

ESRI統計より：景気統計

欠測データの統計的処理について

経済社会総合研究所景気統計部
間 真実

統計調査においては、無回答により一部の調査客体又は調査項目について情報が得られないことが多くある。このような統計調査から作成されるデータは、本来観測・記録されるべき情報の一部が観測・記録されておらず、「不完全データ」と呼ばれる。不完全データを用いた推定には、次の2つの問題がある。

1. 欠測バイアス：

欠測により標本が目標母集団の縮図としての性格を失うことによって推定にもたらされるバイアス

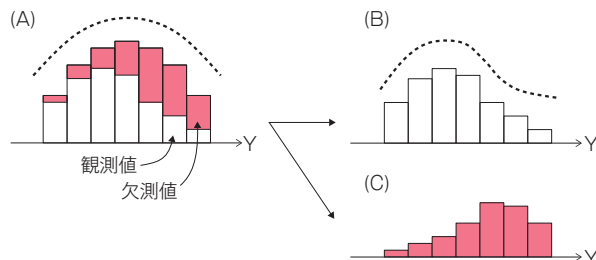
2. 推定精度の低下：

観測値の数が減少することによる標準誤差の増大
信頼区間が小さくてもバイアスのある推定には意味がないので、第1の欠測バイアスのほうが第2の精度低下よりも重要である。また、推定精度の低下に対しては、想定される回答率を考慮して調査客体数を決めることで目標精度を確保することができるのに対して、欠測バイアスへの対処はより困難である。

欠測バイアスが生じるしくみを図1に示す。興味の対象となる変数Yについて標本調査を実施し、一部の調査客体から回答が得られなかったとする。無回答の調査客体についても変数Yの真の値は存在し、それらは特に「欠測値」と呼ばれる。図1(A)は、仮に欠測値も観測されたとしたときの、標本全体における変数Yのヒストグラムである。着色部分が欠測値、無着色部分が観測値である。図1(A)の破線は、仮に欠測が生じなかったとしたときに、標本から推定される変数Yの分布である。標本設計が適切に行われている限り、この推定にはバイアスはない。グラフ(A)は標本全体の真の姿であるが、誰もそれを知ることはできない。図1(B)及び(C)は、標本全体を回答標本及び無回答標本に分離してそれぞれのヒストグラムを描いたものである。不完全データの分析者には、グラフ(B)

の情報だけが利用可能である。この不完全データに基づいて推定された変数Yの分布が(B)の破線である。真の姿(A)と比べて、左方に歪んでいることが分かる。この分布の歪みが、変数Yの平均や分散の推計等のあらゆる推定におけるバイアスの源泉となる。

図1 欠測バイアスをもたらす欠測



ただし、不完全データによる推定には欠測バイアスが必ず生じるというわけではない。欠測バイアスが生じるか否かまたその程度は、欠測の現れ方によって決まる。図1の例では、変数Yの値が大きい階層ほど欠測率が高いという傾向がグラフ(A)からみてとれる。このように、欠測する変数の値と欠測確率との間に正負を問わず相関があるときに欠測バイアスが生じる。

欠測する変数の値と欠測確率との間に相関がない場合を、図2に示す。図2(A)では、欠測する変数Yの値によらず全ての階層で欠測率が等しい。このとき、回答標本は当初の標本よりもサイズが縮小しているものの、母集団の縮図としての性格を失っていないので、観測値の情報のみを用いた推定の結果(図2(B)の破線)は真の姿((A)の破線)と比べて歪むことはない。

図2 欠測バイアスをもたらさない欠測

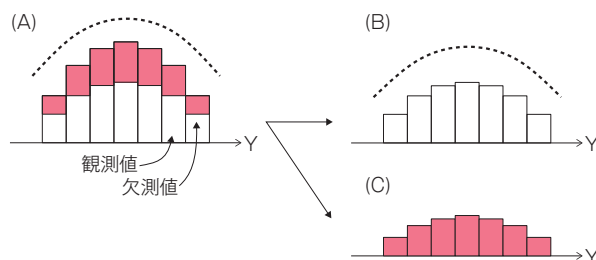


図1のように欠測バイアスが生じる場合、それを補正する処理が必要である。欠測バイアスを補正する統計的処理法には、大きく分けて、(1)単一代入法¹、(2)ウェイト調整法²、(3)多重代入法³、(4)尤度法⁴の4種類がある。全ての手法の共通点は、補助変数の利用

1 単一代入法は、欠測値に適当な値を代入してひとつの疑似完全データを作成する手法である。代入値の決め方に関して、層化平均値代入、回帰代入、マッチング代入等の種類がある。

である。補助変数とは全ての調査客体について観測されている変数であり、標本調査の場合はフレーム（母集団データベース等）に収録されている変数が該当する。

補助変数を利用して欠測バイアスを補正するしくみを図3に示す。興味の対象となる変数Yは一部の調査客体で欠測が生じているが、補助変数Xは全ての調査客体について観測されている。図3のXY座標平面には、標本の散布図が描かれている。散布図では、変数Yの値が観測された調査客体を記号○、変数Yの値が観測されていない調査客体を記号△でそれぞれ表す。Y軸上のグラフは、当該標本において欠測値と観測値を区別した変数Yのヒストグラムであり、図1（A）と同じものである。X軸側のヒストグラムは、変数Yに関する観測の有無で区別した補助変数Xのヒストグラムである。データの分析者には、Y軸上のヒストグラムはみえていないが、X軸側のヒストグラムはみえている。つまり補助変数Xに関しては、歪んでいない分布と歪んだ分布の両方がみえているので、両者の差を知ることができる。変数Xに関する歪みを補正することが、変数Yに関する歪みの補正に連動していれば、欠測バイアスを緩和することができる。この連動は、図3の例では、興味の対象となる変数Yと補助変数との間の相関が（正負を問わず）大きいほど、強い。上述の4種類の統計的手法は、欠測バイアスの補正に資する補助変数を有効活用する点で共通している。

図3 補助変数を用いた欠測バイアスへの対処

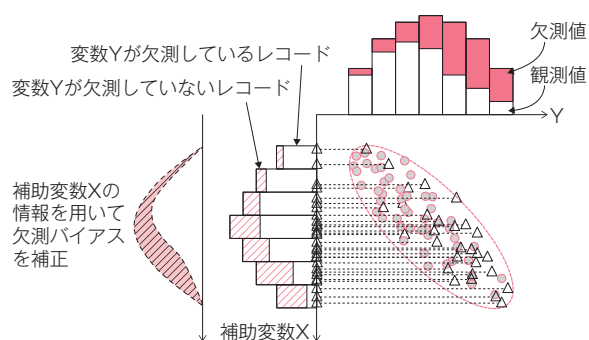


図3の例からも分かるように、標本の調査客体全てについて値が観測されている変数だからといって欠測バイアスの補正に利用できるわけではない（図3の場合、変数XとYが無相関であれば、変数Xは欠測

バイアスの補正に利用できない）。欠測バイアスの補正においては、補正に資する補助変数が標本の全ての調査客体について観測されていなければならない。

これまでの図1～3を用いた議論から分かる通り、不完全データの統計的処理においては、欠測バイアスをもたらす欠測であるか否か、また欠測バイアスをもたらす欠測である場合に欠測バイアスの補正に資する補助変数が観測されているか否かという条件分岐が処理の必要性を決する。この条件分岐により3つの場合が区別される。第1は、欠測バイアスをもたらさない欠測であり、この条件は「完全にランダムな欠測」と呼ばれる。第2は、欠測バイアスをもたらす欠測ではあるがその補正に資する補助変数が観測されている場合であり、この条件は「ランダムな欠測」と呼ばれる。第3は、欠測バイアスをもたらす欠測でありその補正に資する補助変数が観測されていない場合であり、この条件は「ランダムでない欠測」と呼ばれる。これら3条件のいずれが成立しているかによって、統計的処理法ごとの必要性や適性が異なる。完全にランダムな欠測に対しては、特別な処理は必要ない。ランダムな欠測に対しては、補助変数を利用して欠測バイアスを緩和できる。ランダムでない欠測に対しては、観測情報だけでは欠測バイアスを十分に補正できず、モデルの力を借りた補正（尤度法）が必要となる。

注意すべき点として、不完全データが与えられたとき、上記3条件のいずれが成立しているか、すなわちどれが最も適切な統計的処理方法であるかは、不完全データの情報だけでは確かめることができない。それは、図1及び2の（A）や図3の散布図に示される真の姿がみえないからである。この事実、不完全データの統計的処理を限界付けるものである。処理法の適性を決める条件がデータから検証できないということは、換言すれば、処理法ごとの推定結果は検証不可能な前提に基づいているということである。このことから、検証不可能な前提を限界的に変化させていったときに推定結果がどれほど変わるかを確認する作業（「感度分析」と呼ばれる）の重要性が指摘されている。

間 真実（はざま まこと）

- 1 ウェイト調整法は、キャリブレーションや回答傾向スコアにより抽出ウェイトを調整する手法である。セミパラメトリック推定の利点がある。
- 2 多重代入法は、事後予測分布（観測値を与件とした欠測値の事後分布）をモデル化し、事後予測分布から互いに独立に生成する値を代入値として複数の疑似完全データを作成し、それぞれの推定結果を統合する手法である。推定精度の評価に利点がある。
- 3 尤度法は、興味の対象となるデータのデータ生成過程のみならず欠測発生時のデータ生成過程もモデル化して最尤推定を行うことで、欠測確率の推定に資する情報を興味の対象となる推定にも資する情報として活用する。