大規模コホート研究における効率的な多重代入法

古川恭治
放射線影響研究所統計部

多重代入法(MI)は, Missing at randomの下で適用できる柔軟で一般的な欠測データ解析手法であるが，大規模コホート研究などで，欠測を含む変数を異なる部分対象者群に対する多くの解析に用いる場合，いつ，どのようにデータを代入すべきかについては議論の余地がある．例として，放射線被曝が引き起こすさまざまな健康影響を評価するための主要な情報源となっている約十万人の原爆被爆者の寿命調査コホートを考える．このコホートでは，死因別の死亡や臓器別のがん罹患などの解析は，がん登録や死亡票などを情報源にコホート全体が対象となる一方，がん以外の疾患罹患情報やバイオマーカーなどの測定値は臨床健康診断調査対象のサブコホートに限られる．また，放射線量推定値や年齢，性別など，ほぼ欠測がない変数がある一方で，郵便調査などから得られる生活習慣因子の欠測割合は非常に高い．特に，多くの病気の重要なリスク因子である喫煙と放射線の交互作用を評価する場合，欠測を含む対象者を除くなどの単純な解析はバイアスや効率性低下につながりやすい[1]．では，そのような多くの異なる解析に用いられる欠測変数に対しMIを適用する場合，いつ，どのようにデータを代入すべきであろうか？

解析者の立場からは，欠測を含む変数に対し，一度だけMIを行い，コホート研究内の全ての解析に用いるのが最も簡単である．しかし，その場合，すべての将来的なサブ解析に関わるすべての関連が代入モデルに含まれなければならず，それは現実的ではない．MIの妥当性を保証するためには，個別の解析ごとにMIを行うことが望ましいが，小さなサブコホートを対象とする解析では，代入モデル推定のための統計的パワーが十分でなく，結果的にメインの解析での推定効率の減少につながるかもしれない．

本研究では，解析対象以外のコホート全体からの情報も利用して，代入モデルを多段階で推定することにより，メイン解析での推定の一貫性を保ったまま，効率性を向上させる新しいMIの手法を考える．ZとYをサブコホートをターゲットにした解析に含まれる変数とし，Zは欠測を含むが，Yは欠測がないものとする．また，Xをコホート全体に得られる変数で，Zの予測因子となるものとする．そして，Zの代入モデルとして，$f[Z|X,Y] \propto f_1[Z|X] f_2[Y|X,Z]$を考える．ここで，$f_1$はコホート全体，$f_2$は解析対象のサブコートからそれぞれ推定されるものとする．コホート内のより多くの情報を利用することで，代入モデルの予測精度を向上させ，結果的に，メインの解析での推定効率が上がることをシミュレーションと原爆被爆者データへの適用によって示す．大規模コホート研究等で，欠測を含む変数が，異なるサブコホートを対象とする多くの小規模解析で必要となる場合に，本手法は得に有用である．

[1] Furukawa K, Preston DL, Misumi M, Cullings HM. Handling incomplete smoking history data in survival analysis. *Stat Methods Med Res*, 26(2): 707-723, 2017.