交通データにおける無回答バイアスの修正方法

広島大学大学院　学生員　○原田慎也
広島大学大学院　正会員　藤原章正
広島大学大学院　正会員　杉恵頼寧

1. 研究の背景と目的

個人や世帯の交通行動分析には，パーソントリップ調査などのアンケート調査は不可欠である．しかし，アンケート調査には常に“無回答”という問題が生じる．無回答に伴う欠損値を含むデータ（不完全データ）のこれまでの処理方法としては，主として

①不完全データを分析対象から除外する
②回答データに重み付けをする
③回答データの平均値で欠損値を補填する

のようなものがあった[1]．

無回答がランダムに発生する場合，従来の処理方法でもそれほど問題は顕在化しない．しかし，調査への関心度や個人属性などに起因して無回答が生じる場合，回答と無回答の間に偏り，すなわちバイアスが生じ，従来の処理方法では結果が歪められる危険性がある．

したがって，不完全データに含まれる無回答バイアスを修正することは，調査の効率化，データの信頼性向上のために極めて重要である．

そこで本研究はアンケート形式による交通データに含まれる，欠損値を統計的手法に基づいて補填し隠れたバイアスを取り除くことを目的とする．特に回答データを基に複雑なパターンを示す欠損値を補填することのできる Imputation 法に着目し，欠損値を推定値に置き換えてバイアスを除去することの有効性について検討する．

2．無回答の定義

無回答には異なる２つの種類がある．

(1)ある特定の項目のみ無回答(item nonresponse)
ユニット（個人）としては多くの質問項目に答えているものの，一部の質問項目について明らかに誤っていたり答えなかったものがある場合

(2)調査項目全てに無回答(unit nonresponse)
個人が白票で返却，回答の拒否により，個人の回答に関する情報は一切入手できない場合

前者の item nonresponse はほとんどの交通データで発生する問題であり，特に回顧形式の質問や被験者にとってあまり利用したことのない代替案に関する質問などで生じやすい．例えばドイツの大規模交通実態調査である KONTIV において，過去の交通行動に関する質問では 54%の item nonresponse が生じたことが報告されている[2]．

後者の unit nonresponse は，発生率は相対的に低いものの対応は非常に困難であるとされている．調査の主旨，内容，調査方法にも大きく影響を受ける．例えば，広島のパーソントリップ調査では全体として約 12%の unit nonresponse が生じたことが報告されている[3]．

3．EMアルゴリズム

不完全データがある特定の無回答パターンを示す場合には，単純な Imputation 法を用いて欠損値を補填することができる．しかし現実データではこのようなケースはむしろ少なく，無回答パターンは非常に複雑である．そこで無回答パターンに応じて反復計算を行う必要が生じる．EMアルゴリズムはこの反復計算に適した方法の一つであり，不完全データから最尤推定される母数を，完全データから最尤推定される母数と関連づける方法である．概念的にも計算上でも極めて簡単である点が最大の利点である．

本研究では item nonresponse をとりあげ，EM アルゴリズムを利用した Imputation 法によって欠損値を補うことの有効性について，最も単純な平均値法（１で示した③の方法）と比較しながら検討する．

4．分析結果

4．1　アイテム平均値のバイアスの修正結果

９アイテム 1,000 ユニットの仮想データを作成する．９アイテムのうち１つは鉄道と自動車の選択

結果を表す2項データであり，残り8アイテムは選択を説明する要因（費用，乗車時間，アクセス時間，待ち時間）である．アイテム間には各々相関が存在する．このデータを完全データと考え，特定のアイテムの一部を欠損させた場合を不完全データとする．欠損の仕方を変化させて，多様な無回答パターンの下でImputation法の適用効果を測定する．

まず図1に1変数のみ(鉄道の費用)欠損させた場合のバイアスを含む平均値とImputation法により修正した後の平均値を欠損率を変えながら比較した結果を示す．欠損率が高くなればなるほど欠損データの平均値は下がるが，修正後のデータの平均値は完全データに近いことが分かる．

次に複数の変数が同時に欠損した場合について検討する．図2は複数変数が同時に20%ずつ欠損した場合の欠損変数の数とImputation法による平均値の改善率の関係を示したものである．1変数欠損の場合に比べて3変数欠損の場合，改善率は約72～84%まで低下する．アクセス時間(鉄道)と待ち時間(鉄道)の改善率が低いのは，他の変数間に比べてアクセス時間(鉄道)と待ち時間(鉄道)との相関が弱いことが原因として考えられる．通常の交通データでは変数間にある程度の相関があるので，Imputation法によりitem nonresponseに伴うバイアスが修正されることが期待できる．

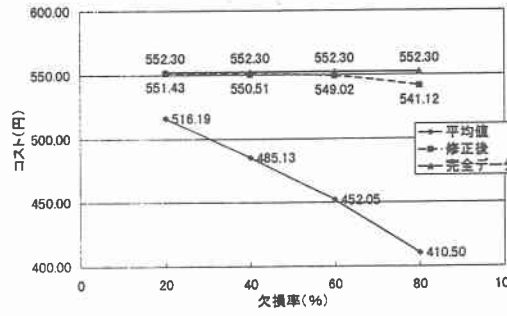

図1　費用(鉄道)の平均値と欠損率との関係

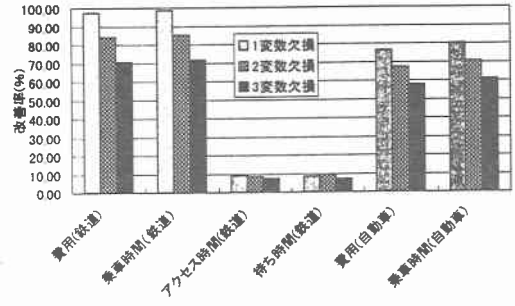

図2　欠損変数の数と平均値の改善率との関係

	完全データ	20%欠損	20%修正後	改善率(%)
乗車時間	−0.5542	−0.0807	−0.5298	94.85
費用	−0.0587	−0.0005	−0.0581	98.98
アクセス時間	−1.3188	−0.2596	−1.2275	91.38
待ち時間	−0.1869	−0.2553	−0.2245	45.00

表1　各変数のパラメータ値とその改善率

4．2　モデルパラメータのバイアスの修正結果

交通需要予測において無回答バイアスのより本質的な問題は予測モデルのパラメータに偏りが生じることである．そこで，交通機関選択の需要予測で頻繁に用いられる非集計ロジットモデルを事例として，不完全データとImputation法による修正後のデータに適用し，推定パラメータの比較を行ってitem nonresponseがモデルパラメータに及ぼす影響について分析する．

表1は，費用(鉄道)が20%欠損している場合の各変数のパラメータ推定値とその改善率を示したものである．欠損した場合，待ち時間を除くパラメータ値は完全データの時に求められる真値よりも絶対値が小さく過小評価となる．しかし，Imputation法による修正により，このようなバイアスは9割以上改善されており，この修正法の有効性が認められた．

5．結論

item nonresponseによるバイアス修正に対してImputation法は非常に有効であることが確認された．今後unit nonresponseに対する修正方法についてさらに十分な検討が必要である．

参考文献

1) Little,R. and D.Rubin (1987): Statistical Analysis with Missing Data, John Willey&Sons.
2) Richardson,A.J., E.S.Ampt, A.H.Meyburg (1995): Survey Methods for Transport Planning, Eucalyptus Press, p.314.
3) 広島都市圏交通計画協議会：昭和63年度広島都市圏パーソントリップ調査報告書－3:現況集計編, p.6.