統 計 技 術 研 究 課

H31 個人企業経済調査の補完について

（経理項目及び従業者数関連項目補完案）

■　補完対象項目

図 1 は、H31 個人企業経済調査（以下「新調査」という。）における補完対象項目である。ここで、*印の［**06** 費用総額］は、新調査の調査項目にはなく、補完作業に使用する中間データとして作成する。

これらの補完項目のうち、ここでは、まず経理項目と従業者数関連項目の補完について取り上げ、設備投資関連項目及び専従者給与については資料を分ける。

図 1．補完対象項目一覧

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**

従業者数関連項目

04	**従業者数**
041	**事業主の家族で無給**
042	**常用雇用者**
043	**パート・アルバイト**
044	**臨時雇用者**

設備投資関連項目

19	**設備投資（新規設備取得）**
20	**うち車両 機械 工具 器具 備品**
201	**最も取得額の多い時期**
21	**設備投資（中古設備取得）**

専従者給与

32	**専従者給与**

これらの項目のうち、**04, 041～044, 05**, **06**, **11～14, 19** は、H28 経済センサス-活動調査において同一項目が存在する。また、項目間には、次のような制約条件が存在しており、補完値はこれらの制約を満たすことが望ましい。

- **04** = **041** + **042** + **043** + **044**
- **06** + **09** = **07** + **08** + **10**
- **10** > Σ(**11**～**18**)

■　作業の基本的な方針

- 欠測について、項目欠測・単位欠測ともに、前回調査時のデータがある項目については、同一企業のデータを、時点の違いを調整した上で補完値として使用する。ただし、従業者数等の人数データは時点の違いの調整は行わない。これは一般に、同時点の他企業データからの推定値よりも、時点が違っても同一企業データの値を使用するほうが精度は高いためである。
- 項目 **10** の内訳である **11～18** の欠測パターンは、一部に数値が入っていれば記入がない部分は 0 補完するため、補完処理では全欠測のみ考慮する。また、**10** が欠測で **11～18** が観測された場合は、**10** は欠測のままとしているため、**10** > Σ(**11**～**18**) という関係性を考慮し、補完処理の対象とする。
- 補完後の単位欠測データの乗率は 1 に修正し、観測データの乗率を調整する形での集計も、併せて行う。

■　H31 個人企業経済調査の補完に使用するデータ

a. H31 個人企業経済調査（H31 新調査）

b. 抽出用母集団名簿（H29 事業所母集団 DB から標本抽出用に作成したもので、以下「H29DB」という）

c. H28 経済センサス-活動調査（以下「H28 センサス」という）

■　補完値の時点の調整について

- 新調査は H31 から実施されるが、補完に使用するデータは、大部分が H28 年時点で、少数の H29 時点も混在する。データ量の多い H28 センサスデータと新調査データの間で、補完項目の時点間の比率を算出し、H29 データにはその比率の 2/3 を適用する。
- 比率の算出は、補完クラス・補完対象項目毎に行う。

■　補完クラスの設定について

補完クラスとは、補完処理を行うデータの単位である。補完クラスの設定は、任意の項目を使用して、補完対象項目について、クラス内の分散が小さく、クラス間の分散が大きくなるように設定することにより、推定精度の高い補完処理を行うことができる。

ホットデックを行う場合、欠測データと同じ補完クラス内でドナーレコードを決めることになる。一般に、調査データを細分化するほど精度は上がる傾向がある一方で、クラス内で補完対象項目に欠測がないドナー候補のレコード数が減るため、ある程度の大きさは必要である。

- 設定作業の時期について

集計スケジュールを勘案すれば、新調査データを用いたクラスの設定作業は現実的ではない。このため、事前に H28 センサスデータを用いて、補完クラスの設定と、各クラ

ス内で欠測のないデータの数が不足した際の統合先の優先順位を決める必要がある。

- 補完クラスの設定に使用する項目の候補(優先順位の高い順)
  [03 主な事業（⇒産業）]，[05 売上金額]，[04 従業者数]

原則として、産業・売上金額・都道府県別で設定された標本抽出の層をベースに検討するが、ただし必要なドナー候補のレコード数が確保できない恐れがあるため、地域別のクラス分けは行わない。

また、同様の理由で、産業分類は、結果表で使用する最も細かいレベルとなる表章中分類までとする。

■ 補完手法の選択について

提示された候補は、H28 センサス準拠のロバスト比率補完であったが、モデルに基づく補完を行う場合、調査実施前に、補完クラスの設定に加えて、使用するモデルが適用対象のデータに合っているかどうかを確認するモデル選択作業を行う必要がある。

H31 新調査の場合、資料 1 の参考 2「平成 28 年経済センサス（個人経営調査票）と平成 31 年個人企業経済調査（新調査票）の対応表」にあるように、新調査の補完対象項目を網羅するデータが存在せず、調査実施前に必要な作業を行うことができない。

このため、既知ではないものを含む補完対象項目のデータ分布や相関を保持し、さらに補完対象となる項目間の制約条件も満たすことのできるような補完方法として、比率ホットデック法を候補とする。この方法の概要は、以下のとおり。

✓ **比率ホットデック法**

負の値をとらない変数 $\boldsymbol{y}_1, \dots, \boldsymbol{y}_m$ について、次のような制約があるとする。

$$\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_m = \boldsymbol{y}_{tot}$$

i 番目のレコード（$y_{i,1}, \dots, y_{i,m}$）について、うち頭から 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、次のような手順でこれらを補完する。

① 欠測値の合計 $r_i = y_{i,tot} - y_{i,1} - \cdots - y_{i,t}$ を計算する
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、これを 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 $r_d = y_{d,j+1} + \cdots + y_{d,m}$ を計算する
④ 下式により、補完値 $\tilde{y}_{i.j}$ を計算し、欠測を補完する

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

出典: The Memobust Handbook, Imputation under Edit Constraints, 2.2.1 Ratio hot deck imputation https://ec.europa.eu/eurostat/cros/content/imputation_en

■　補完の手順

次の A～D のステップを順次行うことで、補完対象の経理項目及び従業者数関連項目の補完を行う。

A)　同一企業データによる補完

① 新調査の項目 **07～10** に欠測がなければ、**07+08-09+10** により **06** を作成する。
② 項目 **041～044** 及び項目 **05** が欠測ならば、H29DB から補完する。
項目 **041～044** の補完値が欠測の場合は、ステップ D において対処する。
③ 項目 **07～10** に欠測があり、①で **06** を作成できない場合、H28 センサスから **06** を補完する。
④ 項目 **07～09** に 1 つ欠測があり、**06** が補完済みで **10** が観測されていれば、**06－10＝07＋08－09** という関係式に基づいてその欠測を補完する。
⑤ 項目 **11～18** が欠測の場合、H28 センサスから **11～14** を補完する。ただし、**10** が観測され、補完値の **Σ(11～14)** が観測値 **10** よりも大きな値になるときは、**11～18** を欠測のままにする。

B)　他企業データを用いた主要経理項目の補完

他調査データにも **05** 以外は欠測が存在するため、ステップ A の補完終了時に、欠測がなくなるのは **05** のみで、このステップで対応すべき欠測の組み合わせは、以下の 26 通り存在する（表 1「ステップ B で扱う全ての欠測パターン」参照）。

- 項目 **07**～**09** が 1 つ欠測、なおかつ **06** が欠測　[a～c]
- 項目 **06** と **10** が両方欠測　[d]
- 項目 **07**～**09** のみに 2 つ以上の欠測がある　[e～g]
- 項目 **07**～**09** が 1 つ欠測、なおかつ **10** が欠測　[h～j]
- 項目 **07**～**09** が 2 つ以上欠測し、なおかつ **06** が欠測　[k～m, u]
- 項目 **07**～**09** が 1 つ欠測し、なおかつ **06** と **10** が両方欠測　[n～p]
- 項目 **07**～**09** がすべて欠測　[q]
- 項目 **07**～**09** が 2 つ以上欠測し、なおかつ **10** が欠測　[r～t]
- 項目 **07**～**09** が 2 つ以上欠測し、なおかつ **06** と **10** が両方欠測　[v～x]
- 項目 **07**～**09** がすべて欠測し、なおかつ **10** が欠測　[y]
- 項目 **06**～**10** がすべて欠測　[z]

さらに、項目 **10** には内訳 **11～18** が存在するので、内訳を含めてこれを **10+**と表記する。ステップ A の補完後に、**10+**の欠測パターンは、以下の 4 通りとなり、このうち項目 **10** が欠測する i と ii の場合についてこのステップ B で処理を行い、iii と iv については、次のステップ C において対処する。

i. **10～18** が全て欠測
ii. **10** が欠測し、**11～14** が補完値で、**15～18** が欠測
iii. **10** が観測され、**11～14** が補完値で、**15～18** が欠測
iv. **10** が観測され、**11～18** が欠測

具体的には、ステップ B で扱う項目 **10** の欠測は、上述の 26 通りのうち[h～s]の 12 通りが該当するが、それぞれの補完処理の際に、**10** の内訳の欠測パターンが ⅰ であれば、**10** の補完と同時に **10** の内訳も併せて補完する。一方で、欠測パターンが ⅱ の場合、**10** の補完時に、**10** の補完値が補完済みの内訳の合計 Σ(**11～14**) よりも大きくない場合は、ドナーが適切ではないので次点の候補をドナーに選び直し、**10**（と同時に補完すべき他の項目）の補完をやり直す。このときは **10** のみを補完して、ⅲ の欠測パターンのレコードとして、ステップ C で内訳の補完を行う。

このステップ B では、比率ホットデック法を使用することにより、同時に補完した範囲について **06+09=07+08+10** 及び **10** ＞ Σ(**11**～**18**)という制約を満たす補完値を作成する。

基本的な手順は、まず上位の制約条件である合計値 **06+09** あるいは **07+08+10** が得られる場合、その値に基づいてドナーと比率を決め、下位レベルの項目を補完する。欠測の存在により、これらの合計値のいずれも作成できない場合は、原則としてできるだけ数値の大きな項目でドナーと比率を定める。

項目 **07**～**09** に 2 つ以上の欠測、かつ **06** に欠測があるパターンについては、まず、**05** の値が最も近いものをドナーとして **06** 値を作成する必要がある。

表 1 ステップ B で扱う全ての欠測パターン

No	パターン	05	06	07	08	09	10
		売上金額	費用総額	期首棚卸高	仕入高	期末棚卸高	経費計
1	a	○	×	×	○	○	○
2	b	○	×	○	×	○	○
3	c	○	×	○	○	×	○
4	d	○	×	○	○	○	×
5	e	○	○	×	×	○	○
6	f	○	○	×	○	×	○
7	g	○	○	○	×	×	○
8	h	○	○	×	○	○	×
9	i	○	○	○	×	○	×
10	j	○	○	○	○	×	×
11	k	○	×	×	×	○	○
12	l	○	×	×	○	×	○
13	m	○	×	○	×	×	○
14	n	○	×	×	○	○	×
15	o	○	×	○	×	○	×
16	p	○	×	○	○	×	×
17	q	○	○	×	×	×	○
18	r	○	○	×	×	○	×
19	s	○	○	×	○	×	×
20	t	○	○	○	×	×	×
21	u	○	×	×	×	×	○
22	v	○	×	×	×	○	×
23	w	○	×	×	○	×	×
24	x	○	×	○	×	×	×
25	y	○	○	×	×	×	×
26	z	○	×	×	×	×	×

今後、H30 までの個人企業経済調査データを用いて全てのパターンについて試算を行い、複数の選択肢が存在する場合は候補を絞り、制約を満たせない可能性があるパターンを特定して対応策を講じる。

C)　経費内訳の補完

ここでは、iii. **10** が観測され、**11～14** が補完値で、**15～18** が欠測する場合と、iv. **10** が観測され、**11**～**18** が欠測する場合の対処を行う。

iii の場合は、**10－Σ(11**～**14)**でドナーと比率を決めて、**15**～**18** を比率ホットデックで補完する。

iv の場合は、**10** でドナーと比率を決めて、**11**～**18** を比率ホットデックで補完する。

D)　従業者数の補完

ここでは、ステップ A－②で従業者数の補完後、項目 **041**～**044** が全て欠測の場合の対処を行う。

補完クラス内の欠測のない新調査データから、[**04** 従業者合計]、[**05** 売上金額]及び[**11** 給与賃金]等の適切な項目によりドナーと比率を決めて、各従業者数 **041**～**044** に比率ホットデックで補完する。

調査項目として従業者数はないため、調査データでは各従業者数から従業者数が算出される。そのため、欠測のパターンはステップ A－②で補完された項目 **04** に値があり、項目 **041**～**044** が全て欠測の場合のみとする。