ソフトウェア開発工数見積もりにおけるカテゴリ変数の扱い

角 田 雅 照[†1]　　天 嵜 聡 介[†2]

本稿では，重回帰分析によるソフトウェア開発工数見積において，データに含まれるカテゴリ変数に対し，ダミー変数化を行った場合，層別を行った場合，階層線形モデルを適用した場合の見積精度を比較した．

Handling Categorical Variables on Software Development Effort Estimation

Masateru Tsunoda[†1] and Sousuke Amasaki[†2]

In this paper, we compared estimation accuracy of software development effort estimation by multiple regression analysis when making dummy variables, stratification, or Hierarchical Linear Model is applied to a categorical variable.

1. はじめに

ソフトウェア開発工数見積もりは，ソフトウェア開発プロジェクトの計画立案，及び進捗管理の基礎となるものであり，プロジェクトマネジメントに必要不可欠な要素の一つである．数学的モデルに基づいてソフトウェア開発工数見積もりを行う場合，過去のプロジェクトにおいて収集，蓄積されたデータを用いて工数見積モデルを構築し，見積もりを行う．見積モデルとして，重回帰分析が広く用いられている．

データに含まれる変数には，順序尺度以上の変数（開発規模など）と，名義尺度のカテゴリ変数（開発言語など）の 2 種類が存在する．カテゴリ変数を重回帰分析で扱う場合，変数の値に基づきデータを層別する場合と，カテゴリ変数をダミー変数化する場合がある．ソフトウェア開発データにはカテゴリ変数が含まれることが非常に多く，モデル構築時に層別を行うかダミー変数化するかの判断を求められることが多い．しかし，どちらの方法でカテゴリ変数を扱ったほうが見積精度が高くなるのかは，これまで明確でなかった．

そこで本稿では，重回帰分析で見積モデルを構築する際に，どちらの方法でカテゴリ変数を扱うべきであるのかを議論するために，カテゴリ変数を層別した場合とダミー変数化した場合の工数見積もりの精度について比較する．さらに，層別されたデータを分析するための手法である階層線形モデル（Hierarchical Linear Model）[4]を適用した場合の見積精度とも比較する．

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
†2 岡山県立大学
Okayama Prefectural University

2. カテゴリ変数の扱い

2.1. データの層別

データの層別とは，カテゴリの変数の値に基づいてデータを分割することであり，分割したデータごとに見積モデルを構築する．データを層別することにより，個別性の高いモデルを構築することができる．開発工数を y，開発規模などの説明変数を $x_1, x_2, \ldots, x_k$ とすると，重回帰分析に基づく見積モデルは以下のようになる．

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (1)$$

ここで，β_0 は回帰定数，$\beta_1, \beta_2, \ldots, \beta_k$ は偏回帰係数，は ε 誤差項である．データを層別した場合，カテゴリ変数の値によって偏回帰係数が異なる，個別性の高いモデルを構築できる．ただし，それぞれのモデル構築に用いるデータ数が少なくなるという問題がある．

2.2. ダミー変数化

ダミー変数化とはカテゴリを数値化する方法であり，ある変数に含まれるカテゴリ数から 1 を減じた個数の変数を新たに定義し，あるカテゴリと一致するなら値を 1，そうでないなら 0 とする．ダミー変数化の場合，個別性の高いモデルは構築できないが，モデル構築に必要なデータ数は層別する場合に比べて少ない．重回帰分析を適切に行う場合，一般に説明変数の 5 倍のデータ数が必要であるといわれており[3]，カテゴリ変数に含まれるカテゴリ数を a，カテゴリ変数以外の説明変数

の数を b とするとき，層別の場合，必要なデータ数は $5ab$ となるが，ダミー変数化の場合 $5\ (a + b - 1)$となる．

2.3. 階層線形モデル

階層線形モデル（Hierarchical Linear Model; HLM）[4]では，重回帰分析では定数となる回帰定数や偏回帰係数を目的変数とした式（ランダム効果）を導入しており，それぞれに以下の式を設定できる．

$$\beta_i = \gamma_i + \mu_{ij} \qquad (2)$$

ここで，γ_iはカテゴリ間の平均，μ_{ij}はカテゴリjの誤差である．HLM では，層別に個別性の高いモデルを得られる．また，モデルの構築時に，データ全体から得られる統計量も利用しているため，単純なデータ層別よりも，見積精度が高まることが期待される．

3. 実験

NASA で収集されたソフトウェア開発プロジェクトのデータ[1]を用い，カテゴリ変数をダミー変数化して重回帰モデルを構築した場合，データを層別して重回帰モデルを構築した場合，HLM により見積モデルを構築した場合の見積精度を比較した．重回帰モデルの構築では，AIC に基づくステップワイズ変数選択を適用した．HLM では変数選択を行わなかったため，比較のため複数の変数の組み合わせ（説明変数が開発規模のみ，変数選択を行わないなど）で実験した．見積精度の評価指標として，*MRE* (Magnitude of Relative Error) と *ABRE* (Absolute Balanced Relative Error)[2]の中央値を用いた．

データには93件のプロジェクトと22個の変数が含まれており，プロジェクト名と開発年を除いた20個の変数を用いた．カテゴリ変数が 3 個含まれていたため．予備実験において最もモデルに影響の大きいカテゴリ変数（アプリケーション種別）を特定し，その変数に基づいて層別及び HLM の構築を行った．モデル構築時に，3-fold cross validation を適用するためデータ数が 3 件未満のカテゴリは除外し，データ数 83 件，アプリケーション種別に含まれるカテゴリ数 8 個となった．

表 1 *MRE*, *ABRE* 中央値の比較

	MRE	*ABRE*
ダミー変数 変数選択あり	**26.1%**	**31.1%**
ダミー変数 変数選択なし	48.6%	51.8%
ダミー変数 開発規模のみ	30.7%	35.3%
層別 変数選択あり	51.4%	58.0%
層別 変数選択なし	58.0%	92.0%
層別 開発規模のみ	**32.7%**	**45.7%**
HLM 切片にランダム効果	31.8%	40.1%
HLM 開発規模の係数にランダム効果	32.8%	39.6%
HLM 切片と開発規模の係数にランダム効果	**32.6%**	**39.0%**
HLM 開発規模のみ 切片にランダム効果	**33.2%**	**34.3%**
HLM 開発規模のみ 係数にランダム効果	31.2%	36.4%
HLM 開発規模のみ 切片と係数にランダム効果	33.3%	37.2%

実験結果を表 1に示す．*ABRE* 中央値に着目すると，ダミー変数化の場合の見積精度が最も高く，HLM は若干精度が低かった．層別の場合の精度が最も低かった．データセットを増やして実験する必要があるが，カテゴリ変数を扱う際には，ダミー変数化が最も適している可能性がある．層別の精度が低かったことから，個別性の高いモデルを構築したい場合，層別よりも HLM のほうが適している可能性がある．なお，紙面の都合上結果を省略するが，カテゴリ毎のデータ数の多寡と精度には関連が見られなかった．

4. おわりに

本稿では，カテゴリ変数の扱いの違いによる見積精度の差を比較した．ワークショップでは，工数見積モデル構築時にカテゴリ変数をどのように扱うべきであるかについて議論したい．

謝辞 本研究の一部は，「次世代 IT 基盤のための研究開発」の委託に基づいて行われた． また，本研究の一部は，文部科学省科学研究補助費（若手 B：課題番号 22700034）による助成を受けた．

参考文献

[1] Boetticher, G., Menzies, T., and Ostrand, T.: PROMISE Repository of empirical software engineering data http://promisedata.org/ repository, West Virginia University, Department of Computer Science (2007).

[2] Miyazaki, Y., Terakado, M., Ozaki, K., and Nozaki, H.: Robust Regression for Developing Software Estimation Models, *Journal of Systems and Software*, vol.27, issue 1, pp.3-16 (1994).

[3] Tan, H. B., Zhao, Y., and Zhang, H.: Conceptual data model-based software size estimation for information systems, *ACM Trans. Softw. Eng. Methodol.*, Vol.19, No.2, pp.1-37 (2009).

[4] 筒井淳也，不破麻紀子：マルチレベル・モデルの考え方と実践，理論と方法，Vol. 23，No. 2，pp.139-149（2008）．