トップエスイー修了制作

一般化線形モデルを用いたソフトウェアプロジェクトデータ分析

所属　メルコ・パワー・システムズ 名前 渡邉　智
Watanabe.Satoshi@zb.MitsubishiElectric.co.jp

開発における問題点

ソフトウェア欠陥を予測する統計モデルとして，一般線形モデルでは確率変数の分布として正規分布を想定していたり、分布の等分散性を前提としているのに対し，ソフトウェア欠陥は0, 1, 2,…とカウントする離散的な計数データである.したがって，ソフトウェア欠陥に一般線形モデルをそのまま適用することは不適切である.

手法・ツールの適用による解決

ポアソン回帰モデルや，負の二項回帰モデルなどの計数データに適用できる回帰モデルを用いる方法がある.これらは，いずれも一般化線型モデル（Generalized Linear Model）に属するモデルであり，分散の均一性や，誤差の正規性といった分析データの性質が一般線形モデルに比べて緩和されている..

分析対象データ

分析対象データには，開発工数（時間），規模（KLOC），欠陥数（件）という3種類の定量データが含まれている.そのほかのデータは定性データであり，(1)仕様化・文書化プロセス，(2)新規の機能性，(3)設計・開発プロセス，(4)テストと手戻り，(5)プロジェクト管理という5分類に分けられた27項目についてVery LowからVery Highまでの5段階の評定値が含まれている. それぞれの項目について，欠陥密度との順位相関係数，欠陥数との順位相関係数，および規模との順位相関係数，ならびに無相関検定の結果を表1に示す.「D1：開発スタッフ経験」は，欠陥密度との相関が強いが欠陥数とも相関がある.一方，「P5：ステークホルダ関与度」及び,「P6：顧客関与度」は，欠陥密度とのあいだに相関があるものの，欠陥数と相関があるとはいえない.

表1　定性データと欠陥密度および欠陥数との順位相関係数

項目名	D1	P5	P6
欠陥密度	-0.693	-0.430	-0.368
欠陥密度P値	0.000	0.020	0.049
欠陥数	-0.424	-0.236	-0.122
欠陥数P値	0.022	0.218	0.528
規模	-0.114	-0.052	-0.028
規模P値	0.555	0.788	0.855

一般化線形モデルの適用

分析対象データに一般化線形モデルに属するポアソン回帰モデル、負の二項回帰モデルを適用して、欠陥数の予測モデルを構築する.

■重回帰モデル

説明変数に規模，D1，P5，P6を指定し，ステップワイズ法により変数選択を実施した結果，決定係数の値は0.8889であった.決定係数は重相関係数の2乗であるため，平方根をとった0.943が重相関係数である.

■ポアソン回帰モデル

ポアソン回帰モデルを用いて構築した欠陥予測モデルの予測値と実測値の重相関係数は0.957であった.ただし，過分散（over-dispersion）の状態であった.

■負の二項回帰モデル

ポアソン回帰で過分散となったデータは，負の二項回帰モデルで分析するのが一般的である. 欠陥予測モデルを構築した結果、サンプルサイズが小さいためか（N=29），統計的有意と判定された説明変数をすべて適用することができなかったが,AICの値はポアソン回帰モデル(AIC=2422.1)と比べ,AIC=390.59と大幅に減少し，過分散の状態も改善されている.重相関係数は0.906であり，ポアソン回帰モデルよりも減少している.

■まとめ

ソフトウェア欠陥という計数データに対して，一般化線形モデルを適用して予測モデルを構築する試みを行った.分析の結果，ポアソン回帰モデルでは，ソフトウェア規模，開発スタッフ経験，ステークホルダ関与度，顧客関与度を説明変数とした利便性の高いモデルが得られた.また，連続変数を前提とした重回帰モデルに比べても妥当なモデルが得られた.しかし，ポアソン回帰により得られたモデルでは過分散の問題が残った.これを解消する為に負の二項回帰を適用した結果，過分散の問題は解消されたが，サンプルサイズの小ささのためか説明力のより高いモデルは得られなかった.今後の課題として，サンプルサイズの大きな実データに対して負の二項回帰モデルを適用し，理論的にも実用的にも有用なモデルが得られることを示していく必要がある.