Semiparametric Inference With Nonignorable Nonresponse Data

森川　耕輔

大阪大学大学院基礎工学研究科

データの欠測値，または欠損値はしばしばプライバシー・倫理的観点による回答拒否から起こるため，近年の情報化社会では深刻な問題となっている．データに欠測値がある場合，完全データに対する通常の統計的推測法は直接適応できない．そのため，しばしば回答が得られない対象者のデータを削除して，完全データを人工的に作成し，完全データに対する統計手法を用いる“リストワイズ削除”という方法が用いられる．しかしこの方法を用いる場合，推定量に対して，(i) 有効性の減少，及び (ii) バイアスが生じる，という２つの問題が起こる．１つ目の問題はサンプルサイズの減少から起こる．また２つ目の問題は，例えば，年収調査を行う場合年収が低い労働者は回答しにくい場合を考える．この場合，観測されたデータから計算される単純平均は，真の平均年収を過大評価してしまう．このような欠測値を含むデータに対して，適切に対処可能な統計的手法の開発を行った．

無視できない欠測値データとは，先ほどの年収の例のように，データが欠測するかどうか（欠測メカニズム）が欠測する応答変数にも依存しているデータである．このようなデータの解析には，欠測メカニズム以外にも応答変数のモデルの特定も要求され，どちらかを誤特定した場合，推定結果にバイアスが生じる．この仮定の強さゆえ，応用上，しばしば欠測メカニズムは無視できるという仮定が置かれている．本発表では欠測メカニズムが無視できない仮定の下で，(1) 応答変数に対するモデルの仮定を必要しないセミパラメトリック推測法を構築する．さらに，無視できない解析に対する既存の解析手法は，モデルの識別性のため，操作変数の存在といった検証不能な仮定が要求される．そのため，(2) データから検証可能な，操作変数を必要となしない識別性条件を提案する．

(1) では２つのセミパラメトリック推定量を提案する．１つ目の推定量は，欠測メカニズム以外にも“観測された応答変数”の分布の特定を要求する．この仮定は，通常の“完全データに対する応答変数”の分布の特定とは異なり，観測されたデータからモデルの検証が可能であり，より客観的な仮定となる．さらに提案する１つ目のセミパラメトリック推定量は，この観測された応答変数の分布が正しく特定されていた場合，セミパラメトリック漸近有効推定量となり，この観測された応答変数の分布を誤特定していても一致性・漸近正規性を有する．また２つ目の推定量は，観測データに対する応答変数の分布に関して，ノンパラメトリックな手法を用いるため，常にセミパラメトリック漸近有効推定量となる．ただ，ノンパラメトリック推定量の性質上，共変量の次元が大きいときはうまく機能しないことが想定される．そのため，状況に応じて提案する２つの推定量をうまく使い分ける必要がある．

また，モデルの識別性条件も提案する．提案する条件は，先行研究とは異なり，観測データから検証可能であり操作変数も必要としない．これまで，操作変数の仮定は必要不可欠なものであり，操作変数をデータが欠測している状況下で発見する手法は未だ提案されておらず，その存在を仮定せざるを得ない状況であった．従って，提案する識別性条件は応用上非常に有用であることが期待される．