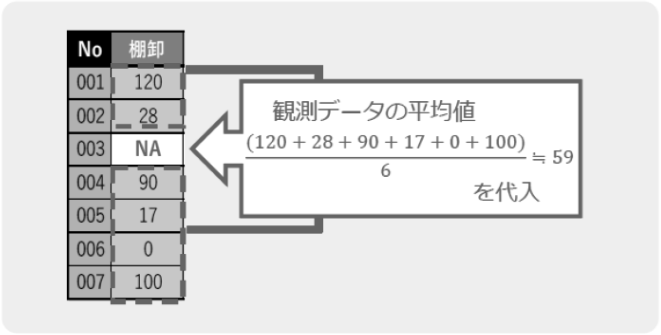
平均値補完の事例

個人企業経済調査における活用事例

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。



第6回評価分科会(令和2年2月19日)
資料５　令和元年度個人企業経済調査～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）より

（統計委員会第５～第９回評価分科会資料から引用）

欠測値補完の留意点

回帰代入の事例

（２）回帰代入（regression imputation）

計算式

回帰補完は、仮定された回帰モデルにより、全てのユニットについて観測されている項目（補助変数）x_1,・・・・x_qから得た予測値yを補完代入するもので、平均値補完や比率補完が一般化されたもの。多くの場合では線形回帰モデルが用いられる。

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon , \quad (3)$$

ここで、α, β_1・・・・・・・β_qは未知のパラメーター、εは誤差項である。すべてのユニットの誤差項は独立的に、平均が0で分散がσ^2である同じ正規分布に従うと仮定されている。

モデル式（３）は項目yの値と補助変数が双方とも観測されているレコードから通常は最小二乗法により推定される。これにより、項目yの予測値は補助変数から以下の通り与えられる。

$$\hat{y} = a + b_1 x_1 + \cdots + b_q x_q , \qquad (4)$$

ここで、a, b_1・・・・・・・b_qは最小二乗法により与えられたα, $\beta 1$・・・・・・・βqの推定値である。補助変数は全て観測されたものであり、項目yの予測値は、無回答の項目に対しても、回答者の項目に対しても与えられる。

Handbook on Methodology of Modern Business Statistics(2017;Eurostat)；Theme: Model-Based Imputation p.4,5

（統計委員会第５～第９回評価分科会資料から引用）

欠測値補完の留意点

回帰代入の事例

活用事例

統計調査名	全数・標本調査の別	調査周期	活用事例
サービス産業動向調査（2012年12月調査まで回帰代入による補完を適用）	標本調査	月次	売上高及び事業従事者数について、1か月目に限り、産業分類別に事業従事者数を説明変数とした対数回帰モデルにより推定した値により補完

サービス産業動向調査における活用事例（2012年12月調査まで回帰代入による補完を適用）

売上高及び事業従事者数についての補完
（1か月目）

産業分類別に事業従事者数を説明変数とした対数回帰モデルにより推定した値を用いる。

$\log(y_i) = \beta_0 + \beta_1 \times \log(x_i)$

y_i：当月の売上高（当月の事業従事者数）
x_i：母集団事業従事者数
ただし、回帰係数β_0，β_1は、回答があった事業所のうち、x_i又はy_iが0又はマイナスのものは除外して計算する。

（2か月目以降）
産業分類別の事業従事者規模別に前月からの変化率の平均値を算出し，前月の売上高（事業従事者数）を乗じて推定する。

平成24年サービス産業動向調査年報　「付録3 調査対象事業所の抽出方法，結果の推定方法及び推定値の標本誤差」より

（統計委員会第５～第９回評価分科会資料から引用）

47

欠測値補完の留意点

最近隣ホットデック法の事例

（４）最近隣ホットデック（nearest neighbor imputation）

計算式

最近隣ホットデックにおいては、補助変数は、補完対象の*i*番目のユニットとドナー候補である*k*番目のユニットの間の距離関数を定義するのに用いられる。*i*番目のユニットの最近隣ユニットは、距離関数が最小となる回答者ユニット*d*と定義される。式としては、

$$d = \arg\min_{k \in obs} D(i,k), \qquad (2)$$

ここで*obs*は項目*y*が観測されているユニットの集合を示している、すなわちドナー候補の集合である。

補完方法の説明に入る前に、上記の式（２）における距離関数の選択肢について簡単に説明しておく。補助変数（ x_1, ・・・・x_q ）は全て量的変数であると仮定すると、良く用いられるのは、以下の式により与えられる距離関数である。

$$D_z(i,k) = \left(\sum_{j=1}^{q} \left| x_{ji} - x_{jk} \right|^z \right)^{1/z} \qquad (3)$$

活用事例

統計調査名	府省	全数・標本調査の別	調査周期	活用事例
個人企業経済調査	総務省	標本調査	年次	仕入金額、経費計、経費計のうち給料賃金について、同一調査年の他のユニットの数値により最近隣ホットデック法により補完

（統計委員会第５～第９回評価分科会資料から引用）

48

欠測値補完の留意点

LOCFの事例

活用事例

統計調査名	府省	全数・標本調査の別	調査周期	活用事例
個人企業経済調査	総務省	標本調査	年次	売上金額について、同一ユニットの過去データを時点調整（回答を得られているユニットの過去からの変化率を乗じる）した数値により補完
商業動態統計調査	経産省	標本調査	月次	全部無回答者の、商品販売額、販売先別商品販売額、商品別手持額について、前月及び当月ともに回答のあった事業所の集計値合計の前月比伸び率を当月無回答者の前月回答値に乗じて補完

（統計委員会第５～第９回評価分科会資料から引用）

49

消費動向指数（CTI）における事例

- ビッグデータ等を活用し、消費動向をマクロ・ミクロの両面から捉える速報性の高い消費指標の体系：**消費動向指数（CTI：Consumption Trend Index）**を新たに開発
- **平成30年１月分**から参考指標として公表開始

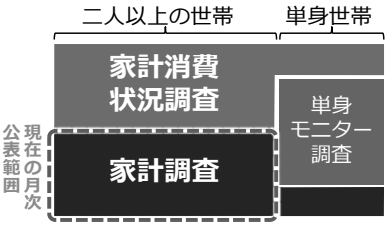
※データソースは、**当初は既存統計を利用**。研究分析・検証を経た後、**ビッグデータを順次活用**

世帯消費動向指数（ＣＴＩミクロ）

世帯の平均消費支出額（10大費目別、世帯類型別など）の月次動向を示す統計指標　※家計調査の上位モデルとなる消費指標

◆ 家計調査（標本規模：二人以上の世帯 約８千、単身世帯 約7百）の結果を、
- □ 家計消費単身モニター調査（標本規模：2千4百）
- □ 家計消費状況調査（標本規模：約３万）

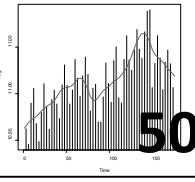
の結果等と**統計的手法によって補正・補強**し、**標本規模を擬似的に拡大、推計精度を向上**



総消費動向指数（ＣＴＩマクロ）

国内経済における個人消費総額（GDPにおける家計最終消費支出）の月次動向を示す統計指標

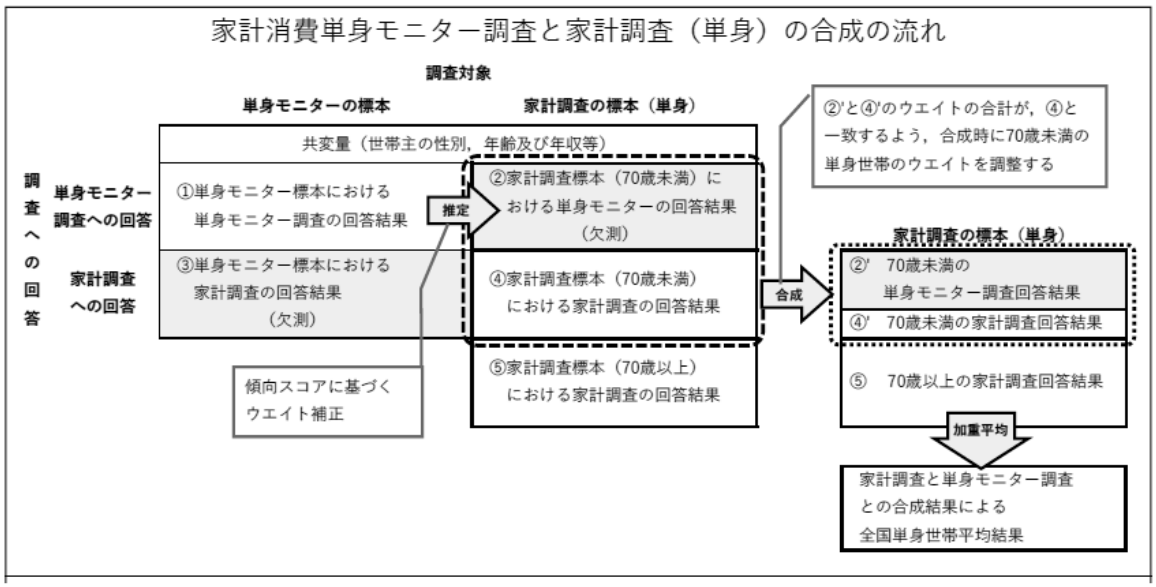
◆ GDP統計（家計最終消費支出）をターゲットとして、最新の動向を推測
◆ GDP統計の四半期別公表値では**観測できない月次の値を時系列回帰モデルによって推計**



（総務省統計局消費動向指数資料から引用）

50

消費動向指数（CTI）における事例



（総務省統計局消費動向指数資料から引用）

51

欠測値補完の留意点

- 統計調査ごとに欠測の発生状況や補完に利用できるデータなどに違いがあり、特定の補完方法の適用など一律の対応は困難
- このため、統計調査ごとの状況を踏まえた適切な対応が重要

（統計委員会第５～第９回評価分科会資料から引用）

52

欠測値補完の留意点

- 調査の内容・対象（世帯・企業、人数・売上など）によって適切な方法が異なる

- 時期・時点によって適切な方法が異なる
  ・以前はLOCF（横置き）が最適
  ⇒震災・金融危機等のショックにより、時期によっては平均値補完の方が適切になった事例など

53

欠測値補完の留意点

- 可能であれば、ミクロデータを用いて、人工的に欠測値を発生させ、各種の手法を適用して精度を評価する

- 補完した結果を表・グラフに表してみて、不自然な動き・傾向がないか（他の統計や過去の数値と比較して）十分に確認をすることが必要
  ⇒組織内・外部への説明に耐えられる内容か？

54