

データをビジネスに活用する実践アナリティクス

<第 14 回> ソフトウェア開発コストの予測モデル

梶山昌之
株式会社ワイハット

前回の連載では基本的な住宅価格の予測モデルを作成しました。分析の結果、住宅価格は専有面積で価格の約 80%は説明できることがわかりました。

価格は専有面積だけではなく、駅からの徒歩分や築年数など、その他の条件で変動します。価格の 20%はこれらの影響により変動すると考えられます。

1. コストドライバ

住宅価格の予測における専有面積のように価格（またはコスト）を説明する要因をコストドライバと言います。この用語は、コスト評価の知識体系である CEBoK™ (Cost Estimating Body of Knowledge) (ICEAA, 2013) にて使用されています。

CEBoK™は、コスト評価（またはコスト見積り）の知識をまとめたもので、16章で構成されている出版物です。この中で第12章は「ソフトウェアコストの見積り」について記載していますので、本連載でも、その知見を引用することにします。

ソフトウェア開発コストに限らず、どの要因がコストの変動に影響するかは、最初からわかるというものではありません。幾つかの要因を説明変数にして試行錯誤するか、または重回帰分析などの手法を使うことになります。本連載では、そのような試行を経て、コストに影響がある要因が発見されたものと理解してください。

2. ソフトウェア開発コスト

本連載の読者の皆様は、IT関連のプロジェクトを担当されている方も多いと思います。そこで、今回はソフトウェア開発コストの予測の問題を取り上げます。

今回のタイトルは「ソフトウェア開発コストの予測モデル」と表現しましたが、これは「ソフト

ウェア開発コストの見積り」と同じです。

また、ソフトウェアの開発コストは労務費が多くを占め、労務費は工数に工数単価を掛けることによって計算されますので、「工数見積り」の課題と考えることができます。

3. コスト見積りとコスト評価

英語の Estimation は「見積り」または「評価」と訳されますので、しばしば混同して使われますが、本連載では「コスト見積り」と「コスト評価」の用語は区別して使用することにします。

例えば、ソースコード行数 (SLOC: Source Line of Code) と開発コストの関係式を作成し、それをプロジェクトの計画段階における見積りに使用する場合とプロジェクト完了後の評価に使う場合では、入力になるデータが異なります。

見積りでは入力となる SLOC は開発する SLOC の推定値（または計画値）であり、実績として計測された値ではありません。

これに対して、プロジェクトの完了後にコストの妥当性を評価する場合は、実績値を用いることができます。また、実行したテストケース数など様々な計測値が評価の入力として使用可能になります。

また、プロジェクトの計画段階では、積算で算出されたコストの妥当性を評価する場合があります。この場合には、過去の実績値で作成された予測モデルによるコスト見積りの値で、積算によるコストを評価します。

この様に、見積りは予測するのが目的ですが、評価は推定された値を用いて良否を判断するのが目的ですので、使い分けが必要です。

4. ソフトウェア開発のコストドライバ

ソフトウェア開発コストの見積りにおけるコスト

ドライバは何でしょうか。見積りは、プロジェクトの計画段階または、要件定義など、プロジェクトの初期の段階で必要になるものですので、その時点で、計測可能または推定可能な量が入力となります。CEBoK™では、ソフトウェア開発コストのコストドライバは以下の3つであるとしています。

- ・規模 (Size)
- ・複雑度 (Complexity)
- ・能力 (Capability)

後日の連載で解説する分析で明らかになります。規模は工数の概ね50%を説明する主要なコストドライバですので、今回は規模に着目します。

CEBoK™では以下の3つをソフトウェアの規模の尺度の例として紹介しています。

- ・ソースコード行数 (SLOC)
- ・ファンクションポイント (FP)
- ・オブジェクトポイント

最も一般的な規模の尺度は SLOC であり、FP は2番目に位置づけられています。オブジェクトポイントは画面、帳票などの実体を計測する方法でプロジェクト初期段階の見積りに有効な方法ですが、現時点では、まだ広く使われていません。

SLOC は古くからソフトウェアの規模尺度として使用されており、現時点でも多くの企業で計測されている尺度です。SLOC は、プロジェクト完了後には自動計測も可能なため、プロジェクト完了後のコストの妥当性を評価するためには有効と考えられます。しかしながら、プロジェクトの計画時点では実体としてのソースコードは存在しないため、見積りに使用するのは困難です。

一方、FP は、ユーザー視点でソフトウェアの持つ機能の量を数値化したものですので、プロジェクトの早い段階で推定することが可能です。ここで、早い段階で「推定する」と表現したのは、FP が実際に計測可能となるのは、概ね外部設計完了後であるためです。従って、プロジェクトの計画段階では、必要となる機能を推定して求めることとなります。

具体的には A の機能を実行するための画面 B が必要となり、その画面では C の値を入力してデータベース D に格納するなどの一連の処理に係る機能を推定することです。

FP 計測の概要については、後日の連載で紹介す

る予定ですが、ここでは、ソフトウェアが持つ機能の量を数値化した値が得られたと考えてください。

尚、本連載は FP の内容を聞いたことがない皆様も対象としています。従って、FP がどのような値をとるものか、感覚がつかめない方もおられるかと思しますので、その規模感について説明します。

ソフトウェアは小さなものから大きなものまでありますが、その規模を FP で表した場合、IPA/SEC(2014)によると、新規開発プロジェクトの平均的な値(中央値)は700FP~800FP かと思いますが、ここでは、切りのよい数字で概ね1000FPと覚えてください。

経済調査会(2015)によれば1FP 当たりの平均的な価格は14万円ですので、1000FP のソフトウェアを開発するコストは概ね1.4億円になります。

ただし、実際には、プロジェクトの種類や特性により、その1/10倍から10倍の変動があります。従って、このコストはFP の規模感を掴むにとどめ、実際のコスト見積りに、そのまま使用することはないようにしてください。

読者の皆様には、FP を使用していないので、その分析も自分には関係がないと思われる方もいらっしゃると思いますが、プロジェクトマネジメント、マーケティング、社会科学、医学などの分野では、FP と同様の性質を持つ測定量が多く存在します。従って、FP をそれらの尺度に置き換えて考えれば、本連載で解説する分析の手法はそのまま応用できると考えていただいて結構です。

5. FP による工数見積り

ソフトウェア開発の見積りで最も著名な方は米国のCapers Jones氏かと思えます。彼はFP を活用したソフトウェアの定量的管理や見積りの分野で多くの著作を出版しています。

日本でもソフトウェアの見積りに関する多くの著作が翻訳、出版されています。彼は、様々な業種におけるソフトウェアメトリクスデータを、コンサルティングの活動で入手しているため、1万プロジェクト以上の信頼性のあるデータを分析対象としています。それ故、信頼性がある知見が得られるため、本連載でも彼の著作物から必要に応じて引用します。

彼の著書「ソフトウェアの定量化手法」(Capers, 1998)にはFP と工数の例として表1に示すデータが紹介されています。

このデータは実例ではないと推測していますが、私の経験によれば、実際の例をよく反映していると思います。

右側の欄には各プロジェクトの生産性を計算していますが、この生産性は FP/工数で計算します。例えば、プロジェクト B では、10 [FP]のソフトウェアを構築するために 0.5 [人月]の工数がかかりましたので、生産性は 20 [FP/人月] です。

同じ仕事を、より少ない工数、例えば 0.4 [人月] で達成した場合、生産性は 25 [FP/人月] となり、生産性が向上したことがわかります。

表 1 FP と工数のデータ

プロジェクト	FP	工数[人月]	生産性
A	10	0.3	33.33
B	10	0.5	20.00
C	12	0.6	20.00
D	15	0.6	25.00
E	30	1.5	20.00
F	50	3.3	15.15
G	60	4	15.00
H	80	5.3	15.09
I	100	6.6	15.15
J	500	100	5.00

(Capers Jones ソフトウェア開発の定量化手法 第2版 177 ページより引用)

さて、前回までの連載で解説した知識を用いて、このデータを分析してみましょう。すると表 2 および図 1 に示す結果が得られます。

表 2 FP と工数の回帰分析

【基本統計量】

特性値	FP	工数	工数/FP
データ数	10	10	10
平均	86.700	12.270	0.068
中央値	40.000	2.400	0.058
標準偏差	148.623	30.905	0.048
ひずみ	2.90	3.13	2.75
とがり	8.78	9.86	8.21

【単回帰分析】 モデル: $y=b_0+b_1*x$

統計量	b0	b1
回帰係数	-5.576	0.206
標準誤差	1.729	0.010
t 値	-3.225	19.707
p 値	0.012	4.576E-08

【分散分析】

変動因	平方和 S	自由度 f	分散 V
回帰 R	8422.810	1	8422.810
残差 e	173.511	8	21.689
全体 T	8596.321	9	
分散比 Fo	388.346		

【その他の統計量】

偏差平方和 Sxx	198800.1
残差標準偏差 se	4.657
寄与率 R ²	98.0%

ここで、「ひずみ」は分布が左右対称であるかどうかを示す統計量です。正規分布の場合はこの値が 0 になりますが、2.90 となっており、分布が左側に歪んでいることがわかります。「とがり」は正規分布であればこれが 0 になりますが、3.13 となっており、正規分布より尖っている分布であることを示しています。

これらの値がいずれも -1.5~+1.5 の範囲にあれば正規分布と見做すことができますが、そうではないため、分布は正規分布に従っていないと判断します。

次に、図 1 を見ますと小規模の FP (100 以下) にデータが集中し、プロジェクト J (500FP) は離れたところにあるように見えます。

このようなデータでも回帰分析の結果を信頼してよいでしょうか。そこで、回帰分析の適用条件を満たしているかどうか検討します。

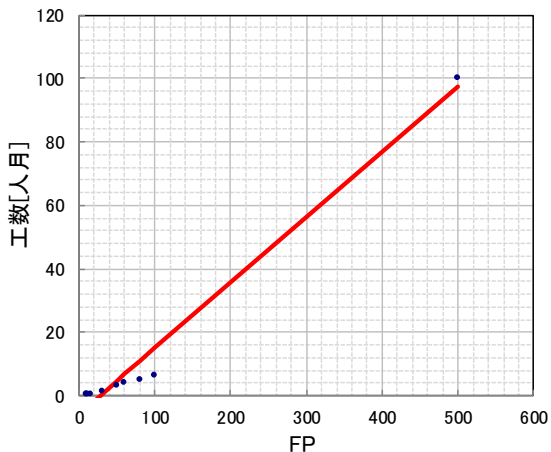


図1 FPと工数の散布図

この適用条件については、前回の連載で解説していますので参照してください。

- ・残差の平均は 0 である (不偏性)

という条件については、残差が 0 になるように回帰直線が計算されていますので、必然的に満たされています。

- ・残差の分散はどの X においても一定 (等分散性)

の条件に現れる X は説明変数のことですが、この場合の X は FP であり、FP のデータは小規模部分に集中していますので、どの X においてもバラツキが等しいかは、それを確認することができません。

- ・残差は正規分布に従う (正規性)

の条件は、任意の X における残差が回帰直線を中心としてばらついている状況を意味しています。

FP と工数の例では、最初は回帰直線より高い値で、その後、低い値となっており、回帰直線の周りで正規分布しているとは言えません。従って正規性の条件は満たしていないと言えます。

- ・任意の残差は他の残差と互いに独立である (独立性)

これは、任意のプロジェクトの残差が、他のプロジェクトの残差に影響を与えるかということですが、プロジェクトは互いに独立であることを前

提とすれば、条件は満たされていると考えることができます。

6. 対数変換について

ところで、図1からわかるようにプロジェクトJ (500FP) の1点で回帰直線がほぼ決まっており、小規模部分の傾向は無視されてしまっています。このような現象が発生する理由は、後の連載で詳しく解説の予定ですが、FPは、前回までの連載で説明した身長例のように正規分布に従う量ではなく、発生頻度が指数的に変化する非対称の分布であるためです。それに対応して、工数も非対称の分布になります。

このような場合には、対数を取って、FPと工数の確認するのが、データ解析の常套手段です。

ここで、対数について簡単に説明します。例えば 1000 は $1000=10^3$ と書けますが、10 の右肩にある数値が「対数」です。ここで 10 は「底」と呼ばれ、10 を底とする対数は常用対数と呼ばれます。従って、1000 の常用対数は 3 です。

ここで、常用対数が 3 の時、元の値 (この場合は 1000) のことを「真数」と言います。この用語は対数から元の値への変換を行うときには必要になりますので、ここでご記憶ください。

底は任意の正の値をとることができます。例えば $8=2^3$ ですが、この場合は、2 を底とする 8 の対数は 3 です。

数値 X の対数は $\log(X)$ と表します。この場合、対数の底を明示する場合は、底を log の右下に添え字で書きます。10 を底とする X の対数は、 $\log_{10}(X)$ となります。

常用対数はわかりやすく、対数から真数への変換が簡単なのでよく使われます。例えば、3 の真数は $10^3=1000$ と計算できます。

しかし、工学および統計学では、対数の底をネイピア数 $e (= 2.718282\dots)$ とするのが慣例になっています。ネイピア数は自然界ではしばしば登場する数値です。高校数学で「ふないっぱちにはち…」と覚えられた方もいるでしょう。

ネイピア数 e を底とする対数を自然対数と言います。 $\log(X)$ のように対数の底の表記を省略した場合、本来は自然対数を意味していたのですが、 $\log(X)$ を常用対数として使用する例も多く、混同を避けるため、自然対数の場合には $\ln(X)$ という表記が使われるようになりました。

\ln は自然対数 (Natural Logarithm) の意味で、読み方は特に定まったものはないのですが、エル

エヌと呼ばれることが多いようです。尚、Excelには自然対数を計算するLN関数があります。また、ExcelにはLOG関数もありますが、底の指定を省略すると常用対数を返しますので注意してください。

今後の解説では、対数変換を活用しますが、対数はすべて自然対数と考えてください。

表3は表1のデータを対数変換した結果です。

表3 FPの対数と工数の対数

プロジェクト	LN(FP)	LN(工数)	LN(生産性)
A	2.303	-1.204	3.507
B	2.303	-0.693	2.996
C	2.485	-0.511	2.996
D	2.708	-0.511	3.219
E	3.401	0.405	2.996
F	3.912	1.194	2.718
G	4.094	1.386	2.708
H	4.382	1.668	2.714
I	4.605	1.887	2.718
J	6.215	4.605	1.609

7. 回帰分析と両対数散布図

表3の対数変換された値に対して回帰分析を実行した結果を表4に示します。

対数変換前のFPと工数の「ひずみ」および「とがり」はいずれも1.5を超えており、正規分布に従っているとは言えないことを示していましたが、対数変換後は、はいずれも1.5以下の値になっています。すなわち、FPおよび工数の対数はいずれも正規分布に従う量であるとみなすことができます。

また、散布図の作成では、この変換された値で散布図を作成することもできますが、ここでは、XとYの両軸が対数になっている散布図(両対数散布図)でFPと工数の関係を表示します(図2)。

この散布図で得られる分布パターンは、対数された値を通常の散布図(真数の軸)にプロットした場合と同じになります。両対数散布図では、個々の点のXおよびYの値を真数への変換の計算を行うことなく読み取れますので、こちらを使うことを推奨します。

この図から、各プロジェクトは概ね直線の周りに分布していることがわかります。大規模の1点は小規模集団からはやや離れていますが、小規模集団が作る直線の近くに大規模の点があり、一つの直線で全体の関係性を表すことができます。

ここで、関係式が直線で表せるという性質が非常

に重要です。何故なら、人間の感覚は直線的な関係は容易に認識できるからです。

この例では、対数を取ることで、直線的な関係があることを見つけましたが、対数を取ればいつでも直線的な関係になるという訳ではありません。変数変換にはいろいろな方法があります。

例えば、Xを2乗してある値Cを加えた値 $X' (=X^2 + C)$ がYと直線的な関係になる場合もあるでしょう。この場合はXの対数とYの対数との関係が直線になるとは限りません。

様々な変数変換の内、XとYの対数変換を試してみたら、直線で表せる関係が得られたと考えてください。もし、直線にならない場合は、他の変換を試みることを推奨します

表4 FPと工数回帰分析(対数変換)

【基本統計量】			
特性値	LN(FP)	LN(工数)	LN(工数/FP)
データ数	10	10	10
平均	3.641	0.823	-2.818
中央値	3.657	0.800	-2.857
標準偏差	1.257	1.724	0.498
ひずみ	0.79	1.07	1.53
とがり	0.38	1.47	4.13

【単回帰分析】 モデル: $y = \exp(b_0) * x^{b_1}$			
統計量	b0	b1	exp(b0)
回帰係数	-4.134	1.361	0.016
標準誤差	0.220	0.057	1.246
t値	-18.805	23.727	
p値	6.607E-08	1.060E-08	

【分散分析】			
変動因	平方和 S	自由度 f	分散 V
回帰 R	26.371	1	26.371
残差 e	0.375	8	0.047
全体 T	26.745	9	
分散比 Fo	562.966		

【その他の統計量】	
偏差平方和 Sxx	14.228
残差標準偏差 se	0.216
寄与率 R ²	98.6%

尚、真数での回帰分析の残差は工数の差だったのですが、対数での回帰分析の場合の残差は工数の差ではない点にも注意してください。このことが示す意味については後日の連載で詳しく解説の予定です。

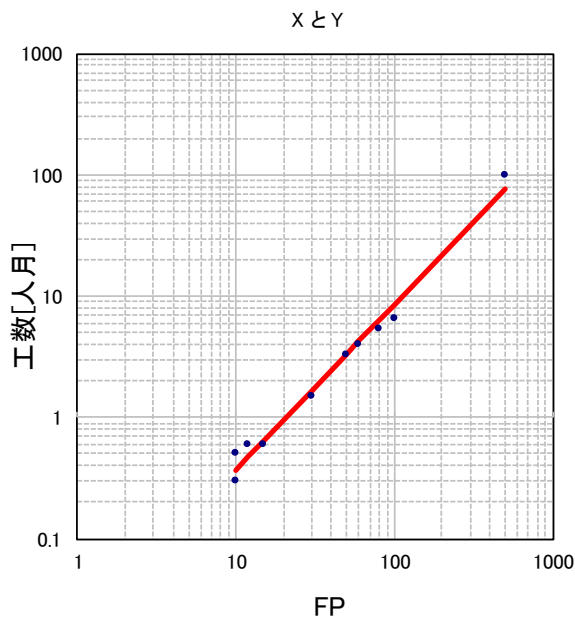


図2 FPと工数の両対数散布図

8. べき乗モデル

FP の対数と工数の対数に直線的な関係がある場合には、これを式で表すと、

$$\ln(Y) = a + b \cdot \ln(X) \quad (\text{ここで, } X:\text{FP, } Y:\text{工数})$$

となります。この関係を Y の式として表すと、

$$Y = A \cdot X^b \quad (\text{ここで, } A = \exp(a))$$

となり、Y は X のべき乗で表せることとなります。べき(冪)とは、X の右肩にある値 b のことです。そこで、この式で表される予測モデルを本連載では「べき乗モデル」と呼ぶことにします。

Capers Jones 氏の事例では、 $A = \exp(-4.13) = 0.016$ 、 $b = 1.361$ (表4参照) になります。ここで、b は 1 より大きい値となっている点に注意しましょう。

生産性 P は FP/工数 ですので

$$P = X/Y = X/(A \cdot X^b) = X^{(1-b)}/A$$

となり、 $b > 1$ の場合は X が大きいほど生産性 P の値が低下するモデルになっています。

これは、100FP のソフトウェアを 1000 万円で構築した実績がある場合、200FP の類似のソフトウェアを構築するコストは、 $b = 1.361$ としたとき 2000 万円ではなく、2569 万円 ($= 1000 \cdot (200/100)^{1.361}$) かかると考えるべきだということ

です。もし、2000 万円で受注した場合は、過少見積りとなり、プロジェクトは失敗します。

回帰係数 b の値がどの程度 1 より大きいかを評価することが重要であり、b が 1 より大きい場合は、「規模の不経済性」と表現する場合があります。

昔からプロジェクトマネジメントの世界では、規模が大きくなるとコミュニケーションのオーバーヘッドが増大するので生産性は低下するといわれていますが、作成されたモデルはその状況に適合しています。

ここで、 $b = 1$ の場合、生産性 P は

$$P = 1/A$$

となり、FP 規模に依存しません。その結果、X と Y の予測式は、

$$Y = A \cdot X$$

となります。Y は X に比例していますので、本連載では、このモデルを「比例モデル」と呼ぶこととします。この様に、べき乗モデルは、その特別な場合として、比例モデルを含んでいます。

分析の対象となったデータでは、100FP を超え 500FP 未満のデータが得られていませんが、予測モデルを使えば、実績データがない範囲での予測も可能になります。例えば、200FP の場合の工数は

$$Y = 0.016 \cdot (200)^{1.361} = 22 \text{ [人月]}$$

となります。

また、実績データがある範囲を超える予測(外挿)も可能です。例えば、類似ソフトウェアの FP 値が 1000 の場合、

$$Y = 0.016 \cdot (1000)^{1.361} = 194 \text{ [人月]}$$

となります。ただし、外挿の場合は予測精度が下がります。妥当な外挿の範囲については第 10 回の連載で解説していますので参照してください。

この事例では、両対数散布図上で、プロジェクトが直線状に並んでいますが、実際のプロジェクトの場合には、この事例の様に、きれいに直線状に並ぶことは稀です。

同じ FP であってもプロジェクト特性の違いにより、工数はバラツクためです。また、バラツキ

が大きい場合には、 b の値が小さくなります。

本連載の第 10 回では「平均への回帰」の現象について解説しましたが、統計学的な現象の結果として b が小さくなることがあるため、FP が大きいほど生産性が高くなるという解釈は本質を捉えているとは限りません。これを、どの様に解釈し取り扱うべきかについては、後日の連載で解説の予定です。

今回は Capers Jones 氏が提示したデータを分析対象とし、FP から工数を予測する式を作成しました。実務でも両対数散布図上で綺麗な直線になる例がみられますが、そのような場合の回帰係数の値は 1 より大きくなる傾向がありますので、本質的には、規模が大きくなれば生産性は低下すると考えてよいと思います。

もし b が 1 より小さい関係式が得られた場合、それは「平均への回帰」の現象が影響している可能性がありますので要注意です。

以上は、ソフトウェアの開発工数の見積りに関しての知見ですが、ソフトウェアの保守工数については、「規模の経済性」が成り立つ場合もあります。

いずれの場合も、決め手となるのは、自社の実績データによる検証です。実績データの蓄積が不十分な場合は、ベンチマークで補完することをお勧めしますが、最終的には、自社実績の分析を目標としていただきたいと思います。

今回は、FP と工数の関係について、相関が高い例を紹介しましたが、実務ではバラツキが大きいデータを扱うことになります。

次回は、FP、工数、生産性の性質について、さらに深く解説する予定です。ご期待ください。

参考文献

- Capers Jones 著, 鶴保証城・富野 壽 監訳 (1998).
ソフトウェア開発の定量化手法第 2 版. 共立出版. ISBN4-320-09722-X.
- ICEAA (2013). コスト評価知識体系 (CEBoK™)
日本語版. 日本コスト評価学会.
- 今泉忠, 田村義保, 中西寛子, 美添泰人 (2015).
日本統計学会公式認定 統計検定 2 級対応 統計学基礎. 東京図書.
- IPA/SEC (2014). ソフトウェア開発データ白書
2014-2015. IPA/SEC.
- 経済調査会 (2015), ソフトウェア開発データリポ
ジトりの分析 2015 年 5 月. 経済調査会.
- 奥野忠一, 久米均, 芳賀敏郎, 吉沢正 (1981).
多変量解析法 (改訂版). 日科技連.

末吉正成, 末吉美喜 (2017). EXCEL ビジネス統計
分析 [ビジテク] 第 3 版. 翔泳社.