# 国際学力調査データの回帰分析による多重代入法の検証

sp11103　板橋宗一　非線形物理学講座

【目的】

統計学は様々なデータを収集して分析する学問であるが、収集したデータには必ずといえるほど欠測値が生じる。統計分析をするなら欠損のないデータを採用するべきではあるが現実的に難しい。そのため生じた欠測値に関して対処が必要になり、多重代入法（Multiple Imputation）はその補完処理のひとつである。多重代入法は理論的には不完全なデータが完全なデータと統計的に妥当になるように補完する欠測値補完法であるが、今回は統計的に有効であるかを検証する。国際学力調査（PISA）のデータを用いて回帰分析の決定係数がどう変化するかを調べ多重代入法を検証する。

ー多重代入法ー

ハーバード大学統計学科のRubin（1978）によって欠測データの対処法として提唱された方法。ベイズ統計学の枠組みで構築され、ベイズ推定に用いられるマルコフ連鎖モンテカルロ法（MCMC）を用いて欠測値を推定し、たくさんのデータを作る。作られた多くのデータを基にして、平均と分散からさらに推定し全体との一体性を確保しつつ値を決める。

【方法】

一般に公開されているPISAのデータを使う。PISAとは国際的な学習到達度に関する調査で、１５歳を対象に数学力、読解力、科学力の三分野について三年ごとに行われるテストのことである。今回は三教科のテストのスコアの線形回帰分析によって多重代入法の検証を行う。多重代入法による補完時には、データを国ごとに分けた上で補完処理を行い、PISAのデータを（親の学歴や出身、学校への帰属意識や忍耐に関することなど）２０個ほどの説明変数にして扱った。それはデータが非常に大きくなると多重代入法の計算量が指数関数的に増大する性質があり処理ができなくなるので、このような処理を行った。多重代入法による決定係数の変化を見るので、多重代入法をする前に回帰分析をした結果を保存し補完処理後に回帰分析をした結果と比べる。

ーPISAデータの欠損状態ー

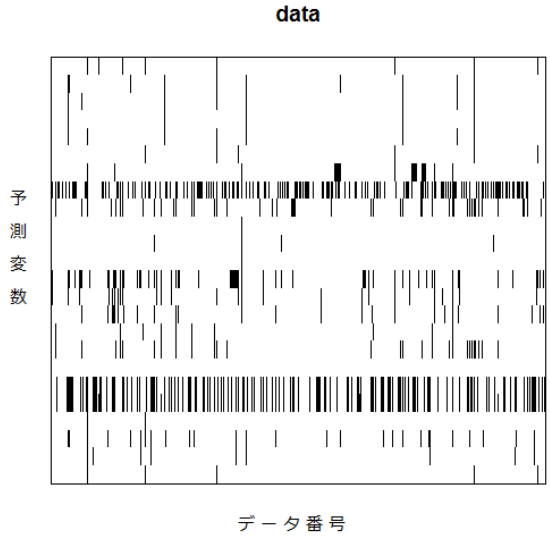


図１　PISAデータの欠損状態
（x軸が学生IDを意味している。y軸は段毎に違う１～24の予測変数を示し、黒い線が欠損している部分を表している。説明変数ごとに欠損の状態が違うことが分かる）

【結果】

表１　ｐ値の変化と使用データ量の変化

| | 処理前 | 処理後 |
|---|---|---|
| 数学 | 0,4077 | 0,4766 |
| 読解 | 0,3705 | 0,4398 |
| 科学 | 0,3780 | 0,4522 |
| 使用データ量 | 100917 | 476436 |
| (未使用データ量) | 384546 | 21289 |

多重代入法によって、使用したデータ量が４７万個に増え廃棄されたのは２万個に減少した。決定係数（説明率）が数学、読解、科学のすべてで有意に７％ほど上昇した。

ー参考文献ー

(1)"About PISA - OECD "＜http://www.oecd.org/pisa/aboutpisa/＞

(2)高橋将宜,伊藤孝之,(統計研究彙報 第71号　2014年3月)様々な多重代入法アルゴリズムの比較 ～大規模経済系データを用いた分析～(P39~82)

(3)Rubin, Donald B. (1978). “Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse,” Proceedings of the Survey Research Methods Section, American Statistical Association (P20~34)