

ソフトウェア開発データを対象とする2変数間の尺度の検討

金平拓生^{†1} 門田暁人^{†1}

概要: ソフトウェア開発データの分析において、2変数間の関係を定量化するために、相関係数がよく用いられる。ただし、2変数間の関係には様々な種類があり、それらを相関係数で区別することは困難である。本発表では、多様な関係を区別するための8つの尺度について検討する。

キーワード: ソフトウェアメトリクス, 相関係数, ソフトウェア開発データ分析

Investigation of Metrics Between Two Variables for Software Project Data

Takumi KANEHIRA^{†1} Akito MONDEN^{†1}

Abstract: The correlation coefficient is commonly used in an analysis of software project data sets for the purpose of quantifying the relationship between two variables. However, while there are various types of relationships between two variables, the correlation coefficient cannot distinguish between these types. This study investigates new metrics between two variables that have potential to characterize relationship types.

Keywords: Software metrics, Correlation coefficient, Software project data analysis

1. はじめに

ソフトウェア開発データの分析において、量的な2変数間の関係を定量化する尺度として相関係数（ピアソンの積率相関係数）がよく用いられている。相関係数は、2変数間に存在する線形な関係の強さを表す統計学的尺度であり、 $[-1, 1]$ の範囲の値をとる。絶対値が1に近いほど2変数間の関係が強いことを示す。

ただし、相関係数によって示すことができるのは、線形な関係の強弱のみである。2変数間に存在する多様な関係を相関係数により区別することは困難である。例えば、ソフトウェア開発データでは、SLOC – 工期のグラフのように、 x 軸側よりも y 軸側に多くの個体が集まっている場合がある一方で、FP (Function Point) – 開発工数のグラフのように、その逆の傾向を示す場合もある。相関係数では、このような違いを数値として表すことができない。そのため、分析者が目で見てその関係の違いを主観的かつ定性的に判断する必要があった。

本研究では、ソフトウェア開発データにおける様々な分布の違いを客観的な尺度により定量的に区別することを目的として、2変数間の関係を表す複数の尺度を提案する。それら尺度を組み合わせることで、2変数間のグラフの形状を定量的に区別することが可能となる。

以降、2章では、ソフトウェア開発データ白書 2018-2019[1]に記載のデータを題材として、典型的にみられる2変数間の関係を示すと同時に、それらを相関係数により区

別することが難しいことを示す。3章では、2変数間の関係を表す8つの尺度を提案する。4章では、提案する尺度の適用事例を示す。5章では関連研究を紹介する。6章はまとめと今後の課題である。

2. ソフトウェア開発データにおける2変数間の関係

ソフトウェア開発プロジェクトデータにおいては、2変数間に多様な関係が存在する。ここでは、典型的な関係を相関係数とともに示す。

2.1 x 軸または y 軸へのデータの偏り

改良開発におけるSLOC規模と工期のグラフを図1に、画面数と工数のグラフを図2に示す。それぞれの相関係数は0.417, 0.480であり同程度である。図1では多くの点が y 軸に近い部分に偏っており、図2では逆に x 軸に近い部分に偏っている。このような分布の違いがあっても相関係数の値からは区別できない。

2.2 左上または右下へのデータの偏り

新規開発における工数と工期のグラフを図3に、FP規模と工数のグラフを図4に示す。いずれも外れ値を除去した後のものである。それぞれの相関係数は0.616と0.706であり、大きな差はない。図3では多くの点が原点から左上

^{†1} 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

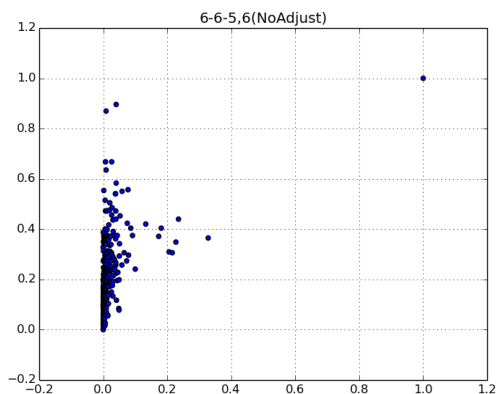


図 1. SLOC 規模と工期（改良開発）

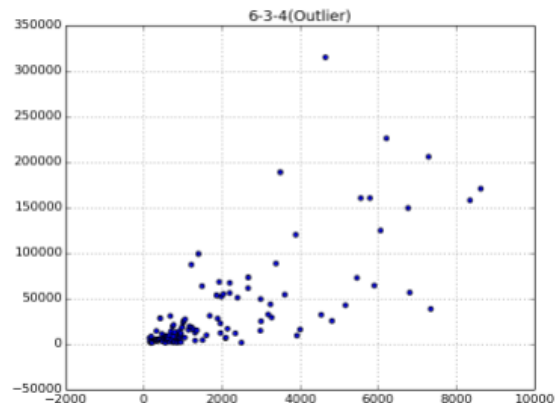


図 4. FP 規模と工数（新規開発）

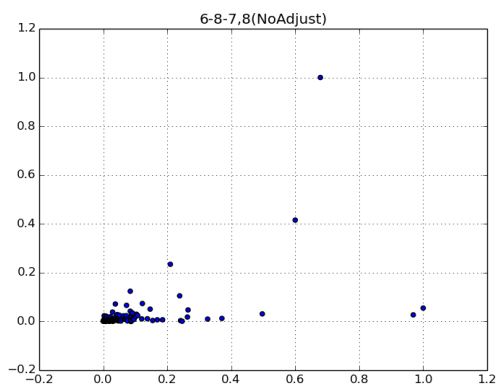


図 2. 画面数と工数（改良開発）

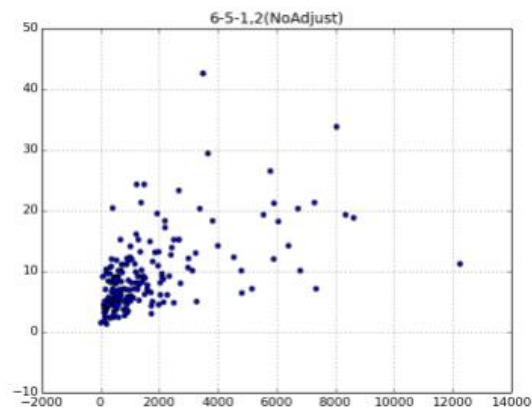


図 5. FP 規模と工期（全開発種別）

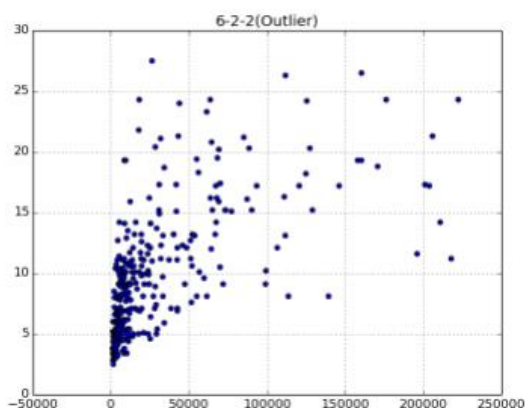


図 3. 工数と工期（新規開発）

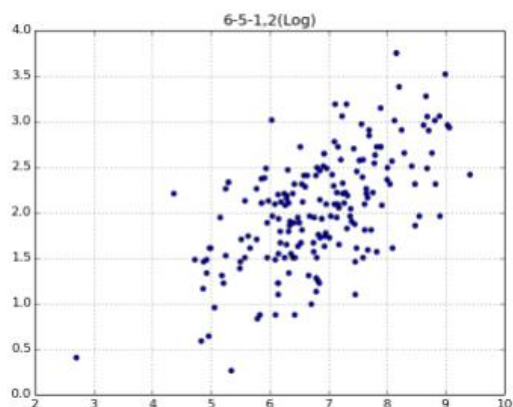


図 6. FP 規模と工数（全開発種別）の両対数グラフ

にかけて分布しており、右下には点が少ない。一方、図 4 では多くの点が原点から右下にかけて分布しており、左上には点が少ない。このような分布の違いがあっても相関係数の値からは区別できない。

2.3 原点または中央へのデータの偏り

全開発種別における FP 規模と工数のグラフを図 5 に、その両対数グラフを図 6 に示す。それぞれの相関係数は 0.52 と 0.61 であり、大きな差はない。図 5 では多くの点が原点付近に分布しているのに対し、図 6 は中央付近に分布

している．このような分布の違いがあっても相関係数の値からは区別できない．

3. 提案する尺度

2 変数間の関係を区別するため，関係を表す尺度として 8 つの尺度を提案する．尺度を算出する前提として，2 変数はそれぞれ値域が $[0,1]$ に正規化されているものとする．変数 A の i 番目の値を a_i ， $\max(A)$ を変数 A の最大値， $\min(A)$ を変数 A の最小値とすると， a_i の正規化された値 a'_i は次式で表される．

$$a'_i = \frac{a_i - \min(A)}{\max(A) - \min(A)}$$

3.1 領域に関する尺度

領域に関する 4 つの尺度を表 1 に示す．これらの尺度は，用いる 2 変数の i 番目の点 (x_i, y_i) がそれぞれの条件を満たしている割合を表しており， $[0,1]$ の範囲の値を取る．また，Upper-left ratio(左上比)と Lower-right ratio(右下比)の和，および，Lower-left ratio(左下比)と Upper-right ratio(右上比)の和は 1 となる．

Upper-left ratio(左上比)と Lower-right ratio(右下比)は，直線 $y = x$ で分割した右下と左上の領域に存在する点の割合をそれぞれ表す．図 7 は図 3 を正規化し，直線 $y = x$ を追記したものである．この例では，Upper-left ratio(左上比) : Lower-right ratio(右下比) = 0.90 : 0.10 となる．このことから，右下よりも左上に値が分布していること定量的に言える．

同様に，Lower-left ratio(左下比)と Upper-right ratio(右上比)は，直線 $y = 1 - x$ で分割した左下と右上の領域に存在する点の割合をそれぞれ表す．図 8 は図 3 を正規化し，直線 $y = 1 - x$ を追記したものである．この例では，Lower-left ratio(左下比) : Upper-right ratio(右上比) = 0.91 : 0.09 となる．このことから，右上よりも左下に値が分布していること定量的に言える．

これら 4 つの尺度から，この 2 変数は，「左上かつ左下を同時に満たす領域」に値が多く分布していることが定量的に言える．

これら尺度の値より，2.1，2.2 節で述べたような， x 軸や y 軸あるいは左上，右下へのデータの偏り具合を定量的に区別できると期待される．

表 1. 領域に関する尺度

尺度名	定義
Upper-left ratio(左上比)	$y > x$ を満たす点の割合
Lower-right ratio(右下比)	$y \leq x$ を満たす点の割合
Lower-left ratio(左下比)	$y > 1 - x$ を満たす点の割合
Upper-right ratio(右上比)	$y \leq 1 - x$ を満たす点の割合

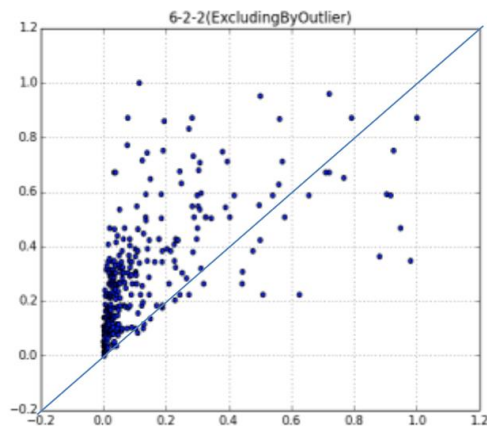


図 7. $y = x$ で分割された領域

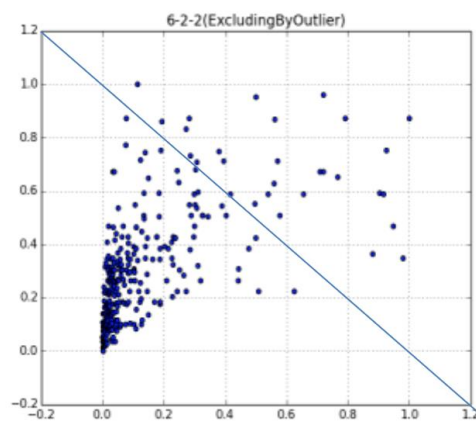


図 8. $y = 1 - x$ で分割された領域

3.2 グラフの四隅からの距離の尺度

グラフの四隅からの距離の尺度を表 2 に示す．4 つの尺度 Lower-leftness(左下度)，Upper-leftness(左上度)，Lower-rightness(右下度)，Upper-rightness(右上度)は，それぞれ四隅の点 $[0,0]$, $[0,1]$, $[1,0]$, $[1,1]$ からのユークリッド距離の平均値を $\sqrt{2}$ で除したものと定義され， $[0,1]$ の範囲の値を取る．これらの尺度は，2 変数の値の分布がグラフの四隅のどこに近いかを定量化できる．図 7 (図 3，図 8 も同様) における 4 つの尺度の値を表 3 に示す．

表 3 からは，原点からの距離の平均値である Lower-leftness(左下度)の値が他と比べ小さいことから，データが原点に近い部分に偏っていることが分かる．また，次に小

表 2. 四隅からの距離に関する尺度

尺度名	定義
Lower-leftness(左下度)	$[0,0]$ からの距離の平均値
Upper-leftness(左上度)	$[0,1]$ からの距離の平均値
Lower-rightness(右下度)	$[1,0]$ からの距離の平均値
Upper-rightness(右上度)	$[1,1]$ からの距離の平均値

表 3. 図 7 における距離の尺度の値

尺度名	値
Lower-leftness(左下度)	0.2429
Upper-leftness(左上度)	0.5319
Lower-rightness(右下度)	0.6727
Upper-rightness(右上度)	0.7951

表 4. SLOC-工期と画面数-工数の尺度の値

尺度名	SLOC- 工期(図 1)	画面数- 工数(図 2)
Upper-left ratio(左上比)	0.9974	0.1504
Lower-right ratio(右下比)	0.0026	0.8496
Lower-left ratio(左下比)	0.9974	0.9774
Upper-right ratio(右上比)	0.0026	0.0226
Lower-leftness(左下度)	0.1501	0.0659
Upper-leftness(左上度)	0.5617	0.7017
Lower-rightness(右下度)	0.7179	0.6532
Upper-rightness(右上度)	0.8948	0.9496

さな値は Upper-leftness(左上度)であることから、原点から左上にかけて値が分布していることが定量的に言える。一方、Lower-rightness(右下度)と Upper-rightness(右上度)の値は大きいことから、右上と右下にはあまり点がないことが定量的に言える。

このように、これら 4 つの尺度を用いることで、データの点がグラフ上の 4 点[0,0],[0,1],[1,0],[1,1]のどれに偏っているかが分かるため、2.3 節で述べたような、原点または中央への偏りについても区別することが期待できる。また、3.1 節の 4 つの領域に関する尺度と合わせることで、より詳細にデータの分布を分析できる。

4. ケースタディ

本実験では、3 章で提案した尺度を、2 章で用いたグラフに適用し、データの分布の違いを区別できるかどうかを検証する。

4.1 x 軸または y 軸への値の偏りの区別

ソフトウェア開発データ白書における、SLOC-工期（改良開発）(図 1) のグラフと画面数-工数（改良開発）(図 2) のグラフについて 8 つの尺度を算出した結果を表 4 に示す。

表 4 から、SLOC-工期 (図 1) のグラフは Upper-left ratio と Lower-left ratio の値が非常に高いため、左上と左下に値が強く偏っていることが分かり、値が y 軸に近い部分に分布していることが読み取れる。さらに Lower-leftness の値が小さく、Upper-leftness と Lower-rightness では Upper-leftness

表 5. 工数-工期と FP 規模-工数の尺度の値

尺度名	工数- 工期(図 3)	FP 規模- 工数(図 4)
Upper-left ratio(左上比)	0.9000	0.1333
Lower-right ratio(右下比)	0.1000	0.8667
Lower-left ratio(左下比)	0.9065	0.9250
Upper-right ratio(右上比)	0.0935	0.0750
Lower-leftness(左下度)	0.2429	0.2071
Upper-leftness(左上度)	0.5319	0.6689
Lower-rightness(右下度)	0.6727	0.5819
Upper-rightness(右上度)	0.7951	0.8380

の方が値が小さいことから、値が y 軸と原点寄りに集中していることが読み取れる。

一方、画面数-工数 (図 2) のグラフは Lower-right ratio と Lower-left ratio の値が高く、右下と左下に値が強く偏っていることが分かるため、値が x 軸に近い部分に分布していることが読み取れる。さらに Lower-leftness の値が極端に小さいため、非常に原点付近に偏ったグラフであると読み取れる。

以上から、SLOC-工期のグラフは値が y 軸付近に偏り、画面数-工数のグラフは値が x 軸側の原点付近に偏っているということが分かり、提案尺度でそれぞれを区別することができた。

4.2 左上または右下への値の偏りの区別

ソフトウェア開発データ白書における、工数-工期（新規開発）(図 3) のグラフと FP 規模-工数（新規開発）(図 4) のグラフについて 8 つの尺度を算出した結果を表 5 に示す。

表 5 から、工数-工期 (図 3) のグラフは左上と左下に値が多く分布していることが分かり、FP 規模-工数 (図 4) のグラフは右下と左下に値が多いことが分かる。そのため、工数-工期のデータは y 軸側にデータが偏り、FP 規模-工数のデータは x 軸側にデータが偏っていることが考えられる。

四隅への距離については工数-工期と FP 規模-工数のいずれについても、Lower-leftness(左下度)の値が小さく、原点付近に値が偏っていると言える。また、Upper-leftness(左上度)と Upper-leftness(左上度)に着目すると、工数-工期では Upper-leftness(左上度)の方がやや小さく、FP 規模-工数では Upper-leftness(左上度)の方がやや小さいことから、工数-工期の方がより左上に点が分布している傾向にあると言える。

このことから、2 つのグラフの値の分布の違いは、領域の尺度によって区別できる。一方、四隅からの距離の尺度では、いずれのグラフも原点に近い部分に値が分布していることが分かるが、それ以外の傾向については、値の違いから明確に区別することは難しい。今後、値の分散に関する

表 6. 工数-工期と FP 規模-工数の尺度の値

尺度名	FP 規模- 工期(図 5)	log(FP 規 模)-log(工 数)(図 6)
Upper-left ratio(左上比)	0.7692	0.2615
Lower-right ratio(右下比)	0.2308	0.7385
Lower-left ratio(左下比)	0.9538	0.3282
Upper-right ratio(右上比)	0.0462	0.6718
Lower-leftness(左下度)	0.1778	0.8112
Upper-leftness(左上度)	0.5893	0.6735
Lower-rightness(右下度)	0.6431	0.8145
Upper-rightness(右上度)	0.8413	0.6291

る尺度を定義することで、区別できるようになる可能性がある。

4.3 原点または中央への値の偏りの区別

ソフトウェア開発データ白書における, FP 規模-工期(全開発種別)のグラフ(図 5)とその両対数変換をしたグラフ(図 6)について 8 つの尺度を算出した結果を表 6 に示す。

表 6 において, 四隅からの距離の尺度に着目すると, 対数変換を行っていない FP 規模-工期のグラフは, Lower-leftness(左下度)が非常に小さいため, 値が原点付近に偏っていることが分かる。一方, 対数変換したグラフでは, 全て 0.6 以上の値となっており, 値が中央付近に分布していることが読み取れる。

また, 領域の尺度に着目すると, 対数変換をしていないグラフでは, Upper-right ratio(右上比)が 0.0462 と非常に小さいことから, 右上にはほとんど点がないことが分かる。また, Lower-right ratio(右下比)も 0.2308 と小さく, 右下には点が少ないことが分かる。一方, 対数変換をしたグラフでは, Lower-right ratio(右下比)と Upper-right ratio(右上比)がやや大きいことから, 前述のように, 四隅からの距離の尺度はすべて 0.6 以上と大きいことと合わせて考えると, 中央付近において, 左上・左下よりも, 右下・右上に値がやや多く分布していることが伺える。

このことから, 8 つの尺度を用いることで, 値の分布が四隅に近いのか否かや, どの領域の分布が多いかを知ることができ, 分布の違いを区別することができた。

5. 関連研究

2 変数間の関係の強さを示す尺度は, 用いる変数が量的変数である場合と質的変数である場合によって異なる。2 変数がともに量的変数である場合には, 相関係数がよく用いられる。相関係数としては, ピアソンの積率相関係数に

加えて, スピアマンの順位相関係数, ケンドールの順位相関係数が知られている。後者の 2 つは, 外れ値を含むデータセットに対してロバストであるという特徴を持つ。また, 変数変換を行うことで, 線形以外の関係の強さを定量化する相関係数の提案もされている[1][3]。ただし, いずれの相関係数も, 2 変数の値の分布を区別する用途には向かない。

一方が量的変数, もう一方が質的変数の場合, その間の関係の強さの尺度として, valiance explained (分散説明率, 因子寄与率) が知られている。また, 両方が質的変数の場合は, ϕ 係数, コンティンジェンシー係数, Cramer's V などが知られている。いずれについても, 関係の強弱を表す尺度であり, 関係の種類を区別することはできない。

2 変数間の多様な関係を考慮した尺度として, maximal information coefficient (MIC)が提案されている。MIC は, 2 変数間の様々な非線形な関係の強さを表せるように工夫されている[4]。MIC は, 2 次元グラフをいくつかの領域に分割することを考えた場合に, 領域間の相互情報量が最大となるような領域分割を探索的に発見し, その時の値に基づいて算出される。MIC そのものは 2 変数間の関係の強弱の尺度であるが, 領域分割と相互情報量を用いることで, 本研究の目的に合うような, データ分布の分類に応用できる可能性がある。

6. おわりに

本研究では, 量的な 2 変数間の関係を区別するための尺度として 8 つの尺度を提案した。ケーススタディにおいて提案尺度をソフトウェア開発データ白書のデータに適用した結果, 2 変数間の様々な関係の区別し得ることが分かった。今後は, 相互情報量を用いた尺度についても検討していく予定である。

謝辞 本研究の一部は, JSPS 科研費 JP20H05706 の補助を受けた。

参考文献

- [1] Breiman, L. and Friedman, J.H., "Estimating Optimal Transformations for Multiple Regression and Correlation," (with discussion) Journal of the American Statistical Association, Vol. 80, pp. 580-619, 1985.
- [2] 独立行政法人情報処理推進機構社会基盤センター, "ソフトウェア開発データ白書 2018-2019," 独立行政法人情報処理推進機構, 2018.
- [3] A. Rényi, A., "On measures of dependence," Acta Math. Hung. 10, pp.441-451, 1959.
- [4] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., Sabeti, P. C., "Detecting Novel Associations in Large Data Sets, Science," Vol. 334, 1518-1524, Dec. 2011.