2. 欠測データの統計的処理

上述のとおり、欠測データが与えられたとき、どのような手法が適切かを決める諸条件のうち、最も重要なものは「欠測データメカニズム」である。第 1.1.1 節では、欠測データメカニズムの種類として MCAR、MAR 及び MNAR の3つがあり、欠測データ処理に伴う問題がこの順に難しくなることを直感的に説明した。本節では、主要な用語と概念を説明したうえで、第 2.1 節～第 2.6 節で欠測データの統計的処理法の主要なものを説明する。

○用語と概念

不完全データの観測されなかった値を「欠測値（missing values）」と呼ぶ。一方、観測された値は特に「観測値（observed values）」と呼ばれる。通常欠測値は、何らかの固定された値をとると考えられる。たとえば、A さんが、あるアンケート調査の調査客体に選ばれ、調査項目のひとつである年齢を回答しなかった場合、この調査から作成された不完全データにおいて、A さんの年齢は欠測値である。しかし、それは観測・記録されていないだけであって、たとえば 30 歳という真の値は存在する。このように、ある不完全データに対して、そのすべての欠測値に値が観測されていれば得られるはずのデータというものが仮想的に存在する。これを当該不完全データの「完全データ（complete data）」と呼ぶ。

欠測データの統計的処理においては、第 2 節の例題にみるとおり、不完全データの標本を層化するために用いることのできる変数が重要な役割を果たす。不完全データのすべてのレコード（※）で値が観測されていて、標本分割の条件付け（層化）に利用可能な変数を特に「補助変数（covariates）」と呼ぶ。欠測データメカニズムは、不完全データのデータ生成過程について定義されるが、補助変数の利用可能性によって再定義できる。その場合、特に実践的には、補助変数の利用可能性が MAR と MNAR の分かれ目を決するとみることができる（第 1.1.1 節 MNAR の説明参照）。このため、欠測データの統計的処理の適性は、適当な補助変数の利用可能性に大きく依存するといえる。

※一般的に、統計調査のデータは「調査客体×調査項目」という2つの次元をもつ。このようなデータは、個人、世帯、企業といった調査客体を行、調査項目を列とする行列で表現される。つまりこの行列の第i行第j列の要素は、第i番目の調査客体の第j番目の調査項目の値を表す。個々の行を「レコード」と呼ぶ。

2.1 完全ケース分析

分析に用いる変数のすべての値が、観測されている調査客体のみを用いて分析を行うことを、「完全ケース分析（complete case analysis）」と呼ぶ。完全ケース分析では、分析に用いる変数の少なくとも1つが欠測となっている調査客体を、分析対象から除外する。この操作を、「リストワイズ削除（list-wise deletion）」と呼ぶ。

ひとつの分析がいくつかの分析に分解できるとき、分解された個々の分析ごとに完全ケース分析を行うことを、「利用可能ケース分析（available case analysis）」と呼ぶ。利用可能ケース分析では、分解された個々の分析ごとに、用いる変数の少なくとも1つが欠測となっている調査客体を分析対象から除外しており、この操作を「ペアワイズ削除（pair-wise deletion）」と呼ぶ。

図2－1－1は、エンゲル係数の推定について、完全ケース分析と利用可能ケース分析の実行例を示したものである。表(A)は、仮に欠測が生じなければ得られるはずの完全データを示す。完全データによると、消費支出の標本総計は 292 万円で、食料品支出の標本総計は 65.7 万円なので、総体のエンゲル係数は 22.5%である。実際には欠測が生じ、表(A)の背景灰色で示した値は欠測値である。

完全ケース分析による総エンゲル係数の推定を表(B)に示す。消費支出と食料品支出のどちらか一方でも欠測となっている調査客体は、分析対象から除外するので、id=1, 2, 3, 4, 5, 6 の6世帯が削除される。残された6世帯については、消費支出の総計が 187 万円で、食料品支出の標本総計が 38.8 万円なので、総体のエンゲル係数は 20.7%と推定される。

利用可能ケース分析による総エンゲル係数の推定を表(C)に示す。利用可能ケース分析による総エンゲル係数の推定は、完全ケース分析による総消費支出（あるいは平均消費支出）の推定と、完全ケース分析による総食料品支出（あるいは平均食料品支出）の推定から成っている。平均消費支出の推定では、id=1, 3, 5 の3世帯が分析対象から除外される。残された9世帯については、消費支出の平均が 241 万円/9 世帯である。他方、平均食料品支出の推定では、id=2, 4, 6 の3世帯が分析対象から除外される。残された9世帯については、食料品支出の平均が 52.6 万円/9 世帯である。これら2つの推定結果を合わせて、総体のエンゲル係数は 21.8%と推定される。

図2－1－1　完全ケース分析と利用可能ケース分析

(A) 完全データ

家計id	消費支出（万円）	食料品への支出（万円）
1	16	5.5
2	16	4.3
3	17	4.4
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	292	65.7

エンゲル係数 $= \frac{65.7}{292} = 22.5\%$

(B) 完全ケース分析

家計id	消費支出（万円）	食料品への支出（万円）
1	削除	
2	削除	
3	削除	
4	削除	
5	削除	
6	削除	
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	187	38.8

エンゲル係数 $= \frac{38.8}{187} = 20.7\%$

(C) 利用可能分析

家計id	消費支出（万円）
1	削除
2	16
3	削除
4	18
5	削除
6	20
7	22
8	26
9	28
10	33
11	36
12	42
合計	241

家計id	食料品への支出（万円）
1	5.5
2	削除
3	4.4
4	削除
5	3.9
6	削除
7	4.8
8	5.5
9	5.8
10	6.7
11	7.6
12	8.4
合計	52.6

エンゲル係数 $= \frac{52.6}{241} = 21.8\%$

利用可能ケース分析は、完全ケース分析よりも削除するレコードの数が少なくなる分だけ、推定精度の低下が抑制されるが、変数ごとに分析対象が異なるため、変数相互間の関係性にゆがみがもたらされるという問題がある。図2－1－2は、図2－1－1と比べて他の条件は一定として、欠測パターンが異なる場合に、完全ケース分析と利用可能ケース分析の実行結果を示したものである。先の図2－1－1では、消費支出についても食料品支出についても、支出額の小さい世帯で欠測が生じやすい状況であった。これに対して、図2－1－2では、消費支出については、支出額の大きい世帯で欠測が生じ、食料品支出については、支出額の小さい世帯で欠測が生じている。つまり、図2－1－2の利用可能ケース分析では、消費支出の平均値は支出額が相対的に小さい世帯の組合せで算出され、食料品支出の平均値は支出額が相対的に大きい世帯の組合せで算出されている。このため、図2－1－2の利用可能ケース分析の結果は、総エンゲル係数の推定値が 28.5%と実際よりもかなり大きい値となっている。エンゲル係数の分母は支出規模の大きい世帯群で計算され、分子は支出規模の小さい世帯群で計算されていることの表れである。

図2－1－2 利用可能ケース分析の問題

(A) 完全データ

家計id	消費支出（万円）	食料品への支出（万円）
1	16	5.5
2	16	4.3
3	17	4.4
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	292	65.7

エンゲル係数 $= \frac{65.7}{292} = 22.5\%$

(B) 完全ケース分析

家計id	消費支出（万円）	食料品への支出（万円）
1	削除	
2	削除	
3	削除	
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	削除	
11	削除	
12	削除	
合計	132	28.8

エンゲル係数 $= \frac{28.8}{132} = 21.8\%$

(C) 利用可能分析

家計id	消費支出（万円）
1	16
2	16
3	17
4	18
5	18
6	20
7	22
8	26
9	28
10	削除
11	削除
12	削除
合計	181

家計id	食料品への支出（万円）
1	削除
2	削除
3	削除
4	4.5
5	3.9
6	4.3
7	4.8
8	5.5
9	5.8
10	6.7
11	7.6
12	8.4
合計	51.5

エンゲル係数 $= \frac{51.5}{181} = 28.5\%$

完全ケース分析は、最も簡単な欠測データ対処法であり、多くの統計処理ソフトウェアで、不完全データを分析する場合には原則的に適用される。しかし、完全ケース分析では、欠測データメカニズムが MCAR でない場合、リストワイズ削除によって分析対象標本が偏ることで、推定にバイアスが生じる。また、回答率が十分に大きくない限り、リストワイズ削除によって、データから失われる情報の量は相当大きい。この失われた情報のために推定の精度は低下する。

本章冒頭で述べたとおり、欠測バイアスとは、不完全データに完全ケース分析を実施したときに推定に生じるバイアスのことである。欠測データの問題を考えるうえで、完全ケース分析が出発点となる。完全ケース分析とは異なる手法でバイアスのない推定を行うのが、欠測データの統計的処理である。

2.2 単一代入法

不完全データのすべての欠測値を、それぞれ適当な規則に基づいて決められた値で置き換えることによって、あたかも欠測のないデータを作成することができる。このように作成されるデータを「疑似完全データ（pseudo-complete data）」と呼ぶ。欠測値に代わる値として代入される値を「代入値（imputed values）」と呼ぶ。疑似完全データを1つ作成し、それに対して分析を適用する方法が「単一代入法（single imputation methods）」である。

前節で説明した完全ケース分析は、欠測を含むレコードを分析対象から除いており、これらの削除されたレコードに含まれる情報が無駄になっているといえる。そこで、欠測を含むレコードに含まれる情報を何らか有効活用できないかという動機が働き、その第1歩として単一代入法を位置付けることができる。単一代入法は、後述の多重代入

法、尤度法、及び IPW 法と比べて、統計的な正当性が弱いものの、MAR のもとでは1次モーメントの点推定に関する限り、欠測バイアスのない推定が可能であり、また処理手順が容易であるため、公的統計で広く用いられてきた。

単一代入法は、代入値の決め方に関していくつかの種類がある。その主要なものは以下のとおりである。

○平均値代入（mean imputations）・層化平均値代入（stratified mean imputations）

観測値の平均値を代入値とする。不完全データを補助変数で層化して、代入値となる平均値を層ごとに計算する方法は、特に「層化平均値代入（stratified mean imputations）」と呼ぶ。

○回帰代入（regression imputations）

欠測が生じた変数を従属変数とし、補助変数を独立変数とする回帰モデルを、観測レコードのすべてを用いて推定し、推定された回帰モデルの理論値を代入値とする。

回帰モデルの関数形の特定（線形か非線形か、どの補助変数を独立変数とするか等）及び推定方法（OLS、GLS、MLE、GMM 等）に関して選択の余地がある。
目的の変数が連続変数である場合は、線形回帰モデルを OLS により推定する方法がよく用いられる。

○確率的回帰代入（stochastic regression imputations）

回帰代入の代入値に、推定された分布から乱数発生させた誤差項を加えた値を代入値とする。誤差項を代入値へ加算するのは、回帰代入では捉えることのできない欠測値のばらつきを捉えることを意図している。

○マッチング代入（matching imputations）

欠測値をもつレコードと、すべての変数が観測されているレコードの間で、互いに補助変数の値が類似しているものを対応付け、後者の観測値を前者の欠測値に代入する。補助変数の値の類似性は、適当に定義された距離によって測る。このマッチングの方法を、「最近傍マッチング（nearest neighbor matching）」と呼ぶ。

マッチングで結び付けられる相手の数の決め方に関しても種類がある。あらかじめ決められた値 k について、近いものから順に k 個の相手を結びつけ、それらのレコードの値の平均値を代入値とする場合は、「k-最近傍マッチング（k-nearest neighbor matching）」と呼ばれる。これに対して、あらかじめ決められた値 c について、距離が値 c を下回る相手を結びつけ、それらのレコードの値の平均値を代入値とする場合は、「キャリパーマッチング（caliper matching）」と呼ばれる。

補助変数の値によって、調査客体間の距離を測定する最近傍マッチングに対して、

補助変数の値で条件付けた観測確率の推定値によって、距離を測定する方法が「傾向スコアマッチング（propensity score matching）」である。

補助変数X_iの値で条件付けた観測確率$p(X_i) \equiv Pr(R_i = 1|X_i)$は、観測指標$R_i$（目的となる変数$Y_i$が観測された場合に値1をとる2値変数）の補助変数$X_i$による「傾向スコア（propensity score）」と呼ばれる。傾向スコア$P(R = 1|X = x)$を言葉で表すと、「補助変数Xが特定の値xをとる場合に、目的となる変数Yが観測される確率」である。補助変数に何を用いるかによって傾向スコアの値は変わる。第 1.1.1 節 MAR の具体例（目的となる変数Yは所得額で補助変数Xは就業状態）では、「学生・無業者は必ず回答し、有業者は 50%の確率で回答する」ので、学生・無業者の傾向スコアの値は 1 であり（$Pr\left(R_i = 1\middle|X_i = \text{学生・無業者}\right) = 1$）、有業者の傾向スコアの値は 0.5 である（$Pr\left(R_i = 1\middle|X_i = \text{有業者}\right) = 0.5$）。第 1.1.1 節 MCAR の極端な具体例（硬貨を投げて表が出れば回答し、裏が出れば回答しない）では、いかなる補助変数を用いても、すべての調査対象について傾向スコアの値は 0.5 である。

通常、観測指標の傾向スコアの値は知られていないので、データから推定する必要がある。傾向スコアの推定値は、観測指標を従属変数とし、補助変数を独立変数とする2項モデルの推定によって得られる。

傾向スコアマッチング代入法は、傾向スコアの推定値のみを補助変数とする最近傍マッチング代入法であるといえる。

○LOCF（last observation carried forward）・LVCF（last value carried forward）

同一の標本について複数時点にわたって変数の値を観測・記録することで得られるデータを「パネルデータ」と呼ぶ。パネルデータが作成される統計調査では、直近の観測値を代入値とすることが可能である。この方法は LOCF あるいは LVCF と呼ばれる。

2.2.1 各単一代入法の処理手順

各単一代入法の処理内容を理解するための例を、図2－2－1～2－2－8に示す。これらの図では、同一の不完全データに対して、それぞれの手法を適用している。例示に用いた不完全データは人工的に作成したものであり、レコード数が 20、変数の数が3（欠測指標を含めると4）である。実感をもたせるために例として、調査客体単位を個人とし、3つの変数は、変数yを今月末の対前月末体重変化分（kg）、変数x_1を前月末の対前々月末体重変化分（kg）、変数x_2を今月の対前月1日当たり運動量変化分（時間/日）とする。今月末の体重変化分yに欠測が生じ、前月末の体重変化分x_1及び今月の運動量変化分x_2には欠測が生じないとする。

〇完全ケース分析

図2－2－1は、完全ケース分析の実行例を示したものである。20 人中7人で今月末の体重変化分yが観測されておらず、淡灰色で示している。これらの7人を除き、残った 13 人のみを用いて分析を行うのが完全ケース分析である。データセット2列目の変数y^*は、体重変化分yの真の値であり、欠測となった 7 人の値も入っているが、実際は観測されていない。

ここで、人工的に作成した不完全データの性質をいくつか指摘しておく。第1に、今月末の体重変化分（の真の値）y^*と前月末の体重変化分x_1の間には正の相関、今月末の体重変化分（の真の値）y^*今月の運動量変化分x_2の間には負の相関がある。つまり、すべての調査客体について値が観測されているx_1とx_2は、y^*の値を予測するのに有用な情報を含んでいる。完全ケース分析は、このような有用な情報を一切用いておらず無駄にしているということができる。この無駄に対する「もったいない」という意識が、補助変数を用いた層化平均値代入法、回帰代入法、マッチング代入法等の手法の動機となっている。

〇平均値代入法

図2－2－2は、平均値代入法の実行例を示したものである。今月の体重変化分yについて値が観測されているレコードのみで平均値を求め、その値を 7 人分の欠測レコードの代入値とする。上述のとおり、この不完全データは MNAR 条件のもとで作成しているため、代入値となる平均値は、yの値が比較的小さい人たちの平均値であり、その小さい値を、実はyの値が大きい人たちに代入している様子が分かる。もしこの不完全データを MCAR 条件のもとで作成していれば、代入値となる平均値の算出では、体重変化分yの値が小さい人たちの側にも大きい人たちの側にも偏ることはなく、真の平均値に近い値が代入値となる。最後に、この平均値代入は、完全ケース分析と同様に、欠測データに含まれる有用な情報を一切用いていないことが分かる。

〇層化平均値代入法

図2－2－3は、層化平均値代入法の実行例を示したものである。ここでは、運動量変化分x_2の値の4分位により標本を4分割している。変数 class_x_2は、x_2の値が第何分位に含まれるかを表している。4分割された部分標本のそれぞれについて、観測されているレコードのみで平均値を求め、その値を欠測値に代入する。

図の右側には4つの部分標本が示されている。第1の部分標本は、前月と比べて今月の運動量を大きく減らした人たちのグループであり、運動不足により今月は前月よりも体重が増えている可能性が高い。第2の部分標本は、運動量を減らしたものの、第1のグループほどではない人たちのグループであり、運動不足により体重が増えているかもしれないが、その程度は第1のグループほどではないことが期待される。第3の部

分標本は、運動量を大きくは変化させなかった人たちであり、体重にも大きな変化はない可能性が高い。第4の部分標本は、運動量を増やした人たちであり、体重は減っていると予想される。そして、たまたま第4の部分標本では、今月の体重変化分yの値が欠測となっている人はいないため、運動量変化分にもとづく上述の予想を確かめることができる。確かに、第4の部分標本では概ね体重が減少している。

層化は、似た者同士を同じグループにまとめるので、欠測値への代入値が似た者同士で似た値となる。このとき、何に関して似ているかが重要である。欠測する変数と相関の高い変数に関して似ているほうが、代入値は真の値に近いものとなる。運動量変化分x_2や前月の体重変化分x_1は、今月の体重変化分（の真の値）y^*と相関があるものの例だが、たとえば、「各人の連絡先電話番号の最後の3桁の数字」といった変数の値で層化すると、体重変化分とは無関係なことに関して似た者同士に似た値を代入することになる。欠測となった7人が、体重変化分に関して意味のある区別をされずに分割されるので、たとえばid=19と20の人を互いに似た者同士とはせずに、id=3と20の人を互いに似た者同士とする余地がでてくる。

この例では、今月の運動量変化分x_2の値に基づいて層化しているが、もちろん、前月の体重変化分x_1を層化に用いてもよいし、また前月の体重変化分x_1と運動量変化分x_2の両方を層化に用いてもよい。理屈としては、MNAR に対しては、欠測を生じる変数の真の値y^*に対する予測力が最も大きい変数ないし変数の組合せを層化に用いるのがよい。最後に、層化平均値代入法では、完全ケース分析や平均値代入法で無駄にされていた情報が活用されているといえる。

○回帰代入法

図2－2－4は、回帰代入法の実行例を示したものである。まず不完全データに対して、今月の体重変化分yを被説明変数とし、前月の体重変化分x_1及び運動量変化分x_2を説明変数とする回帰分析を（完全ケース分析により）行う。そこで推定された回帰モデル（図の例では線形回帰モデル）にもとづいて、欠測を出した7人についてx_1及びx_2の値からyの理論値を代入値とする。層化平均値代入法と同様に回帰代入法では、完全ケース分析や平均値代入法では無駄にされていた情報（x_1及びx_2）が、代入値を回帰モデルの理論値として算出する段階で活用されている。

図中右上部分のグラフは、横軸に回帰分析の理論値、縦軸に観測値をとった散布図である。灰色の点は今月の体重変化分yが観測された13人を表し、黒色の点は（の真の値）y^*が観測されなかった7人を表している。この7人については、縦軸の座標が分からないので、回帰分析の理論値を観測値の代わりに縦軸座標としている。完全ケースの 13 人については、縦軸は観測値を表し、欠測となった7人に関しては、縦軸は横軸と同じく理論値であるから、欠測となった7人の点はグラフ中の点線で示した45度線上に位置する。回帰モデルのパラメータの推定値は、灰色の点について 45 度線か

らの垂直距離の平方和を最小化する値である。こうして得られたパラメータ推定値に基づいて欠測となった7人の体重変化分yの理論値が計算される、つまり黒色点の 45 度線上での位置が決まる。

回帰代入法も似た者同士には似た値を代入値とするという点で、上述の層化平均値代入法と同じである。回帰代入の場合、何に関する類似性かといえば、説明変数に関する類似性である。

〇確率的回帰代入法

図2－2－5は、確率的回帰代入法の実行例を示したものである。図2－2－4の回帰代入法と同様に、まず不完全データに対して、今月の体重変化分yを被説明変数とし、前月の体重変化分x_1及び運動量変化分x_2を説明変数とする回帰分析（完全ケース分析により）を行う。そこで推定された回帰モデル（図の例では線形回帰モデル）にもとづいて、欠測を出した7人についてx_1及びx_2の値からyの理論値を求める。確率的回帰代入法では、推定された回帰モデルの誤差項が従う分布から乱数を発生させ、得られた値を回帰モデルの理論値に加えたものを代入値とする。この点が、回帰代入法と異なる。

〇マッチング代入法（最近傍マッチング）

図2－2－6は、最近傍マッチング代入法の実行例を示したものである。最近傍マッチングでは、変数yが欠測している調査客体i、及び変数yが観測されている調査客体jとの間の類似性を、調査客体iの補助変数の値x_i及び調査客体jの補助変数の値x_jの間の距離によって測り、距離が十分小さいもの同士を結び合わせる。そして、結び合わされた観測レコードと欠測レコードで、観測レコードの値を欠測レコードの欠測値に代入する。マッチング代入法でも、完全ケース分析や平均値代入法では無駄にされていた情報（x_1及びx_2）が、回答者と無回答者の間の距離を算出する段階で活用されている。

前月の体重変化分x_1と運動量変化分x_2という2つの補助変数があるので、2次元実数ベクトル空間で、今月の体重変化分yの値が欠測となった7人と、欠測とはならなかった 13 人の間の距離を総当たりで測る。つまり、7×13＝91 通りの組み合わせについて両者間の距離を計算する。図の例では距離概念の定義としてマハラノビス距離を用いている。欠測となった7人のそれぞれに組み合わせの相手となる候補が 13 人いるが、そのなかで最も距離の近い者と組み合わされる。計算の結果、たとえば id=3 の無回答者には id=8 の回答者が最も近い者、つまり最も似ている者として結び合わされている。両者間の距離はマハラノビス距離で 0.67 である。結ばれたもの同士の距離で両者の類似性を測るので、たとえば、id=16 の無回答者と id=12 の回答者の類似性よりも、id=13 の無回答者と id=14 の回答者の類似性のほうが大きいということも分かる（距離

の値 1.44 と 0.43 の比較)。

マッチング代入法で、欠測を出したレコードと結びついて代入値を提供するレコードを「ドナー」と呼ぶ。id=14 の回答者が id=10 の無回答者にも id=13 の無回答者にも結ばれているが、このように、同じ回答者が複数の無回答者のドナーとなることがある。また、ここでの例のようにひとりの無回答者のドナーとなる回答者の数はひとりと限る必要はなく、複数のドナーを結びつける場合は、それらの平均値を代入値とするなどの処理も考えることができる。

○マッチング代入法（傾向スコアマッチング）

図2－2－7は、傾向スコアマッチング代入法の実行例を示したものである。傾向スコアマッチング代入法では、まず標本全体を用いて傾向スコアを推定する。傾向スコアは、補助変数の値によって条件付けた観測確率である。標本全体、すなわち 20 人分すべてのレコードを用いて観測確率を2つの補助変数(前月の体重変化分x_1と運動量変化分x_2)で説明する2項回帰モデルを推定し、全員についてそれぞれの傾向スコアを得ることができる。

たとえば、図2－2－7によると、id=1 の人の傾向スコアは 0.99 であるが、これは、id=1 の人の前月の体重変化分x_1(約 0.66kg 減)と運動量変化分x_2(約 0.60 単位増)から推定した結果、id=1 の人は約 99%の確率で、今月の体重変化分(の真の値)y^*を回答する、ということを意味している。そして実際、id=1 の人は回答している。一方、id=3 の人は、約 52%の確率で体重変化分(の真の値)y^*を回答するとの推定結果であったにもかかわらず(それほど低くはない確率で回答する条件を備えていた人であったが)、結果的には回答してくれなかった。

次に、7人の無回答者と 13 人の回答者の間で傾向スコアの差の絶対値を総当たりで計算し、7人の無回答者それぞれについてその値が最も小さい回答者を結びつける。たとえば id=3 の無回答者は、回答者の中では id=14 の人と傾向スコアの値が最も近く、その差は約 18%ポイントである。つまり傾向スコアマッチング代入法では、観測確率に関して似た者同士を結び合わせている。無回答者の欠測値には、結ばれた相手の回答者の値を代入値とする。

○LOCF

図2－2－8は、LOCF の実行例を示したものである。LOCF は、他の単一代入の手法と異なり、パネルデータにのみ適用できる手法である。この例では、補助変数の中に欠測を生じる変数yの1期前の値x_1がすべての人について観測されているため、欠測値には 1 期前の値を代入値とする。たとえば、id=3 の調査客体は、今月の体重変化分yは観測されていないが、前月の体重変化分x_1が観測されているので、LOCF ではその値(0.59kg)を今月の体重変化分yに代入する。

今月の体重変化分の値が大きい人ほど欠測となっているので、欠測バイアスは今月の体重変化分の値が小さい側へのバイアスとなるが、LOCF がこの欠測バイアスを緩和するのは、前月と今月の体重変化分が正の相関をもつときだけである。これは他の単一代入法にはない LOCF 独自の性質である。

◇まとめ

各単一代入法を比較すると、第1に、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法は、完全ケース分析や平均値代入法では無駄に捨てられていた情報の活用が図られている。パネルデータにおいて、直近に観測された値を代入値とする LOCF も、完全ケース等と比べて情報の活用が図られている。

第2に、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法は、「似た者同士は似た値をとる」という発想になじんでいる。そこでは、補助変数に関して類似していれば欠測する変数に関しても類似しているはずであるという推定原理が働いている。この推定原理は明らかに、補助変数と欠測する変数の相関が高いほど正しい推定を導く。LOCF については、「似た者同士」の相手が自分自身ということになるが、他の単一代入法にはない LOCF 独自の問題（上述）を招かないためには今期と前期が似た状況であることが重要である。

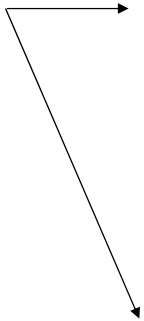
第3に、確率的回帰代入法は回帰代入法にともなう推定精度の過大評価及び1次よりも大きいモーメントの推定における欠測バイアスを緩和する。この点については、後でさらに詳しくみる。

欠測レコードの情報を無駄にせず、また補助変数の説明力を活用するという点は、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法の完全ケース分析及び単純な平均値代入法に対する優位性である。

図２－２－１　完全ケース分析の処理手順

不完全データ

id	y*	y	missing	x_1	x_2
1	-1.49	-1.49	0	-0.66	0.60
2	-1.25	-1.25	0	0.30	1.93
3	-0.83		1	0.59	0.31
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
10	0.15		1	0.18	-0.40
11	0.18	0.18	0	-0.01	0.57
12	0.80	0.80	0	0.70	-0.83
13	0.81		1	0.33	-0.01
14	0.81	0.81	0	0.61	-0.28
15	0.86	0.86	0	-0.66	-1.40
16	1.05		1	1.78	-0.68
17	1.09	1.09	0	0.00	0.20
18	1.33		1	0.16	-1.86
19	1.41		1	0.62	-1.61
20	1.43		1	0.97	-1.12



分析に用いる情報

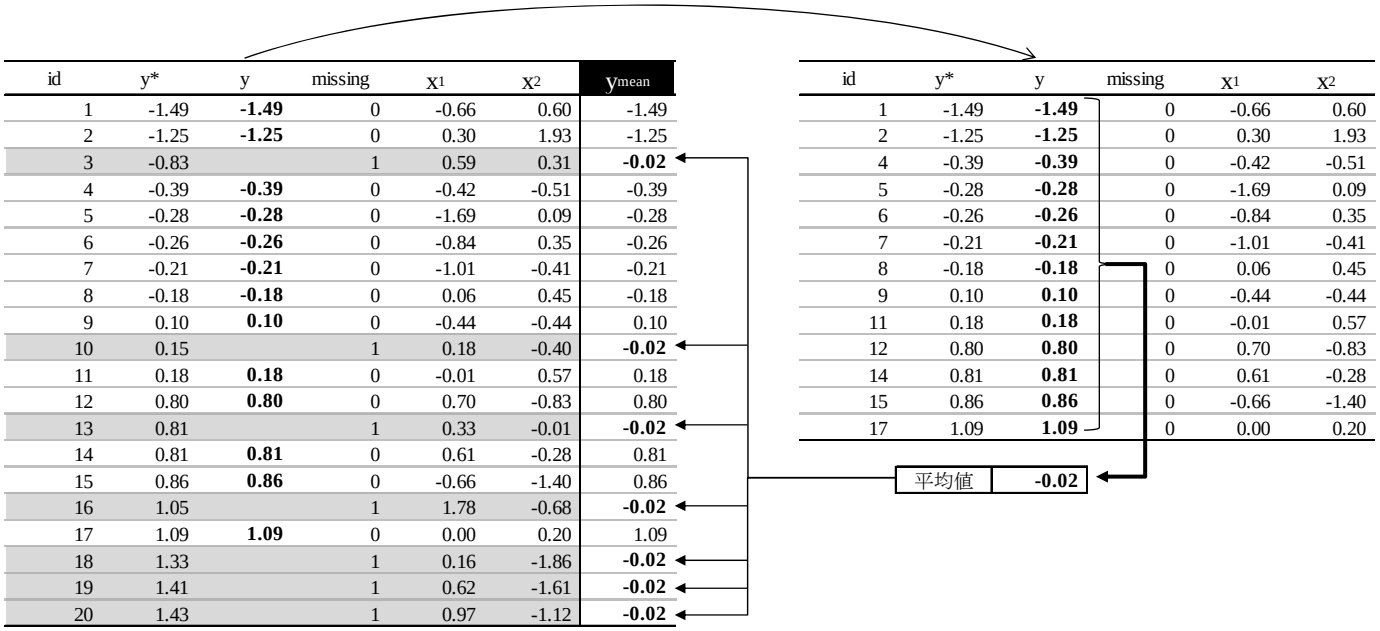
id	y*	y	missing	x_1	x_2
1	-1.49	-1.49	0	-0.66	0.60
2	-1.25	-1.25	0	0.30	1.93
4	-0.39	-0.39	0	-0.42	-0.51
5	-0.28	-0.28	0	-1.69	0.09
6	-0.26	-0.26	0	-0.84	0.35
7	-0.21	-0.21	0	-1.01	-0.41
8	-0.18	-0.18	0	0.06	0.45
9	0.10	0.10	0	-0.44	-0.44
11	0.18	0.18	0	-0.01	0.57
12	0.80	0.80	0	0.70	-0.83
14	0.81	0.81	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

分析には用いない情報

id	y*	y	missing	x_1	x_2
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

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数

図２－２－２　平均値代入法の処理手順



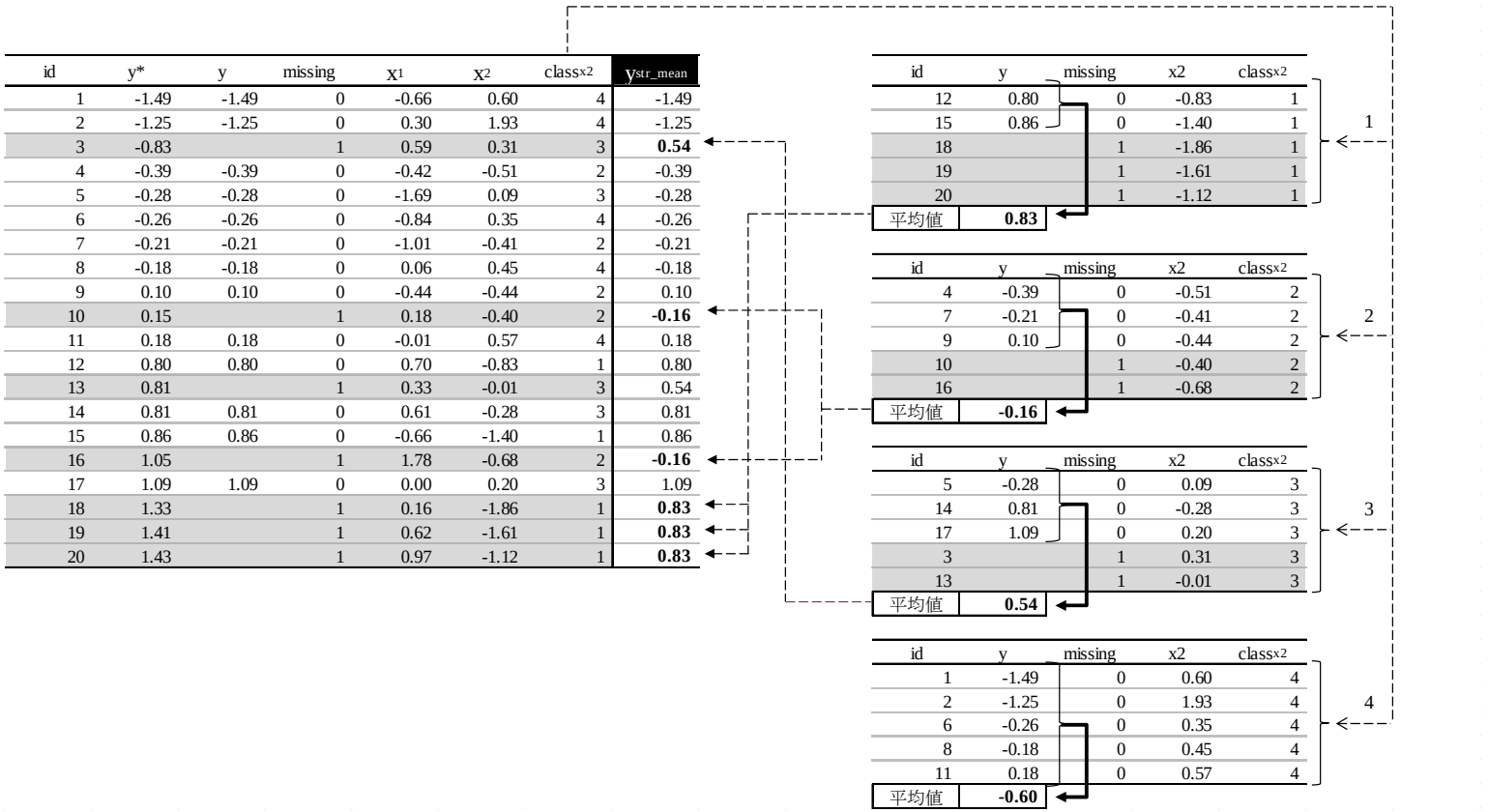
id	y*	y	missing	x_1	x_2	y_{mean}
1	-1.49	**-1.49**	0	-0.66	0.60	-1.49
2	-1.25	**-1.25**	0	0.30	1.93	-1.25
3	-0.83		1	0.59	0.31	**-0.02**
4	-0.39	**-0.39**	0	-0.42	-0.51	-0.39
5	-0.28	**-0.28**	0	-1.69	0.09	-0.28
6	-0.26	**-0.26**	0	-0.84	0.35	-0.26
7	-0.21	**-0.21**	0	-1.01	-0.41	-0.21
8	-0.18	**-0.18**	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	**-0.02**
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	**-0.02**
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	**-0.02**
17	1.09	**1.09**	0	0.00	0.20	1.09
18	1.33		1	0.16	-1.86	**-0.02**
19	1.41		1	0.62	-1.61	**-0.02**
20	1.43		1	0.97	-1.12	**-0.02**

id	y*	y	missing	x_1	x_2
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

平均値	**-0.02**

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{mean}: 平均値代入による代入値

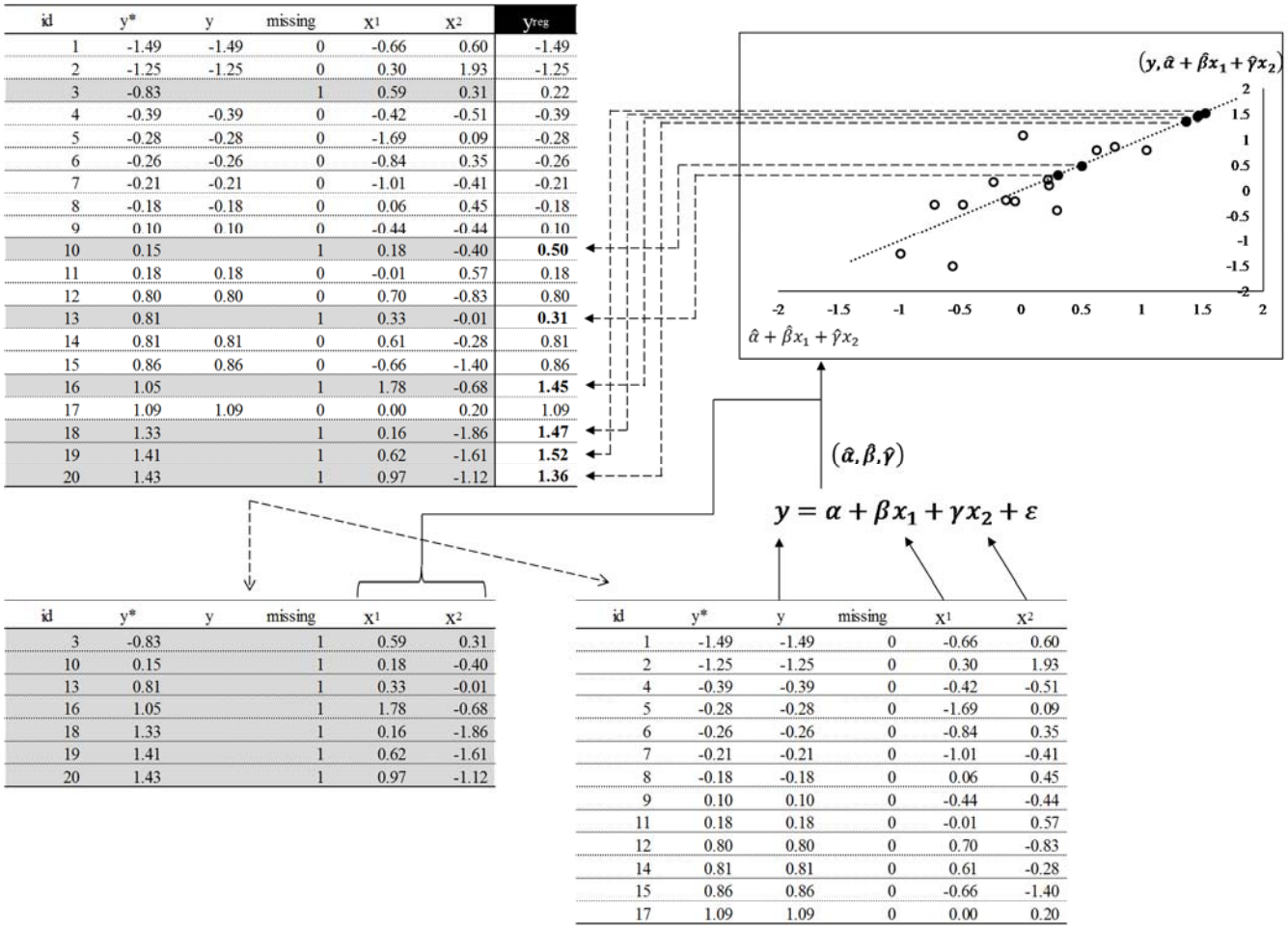
図２－２－３ 層化平均値代入法の処理手順



id	y*	y	missing	x_1	x_2	$class_{x2}$	y_{str_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**

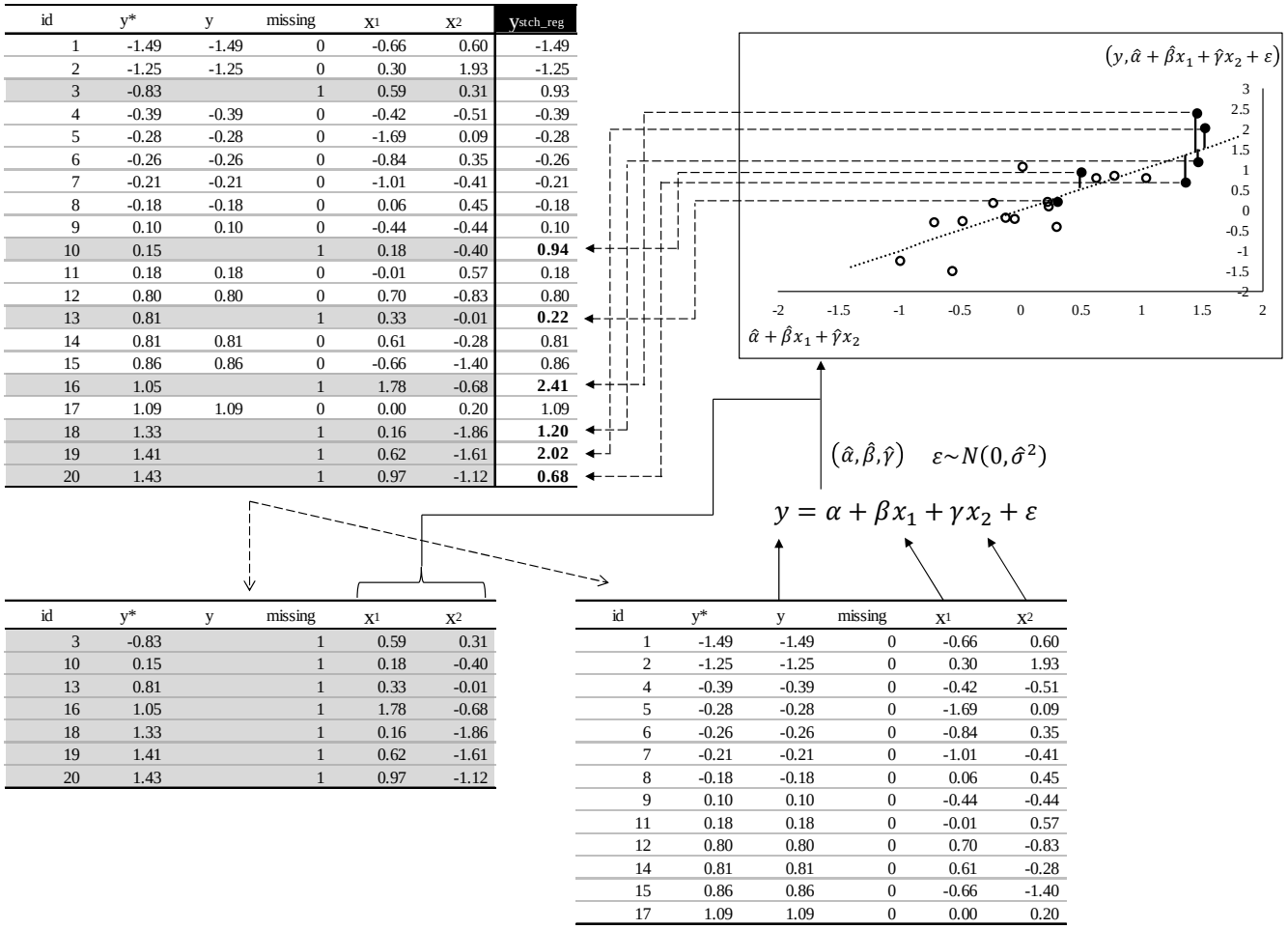
y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、$class_{x2}$: 補助変数 x_2 の４分位階層、y_{str_mean}: 補助変数 x_2 にもとづく層化平均値代入による代入値

図２－２－４ 回帰代入法の処理手順



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

図２－２－５ 確率的回帰代入法の処理手順



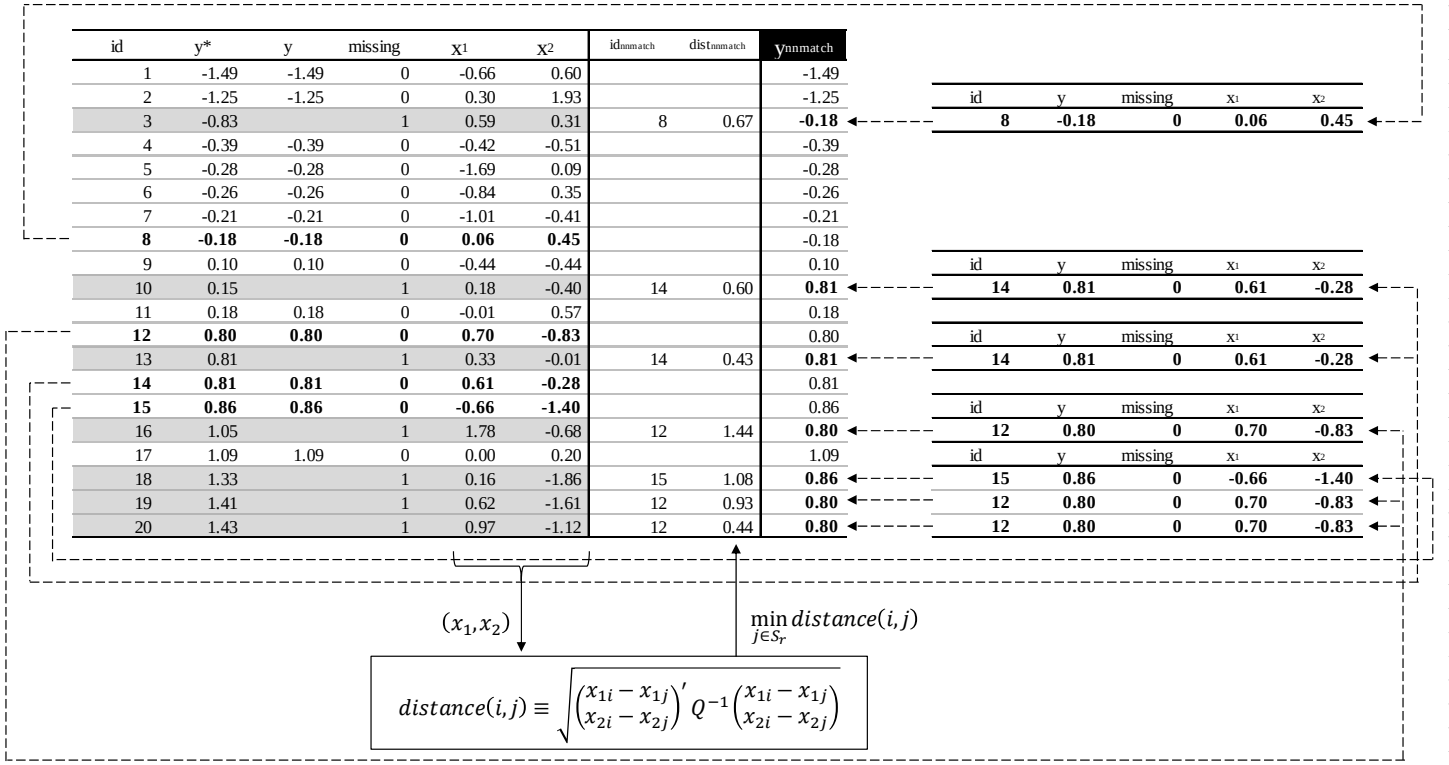
id	y*	y	missing	x_1	x_2	y_{stch_reg}
1	-1.49	-1.49	0	-0.66	0.60	-1.49
2	-1.25	-1.25	0	0.30	1.93	-1.25
3	-0.83		1	0.59	0.31	0.93
4	-0.39	-0.39	0	-0.42	-0.51	-0.39
5	-0.28	-0.28	0	-1.69	0.09	-0.28
6	-0.26	-0.26	0	-0.84	0.35	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	-0.21
8	-0.18	-0.18	0	0.06	0.45	-0.18
9	0.10	0.10	0	-0.44	-0.44	0.10
10	0.15		1	0.18	-0.40	**0.94**
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	**0.22**
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	**2.41**
17	1.09	1.09	0	0.00	0.20	1.09
18	1.33		1	0.16	-1.86	**1.20**
19	1.41		1	0.62	-1.61	**2.02**
20	1.43		1	0.97	-1.12	**0.68**

id	y*	y	missing	x_1	x_2
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	x_1	x_2
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

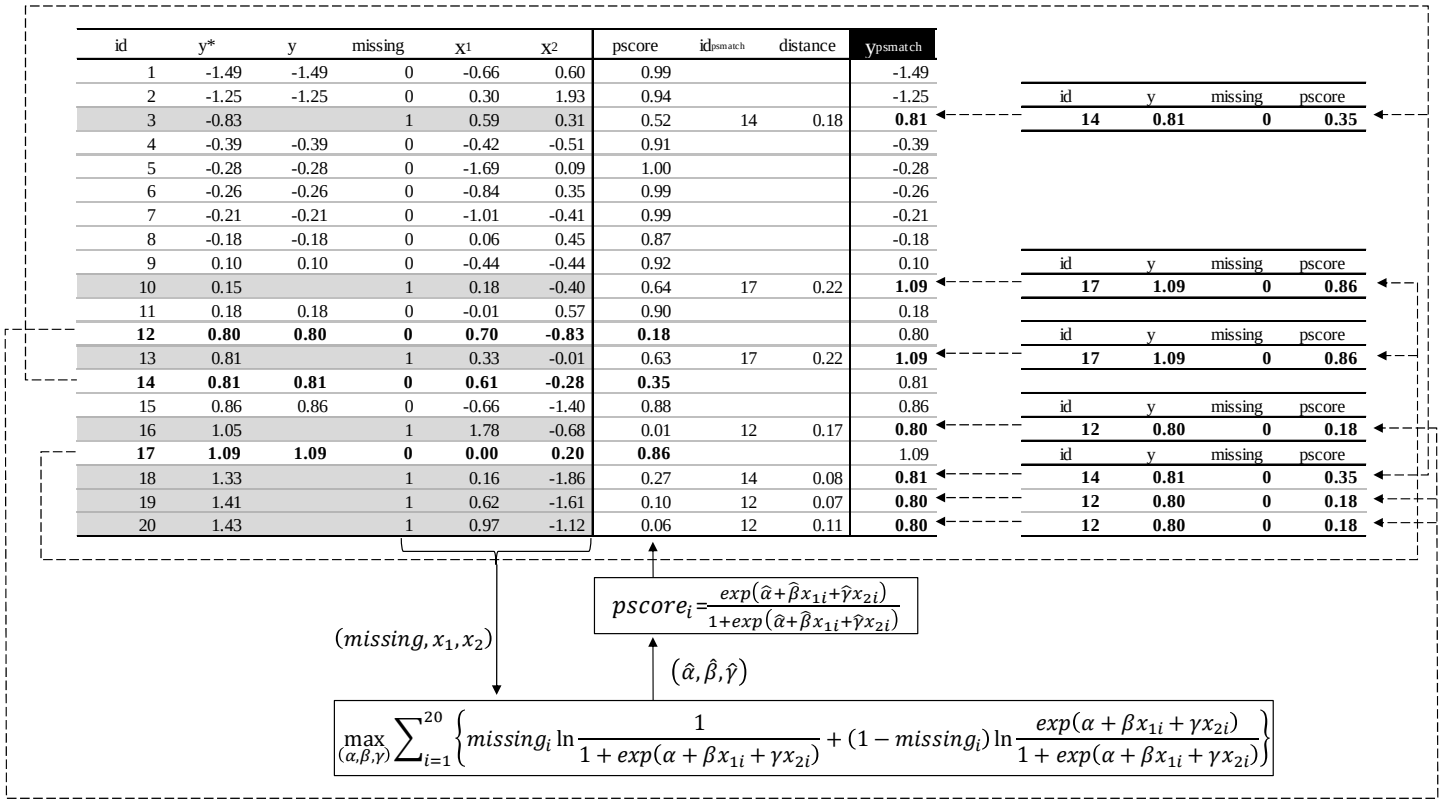
y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{stch_reg}: 確率的回帰代入による代入値

図２－２－６ 最近傍マッチング代入法の処理手順（図は１対１マッチング）



y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、$id_{nnmatch}$: 最近傍マッチングによって合された相手レコードの id、$dist_{nnmatch}$: マッチングの相手との間の距離（行列 Q により距離概念を定義する）、$y_{nnmatch}$: 最近傍マッチング代入による代入値

図２－２－７ 傾向スコアマッチング代入法の処理手順（図は１対１マッチング）



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、$id_{psmatch}$: 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）、$y_{psmatch}$: 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）

図２－２－８ LOCF の処理手順

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、y_{LOCF}: LOCF による代入値

2.2.2 各単一代入法の特徴

次に、各単一代入法の特徴を理解するために、不完全データの数値例を図２－３－１に示す。目標母集団から単純無作為抽出した 1,000 人の標本について、各調査客体の体重の前月差の値を２時点にわたって収集する調査を考える。第１時点の体重の前月差（kg）を W1、第２時点の体重の前月差（kg）を W2 とする（実は図２－２－１～８で用いた数値例のデータを発生させたモデルと同じである）。幸運にも第１時点の値はすべての調査客体について観測されたが、第２時点の値は一部の調査客体について欠測が生じたとする。この調査から作成される不完全データに、各単一代入法を適用する（図では参考までに、第 2.5 節で取り上げる多重代入法の適用についても示している。）。この場合、第２時点の体重前月差 W2 が目的の変数、第１時点の体重前月差 W1 が利用可能な補助変数となる。ここでは、目的となる変数と補助変数との間に正の相関がある場合を考える。

パネル(A)の上図は、第２時点の体重前月差 W2 が観測されなかった調査客体についても、真の値が得られたときに作成される W2 のヒストグラムである。ヒストグラムでは実際に観測されたデータの部分は淡灰色、欠測となったデータの部分は濃灰色で示し区別している。パネル(A)の下図は、第２時点と第１時点の体重前月差の散布図である。散布図では、W2 が実際に観測された調査客体は記号〇、欠測となった調査客体は記号△で示し区別している。誰もパネル(A)のような真の姿を知ることはできない。

パネル(A)をみると、第２時点の体重前月差が大きい調査客体ほど欠測する割合が高い。この点だけを考えると、欠測の割合が欠測する変数に依存しているので、欠測データメカニズムは MNAR である。しかしここでは、第２時点の体重前月差が、補助変数として利用可能な第１時点の体重前月差と正の相関を示している。このため、この補助変数で層化すれば、層ごとの欠測割合は欠測する変数の値に依存しない。つまり、条件付けに用いることで、欠測する変数の値と欠測確率との相関を消すことができるという性質をもつ変数（第１時点の体重前月差）が補助変数として利用可能なので、欠測データメカニズムは MAR とみなせる。

〇平均値代入法

パネル(A)に示す不完全データ（淡灰色部分）に平均値代入法を適用した結果を、パネル(B)に示す。第２時点の体重前月差 W2 が大きい調査客体ほど欠測する割合が高いため、W2 が小さい側へ偏った部分標本によって計算される平均値が代入値となる。このため疑似完全データの標本平均による母集団平均の推定

には下方バイアスが生じる。

また、パネル(A)のヒストグラムで濃灰色に示されたレコードは、観測されたW2 の平均値という一点に集められ、パネル(B)のヒストグラムのように高頻度帯（点）を形成する。これが、１次よりも大きい母集団モーメントの推定における過小バイアス及び平均値代入法による推定精度の過大評価の原因となる。

パネル(B)の散布図を真の姿であるパネル(A)の散布図と比べると、欠測となった調査客体について第２時点の体重前月差を一律に一定値（観測値の平均値）とすることの副作用がみてとれる。真の姿では、この目標母集団は、第１時点と第２時点の体重前月差に比較的高い正の相関がある（パネル(A)の散布図）。しかし、W2 の値が大きい調査客体でより多くの欠測が発生したため、平均値代入法による疑似完全データでは、W1 が最も大きい階層に属する調査客体で、W2 の代入値が実際よりも小さい値となっている。平均値代入法では、W2 の分布をゆがめるだけでなく、W2 と W1 との関係性までもゆがめてしまう。

〇層化平均値代入法

パネル(B)にみるような平均値代入法の問題点は、標本を適当な補助変数により層化することで緩和される。パネル(A)の不完全データに層化平均値代入法を適用した結果を、パネル(C)に示す。これは、第１時点の体重前月差 W1 の４分位点を境に標本を４層分割し、各層ごとに層内の観測データを用いて平均値代入を行ったものである。層が４つに分かれたため、パネル(C)とパネル(B)のヒストグラムを比べると、パネル(B)にみられる高頻度点は１つから４つに分散している。この分散の効果は、層の数を増やすほどより大きくなる。

〇回帰代入法

補助変数による標本の層化からさらに進んで、補助変数の値ごとにモデルに基づく推定値を代入する方法が回帰代入法である。回帰代入法の結果を、パネル(D)に示す。この例では、W2 が観測されたレコードについて、W2 を W1 へ回帰するモデルを推定し、推定されたモデルに基づく欠測レコードの理論値を代入値としている。平均値代入の場合にヒストグラムに出現していた高頻度点の問題は、解消しているようにみえる。

散布図については、パネル(B)や(C)よりは真の姿に多少近づいているようにみえるとはいえ、欠測データに対応するレコードが、疑似完全データでは回帰直線上に固定されており、回帰直線からの乖離が無視されていることが分かる。少なくとも、このばらつきが取り除かれている分だけでも、推定精度が過大評価されることになる。そこでこの点を考慮に入れた手法として、次のパネル(E)に示す

確率的回帰代入法を考えることができる。

〇確率的回帰代入法

パネル(E)には、確率的回帰代入法の結果を示す。ここでは、パネル(D)の代入値に乱数発生させた誤差項を加えている。誤差項は正規分布に従い、その標準偏差は回帰推定の残差から推定した値を用いている。パネル(D)と比べて、パネル(E)では回帰直線からランダムに乖離するので、代入値のばらつきが大きくなり、ヒストグラムも散布図も真の姿により近づいているようにみえる。このように、誤差項を代入値に加算することで、代入を施した変数の分布のばらつきが維持されるため、確率的回帰代入は１次よりも大きいモーメントの推定について回帰代入よりも優れている。

〇マッチング代入法

パネル(F)及び(G)には、それぞれ最近傍マッチング代入法及び傾向スコアマッチング代入法の結果を示す。マッチング代入の結果は、「（補助変数に関して）似た者同士は（欠測値でも）似た値をとる」という回帰代入の性質を共有しつつ、回帰代入の結果が回帰直線上に集中するのと比べて、マッチング代入の結果はばらつきを保っている。しかし一部に平均値代入の結果でみられた高頻度点がみられる。欠測が起こる変数 W2 と補助変数 W1 に正の相関がある条件の下で、W1 と W2 が共に大きい値をとる領域（散布図の右上側領域）で欠測率が高くなっており、この領域では代入値を求める欠測レコードに対して、代入値を与えてくれるマッチングの相手が希少になっている。このため当該領域では、特定の観測レコードの値が代入値として頻繁に利用され、散布図及びヒストグラムにみられるとおり、代入値の高頻度点が生じることになる。

〇LOCF

パネル(H)には、LOCF の結果を示す。すなわち、欠測した W2 には W1 の値を代入している。この例では、W1 と W2 の相関が比較的に高いため、LOCF によって作成されるヒストグラムも真の姿に近いとみえる。系列相関が負となるような例（後述の図２－３－２）だと、逆の結果をもたらす。すなわち、ヒストグラムの欠測部分は LOCF によって反転する（大きい欠測値には小さい代入値、小さい欠測値には大きい代入値）。また、相関図は当然ながら、欠測データの部分は 45 度線上に固定される。

LOCF は、回帰代入法をパネルデータに適応した場合の特殊形である。パネル(H)LOCF に示す結果は、パネル(D)回帰代入法において定数項 0 及び補助変数の係数 1 という極めて厳しい線形制約を課したものとみることもできる。

○補助変数との相関が負の場合

図２－３－１では、第１時点の体重前月差 W1 と第２時点の体重前月差 W2 が正の相関関係にある場合の例となっているが、W1 と W2 が負の相関関係にある場合の例を図２－３－２に示す。補助変数との相関が正である場合と負である場合で疑似完全データの分布を比較すると、LOCF を用いた場合に違いがみられる。他の単一代入法については、W1 と W2 との相関の正負にかかわらず図２－３－１で指摘した点が成り立つ。つまり、目的となる変数との相関が正であれ負であれ、補助変数が観測確率に説明力をもてば、その補助変数を利用した単一代入法によって MAR の下での推定における欠測バイアスを緩和できる。

LOCF については、W1 と W2 が負の相関関係にある場合、欠測レコードで代入値と欠測値とでレコード間の大小関係が反転するため、図２－３－２パネル(A)の完全データの分布にみられる欠測率と当該変数の値との正の相関が、同パネル(H)の疑似完全データではみられなくなっている。この点で LOCF は他の手法と異なり、負の系列相関や一時的共通ショックを特徴とする変数に適用すると推定のバイアスをより大きくする。

○補助変数との相関がない場合

図２－３－３は、第１時点の体重前月差 W1 と第２時点の体重前月差 W2 に相関がない場合について、各単一代入法の性質を示した。この場合、層化平均値代入法（パネル(C)）は平均値代入法（パネル(B)）と似たような結果をもたらす。MNAR のもとで W1 が W2 と無相関であれば（正確には、補助変数が欠測確率に説明力をもたなければ）、当該補助変数 W1 による層化のメリットはない。回帰代入（パネル(D)）もまた平均値代入法（パネル(B)）と似たような結果をもたらす。補助変数に欠測を説明する力がなければ回帰代入にもメリットはない。事実、補助変数と目的となる変数が無相関の場合、層化平均値代入法及び回帰代入法は、平均値代入法と同値である。したがってこの場合の確率的回帰代入の結果は、平均値代入の結果に誤差項のばらつきを与えたに過ぎないものとなる。マッチング代入法（パネル(F)及び(G)）も、代入値にもっともらしさを与えるマッチングではなくランダム・マッチングとなり、結果は確率的回帰代入と同様にほぼ意味のない代入となる。LOCF（パネル(H)）も（LOCF が非常に厳しい制約下での回帰代入法であると考えると）回帰代入法と同様に、真の姿（パネル(A)）を再現できない。

用いる補助変数が欠測確率に説明力をもたなければ、どの単一代入法も欠測バイアスを緩和できない。それぱかりではなく、作成される疑似完全データにおいて、変数相互間の関係が代入によってゆがめられるという害をもたらすので、このように不適当な補助変数を用いた単一代入法は避けなければならない。