未観測交絡因子が存在する場合の因果探索における欠測データ解析

宮林 剛士

因果探索とは，探索的データ分析の文脈でそもそもどんな因果関係が起きているかを調べるための手法である。Shimizu et al.(2006)の LiNGAM(Linear Non Gaussian Acyclic Model)は因果探索手法の中でも代表的な手法であり，実社会のデータに応用する事も含めて，今後もそのニーズは高まっていくことが予想される。しかしながら実際のところ実証的研究のプロセスにおいては，研究者が当初計画していたような完全な形でデータが得られない可能性があるのも事実である。欠測値に対する処理は様々な応用統計分野で研究がなされている。因果探索の分野でも欠測データを扱う研究が報告されているが，未観測交絡因子を仮定したモデルでの欠測データ解析についてはまだ十分な検討がなされていない。そこで本研究では，多重代入法の１つである予測平均マッチング法(predict mean matching)を用いて，未観測交絡因子が存在する場合の LiNGAM で欠測データを扱う際の手法を提案した。

未観測交絡因子を考慮した因果探索モデルとして本研究で使用したのは，Wei&Ruichu(2022)の MLCLiNGAM(LiNGAM with Multiple Latent Confounder)である。この手法は，回帰分析と独立性の検定を繰り返して因果方向を推定し，最大クリークから未観測交絡因子の所在を発見する，DirectLiNGAM(Shimizu et al.,2011)の拡張手法である。

シミュレーションでは，MCAR(Missing Completely At Random)と MAR(Missing At Random)という２種類の欠測データのメカニズムを扱った。設定した実験要因(係数の大小・サンプルサイズの大小・欠測割合大小)の水準の組み合わせごとに 2 つのメカニズムの欠測データを発生させ，平均値代入法とリストワイズ削除法，及び提案手法である予測平均マッチング法による方法を各々に適用した分析結果の正答率を算出した。結果として，欠測データメカニズムが MCAR の場合はどの手法もサンプルサイズが十分であれば，３つの手法において同程度の正答率が確認できた。しかしデータが MAR の場合には，平均値代入法とリストワイズ削除法の正答率は著しく低下し，提案手法の予測平均マッチング法による方法のみ MCAR の時とほぼ同水準の正答率を維持する結果となった。このことから，未観測交絡因子が存在する場合の因果探索においての，予測平均マッチング法の有効性が示された。

今後の課題としては，補完と分析の２段階ではなく１段階で欠測値を分析する手法や，因果探索と他の多重代入法の組み合わせも検討していく必要があるだろう。(行動統計科学)