「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

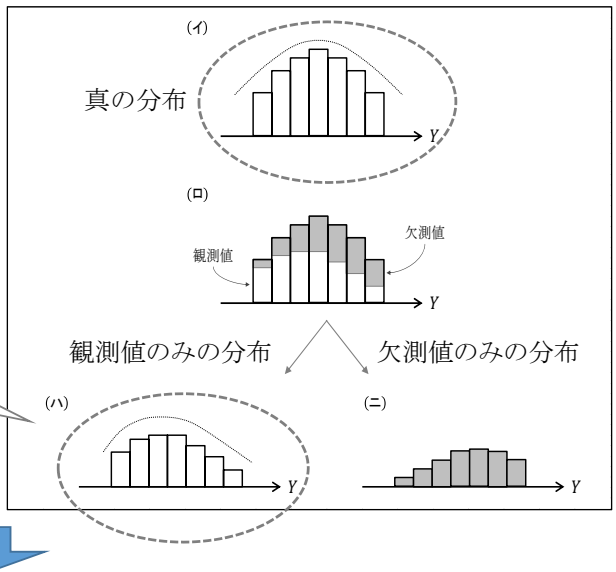
内閣府経済社会総合研究所
景気統計部

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

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

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）

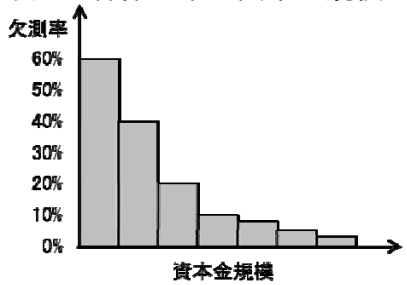
何らかの理由（業績悪化等）により、一度欠測すると欠測が継続する可能性（E）

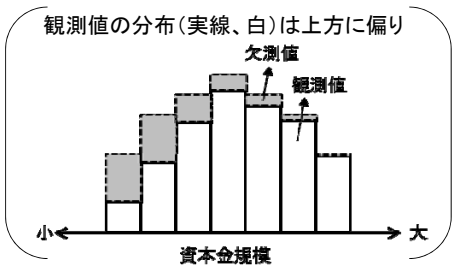
欠測データの処理手順（Step2）

● 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

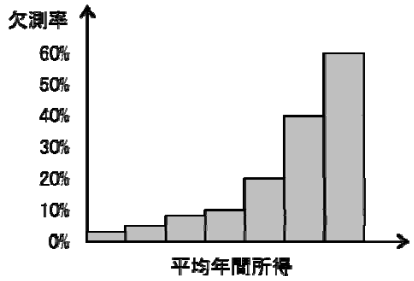
＜例2＞客体企業の資本金規模ごとの欠測率を見てみる

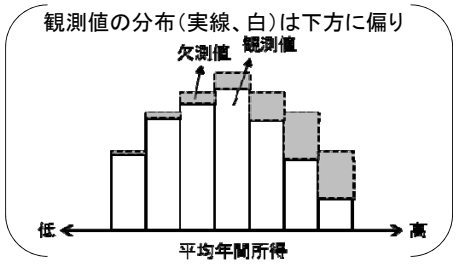




中小企業ほど欠測しやすい傾向：資本金規模が補助変数の候補となり得る

＜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：欠測データメカニズム、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

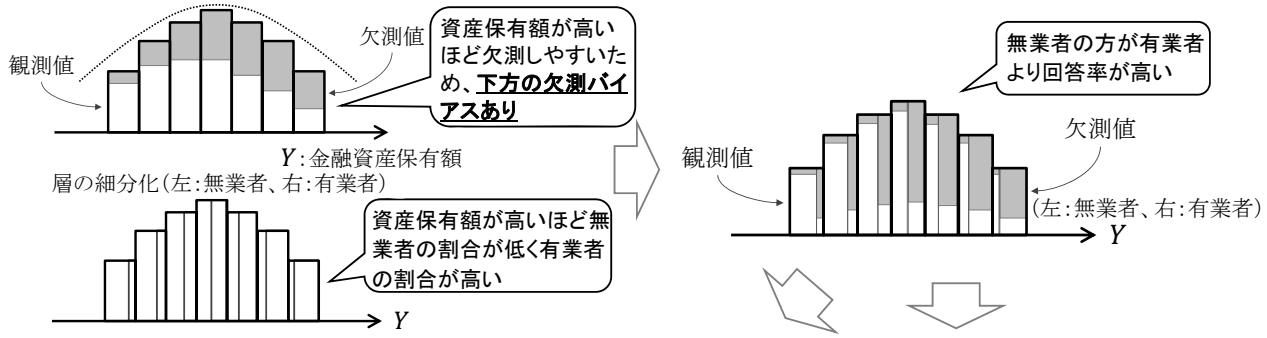
③：**観測情報では欠測バイアスの緩和が不可能：モデル化が必要**

欠測データの処理手順（Step3）

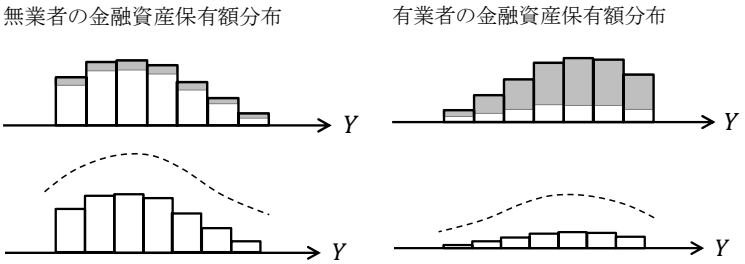
ランダムな欠測（MAR）の下での欠測バイアスの緩和の例

＜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

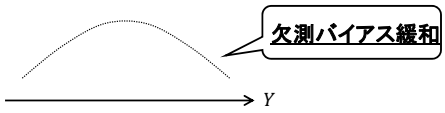


(2) 補助変数の値ごとの分布を作成



(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

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】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

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

【参考】 主な単一代入法の実施手順（Step6）

- 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】 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

id	y*	y	missing	x1	x2	classx2	ystr_mean
1	-1.49	-1.49	0	-0.66	0.60	4	-1.49
2	-1.25	-1.25	0	0.30	1.93	4	-1.25
3	-0.83		1	0.59	0.31	3	**0.54**
4	-0.39	-0.39	0	-0.42	-0.51	2	-0.39
5	-0.28	-0.28	0	-1.69	0.09	3	-0.28
6	-0.26	-0.26	0	-0.84	0.35	4	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	2	-0.21
8	-0.18	-0.18	0	0.06	0.45	4	-0.18
9	0.10	0.10	0	-0.44	-0.44	2	0.10
10	0.15		1	0.18	-0.40	2	**-0.16**
11	0.18	0.18	0	-0.01	0.57	4	0.18
12	0.80	0.80	0	0.70	-0.83	1	0.80
13	0.81		1	0.33	-0.01	3	0.54
14	0.81	0.81	0	0.61	-0.28	3	0.81
15	0.86	0.86	0	-0.66	-1.40	1	0.86
16	1.05		1	1.78	-0.68	2	**-0.16**
17	1.09	1.09	0	0.00	0.20	3	1.09
18	1.33		1	0.16	-1.86	1	**0.83**
19	1.41		1	0.62	-1.61	1	**0.83**
20	1.43		1	0.97	-1.12	1	**0.83**

id	y	missing	x2	classx2
12	0.80	0	-0.83	1
15	0.86	0	-1.40	1
18		1	-1.86	1
19		1	-1.61	1
20		1	-1.12	1
平均値	**0.83**			

1

id	y	missing	x2	classx2
4	-0.39	0	-0.51	2
7	-0.21	0	-0.41	2
9	0.10	0	-0.44	2
10		1	-0.40	2
16		1	-0.68	2
平均値	**-0.16**			

2

id	y	missing	x2	classx2
5	-0.28	0	0.09	3
14	0.81	0	-0.28	3
17	1.09	0	0.20	3
3		1	0.31	3
13		1	-0.01	3
平均値	**0.54**			

3

id	y	missing	x2	classx2
1	-1.49	0	0.60	4
2	-1.25	0	1.93	4
6	-0.26	0	0.35	4
8	-0.18	0	0.45	4
11	0.18	0	0.57	4
平均値	**-0.60**			

4

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

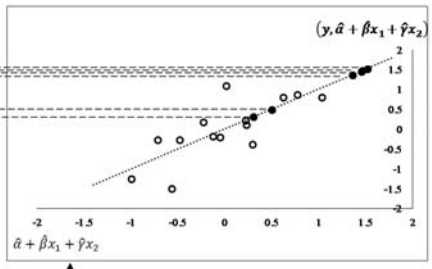
【参考】 主な単一代入法の実施手順（Step6）

● **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	**0.50**
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	**0.31**
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	**1.36**



$(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$

$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$

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**

（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

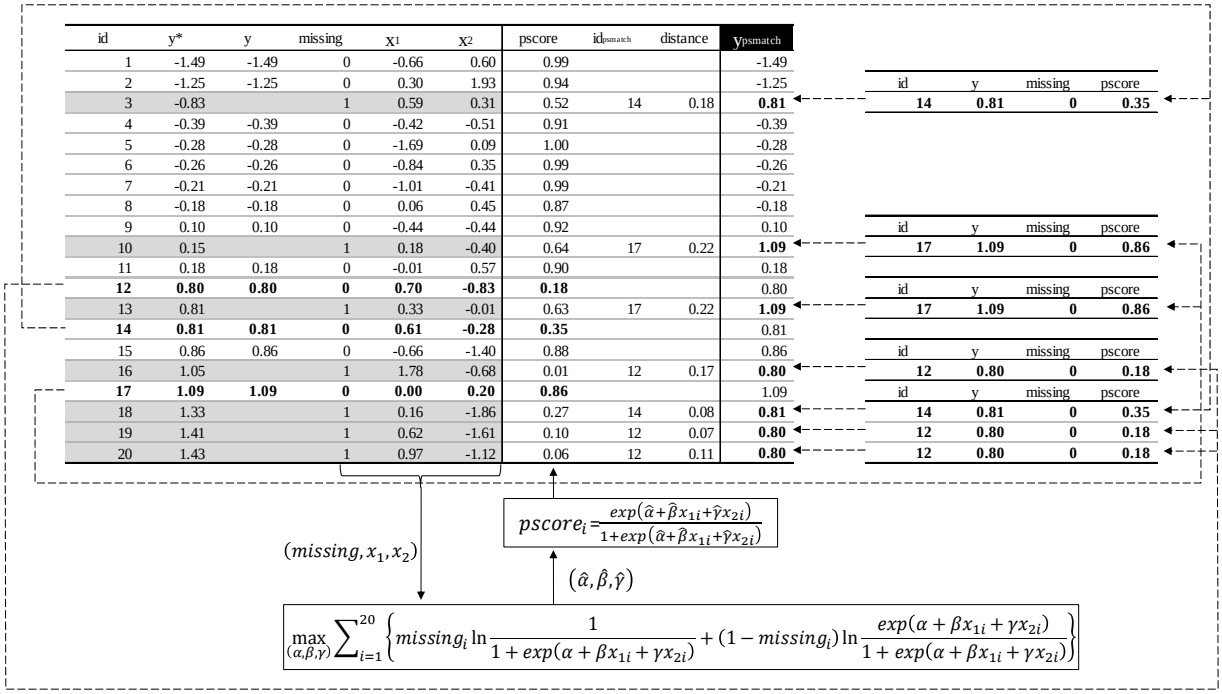
※傾向スコア 0. 99は「補助変数（x_1, x_2）から推定した結果、約99%の確率で回答する」意味

（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）

● **傾向スコアマッチング代入法（続き）**



id	y*	y	missing	x_1	x_2	pscore	$id_{psmatch}$	distance	$y_{psmatch}$
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	0.87			-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
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**

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

$(missing, x_1, x_2)$

$(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$

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

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）

● **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	x_1	x_2	y_{LOCF}
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

まとめ

- 適切な欠測データの処理方法は、欠測データメカニズム（＝欠測が生じるしくみ）の種類や統計調査の推定目標、欠測の発生状況等により異なる
- まずは欠測の発生状況に着目し、欠測しやすさを説明する説明変数になり得る変数（補助変数）が観察されているかを確認することが重要
- 欠測データメカニズムが**完全にランダムな欠測（MCAR）**の場合、**欠測値の処理を行わず、観測値のみを利用した分析**により欠測バイアスは生じない
- **ランダムな欠測（MAR）であり、かつ平均・総計等の推定**の場合、適切な補助変数を利用することで、**各種単一代入法（層化平均値代入、回帰代入、マッチング代入、横置き代入等）やウエイト調整法**により欠測バイアスの緩和が可能
  ※ただし、分散・推定値の標準誤差等の推定等の場合には、より高度な手法（多重代入法 やIPW法）を用いる必要
- **ランダムでない欠測（MNAR）**の場合、**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**が望ましい