ソフトウェア開発データにおけるデータ欠損の欠損メカニズム特定の検討

柿元 健[1]　安丸 怜那[1]

概要：本稿では，ソフトウェア開発データに含まれるデータ欠損について，その欠損メカニズムを特定する方法について検討する．

1. はじめに

定量的ソフトウェア開発管理の適用において，入力となるソフトウェア開発データに含まれるデータ欠損がしばしば問題となる．ソフトウェア開発データにデータ欠損が含まれると，定量的手法が利用できない，もしくは，利用できても定量的手法が有効に働かないことがある．

データ欠損が含まれるデータに対して定量的管理技法を適用するために，データ欠損を補完，もしくは削除する欠損値処理 [1] が適用される．欠損値処理を用いてソフトウェア開発データからデータ欠損を無くすことで，回帰分析のようなデータ欠損に対応できない定量的手法に適用できる，もしくは Analogy 手法のようなデータ欠損に対応している定量的手法であってもより有効に適用できるようになる．

多くの欠損値処理が提案されているが，既存の欠損値処理はデータ欠損に対して一律に補完や削除を行う．そのため，様々な要因によって生じるデータ欠損が含まれるソフトウェア開発データにおいては，既存の欠損値処理は有効に働かない場合がある．

一方で，任意のデータ欠損を含むソフトウェア開発データを作成するために，疑似的にデータ欠損を与える欠損メカニズム [2] が提案されている．欠損メカニズムはデータ欠損が生じる状況を考慮している．しかし，欠損メカニズムはデータ欠損を含むデータの作成に用いられる [4] だけで，データ欠損の分析には用いられていない．

ソフトウェア開発データに含まれるデータ欠損について，データ収集に関わった人物へのインタビュー等を行わない限り，データ欠損がどのような要因によって生じたかという「素性」を明らかにすることは困難である．そこで，欠損メカニズムを特定することでデータ欠損の「素性」を明らかにできるという仮説に基づいて，定量的手法の精度向上を目指す．

本稿では，データ欠損を含むソフトウェア開発実績データにおいて，ソフトウェア開発データに含まれるデータ欠損について欠損メカニズムの特定に向けた議論を行う．ソフトウェア開発データの欠損メカニズムを特定することができれば，欠損メカニズムに応じた欠損値処理を行い定量的管理手法をより有効に適用可能となることが期待される．

2. 欠損メカニズム

欠損メカニズムは，統計分野で提案されたもので，以下の 3 種類のメカニズムが提案されている．

MCAR ある値がデータ欠損となる確率は，データ中のいずれの値に依存しない．すなわち，ランダムにデータ欠損が生じる．例えば，不注意によってメトリクスの記録漏れが生じデータ欠損となった場合，データ中のいずれの値にも依存せずデータ欠損しているため，欠損メカニズムは MCAR といえる．

MAR ある値がデータ欠損となる確率は，データ欠損するメトリクス以外のあるメトリクスの値の大きさに依存する．例えば，ソフトウェアの規模をあらわす尺度（ファンクションポイント等）とコードレビューにおける発見バグ数が含まれるデータを考えた時，規模が小さなプロジェクトほどコードレビューが省略されやすく発見バグ数がデータ欠損となりやすいとする．この場合，コードレビューにおける発見バグ数がデータ欠損となる確率は規模の大きさに依存しているため，欠損メカニズムは MAR といえる．

NM ある値がデータ欠損となる確率は，その値自身の大きさに依存する．例えば，ソフトウェアの規模を表す尺度がメトリクスとして含まれるデータを考えた時，

[1] 香川高等専門学校電気情報工学科
National Institute of Technology, Kagawa College

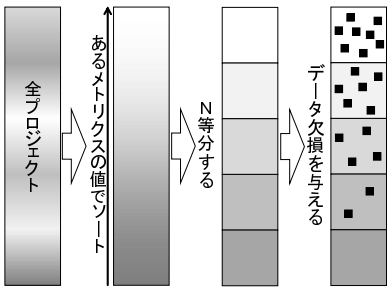

図 1 MAR，NM でデータ欠損与える例

規模が小さいプロジェクトほど，人員的余裕がないため規模自体が記録されにくく，データ欠損となりやすいとする．この場合，規模がデータ欠損となる確率は規模自体の大きさに依存しているため，欠損メカニズムは NM といえる．

3. 欠損メカニズムの特定

本節では，欠損メカニズムの特定方法について検討を行う．まず，欠損メカニズムの特定におけるデータの前処理について検討したうえで，ソフトウェア開発データから欠損メカニズムを特定する方法を検討する．

3.1 データの前処理

欠損値補完 本来，欠損値の補完のために欠損メカニズムの特定を行うことが目的であるが，欠損メカニズムの特定には欠損値の補完が必要である．

NM ではデータ欠損している箇所に本来入るはずの値に依存してデータ欠損が生じるため，データ欠損している箇所の補完値の算出は必須である．また，MAR においてもデータ欠損する確率が依存するメトリクスがデータ欠損している可能性もあるため，補完値の算出が必要な場合がある．そのため，データ欠損箇所について欠損値補完法を用いてデータ欠損箇所の推定値をあらかじめ算出しておき，欠損メカニズムの特定時にはその推定値を用いる．

欠損値補完を行う手法についても検討が必要である．

対数変換 ソフトウェア開発データの分析において，対数変換を行い，データ分布を線形に近づけることで精度が向上する場合がある [3]．MAR，NM では値の大きさに依存してデータ欠損が生じるため，対数変換を行うことで特定が容易となる可能性がある．

正規化 ソフトウェア開発データでは，メトリクスによって値域が大きく異なることが多い．そのため，正規化を行うことでメトリクスの値域による影響を除くことで特定が容易となる可能性がある．

3.2 特定方法

既存の欠損メカニズムは，MCAR はランダム要素，MAR，

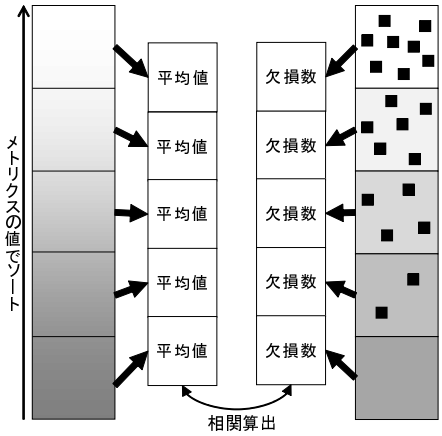

図 2 欠損メカニズムの特定方法の例

NM は何らかの値に依存した確率要素が含まれるため，データ欠損を与える手法を逆順で実行して特定することはできない．そこで，図 2 の様に，依存すると仮定する各メトリクスと欠損メカニズムを特定するメトリクスとについて，依存メトリクスの値でソートしたうえで n 等分し，分割ごとに依存メトリクスの平均値と特定メトリクスの欠損値数を算出し，何らかの相関を算出するといった方法が考えらえる．

ただし，相関係数はあくまでも相対的な関係にしか意味がないため，算出された相関係数の値から欠損メカニズムを特定する方法については検討が必要である．また，どのような相関を用いるのかや，分割数の設定についても検討が必要である．

4. おわりに

本稿では，ソフトウェア開発データのデータ欠損について欠損メカニズムの特定方法について，単独の欠損メカニズムの特定に限って検討を行った．ワークショップでは，特定した結果を示したうえで，特定方法だけでなく，既存の欠損メカニズム以外に，ソフトウェア開発データで考えられる欠損メカニズムについても議論したい．

参考文献

[1] Kromrey, J. and Hines, C., :Nonrandomly Missing Data in Multiple Regression: An Empirical Comparison of Common Missing-data Treatments, Educational and Psychological Measurement, vo.54, no.3, pp.573–593, (1994).

[2] Little, R.J.A., and Rubin, D.B., Statistical Analysis with Missing Data, 2nd edition, John Wiley and Sons, New York, (2002).

[3] 門田暁人，小林健一：線形重回帰モデルを用いたソフトウェア開発工数予測における対数変換の効果，コンピュータソフトウェア Vol. 27, No. 4, pp. 234-239, (2010).

[4] Strike, K., El Eman, K., and Madhavji, N. :Software Cost Estimation with Incomplete Data, IEEE Transaction of software engineering, vol.27, no.10, pp.890-908, (2001).