原点を通る線形回帰モデルと統計ソフトウェア

札幌学院大学　中　村　永　友
城西大学　土　屋　高　宏

概要：本報告は，原点を通る線形回帰モデルを統計ソフトウェアで推定するとき，その出力がソフトウェアによって異なる数値が出力されることの例示や，このモデルの特性と注意点を指摘するものである．

キーワード：原点回帰モデル，決定係数，自由度修正済み決定係数，統計モデル，統計ソフトウェア

1　はじめに

著者らは，共通の定点を通る回帰直線群のモデルである「焦点回帰モデル」を提案した（中村・土屋，2007）．焦点回帰モデルは原点を通る回帰モデル（以下，原点回帰モデル）に帰着され，同論文ではこのモデルの諸統計量が特種なことを示した．さらに，Nakamura and Tsuchiya（2013b）では，原点回帰モデルを分析手法として用いる場合には，初学者や統計解析に慣れていないユーザはその判断を誤る可能性があることも指摘した．その主な問題点は，数学的に整合のある統計量である決定係数の値が，どんなデータに当てはめても 1 にかなり近い値となり，通常の回帰モデルの決定係数とは同じ土俵で比較ができないというものである．本論文ではこの点に注目し，原点回帰モデルが統計関連ソフトウェアでどのような振る舞いや出力をするのか具体的に指摘するものである．コンピュータがあれば誰でもデータ解析ができる時代になり，とくにデータ解析を専門としない，あるいは統計学の初学者などに対して注意を喚起するという視点で論じていく．

2　データと統計モデル

まずはじめに，以下の 2 つの問題を考えてみる．

1. データが取得される現象の背景にある（制約）条件とあてはめるモデルの問題，
2. 考え得るモデルを複数立ててモデル選択して，制約条件のないモデルが選ばれたときの問題，

である．後者は仮説検定の方法も含むとする．

まず，前者であるが，データ解析するときには最大限そのデータの取得状況や現象の背景を考慮すべきであ

る．それはモデル構築には必須だからである．しかし，場合によっては詳細な状況を知らずに（あるいは，知らされずに／無視して）モデルを作ることもあるだろう．データを採取（抽出，観測）したのが解析者本人であれば，ある程度の条件や状況を知りうるであろうが，そうでない場合にはこのようなことが起こりうるであろう．

モデリングの目的は大きく分けて2つある．それは当該現象を説明するため，すなわち帰納的な推論によって未知の現象をなんらかのモデルで説明しようとすることである．もう一つは予測である．当該データを通して構築されたモデルを前提に，将来の任意の入力値に対してどのような出力がありうるのかを知りたいということである．この場合，予測精度を上げるという視点でのモデリングが考えられ，これはデータの状況によっては制約条件を課さないモデルが良いモデルとなることもありうる．

以上のように，観測データからの統計モデル構築は，その目的によって使い分けされているのである．ここから見え隠れするのは，統計科学のある種の限界と矛盾である．それは統計学が帰納推論による科学的手法であるがゆえの抱える根本問題であろう．

次に，後者のモデル選択についてである．考え得る統計モデルを複数作り，情報量規準などでモデル選択する方法は今や現代的で標準的な科学の手法である（仮説検定の枠組みでもほぼ同様な事はできるが，理論は複雑になりがちである）．この枠組みで，単純にモデル選択したとき，当該制約条件を入れないモデルが選択されることがある．つまり，データが内在する情報では，その現象を説明しきれないということである．（ここでは単純にデータから現象を説明しようとするモデルを想定している．予測分布や事前分布などを考慮しない単純な統計モデルを考えている．）この時の取り得る判断は，

1. 解析の放棄，
2. さらにデータを取得する(同じ条件下である必要がある)，
3. 制約条件を入れた統計モデルの範囲で考える，
4. データ取得できなかった状況も含めた複雑なモデリングをする，

等が考えられる．1つめは，最ももったいない判断であるが，そのデータが実験計画通りに採取されているのであれば，計画自体が誤っていたと考えるべきであろう．2つめは現実的に困難な場合が多く，不可能な場合が多いのではないか．実験計画を根本から考え直さねばならないであろう．以上の2つは実験計画に係わる根本的な問題である．しかし初学者が実験計画なしにデータ採取していることも容易に想像でき，対応が困難な案件かも知れない．

3つめは，制約条件付きモデルを選択したときのモデルの解釈が難しくなることがあるかも知れない．統計学･データ解析の初心者はとくに混乱する選択となるかかもしれない．(つまり情報量規準(あるいは仮説検定)という道具が万能であるかのような錯覚・幻想から，非制約モデルを選択してしまうかもしれない.)

4つめはより高度なモデリングとなるため，統計の専門家との共同研究が必要となる選択である．

本論文の立ち位置は，解析者が制約条件（原点を通るべきデータであること）を知っていて，この下で解析しなければいけない，とする．より具体的には，制約条件として線形回帰モデルが原点を通るという状況に話題を絞って次節で議論する．

3　原点回帰モデルをあてはめる状況と問題の指摘

入力がないときに出力もないという状況を含む現象において観察されるデータ（原点を通るべきデータ）に

対する回帰モデルは,「原点回帰モデル」を使うべきであろう．観察される事象の背景を考慮したモデルの典型的事例である．しかし，例えば身長と体重のデータに対しては原点回帰モデルが妥当であり，このモデルをあてはめるべき事も知りながら，実際には切片（定数項）のある通常の単回帰モデルをあてはまることがほとんどであろう．加えた力とバネの伸び（フックの法則）の関係も同様の典型であり，線形回帰モデルの入門段階でよく用いられる例でもある．しかし，線形回帰モデルの入門時の例としては，ある意味わかりやすく適切な例であろうが，このようなデータを厳格で厳密な分析しようとする場合には考慮すべきであろう．

原点を通るべきデータに対して，あてはめが可能な回帰モデルの選選択肢は，次の 2 つであろう．

1. 強制的に切片を 0 とする原点回帰モデルをあてはめる．
2. 通常の回帰モデルをあてはめて切片が 0 であるという検定を行う．

前者は当然のモデルなので，このときモデルのあてはまりの良さをみる必要がある．しかし，原点を通るという条件を入れたモデルでは流通する統計ソフトウェア等では $R^2 \fallingdotseq 1$ となる傾向が強く，通常の回帰モデルとの比較という意味では，あてはまりの程度がよくわからないという問題が指摘できる．

後者は，切片が 0 であるという検定で（H_0：切片＝0)），帰無仮説 H_0 が棄却されなければ，原点を通ると考えても良い，という結論になるであろう．しかし，帰無仮説 H_0 が棄却されず，もし切片 $\neq 0$ という結論が出たときに，その後どう分析すればよいのか深刻な判断を迫られる．

あるいは，モデル選択の観点から，

1. 原点を通るという制約条件を入れたモデル（RTO モデル），
2. 通常の回帰モデル（OLR モデル），

の 2 つに対して情報量規準でこの 2 つのモデルから選択することができる．しかし，後者のモデルが選ばれた場合，このデータに対するモデリングをどう考えて，どう解釈すればいいのだろうか？　モデル選択の観点から，情報量規準 AIC で 2 つのモデルを比較して通常の回帰モデルが選択されたとき，この方法は切片が 0 であるという検定は棄却されたことと同等になる．

このような事情を知らないデータ解析者が多いことは容易に想像でき，この性質（決定係数がほとんど 1 に近いために，あてはまりの良さを解釈する際に解釈不能であるなどの性質）を持つデータに対して原点回帰モデルを用いることはそう多くないのではないだろうか？　つまり，制約条件のある線形回帰モデルが視野にない分析者が少なからずいると考えられる．このように，原点回帰モデルにおいては，考慮しなければいけない事項や状況があるという事である．

では，なぜ原点回帰モデルがこのような特殊なモデルであるのか，次節で説明をする．

4　原点回帰モデルと諸統計量

原点回帰モデルを適用すべき現象は，生物学，化学，様々な工学分野などの様々な分野・領域に数多く存在する．そのために，標準的な統計ソフトウェアに標準装備されていて，ポピュラーである．統計学的な特性を指摘するならば，必ず原点を通るということによって，(1)データの重心を必ずしも通らない，(2)一般的な平方和分解が成り立たない，(3)決定係数が大きくなる傾向がある，という理由によって，通常の線形回帰モデルと単純な比較はできないのである．以下にその理論的背景を示す．

原点回帰モデルは，

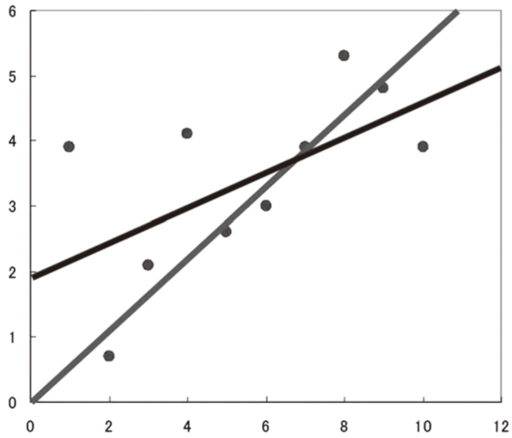


図１：テストデータの散布図と２本の回帰直線

原点を通る直線が原点回帰モデル（$y=0.5478x$）．もう一方の直線が通常の回帰モデル（$y=0.2697x+1.9467$）．

$$y=ax+\varepsilon,\quad \varepsilon \sim N(0, \sigma^2)$$

となり，傾きの最尤推定量，あるいは最小自乗推定量は

$$\hat{a}=\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

となる．

一方，通常の回帰モデルや分散分析などで用いられる目的変数の平方和分解は，

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{1}$$

で定義され，データの分解と対応する自由度は $(y_i-\hat{y})=(\hat{y}_i-\overline{y})+(y_i-\hat{y}_i)$，$(n-1)=(2-1)+(n-2)$ となる．しかし，原点回帰モデルのための非常に奇妙な平方和分解は

$$\sum_{i=1}^{n} y_i^2=\sum_{i=1}^{n} \hat{y}_i^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2}$$

で定義され (Eisenhauer, 2003)，データの分解と対応する自由度は $y_i=\hat{y}+(y_i-\hat{y}_i)$，$n=1+(n-1)$ となる．この平方和分解が用いられる理由は，このモデルがデータの平均(重心)を必ずしも通らないからである．すなわち(1)式において $\overline{y}=0$ とおいているのである.「平方和分解の特異性」と表現しているのは，(1)式の分解はデータ集合の内部にある重心からデータの変動を測っているのに対して，(2)式は(ほとんどの場合)データの外部にある原点から変動を測っていることである．

この(2)式の平方和分解から定義される決定係数は，

$$R_c^2 \equiv \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{3}$$

となり，ほとんどの統計ソフトウェアでは，この定義式に基づいてその値を計算している．さらにこれを基礎として，自由度調整済み決定係数は，

$$\overline{R}_c^2 \equiv 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-1)}{\sum_{i=1}^{n} y_i^2/n} \tag{4}$$

で定義されている．

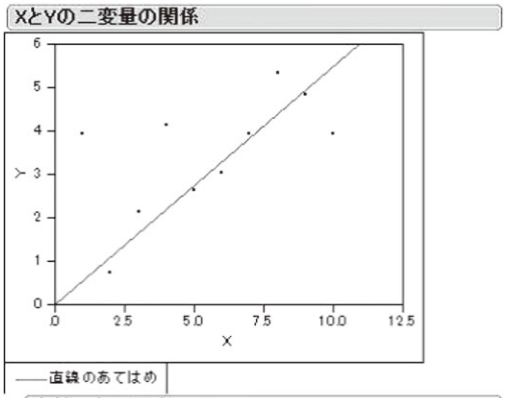
5　ソフトウェアでの原点回帰モデル

本節では，統計ソフトウェアで原点回帰モデルの出力がどのようになっているか示していく．この出力を見ることで，公開されていない統計量の計算方法や，このモデルに対するソフトウェア開発者(社)の意図もある程度伺うことができる．

テストデータを図1に示す．原点を通っている右上がりの直線が原点回帰モデルの回帰直線，もう一つの直線が通常の線形回帰モデルである．この原点回帰モデルを各種ソフトウェアでの出力を図2～図5に示す．

5.1　JMP

図2は統計ソフトウェアJMP6以降のヴァージョンの出力である．このソフトウェアの出力の特徴は，まず，決定係数である〝R2乗〟と自由度調整済み決定係数である〝自由度調整R2乗〟の値が出力されていないことである．また，〝パラメータ推定値〟欄の切片で，「ゼロに固定」と明示されている点が特徴である．



直線のあてはめ

Y = 0 + 0.5477922 X

あてはめの要約

R2乗	.
自由度調整R2乗	.
誤差の標準偏差(RMSE)	1.441474
Yの平均	3.43
オブザベーション(または重みの合計)	10

分散分析

要因	自由度	平方和	平均平方	F値
モデル	1	115.52938	115.529	55.6005
誤差	9	18.70062	2.078	p値(Prob>F)
全体(修正済み)	10	134.23000		<.0001*

縮小モデルで検定: Y=0

パラメータ推定値

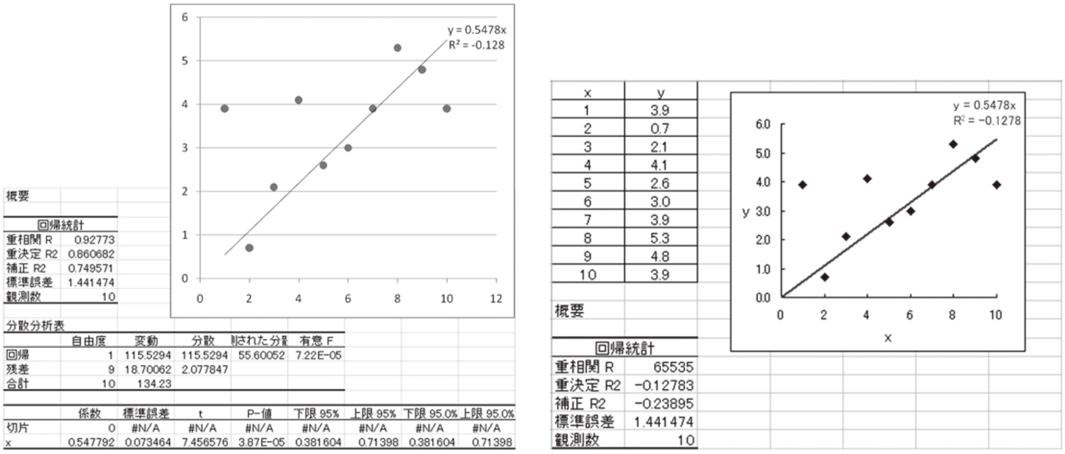
項	推定値	標準誤差	t値	p値(Prob>|t|)
切片　ゼロに固定	0	0	.	.
X	0.5477922	0.073464	7.46	<.0001*

図2：JMP6の出力

5.2　Excel 2003以降と2002以前

図 3 (左) は表計算ソフトウェア Excel2003 以降のヴァージョンでの出力である．これは,「データ分析」⇨「分析ツール」⇨「回帰分析」からの出力と，散布図を描いた後で「近似曲線の追加」から得られる結果である．特徴は，決定係数である〝重決定 R2〞の値と，散布図で表示される R^2 の値が異なると同時に，後者の値が負の値である点が挙げられる．そもそもこの「重決定」という単語を見ても統計学への無知さが明白である．

図 3 (右) は表計算ソフトウェア Excel2002 以前のヴァージョンでの出力である．「データ分析」⇨「分析ツール」⇨「回帰分析」からの出力と，散布図を描いた後の「近似曲線の追加」から得られる結果である．これも Excel2003 以降のヴァージョンと同様な指摘ができ，計算結果と散布図上の決定係数(重決定 R2)の値がともに負である．さらに図 3 （右）の「重相関R」の値がとんでもなく大きな値である．これは明らかなバグである．



概要

回帰統計	
重相関 R	0.92773
重決定 R2	0.860682
補正 R2	0.749571
標準誤差	1.441474
観測数	10

分散分析表

	自由度	変動	分散	則された分散	有意 F
回帰	1	115.5294	115.5294	55.60052	7.22E-05
残差	9	18.70062	2.077847		
合計	10	134.23			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
x	0.547792	0.073464	7.456576	3.87E-05	0.381604	0.71398	0.381604	0.71398

x	y
1	3.9
2	0.7
3	2.1
4	4.1
5	2.6
6	3.0
7	3.9
8	5.3
9	4.8
10	3.9

概要

回帰統計	
重相関 R	65535
重決定 R2	-0.12783
補正 R2	-0.23895
標準誤差	1.441474
観測数	10

図 3 ：Excel 2007（左）と2002（右）の出力

5.3　Mathematica

図 4 (左) は数式処理ソフトウェアの Mathematica6 の出力である．このソフトウェアの特徴は，コマンドを投入した後にすぐに警告を出していることである．このソフトウェアの特徴は，コマンドを投入した後にすぐに警告を出していることである．一方，図 4 (右) は Mathematica7 以降のヴァージョンの出力である．これ以降のヴァージョンでは警告が出なくなり，各種統計量の出力がプロパティを指定することで得られるようになっている．これはフリーソフトウェア R をかなり意識した改変である．

決定係数は〝RSquared”，自由度調整決定係数は〝AdjustedRSquared〞で示されていて，これまで説明したソフトウェアの中ではじめてこの 2 つまともな値である．

5.4　SPSS

図 5 は統計ソフトウェアの SPSS の出力である．特徴は，第 1 段目「投入済み…」の c で原点を通ることを明示していること，第 2 段目「モデル集計」の a で原点回帰モデルにおける決定係数の注意が出力されている．これはかなり丁寧な扱いである．第 3 段目「分散分析」の b では自由度調整をしていないことの注意が促されているが，もう少しわかりやすい表現をしてほしいものだ．

```
In[3]:= Needs["LinearRegression`"]

In[4]:= regress = Regress[rtodata, {x}, x, IncludeConstant → False]

  DesignedRegress::tsos :
   Warning: the total sum of squares in the ANOVATable is uncorrected (not centered on the
      response mean) when there is no constant term in the model; it is designated U Total. >>

  DesignedRegress::rsqr :
   Warning: the RSquared and AdjustedRSquared diagnostics are redefined when
      there is no constant term in the model. >>

Out[4]= {ParameterTable →   |Estimate SE        TStat   PValue
                           x|0.547792 0.0734643 7.45658 0.0000386526 ,
  RSquared → 0.860682, AdjustedRSquared → 0.845202, EstimatedVariance → 2.07785,
                         DF  SumOfSq  MeanSq   FRatio   PValue
  ANOVATable →  Model    1   115.529  115.529  55.6005  0.0000386526
                Error    9   18.7006  2.07785                        }
                U Total  10  134.23
```

```
In[2]:= regress = LinearModelFit[rtodata, {z}, z, IncludeConstantBasis → False]
        regress["ParameterTable"]
        regress["ANOVATable"]
        regress["AdjustedRSquared"]
        regress["RSquared"]

Out[2]= FittedModel[ 0.547792 z ]

Out[3]=    Estimate  Standard Error  t Statistic  P-Value
        z  0.547792  0.0734643       7.45658      0.0000386526

               DF  SS       MS       F Statistic  P-Value
Out[4]= z      1   115.529  115.529  55.6005      0.0000386526
        Error  9   18.7006  2.07785
        Total  10  134.23

Out[5]= 0.845202

Out[6]= 0.860682
```

図４：Mathematica 6（左）と Mathematica 7（右）の出力

投入済み変数または除去された変数[b,c]

モデル	投入済み変数	除去された変数	方法
1	X[a]		投入

a. 必要な変数がすべて投入されました。
b. 従属変数: Y1
c. 原点を通る線型回帰

モデル集計

モデル	R	R2 乗[a]	調整済み R2 乗	推定値の標準誤差
1	.928[b]	.861	.845	1.4415

a. 原点を通る回帰 (切片のないモデル) に対して、R2 乗は回帰によって説明された原点についての従属変数の可変性の比率を測定します。これを切片を含むモデルに対する R2 乗と比較することはできません。
b. 予測値: X。

分散分析[c,d]

モデル		平方和	自由度	平均平方	F 値	有意確率
1	回帰	115.529	1	115.529	55.601	.000[a]
	残差	18.701	9	2.078		
	全体	134.230[b]	10			

a. 予測値: X。
b. 原点を通る回帰では定数が 0 であるため, 全体の平方和は定数に対して訂正されていません。
c. 従属変数: Y1
d. 原点を通る線型回帰

係数[a,b]

モデル		非標準化係数 B	非標準化係数 標準誤差	標準化係数 ベータ	t	有意確率
1	X	.548	.073	.928	7.457	.000

a. 従属変数: Y1
b. 原点を通る線型回帰

図５：SPSS 14.0J の出力

5.5　他のソフトウェア

これら以外のソフトウェアとして，MINITAB R14, R 2.15.0, SYSTAT 11, TSP 5.0 を調べているが，際立った特徴がないので，その出力値を表１に示す．

6　決定係数の比較

調査した統計ソフトウェアの重相関係数，決定係数，自由度調整済み決定係数等の出力を表１に示す．

では，改めて図１に示したデータに対する決定係数と自由度調整済み決定係数のソフトウェアによる違いに

表 1 ：各種統計ソフトウェアの出力

ソフトウェア名	重相関係数 R	決定係数 R^2	自由度調整済み R^2 $\overline{R}^2$	警告出力
EXCEL2002 以前	65535	−0.128	−0.239	なし
EXCEL2003 以降	0.928	0.861	0.75	なし
JMP6.0.3J	—	—	—	なし
Mathematica 6	—	0.861	0.845	あり
MINITAB R14	—	—	—	なし
R 2.7.1	—	0.861	0.845	なし
SPSS 14.0J	0.928	0.861	0.845	あり
SYSTAT J11	0.928	0.861	0.861	なし
TSP 5.0	—	0.362	0.362	なし
通常の回帰モデル	0.602	0.362	0.282	
整合性のある定義による出力	0.928	0.861	0.845	

注目する．整合性のある定義に基づく各係数の値は，表 1 の最下段に示す．この値と同じ値のソフトウェアはある意味正常で信頼性があるといえるが，値を出力しないのは，それなりの見識の下でそうしているものと思われるが，反対の見方をすれば全く出力がないというのもいかがなものであろうか．正確な意図を知りたいものである．一方，その値が全く違うものを出力したり，そもそも決定係数が負の値になるものは，言語道断であろう．この広く使われている表計算ソフトウェアは，明らかな間違いがユーザから指摘され，問い合わせがあっても，直す意志がないことが多くのユーザーから指摘されている（青木，2005；Heiser, 2012 など）．

7　原点回帰モデルの特異性

原点回帰モデルに関する諸統計量が通常の線形回帰モデルと比べて特異であることを示す．

まず，決定係数と図 1 で示したデータの決定係数の値は $R^2=0.861$，自由度調整済み決定係数は $\overline{R}^2=0.845$であった．このデータに，それぞれ x，y に100を加えて平行移動させて，各係数を計算すると，前者が0.9995，

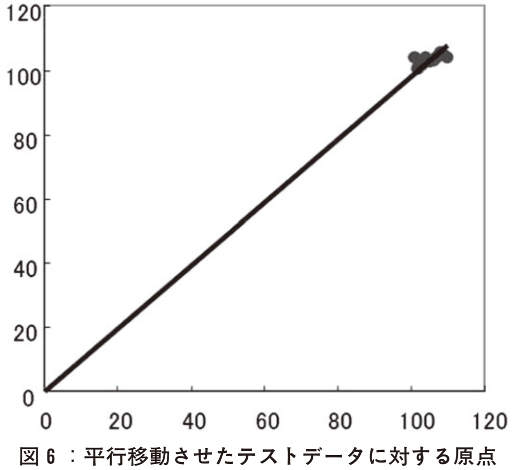


図 6 ：平行移動させたテストデータに対する原点回帰モデル

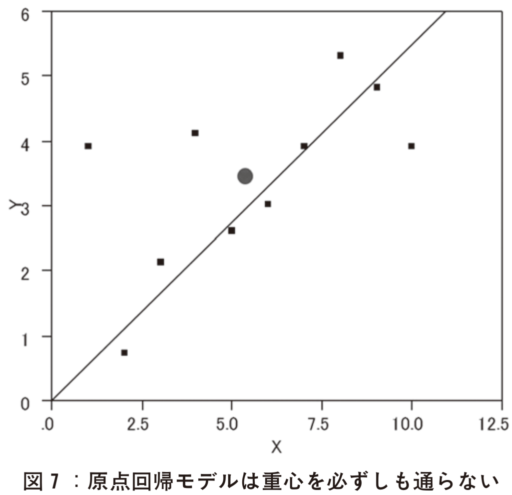


図 7 ：原点回帰モデルは重心を必ずしも通らない（大きな黒丸はデータの重心）

後者が0.9995であった(図6参照)．通常の回帰モデルではデータを平行移動しただけでは，これらの統計量は不変であるが，原点回帰モデルでは不変ではないということがこの簡単な例で確認できる．

次に，このモデルの特異なことは，図7に示すように，推定されたモデルが必ずしも重心を通らないということが挙げられる．これはモデルの推定方法にも関係することである．

最後に，再指摘となるが，目的変数の平方和分解が特異なことである．この特異な関係を基礎として(3)式の決定係数や(4)式の自由度調整済み決定係数が定義されているために，Excelにおける負の決定係数が誘発されているのである．

8 決定係数の特異性の原因

このような特異な統計量が出てくる原因を探っていく（中村・土屋，2007；Nakamura and Tsuchiya, 2013a)．通常の平方和分解である(1)式の左辺と右辺の第1項から，この分解は目的変数の平均からの変動を測ることを目的としたものであることがわかる．一方，(2)式の分解は，(1)式と同様の解釈をすると，原点からの変動を測ることを目的とした分解であることが理解できる．

このことから，前節でも指摘した原点回帰モデルの統計量の不変性がないことは，(2)式の左辺と右辺第1項の平方和が大きくなるために決定係数等が大きくなるということがこの式を見れば理解できる．ではなぜ，(2)式の平方和分解が使われるのか．それは次のように説明できる．原点回帰モデルを前提として，通常の平方和分解を行うと，次の式となる：

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2n\overline{y}(\overline{\hat{y}}-\overline{y}). \tag{5}$$

ここで，

$$\overline{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\frac{1}{n}\sum_{i=1}^{n}\hat{a}x_i=\hat{a}\overline{x}$$

である．(5)式には通常の平方和分解では出てこない右辺第3項がある．通常の回帰モデルでは $\overline{\hat{y}}=\overline{y}$ となるので，この項は通常消えてしまうが，原点回帰モデルでは $\overline{\hat{y}}\neq\overline{y}$ なのである．つまり，余剰項のために原点回帰モデルでは，数学的に整合させるために(2)式の平方和分解を使わざるを得ないのである．

では，この平方和分解に基づいて，次のように2種類の決定係数が定義できる．

$$\begin{aligned}R^2_{c_1}&=\frac{\text{回帰平方和}}{y\text{ の偏差平方和}}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}\\&=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}-\frac{2n\overline{y}(\overline{\hat{y}}-\overline{y})}{\sum_{i=1}^{n}(y_i-\overline{y})^2},\end{aligned}$$

$$\begin{aligned}R^2_{c_2}&=1-\frac{\text{回帰平方和}}{y\text{ の偏差平方和}}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}\\&=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}+\frac{2n\overline{y}(\overline{\hat{y}}-\overline{y})}{\sum_{i=1}^{n}(y_i-\overline{y})^2}.\end{aligned}$$

これらの定義に基づくと，$R^2_{c_1}$ と $R^2_{c_2}$ が正当な値 ($0<R^2_{c_1}<1,\ 0<R^2_{c_2}<1$) であるのは，

$$\frac{2n\overline{y}(\overline{\hat{y}}-\overline{y})}{\sum_{i=1}^{n}(y_i-\overline{y})^2}>0 \quad \text{すなわち} \quad \overline{y}\overline{\hat{y}}>\overline{y}^2$$

のときである．一方，不当な値 ($R^2_{c_1}>1$ や $R^2_{c_2}<0$) となるのは，

$$\frac{2n\overline{y}(\overline{\hat{y}}-\overline{y})}{\sum_{i=1}^{n}(y_i-\overline{y})^2}<0 \quad \text{すなわち} \quad \overline{y}\overline{\hat{y}}<\overline{y}^2$$

のときである．Excel の決定係数が負の値になるのは，この定義によると考えられる．

9 おわりに

本稿では，原点回帰モデルの統計量における特異性を確認した．このモデルは諸分野でよく使われるモデルであるので，著名なソフトウェアにおけるこれら統計量の計算結果の不一致は，非常に大きな問題であると考える．広く文献やウェブページを調べたが，この点に関して指摘した研究はほとんどないようである．参考文献には本文で引用したものも含めて，このモデルに関係する論文を示しておく．

今後の検討課題としては，原点回帰モデルにおけるソフトウェアの出力の注意喚起をすることと，平方和分解の新たな考え方による定義式を検討することである．

謝辞　本研究は札幌学院大学　研究促進奨励金　SGU-S10-198007-01の補助を受けています．

参考文献

青木繁伸 (2005).「Excel は，コンピュータ・ソフトウェアの三種の神器のようになっていますが，とんでもないこともあるというお話.」，URI；http://aoki2.si.gunma-u.ac.jp/Hanasi/excel/index.html，(同氏開設のウェブページ．最終更新の日付).

Bissell, A. F. (1992). Lines through the origin — is NO INT the answer?, *J. Appl. Statis.*, **19**(2), 193-210.

Casella, G. (1983). Leverage and regression through the origin, *Amer. Statis.*, **37**(2), 147-52.

Eamonn, M. (1982). Letters to the Editor, *Statistician*, **31**(3), 267-268.

Eisenhauer, C. C. (2003). Regression through the origin, *Teaching Statistics*, **25**(3), 76-80.

Gordon, H. A. (1981). Errors in computer packages. least squares regression through the origin. *Statistician*, **30**(1), 23-29.

Griffiths, D. (1982). Letters to the Editor, *Statistician*, **31**(3), 268-270.

Hahn, G. J. (1977). Fitting regression models with no intercept term, *J. Qual. Tech.*, **9**(2), 56-61.

Heiser, D. A. (2012). Microsoft EXCEL 2000, 2003 and 2007 faults, problems, workarounds and fixes, URI: http://www.daheiser.info/excel/frontpage.html.

Hawkins, D. M. (1980). A note in fitting a regression without an intercept term, *Amer. Statist.*, **39**, 233.

中村永友，土屋高宏 (2007)．焦点をもつ回帰直線群の推定とその周辺，応用統計学，**36**(1)，31-50．

Nakamura, N. and T. Tsuchiya (2013a). A model of regression lines through a common point: Estimation of the focal point in wind-blown sand phenomena, *J. Appl. Statis.*, **40**, DOI: 10.1080/02664763.2013.772570.

Nakamura, N. and T. Tsuchiya (2013b). An alternative definition of R-square for regression through the origin and its asymptotic properties, submitted.

The Linear Regression Model through the Origin and the Statistical Data Analysis Software

by

Nagatomo NAKAMURA[1] and Takahiro TSUCHIYA[2]

Abstract

This report aims to reveal the characteristics and important notes for the model of linear regression passing through the origin. Because there exist many statistical data analysis software which output different results of statistics for an arbitrary dataset.

Keywords: Linear Regression Model through the Origin, the Coefficient of Determination, Coefficient of Determinationadjusted for Degree of Freedom, Statistical Model, Statistical Data Analysis Software

[1]Department of Economics, Sapporo Gakuin University, nagatomo@sgu.ac.jp
[2]Department of Mathematics, Josai University, takahiro@math.josai.ac.jp