生産性と品質データの解析手法についての提案

A Proposal for Analyzing Productivity and Quality Data

岡野 麻子* 矢田部 学*
Asako Okano, Manabu Yatabe

現在、我々は生産性・品質データの解析手法の見直し作業を行っている。本稿では、その作業をとおして得られた解析手法を紹介する：想定された４タイプの回帰モデルの候補に対して回帰分析を行い、次に、それらの決定係数を用いて適切な回帰モデルを判定する。その手法の適用例として、生産性・品質データの解析結果を示す。

We have been reviewing approaches for analyzing productivity and quality data. This paper mentions a method obtained through the review: First four types of candidate model are arranged for regression analysis, and then a suitable model is selected from these four candidates by using the coefficients of determination. Moreover, we show some examples of analyzing the productivity and quality data with this method.

1. まえがき

鎌倉事業部では、ソフトウェアプロセスアセスメントとして、グループ共通の標準プロセスに基づく診断を継続して実施している。ここ数年で、ある一定のレベルに達したという判断のもと、次へのステップとして、CMMI®（Capability Maturity Model Integration：能力成熟度モデル統合、米国Carnegie Mellon大学の登録商標）[1]の上位レベルをモデルに改善を進めることとなった。CMMI®には５段階の成熟度レベルがあり、レベル４からは統計的・定量的手法を用いてデータを解析することが要求されている。

そこで、これまで生産技術部門において実施してきた作業方法を以下のとおり整理した：

- IPA（Information-Technology Promotion Agency：情報処理推進機構）から発行されているデータ白書[2]、既存のプロジェクト診断支援ツールなどを参照して、生産性や品質データの解析を実施
- 収集したデータを層別し、データ白書の解析結果を利用したベンチマーキングを実施
- 収集したデータを層別し、当事業部内の業務分野ごとに比較を実施

これらの解析のために当時使用していた既存ツールでは、標本数に制約があり、ツールで指定された標本数以下となるクラスターは解析対象外とした。ただし、その制約の理由などは不明であり、解析作業や結果の考察にも限界があると考えられた。

統計的な手法を用いてプロジェクト管理や組織目標を達成するには、それらを支配する要因を探らなければならない。その手法として、回帰分析をしてデータの相関を求める方法がある。これまで、当部門においては、標本数の不足、層別の困難さ、散布図プロットで相関が見えないという問題などから、その要因を見つけるのが困難であった。

以上の点から、統計学の基礎からデータの解析の手法を再検討することにした。その結果、さまざまな分野、背景因子、標本数に関する生産性・品質データの解析を行う際の道筋が見えてきた。

本稿では、今回の生産性・品質データの解析手法の再検討作業をとおして得られた手法の概要について述べるとともに、その適用例を紹介する。

2. 生産性・品質データの現象論

ソフトウェアの生産性や品質に関する現象が独立変数（説明変数）xと従属変数（被説明変数）yの関係式で表されると仮定する。例えば、関係式としては生産量xと工数y、生産量xと混入誤り数y、流用率xと生産性y、開発者の力量（経験年数、開発量、対象システムや工学理論の理解度など）xと生産量yの関係などがあげられる。各変数においては、yはxの関数であり、生産性の定義は、生産量yを工数xで微分したdy/dxとして捉えることができる。このことから、微

＊鎌倉事業部　生産技術部

分量にも着目する。

これらの間に成り立つ現象論的なモデルを構築するために、微分量dy/dxと変数xおよびyの簡単な関係式から出発する。この関係式として以下の３タイプを考える（数式中のa, b, c, kは定数でタイプごとに独立）。

・dy/dx が定数k：

$$\frac{dy}{dx}=k$$

これを解くと

$$y=kx+c \qquad (1)$$

・dy/dx がyに比例：

$$\frac{dy}{dx}=ky$$

これを解くと

$$y=c\exp(kx)$$

あるいは$e^k \equiv a$とおき

$$y=ca^x \qquad (2)$$

・dy/dx がxに比例：

$$\frac{dy}{dx}=k\,x$$

これを解くと

$$y=\frac{1}{2}kx^2+c$$

ここで、定数cを右辺に移項した$y-c$はcを基準にして測るということなので、$c\equiv 0$（基準値をゼロ）としても一般性を失わない。さらに$k/2 \equiv a$とおき、べき乗を一般化（2→b）して

$$y=a\,x^b \qquad (3)$$

とする。これは微分方程式$dy/dx=abx^{b-1}$の解である。

一見するとこれらの式（1)、(2)、(3）は異なった表現に見えるが、以下の共通した形式に帰着する。

$$Y=AX+B \qquad (4)$$

すなわち、式（1）では$Y\equiv y$、$A\equiv k$、$X\equiv x$、$B\equiv c$であり、式（2）では両辺の常用対数をとり、$Y\equiv\log y$、$A\equiv\log a$、$X\equiv x$、$B\equiv\log c$とする。同様に、式（3）では$Y\equiv\log y$、$A\equiv b$、$X\equiv\log x$、$B\equiv\log a$と置き換える。換言すれば、xとyの関係を直接$x-y$で解析するのが式（1）のモデル、片対数$x-\log y$で解析するのが式（2）のモデル、そして両対数$\log x-\log y$で解析するのが式（3）のモデルである。

以上より、xとyの関係を見出すとき、式（4）の線形関係を仮定してデータを回帰分析[3][4]することが可能になる。なお、「回帰」という意味は独立変数Xと従属変数Yは対等ではなく、XがYを決定するという考え方である。データ解析モデルとしてはXには誤差を含めずにYに誤差を含める。

3. 解析手法ー回帰モデルのタイプとその採択

２節の現象論的考察から得られたデータのタイプは（a）$x-y$、（b）$x-\log y$、（c）$\log x-\log y$であるが、変数の対称性を考慮に入れ、（b）に対称な独立変数と従属変数の組み合わせ$\log x-y$を追加した以下の４タイプを考える。

1. $x-y : y=Ax+B$
2. $x-\log y : \log y=Ax+B$（$y=10^{Ax+B}$）
3. $\log x-y : y=A\,\log x+B$
4. $\log x-\log y : \log y=A\log x+B$（$y=10^B x^A$）

観測された生産性・品質データはこれら４タイプの何れかに属すると仮定する。データの属するタイプを決定するために回帰分析の決定係数[3]（付録Aを参照）を利用する。解析すべきデータ（x_i, y_i）（$i=1, 2, \cdots, n$）をこれら４タイプに変換後、式（4）の線形モデルを適用して回帰分析を行う。それぞれのタイプの決定係数[2]（$0 \le r^2 \le 1$）を算出し、それが最大となるタイプを（x, y）の回帰モデルとして採択する。

上記のように、最適なものが一つとは限らず、複数のタイプで決定係数が高い値をとることも想定される。その場合は$x-y$のプロット結果などと照らし合わせて相関の有無を判断し、最適なものを採択する。また、すべてのタイプの決定係数が低く、何れのタイプでも相関が無いと判断された場合は、観測された生産性・品質のデータ（x_i, y_i）で成り立つ意味のある関係は無いと結論する。

これまでの生産性・品質データの解析ではデータの確率分布を仮定して回帰分析が行われている。例えば、参考文献（2）ではソフトウェア開発プロジェクトのデータは対数正規分布[3]に従うことを仮定して、本稿のタイプ４のモデルを用いて回帰分析を行っている。本節で示した４つのタイプの決定係数を用いる手法の利点はデータの確率分布を仮定する必要が無いことである。

4. 手法の適用例

３節で述べた手法を適用する前に、データの層別という作業が必要である。解析する目的を設定し、背景因子を考慮して、クラスター化したデータごとの特性を検討しながら層別する：

1. 散布図を描き、データのばらつき度合を確認
2. ２節で述べたような観点から、仮説を立て、層別
3. 層別したデータを、３節で述べた手法により、決定係数を参考にして４タイプの何れのモデルに属するかを判断

以上のように層別したデータを用いた適用例を、4.1節と4.2節に示す。ここでは、実際のデータを公開することができないため、ある事業部門のデータを加工したものを用い、グラフのスケールは非表示にしている。

4.1　生産量と工数

生産量xと工数yのデータに関して、タイプ１からタイプ４のモデルに対して回帰分析を行った結果を図１に示す。モデルの当てはまりの良さを測る指標である決定係数はタイプ４→タイプ１→タイプ３→タイプ２の順に小さくなる。

決定係数から判断すると、タイプ４が採択すべき回帰モデルということになるが、この例ではタイプ１とタイプ４の決定係数が共に大きい（$r^2 \geq 0.8$）ので、何れを採択するかは当該部門のこれまでの状況（データの信頼性など）を考慮して解析結果から品質管理担当が判断する必要がある。

図１において、タイプ２、３、４は対数をとったデー

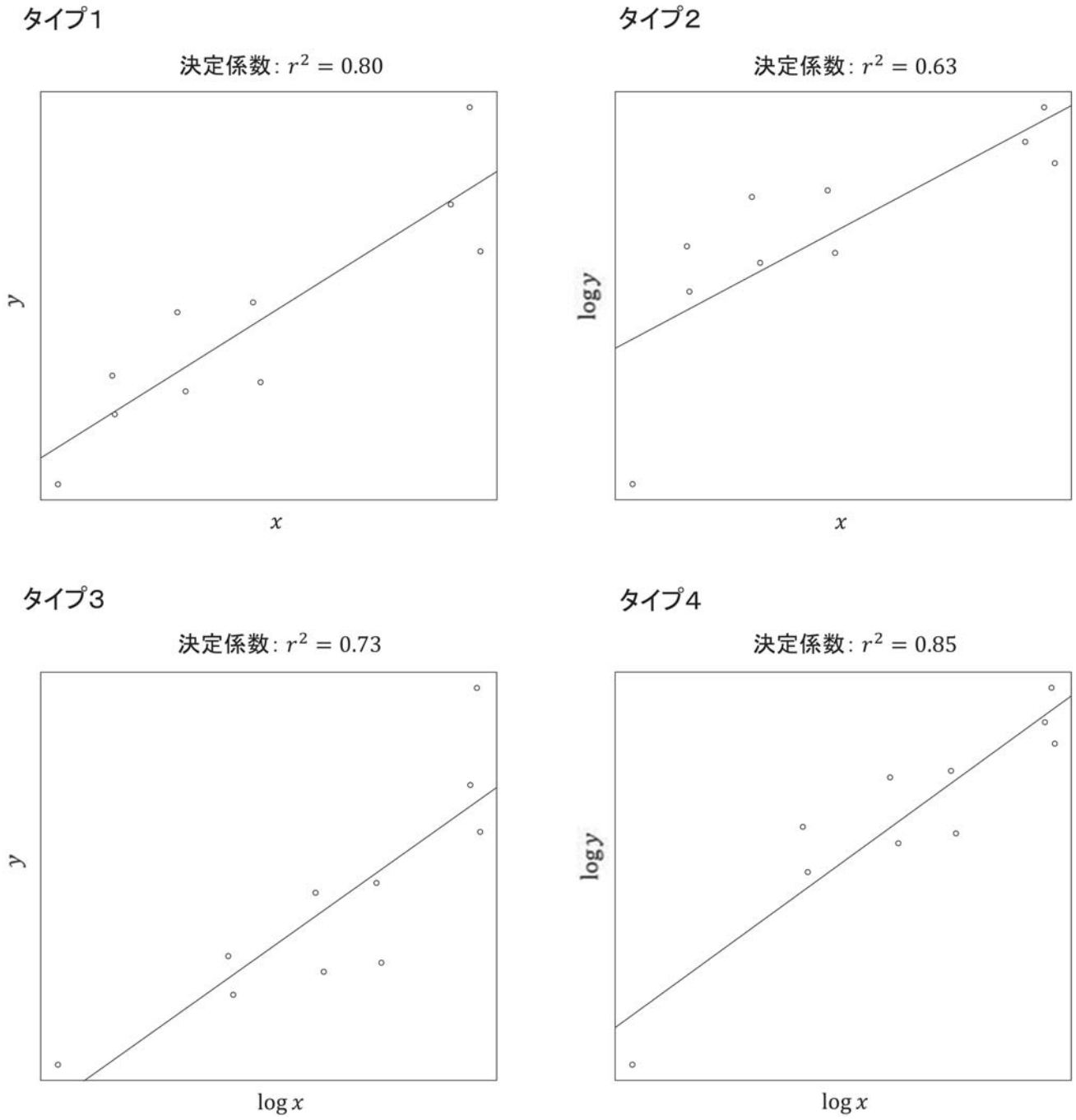

図１　各タイプの散布図と決定係数

タに対して回帰分析を行っている。このままでは生産量xから工数yを予測するモデルとしては使いづらいので、リニアスケールに戻す。リニアスケールで表現した回帰曲線（実線）と90%予測区間（破線）を図２に示す（予測区間については付録Bを参照）。

予測区間[4][5]はこれまで得られたデータの回帰分析に基づいて、将来得られるデータの入る範囲を予測するものである。ここで示した例では、受注した作業の生産量xを見積もったとき、予測される作業の工数yが90%の確率（100回見積もると90回の割合）で予測区間に入るということを意味する。

決定係数およびデータ分布に対する回帰曲線と予測区間から総合的に判断すると、タイプ１またはタイプ４が予測モデルとしては適切であると考えられる。先に述べたように、何れを採択するかの判断は、当該部門の過去の状況に精通した品質管理担当が行う必要がある。その結果を予測モデル（ここでは工数の予測）として用いる。活用方法としては、類似のプロジェクトの見積もり

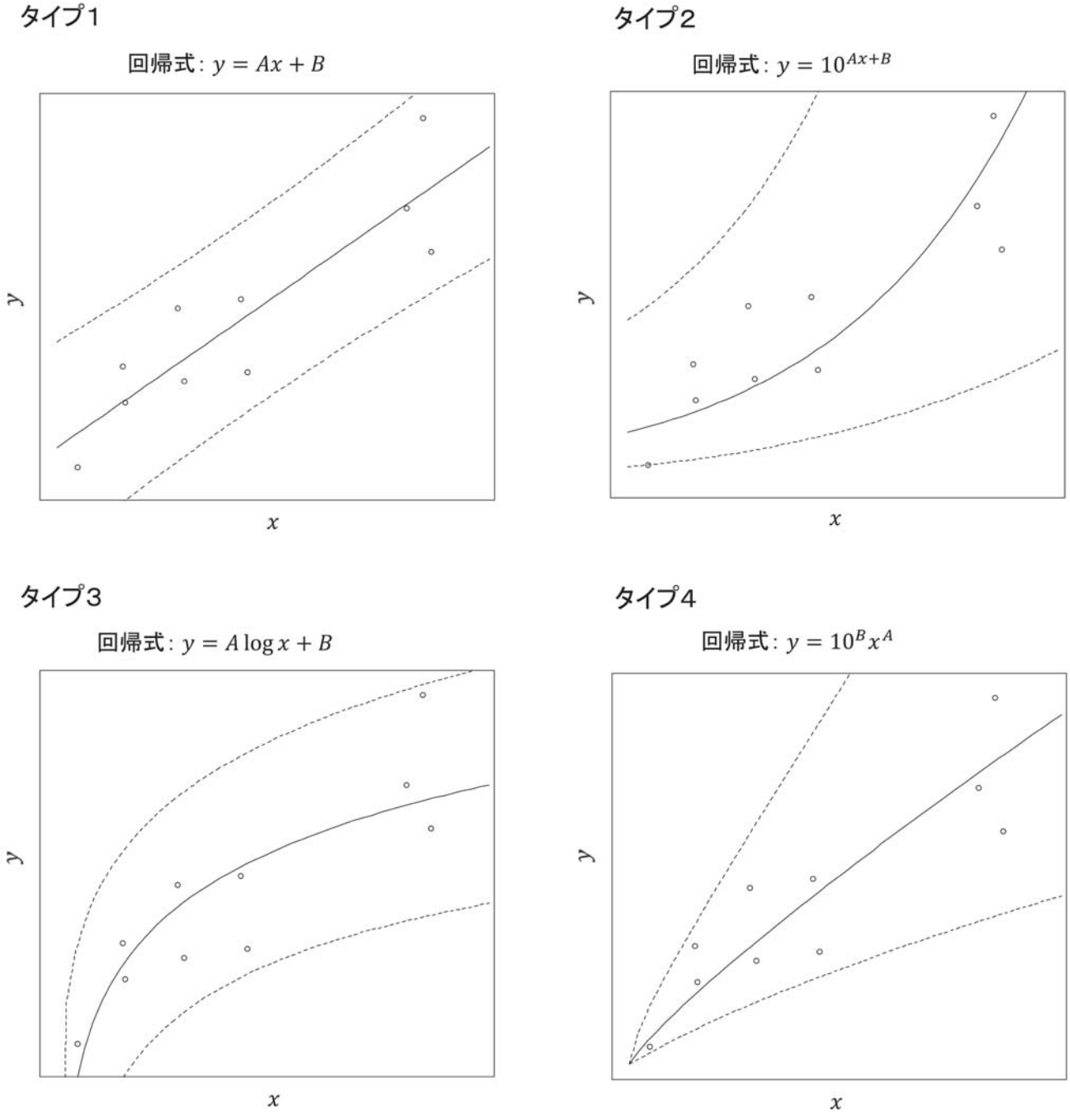

図２　各タイプの回帰曲線と90%予測区間

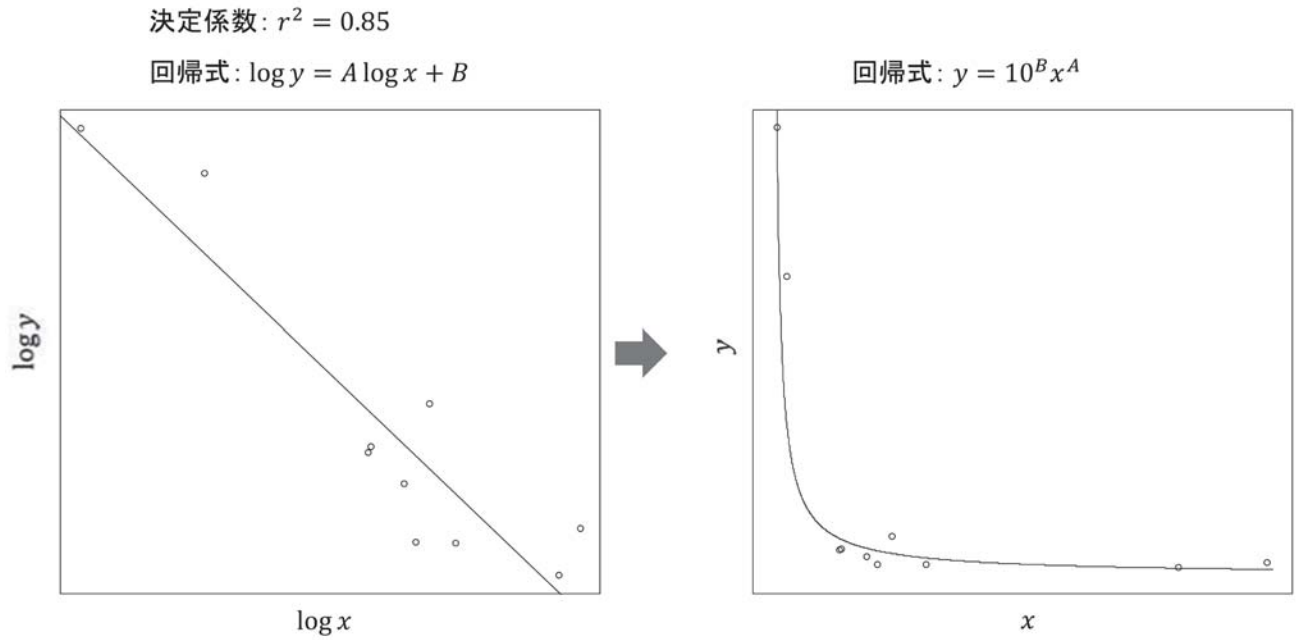

図３　生産性 x と累積誤り検出率 y

の参考や、見積値の妥当性を判断する材料とすることが挙げられる。このことで、見積ミスを防ぐということも期待できる。

なお、数学的特性によりタイプ４以外の回帰曲線は原点を通らない。これは、同じデータを使って解析しても表現する空間により見え方が異なるということである。

4.2　生産性と累積誤り検出率

生産性 x と累積誤り検出率 y のデータの関係性を求めるため、4.1節と同様の手順に従って回帰分析を行った結果、タイプ４の決定係数（r^2=0.85）が一番大きくなり、このモデルを採択する結果になった。

回帰分析の結果を図３に示す。図の左側は両対数 $\log x - \log y$ で表示したグラフで、右側はそれをリニアスケール $x-y$ に変換したものである。図３の右側のグラフが示すように回帰式の当てはまりはかなり高いと言える。このことより、ここで用いたデータの性質はほぼ一貫しており、データを層別した領域が適切であると考えられる。

適切な層別を行ったデータに対して、ここで述べた解析手法を適用することで、これまで相関が無いと思われていたデータにも相関があることが分かった。

5. むすび

本稿では、生産性・品質データの解析手法について一つの提案を行った。

２節および３節で述べた手法を用いて、仮説の検証やモデルの事例を蓄積することができる。例えば、開発言語、作業者の力量、ソフトウェア製品区分（組込み、エンタープライズ、科学技術計算）などの背景因子の中から層別したデータに、どのモデルを当てはめ、どのような目的・用途で使用したかという事例である。これは、精度の高い予測モデルを構築するためには重要な事項である。

一方、作業を行う中で、何らかの関連があると判断したデータでも、回帰分析を行ってみると、決定係数が小さく、相関が見られないことも多々あった。ソフトウェア開発分野で一般的に使用されている生産性・品質管理データの解析手法に当てはまるケースと当てはまらないケースがあるということである。当てはまらないケースとしては、例えば、要求分析などで、工学や数学・物理の基礎知識が必要なフェーズがあるということである。このような業務では、単純にソフトウェアの規模と工数というような指標で測ることは難しい。単位や指標などに新たな概念や工夫を取り入れることを視野に入れていく必要がある。

ここに述べたような作業を繰り返し、データの精度が上がり、標本数が増えると、予測区間が的確な領域を示すことになり、見積もり精度が向上していく。このようなことから、定量的プロジェクト管理の精度向上には、意味のあるデータを見極め、関係性を導き出し、我々に気付きをも与えてくれる統計解析の基礎を固めることは必須である。

以上を踏まえ、ここで述べた手法を改良していきたいと考えている。

参考文献

⑴ Japanese Language Translation of CMMI for Development, http://cmmiinstitute.com/resource/japanese-language-translation-of-cmmi-for-development-v1-3/, CMMI Institute, 2012

⑵ ソフトウェア開発データ白書2012-2013, 情報処理推進機構, 2012

⑶ 東大教養学部統計学教室編, 統計学入門, 東大出版会, 1991

⑷ 井原俊英・新重光, ようこそ化学標準物質の不確かさへのいざない（回帰分析）, 産総研 https://staff.aist.go.jp/t.ihara/reg.html

⑸ 林　岳彦, おっと危ない：信頼区間と予測区間を混同しちゃダメ http://takehiko-i-hayashi.hatenablog.com/entry/20110204/1296773267

付録

A. 決定係数

式（4）の回帰モデルを想定して、観測データ$(X_i, Y_i)(i=1, 2, \cdots, n)$に最小二乗法を適用すると

$\hat{Y} = \hat{A}X+\hat{B}$

が得られる（ハット∧は推定値）。この回帰方程式の決定係数はこの$\hat{Y}_i$との平均値$\overline{Y} = \sum Y_i/n$を用いて以下の式で定義される。総和は$i=1, 2, \cdots, n$についてとるものとする。

$$r^2 \equiv 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2}$$

$\sum(Y_i - \hat{Y}_i)^2$は回帰後の残差、$\sum(Y_i - \overline{Y})^2$は観測データのばらつきを表す。

回帰方程式に基づいてX_iがY_iを完全に説明するとき、$Y_i = \hat{Y}_i$となり、$r^2=1$である。他方$\hat{Y}_i = \overline{Y}_i$のとき$r^2=0$となり、回帰方程式に基づいて$X_i$が$Y_i$を完全に説明できない。これより、決定係数は$0 \leq r^2 \leq 1$の値をとる。

B. 予測区間

観測データ$(X_i, Y_i)(i=1, 2, \cdots, n)$（母集団から抽出した大きさ$n$の標本）に対して回帰分析を行い、式（4）の形の回帰方程式

$\hat{Y} = \hat{A}X+\hat{B}$

が得られているとする。この回帰方程式に基づいて、次に行われる新たな観測（母集団から新たな１つの標本を抽出）のデータが入る範囲を見積もったものが予測区間である。これは回帰分析の信頼区間に新たな観測の誤差が加わったものと解釈できる。

３節で述べたそれぞれのタイプのモデルに対する予測区間$y(-) \leq \hat{y} \leq y(+)$の上限・下限$y(\pm)$を以下にまとめる。ここで、$t_{\alpha/2}(n-2)$は自由度$n-2$の$t$分布の上側確率100（$\alpha/2$）%のパーセント点である。

1. $\hat{y} = \hat{A}x+\hat{B}$のタイプ
   $y(\pm) = \hat{A}x+\hat{B} \pm t_{\alpha/2}(n-2) \cdot s_1 \cdot r_1$
2. $\log \hat{y} = \hat{A}x+\hat{B}$のタイプ
   $y(\pm) = 10^{K(\pm)}$
   $K(\pm) = \hat{A}x+\hat{B} \pm t_{\alpha/2}(n-2) \cdot s_2 \cdot r_1$
3. $\hat{y} = \hat{A} \log x+\hat{B}$のタイプ
   $y(\pm) = \hat{A} \log x+\hat{B} \pm t_{\alpha/2}(n-2) \cdot s_1 \cdot r_2$
4. $\log \hat{y} = \hat{A} \log x+\hat{B}$のタイプ
   $y(\pm) = 10^{K(\pm)} x^{\hat{A}}$
   $K(\pm) = \hat{B} \pm t_{\alpha/2}(n-2) \cdot s_2 \cdot r_2$

ただし、各タイプ共通で

$$s_1 = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$$

$$s_2 = \sqrt{\frac{1}{n-2}\sum(\log y_i - \log \hat{y}_i)^2}$$

$$r_1 = \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

$$r_2 = \sqrt{1 + \frac{1}{n} + \frac{(\log x - \overline{\log x})^2}{\sum(\log x_i - \overline{\log x})^2}}$$

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \overline{\log x} = \frac{1}{n}\sum \log x_i$$

執筆者紹介

岡野　麻子

1997年入社。入社以降、鎌倉事業部で防衛分野に従事。2005年4月より品質保証に従事。2012年4月より生産技術部門としてプロセス改善に従事。

矢田部　学

1986年入社。つくば事業部で宇宙分野の解析や金融工学に従事。2004年11月より鎌倉事業部で宇宙・防衛分野のモデリングや統計解析に従事。博士（理学）。