

自主研究

生産性に基づくソフトウェア開発工数 予測モデル

生産性に基づくソフトウェア開発工数予測モデル

門田 暁人 松本 健一 奈良先端科学技術大学院大学 情報科学研究科
大岩佐和子 押野 智樹 一般財団法人 経済調査会 調査研究部 第二調査研究室

1 はじめに

ソフトウェア開発プロジェクトにおいて、開発工数（人月または人時）の見積もりは、必要な資源の確保やスケジュール立案を行う上で必要である。本稿では、一般財団法人経済調査会において2007～2011年度に収集されたソフトウェア開発プロジェクトデータを用い、式(1)の工数予測モデルの構築を行う。

$$\left. \begin{aligned} \text{開発工数} &= \text{開発規模} \times \text{生産性} \\ \text{生産性} &= \sum_{j=1}^n k_j \times \text{プロジェクト特性}_j + C \end{aligned} \right\} (1)$$

このモデル式では、開発工数を直接予測するのではなく、生産性（ファンクションポイントで表す開発規模あたりの工数）をまず推定し、その後、開発規模（ファンクションポイント：以降はFPとする）に生産性を乗じて工数を導出する点に特徴がある。生産性は、複数のプロジェクト特性（適用業種、発注要件の明確度・安定度、開発スケジュール要求など）の線形結合として表され、係数 k_j および定数 C はステップワイズ重回帰分析により推定される。

このモデルの特長は、次の点が分かりやすいことにある。

- 各プロジェクト特性が生産性にプラスに働くのか、マイナスに働くのかが明らかとなる。（係数 k_j の符号を見ることにより）
- 各プロジェクト特性が生産性に寄与する度合いが明らかとなる。（係数 k_j の標準化された値（標準化偏回帰係数）を見ることにより）
- 生産性に寄与しないプロジェクト特性も明らか

※本稿においてlogは、eを底とする自然対数を用いる。

となる（ステップワイズ変数選択により、除外されるため）。

なお、本稿では、精度向上のための実用上の工夫として、生産性を予め対数変換*してから重回帰分析を行う。したがって、式(1)の後半部分は、

$$\log(\text{生産性}) = \sum_{j=1}^n k_j \times \text{プロジェクト特性}_j + C \quad (2)$$

となることから、

$$\text{生産性} = \exp\left(\sum_{j=1}^n k_j \times \text{プロジェクト特性}_j + C\right) \quad (3)$$

と表される。

以降、2章では、モデル構築の前処理として、プロジェクトの選定、説明変数の選定、外れ値除去、カテゴリ変数の取り扱い、プロジェクト特性値の対数変換について述べる。3章では、得られたモデルと予測結果を述べる。4章では、関連研究を紹介する。5章は本稿のまとめである。

2 モデル構築の前処理

2.1 プロジェクトの選定

予測モデルの構築にあたっては、欠損値のないデータセットが必要である。そこで、次の変数が欠損していないプロジェクトを選定した。

- 実績工数（人月）
 - 実績開発規模（FP）
 - 実績開発期間（月）
 - 生産性変動要因（機能性、信頼性など10変数）
- さらに、以下の開発5工程すべてにおいて作業が実施されているものを選定した。

- 基本設計（基本設計Aまたは基本設計Bにおいて作業が実施されている）
- 詳細設計
- PG設計・製造
- 結合テスト
- 総合テスト（ベンダ確認）

この結果、プロジェクト数は204件となった。

2.2 説明変数の選定

生産性を推定するための説明変数としては、次のものを選定した。各変数の定義については、経済調査会のレポートを参照されたい[7]。

- 生産性変動要因(10変数)
 - 発注者側の要因6変数（機能性、信頼性、プラットフォーム、開発スケジュール要求、発注要件の明確度と安定度、ユーザの参画度合い）＜生産性の影響度合いを5段階に区分＞
 - 受注者側の要因4変数（先行モデルの流用と標準モデルの採用、プロジェクト管理者の経験と能力、アナリストの経験と能力、SE・プログラマの経験と能力）＜生産性の影響度合いを5段階に区分＞
- 開発スピード（開発規模÷開発期間）
 - 「開発期間」ではなく「開発スピード」を採用した理由は、開発スピードの方が生産性に寄与すると考えられるためである。なお、ここでの開発期間は、本来は見積もり値を用いるべきであるが、欠損値が非常に多かったために実績値を用いている。
- 発注者分類
 - 民間企業と公共機関では生産性に差がある可能性がある。
- 開発規模(FP)
 - 式(1)では、開発規模×生産性によって工数を求めるが、開発規模自体も生産性に影響するため、説明変数として採用した。
- 適用業種
 - 一般に、金融系は（高い信頼性が要求されるため）他の業種よりも生産性が低い[3]。

2.3 外れ値の除去

工数予測モデルは、通常の開発において、どの程度の工数がかかるかを見積もるためのものである。そのため、特異なプロジェクト（失敗プロジェクトなど）は、予測モデルに組み込むべきではない。

本稿では、生産性の外れ値に着目する。生産性が外れ値となっているプロジェクトは、何らかの理由により、小規模であるのに莫大な工数がかかっていたり、わずかな工数で大規模開発を成し遂げているものであり、特異なプロジェクトであるといえる。

なお、工数予測においては、工数の外れ値に着目するのが一見自然であるが、工数が外れているプロジェクトというのは、単に工数が小さい（小規模な）もしくは大きい（大規模な）プロジェクトであり、それだけでは特異なプロジェクトであるとはいえない。仮に、

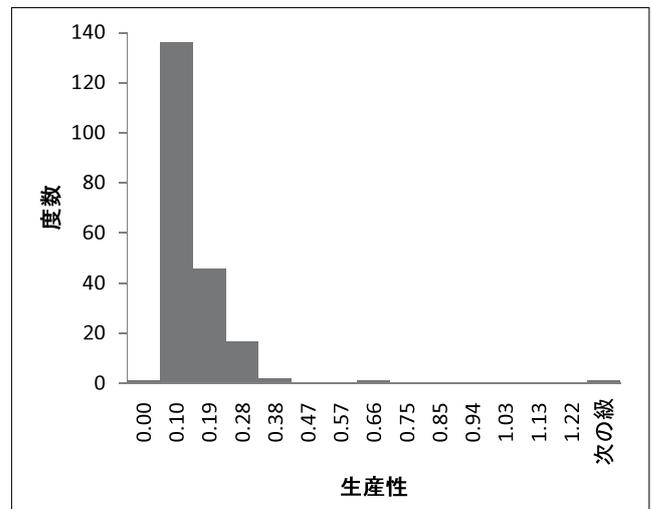


図1 生産性の分布

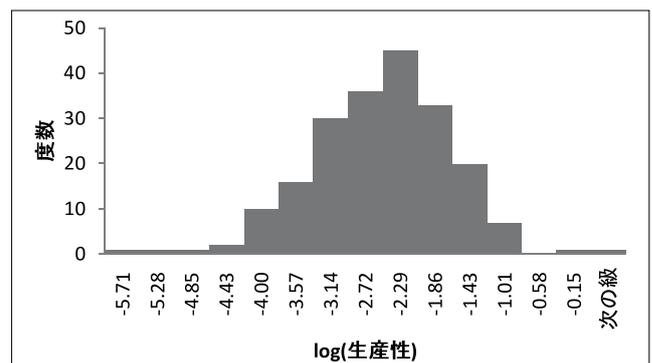


図2 log(生産性)の分布

そのようなプロジェクトを除外すると、逆に予測性能の低下を招く。なぜならば、工数が小さいものから大きなものまでを精度よく予測するためには、モデル構築時にもそのようなプロジェクトを含んだデータセットが必要となるためである。

生産性の分布を図1に示す。図1より、生産性は右側に広がる対数正規分布の形をしているため、対数変換してから外れ値を除去することにする。対数変換後の生産性の分布を図2に示す。本稿では、図2の分布において、平均値±3σより外側の値を外れ値として除外する。ここで、σは標準偏差であり、平均値±3σより外側の値を除外することは、正規分布における生起確率が0.3%未満のプロジェクトを除去することを意味する。本稿のデータセットでは、3件が外れ値に該当した。

2.4 カテゴリ変数の取り扱い

2つの説明変数「発注者分類」と「適業業種」はカテゴリ変数である。これら変数では、該当プロジェクトが著しく少ないカテゴリが存在するため、複数のカテゴリを集約した。集約されたカテゴリは下記の通りである。

- 発注者分類(2カテゴリに集約)
 - 民間企業
 - 民間企業以外 ⇒ 公共機関(府省庁、独立行政法人、自治体)、その他
- 適業業種(8カテゴリに集約)
 - 製造業
 - 電気・ガス・熱提供・水道業
 - 情報通信業
 - 運輸業
 - 卸売・小売業
 - 金融・保険業
 - 公務
 - その他

なお、欠損値は「その他」に含めている。

これらの変数をモデル構築に用いるにあたっては、ダミー変数化(2値変数化)を行った。発注者分類については、「民間企業ならば1そうでないならば0」とな

る2値変数に変換した。同様に、適業業種についても、「製造業ならば1そうでないならば0」、「電気・ガス・熱提供・水道業ならば1そうでないならば0」など、カテゴリごとに2値変数に変換した。

2.5 対数変換

説明変数または目的変数の値の分布が図1のように右側に広がる対数正規分布の形をしている場合、対数変換して正規分布に近づけることで、性能の良い線形重回帰モデルを構築できることが経験的に知られている[2]。

本稿では、説明変数のうち、開発スピード(開発規模÷開発期間)と開発規模(FP)の2変数について、対数変換してからモデル構築した。

3 予測モデルと予測結果

3.1 予測モデル

得られた予測モデルは、式(4)のとおりである。

開発工数 = 開発規模 × 生産性

$$\text{生産性} = \exp\left(\sum_{j=1}^n k_j \times N_j + C\right)$$

プロジェクト特性 N_j	N_j の値	k_j または C の値
N_1 : 発注者分類(民間企業)	true:1 false:0	0.323
N_2 : 適業業種(製造業)	true:1 false:0	-0.573
N_3 : 適業業種(金融・保険業)	true:1 false:0	0.319
N_4 : log(開発スピード)	log(開発スピード)	-0.636
N_5 : log(開発規模)	log(開発規模)	0.375
N_6 : 信頼性	1~5の値	-0.175
N_7 : 開発スケジュール要求	1~5の値	-0.157
N_8 : 発注要件の明確度・安定度	1~5の値	-0.087
C : 定数	-	-1.230

(4)

ここで、 k_j の符号に着目すると、「発注者分類(民間企業)」「金融・保険業」「開発規模(FP)」の3変数はプラスである。ここでの生産性の定義は「開発規模あたりの工数」であるので、値が大きいほど1 FPあたりより多くの工数を要する、つまり、生産性が悪いことを意味する。したがって、「発注者分類(民間企業)」、「金

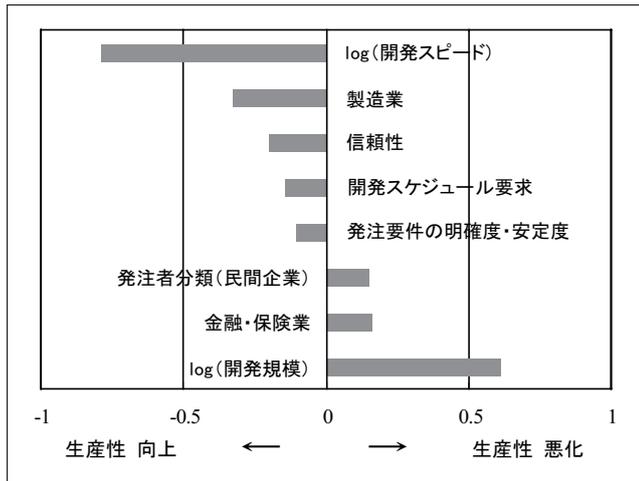


図3 各説明変数の標準化偏回帰係数

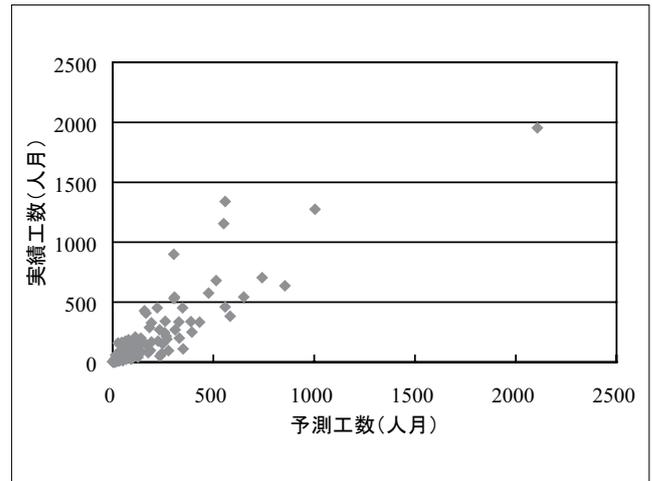


図4 予測結果

融・保険業」「開発規模 (FP)」の3変数は、生産性を悪化させる要因である。このことは、①公共よりも民間からの委託の方が生産性が悪化する傾向にある、②金融・保険業は他の業種よりも生産性が悪化する傾向にある、③開発規模が大きいほど生産性が悪化する傾向にある、ことを示している。いずれも直感に合う結果である。

一方、「製造業」「開発スピード」「信頼性」「開発スケジュール要求」「発注要件の明確度・安定度」については生産性を向上させる要因となった。この解釈としては、①製造業は他の業種よりも相対的に生産性が良い、②開発スピードが速いほど生産性が良い、③信頼性要求が低いほど生産性が良い、④開発スケジュール要求が緩いほど生産性が良い、⑤発注要件が明確・安定であるほど生産性が良い、ことを示している。これらについても、直感に合う結果となった。

一方、選択されなかった変数、つまり、生産性にほとんど寄与しなかった変数は次の通りである。

- 生産性変動要因(発注者):機能性
- 生産性変動要因(発注者):プラットフォーム
- 生産性変動要因(発注者):発注者の参画割合
- 生産性変動要因(受注者):先行モデルの流用と標準モデルの採用
- 生産性変動要因(受注者):プロジェクト管理者の経験と能力
- 生産性変動要因(受注者):アナリストの経験と能力

- 生産性変動要因(受注者):SE・プログラマの経験と能力
- 適用業種:電気・ガス・熱提供・水道業
- 適用業種:情報通信業
- 適用業種:運輸業
- 適用業種:卸売・小売業
- 適用業種:公務

上記を見ると、生産性変動要因のうち受注者側の要因はすべて選択されなかったことになる。このことは、開発の生産性は、主に発注者側の事情で決まることを意味している。また、適用業種については、「金融・保険業」と「製造業」の2業種は生産性に寄与するが、他の業種は生産性に大差がないことが分かった。

次に、式(4)の各プロジェクト特性について、生産性への寄与の度合いを見るために、標準化偏回帰係数を導出した。結果を図3に示す。図3において、標準化偏回帰係数の値が大きいほど(グラフの右に行くほど)生産性が悪くなる方に寄与していることを示す。図3より、生産性の悪化にもっとも影響する要因は開発規模である。逆に、生産性をもっとも高める要因は開発スピードである。開発スピードが大きい(速い)プロジェクトは、短期間で大規模開発を行ったことを意味するが、どのようなプロジェクトで開発スピードが速くなっているのかについては、現時点では明らかでない。おそらく、テストをあまりしなくてよいといった何らかの事情のあるプロジェクトが含まれていると推測される。今後、さらなる分析が求められる。

3.2 予測結果

前節で得られたモデル式を用いた予測結果のグラフを図4に示す。横軸が予測工数、縦軸が実績工数である。相対誤差平均は51.7%、相対誤差中央値は35.0%であった。多組織から計測されたデータを用いた予測モデルとしては、悪くない精度である。

ただし、図4の結果は、モデル構築用データへの当てはめの結果であることに注意されたい。今後の課題として、未知データに対する予測性能の評価を行うことが求められる。

4 関連研究

従来、開発プロジェクトの実績データを用いて、開発工数を予測する研究が盛んに行われてきた。予測方法としては、線形重回帰モデル、log-log重回帰モデル[4]、アナロジーベース法[6]などが用いられてきた。ところが、これらの予測方法は、ソフトウェア開発企業において必ずしも活用されていない。その原因の一端は、(たとえ予測精度が高いとしても)得られた予測モデルに説得力がない、もしくは、理解しにくいことにあると考えられる。

一方、COCOMOのような予め定義された工数予測モデルを、開発組織に合わせてパラメータ調整して用いるというアプローチの方が、企業において採用されている[1]。その理由は、おそらく、COCOMOの「工数=規模^c×生産性調整係数(工数変動要因)」という式が直感に合うのだと思われる。近年注目されているCoBRA法[5]も、「工数=α×規模×(1+Σコスト変動要因)」という、規模をベースにした式を採用している。本稿もこれに近い式を採用しており、生産性に影響する要因が直感的に分かりやすいことから、開発現場で試用いただくことを期待している。

5 まとめ

本稿では、生産性に着目した開発工数予測モデルを構築し、その予測精度を評価した。特に、モデル構築の前処理について詳しく説明するとともに、得られたモデルの解釈について述べた。

結果のまとめは以下のとおり。

- ・式(4)の開発工数予測モデルが得られた。
- ・上記式の予測精度の評価としては、多組織から計測されたデータを用いた予測モデルとしては、悪くない精度であった。
- ・「発注者分類(民間企業)」、「金融・保険業」、「開発規模(FP)」の3変数は、生産性を悪化させる要因であることが分かった。
- ・「製造業」、「開発スピード」、「信頼性」、「開発スケジュール要求」、「発注要件の明確度・安定度」については生産性を向上させる要因であることが分かった。
- ・生産性の悪化にもっとも影響する要因は開発規模である。逆に、生産性をもっとも高める要因は開発スピードであることが分かった。

今後の課題としては、開発スピードに影響する要因の分析、未知データに対する予測性能の評価を行うことが求められる。また、ファンクションポイントの代わりにソースコード行数を用いたモデル化なども挙げられる。

【参考文献】

- [1] 独立行政法人情報処理推進機構ソフトウェア・エンジニアリング・センター編, “ソフトウェア開発見積りガイドブック: ITユーザとベンダにおける定量的見積りの実現”, オーム社、2006.
- [2] B. Kitchenham, E. Mendes, “Why comparative effort prediction studies may be invalid”, Proc. 5th International Conference on Predictor Models in Software Engineering, May 2009.
- [3] M. Tsunoda, A. Monden, H. Yadohisa, N. Kikuchi, and K. Matsumoto, “Software Development Productivity of Japanese Enterprise Applications”, Information Technology and Management, Vol.10, No.4, pp.193-205, Dec. 2009.
- [4] 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, Nov. 2010.
- [5] 中村宏美、“CoBRA法に基づく見積もり支援ツール”、SECジャーナル、Vol. 5, No. 6. Dec. 2009.
- [6] M. Shepperd and C. Schofield, “Estimating software project effort using analogies”, IEEE Transactions on Software Engineering, vol.23, np.11, pp.736-743, 1997.
- [7] 財団法人経済調査会、“平成22年度ソフトウェア開発に関する調査票(受託者向け)集計結果その1”, 2011.

