多変量予測モデルにおける目的変数の選択について

門田暁人†

ソフトウェア開発工数の予測では「開発工数」を目的変数とし，ソフトウェアの残存バグ数の予測では「バグ数」を目的変数とすることが自然である．しかし，生産性（規模あたりの工数）を予測してから規模を乗じて工数を導出することも可能であるし，平均要員数を予測し，開発期間を乗じて工数を導出することも可能である．また，バグ密度を予測し，規模を乗じてバグ数を導出することも可能である．いずれの方法を採用しても，一見，同じ結果が得られるように思われるが，実際には異なる結果が得られる．本論文では，問題提起とケーススタディを示す．

On Selection of Objective Variables in Multivariate Predictor Models

Akito Monden†

To estimate the software development effort, it seems clear that the objective variable is the effort. But is it really clear? We could estimate the productivity; then, derive the effort by multiplying it by the product size. Also, we could estimate the average number of staffs; then, multiply it by the project duration to derive the effort. Do these yield a same result? The answer is no. Similarly, to predict the number of bugs in a software module, the objective variable can be either the number of bugs or the bug density. This paper presents a problem statement and a case study.

1. はじめに

ソフトウェア開発において，プロジェクトマネジメントを効果的に行うためには，QCD（Quality, Cost, Delivery）に関する値を予測もしくは見積もることが必須である．特に，開発工数と残存バグ数は多くの開発現場において重要な見積もり対象となるため，数多くの予測モデルが提案され，用いられてきた．

予測モデルでは，当然のことながら，予測したい値を目的変数に設定する．ソフトウェア開発工数の予測では「開発工数」を目的変数とし，残存バグ数の予測では「バグ数」を目的変数とすることが自然である．ところが，現実には，生産性を予測してから規模を乗じて工数を導出することも可能であるし，平均要員数を予測し，開発期間を乗じて工数を導出することも可能である．また，バグ密度を予測し，規模を乗じてバグ数を導出することも可能である．

このように，予測したい値を間接的に導出したとしても，同じ結果が得られるかというと，実はそうではない．筆者による予備実験では，ステップワイズ重回帰分析を用いてモデル構築した場合，「バグ数」の予測と「バグ密度」の予測では，異なる説明変数が選択される場合があり，全く異なるモデルが得られた．異なるモデルが得られるということは，モデルの適合度，残差分布，ロバスト性などに違いがみられるということであり，いずれのモデルが優れているかは自明でないことになる．つまり，間接的な導出が悪いとは一概にいえず，実際に両方のモデルを構築してみて，統計的に良いと判断できる方を選択すべきであると思われる．しかし，従来，このような間接的な導出についてはほとんど検討されておらず，間接的な導出を検討する価値があるかどうかも不明であった．

以上の問題提起から，本論文では，予測したい値を直接導出した場合と，間接的に導出した場合について，どの程度予測精度が異なるのかを分析する．本論文では，特に，ソフトウェア開発工数の予測を対象とし，モデリング手法として，線形重回帰分析とランダムフォレストを用いた事例について述べる．

2. 実験

2.1. 概要

Desharnais データセット[1]を用いて，開発工数を説明変数するモデルを構築した場合と，開発工数を間接

†奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

的に導出するモデルを構築した場合について，モデルの残差，選択される変数，各変数の重要度などを比較する．予測モデルとしては，線形重回帰モデル，および，ランダムフォレストを用いる．

2.2. 条件

Desharnais データセットに含まれる 77 プロジェクトを全て用いてモデルを構築する．説明変数として，AdjFPs（調整済ファンクションポイント），Duration（開発期間），ExpEquip（開発チーム経験年数），ExpProjMan（プロジェクト管理者の経験年数），DevEnv（開発言語）を用いた．DevEnv は 2 値変数化して DevEnv1 と DevEnv2 の 2 変数を設けた．

目的変数として，(1)開発工数，(2)開発工数／ファンクションポイント（＝規模あたり工数＝生産性），(3) ファンクションポイント／開発工数（＝生産性の逆数）の3つについてモデル構築した．(2)については，予測後に AdjFPs を乗じて開発工数を導出した．(3)については，予測後に，予測値の逆数に AdjFPs を乗じて開発工数を導出した．なお，AdjFPs，Duration，ActualEffort の 3 変数については，あらかじめ対数変換を行った．工数予測の精度評価も，対数変換後の値に対して行った．

モデル構築には統計ツール R を用い，線形重回帰モデルに対しては変数選択法として， AIC(Akaike's Information Criteria)による変数増減法を用いた．

2.3. 結果

実験結果を表1に示す．表中，MMRE，MMER，MAE はそれぞれ，Mean Magnitude of Relative Error（相対誤差平均），Mean Magnitude of Error Relative to Estimate（分母を予測値とする相対誤差平均），Mean Absolute Error（絶対誤差平均）である．

表1より，重回帰モデルでは，「FP／工数」モデルの残差がもっとも小さくなった．一方，ランダムフォレストでは，「工数／FP」モデルの残差がもっとも小さくなった．このことから，開発工数を直接予測するのではなく，間接的に導出することが，場合によっては効果的であることが示唆された．

選択された説明変数を調べたところ，3つの重回帰モデルでは同じ変数セットが選択された（AdjFPs, Duration, DevEnv1, DevEnv2）．一方，ランダムフォレストについては，各モデルにおける各説明変数の重要度の尺度である IncNodePurity を分析した（表 2）．表 2 では，各モデルにおける説明変数間の IncNodePurity の相対的な差に着目されたい．表より，いずれのモデルもファンクションポイント（AdjFPs）の重要度が最も高く，次いで，Duration（開発期間）の重要が高かった．ただし，

表 1. 実験結果

	目的変数	工数の残差		
		MMRE	MMER	MAE
重回帰モデル	工数	0.0372	0.0371	0.3030
	工数／FP	0.0376	0.0376	0.3064
	FP／工数	0.0371	0.0371	0.3029
ランダムフォレスト	工数	0.0290	0.0287	0.2318
	工数／FP	0.0276	0.0273	0.2206
	FP／工数	0.0276	0.0276	0.2228

表 2. ランダムフォレストにおける各変数の重要度

	モデルの目的変数		
	工数	工数／FP	FP／工数
Duration	13.35	0.171	0.0388
ExpEquip	3.39	0.125	0.0249
ExpProjMan	3.10	0.102	0.0214
AdjFPs	19.62	0.463	0.0908
DevEnv1	3.28	0.123	0.0362
DevEnv2	2.21	0.101	0.0292

3番目以降の重要度の順序は，各モデルで異なっている．このことから目的変数を変えると（ある意味当然ではあるが）異なるモデルが得られることを確認した．

3. おわりに

本稿では，(1)開発工数予測モデル，(2)規模あたりの工数を予測してから規模を乗じて工数を求めるモデル，(3)工数あたりの規模を予測してから，その逆数に規模を乗じて工数を求めるモデル，の 3 つを構築し，これらのモデルの予測結果が等しくならないことを示した．また，開発工数を直接予測する(1)のモデルが，必ずしも最適な（すなわち，もっとも残差の小さい）モデルとなるとは限らないことを示した．

今後は，より多くのデータセットを用いて，工数を間接的に導出する他のモデル（例えば，平均要員数を予測し，開発期間を乗じて工数を導出するモデル）についても評価していく予定である．また，バグ数の予測についても実施する予定である．

謝辞

本研究の一部は，文部科学省科学研究費補助金（基盤研究（C）：課題番号 22500028）に基づいて行われた．

参考文献

[1] Desharnais, J.M., "Analyse statistique de la productivitie des projets informatique a partie de la technique des point des function", Masters Thesis, University of Montreal, 1989.