回帰分析における整合性評価の一考察

A Note on the Necessity of the Consistency Evaluation in Linear Regression Analysis

田　中　浩　光
Hiromitsu TANAKA

和文要旨：
本稿では、説明変数が２個とする回帰分析における整合性評価について考察する。回帰分析の結果に対する従来の評価方法では、外れ値あるいは説明変数間の冗長性などを対象とする。通常、評価基準には平均平方誤差、予測平均平方誤差など期待的効果が採用されているが、所与の回帰データに対し、必ずしも整合的な解釈を与えるものでない。本稿では、回帰標本上において、解釈上の整合性を確保することの必要性について確認する。

英文要旨：
We consider at the consistency of regression results with two explanatory variables. This purpose of conventional consistency evaluation methods are the detection of outliers or the redundancy of explanatory variables. We ordinarily have the mean square errors or predictive mean square errors as evaluation criteria.However,these criteria do not imply a consistent explanation for given regression data.In this article, we confirm the necessity of consistency evaluation for regression results.

和文キーワード：回帰標本、相関係数、準共線性、抑制、Schey のデータ
英文キーワード：regression sample, correlation coefficients, near collinearity, suppression, Schey's data

目　　次

1. はじめに

最小二乗法に基づく推定回帰式には、解釈上の問題が指摘されている。説明変数値間の相関係数の影響を受ける所謂多重共線性問題（あるいは準共線性問題）が代表的である。技法上の対策として統計的性能の改良を意図する主成分回帰・リッヂ回帰とする縮小回帰法など多様な接近が提示されているが、これらは平均平方誤差・予測平均誤差など期待的効果の向上を目的とするが、所与の回帰データに対し必ずしも整合的な解釈を与えるものでない。一方、解釈上の問題に留意して、回帰標本上での整合性を確保することを第 1 として、その上で回帰性を得る推定回帰式の構成が考えられる（田中（2011、2012））。

本稿では、説明変数が 2 個の場合に限定する。回帰分析の実施・結果において、整合性評価が要請されなければならないことについて、若干な数値例と併せて考察する。本稿では、首尾一貫、推定回帰式の適切性を得るためには、回帰データと推定回帰式が固有知識など解釈上の整合性の意味で合致することが望ましい立場を採る。

2. 回帰標本と観測過程

本稿では、目的変数 y を説明変数 x_1、x_2 で説明する重回帰分析について考察する。n 個の観測個体に基づく観測値からなる回帰標本（Y、X_1、X_2）に基づいて回帰分析を実施する。回帰標本は、主題の統計的咀嚼から、母集団の規定・標本の抽出・観測変数の選定・観測定義域の特定を伴う観測の実施など一連の作業からなる観測過程に基づいて得られる。観測過程での各作業の崩れは、回帰標本に影響を及ぼすことになる。代表的な難題の 1 つである、外れ値の問題は、標本の抽出の影響を強く受ける。説明変数間の相関から惹起する多重共線性の問題は、主として説明変数の選定に起因する。他に、規定誤差の問題、回帰性の問題についても説明変数の選定の影響が大きいと考えられる。図 1 に、統計的データ解析の過程、とくに回帰標本の崩れを惹起する原因と考えられる標本抽出、説明変数の選定に着目して、推定回帰式の適切性を得るための問題点を簡易な図で整理する。

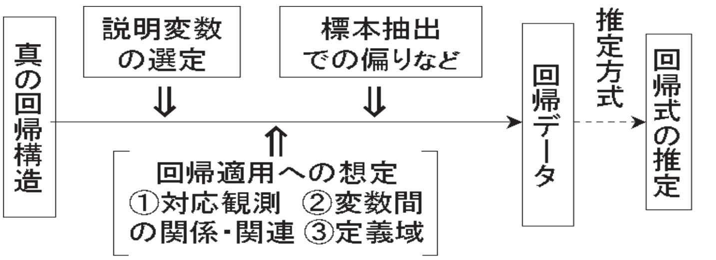

図 1. 回帰データと推定

3. 関連する統計値の要約

本報告では、目的変数 y を設定したとき、二つの説明変数 x_1、x_2 で、被説明変数である目的変数を記述することを考える。本報告では、標本のみに着目して考察を進める。標本サイズは n とする。また、計算の便宜上、切片項を省略し、相関形式のデータとする。すなわち、. y 、x_1、x_2 については、それぞれゼロとする。また、

$$\Sigma x_1 = 1 \quad 、\Sigma x_2 = 1$$
$$\Sigma y = 1$$

とする。

y と x_1、y と x_2、x_1 と x_2 の相関係数をそれぞれ r_{y1}、r_{y2}、r_{12} とするとき、x_1 の影響を取り除くときの Y と x_2 の偏相関係数は、次式で与えられる。

$$r_{y2\cdot1} = (r_{y2} - r_{12}r_{y1}) / \{(1-r_{y1}^2)(1-r_{12}^2)\}^{1/2} \tag{1}$$

重相関係数の二乗は、次式で与えられる。

$$R^2 = r_{y1}^2 + (1-r_{y1}^2)\ r_{y2\cdot1}^2 \tag{2}$$

Y を X_1 で説明すときの寄与率（決定係数）は r_{y1}^2 で表される。同様に、Y を X_2 で説明すときの寄与率（決定係数）は r_{y2}^2 で表される。

単回帰係数の推定値 b_1、b_2 を、重回帰係数の推定値 $b_{1\cdot[1,2]}$、$b_{2\cdot[1,2]}$ を相関係数で表すことができる。

$$b_1 = r_{y1} \tag{3}$$

$$b_2 = r_{y2} \tag{4}$$

$$b_{1\cdot[1,2]} = (r_{y1} - r_{y2}\,r_{12}) / (1 - r_{12}^2) \tag{5}$$

$$b_{2\cdot[1,2]} = (r_{y2} - r_{y1}\,r_{12}) / (1 - r_{12}^2) \tag{6}$$

4. 回帰標本の崩れの原因と影響

本節では、回帰データの観測過程に基づき、回帰標本の崩れによる最小二乗推定値への影響を考察する。あわせて、回帰標本の崩れの原因を整理する。

（1）推定回帰式への影響

- i　予測が不十分、説明力が不十分
- ii　回帰係数の推定値と相関係数値で符号の意味で不一致
- iii　回帰係数の推定値と固有知識の乖離

（2）回帰標本の崩れの原因

- i　モデルの前提との乖離
  誤差の諸問題、非線形性
- ii　標本抽出
  異常値（外れ値）の混入、部分（偏）標本の抽出
- iii　説明変数の選定
  モデルの規定の誤り（規定誤差）、説明変数の不足（回帰性が小さい）、過剰な説明変数集合、高相関な説明変数対（準共線性）

回帰診断の立場に沿い、回帰標本の崩れの形態を症状、崩れの原因を疾患とみなして、その対策を治療とする。通常は、崩れの形態を対立仮説とする検定の枠組みのなかで診断が行われる。対策としての治療には、データの追加・削除と層別などの根治療法とデータの変換など症状に適応する対処療法に分けることができる。詳細は表1を参照（田中（2003））。

5. 整合性評価の必要性

本節では、回帰分析における、とくに推定回帰式において、整合性評価の必要性と留意点、そして整合性基準について整理する。

表1. 診断と治療

崩壊原因（疾患）	崩壊形態（症状）							原因対策（根治）療法
	モデルの前提条件崩壊	準特異（X）	はずれ値	回帰寄与小	モデル不適合	過度のあてはめ	現象と不整合	
偏サンプリング	*	*	−	*	*	*	*	計画された追加観測
拡大サンプリング	*	−	*	*	*	−	*	背景因子による観測値集合の層別
分布の端値	*	*	*	*	*	−	*	観測値の削除
標本サイズ小	*	*	*	*	*	*	*	追加観測
説明変数過多	−	*	−	−	−	*	*	説明変数の選択
説明変数不足	*	−	*	*	*	−	*	説明変数の追加
異質標本の併合	*	−	*	*	*	−	*	観測値集合の分割
診断方法	分散均一性検定 無相関性検定 正規性検定	多重共線性検定	はずれ値検定	回帰性検定 回帰寄与率	モデル偏検定	偏係数有意性	解釈可能性	
個別対策（対処）療法	データ変換 観測地の削除	変数選択 リッジ回帰 主成分回帰 観測値の追加	データ変換 観測値の削除	データ変換 変数の追加 交互作用 観測値の追加	変数の追加 交互作用	変数選択 リッジ回帰 主成分回帰	変数選択 主成分回帰	

注）表中の*、−はそれぞれ疾患が症状に影響を及ぼす、あるいは及ぼさないことを表す。

(1) 整合性評価の必要性

① 推定結果は手許のデータ、固有知識において、解釈可能性を有することが望ましい。統計的基準では、期待(的)効果をみるが、手許のデータ(標本)ではない。

② 推定結果において、すくなくとも合致してほしい最小限の確認事項を満足する。その上で、高級な(複雑な)解析手法の適用に対しても受容できる。

③ 回帰診断の前段階として捉えることもできる。

(2) 整合性評価での留意点

辻褄をあわせる対照に留意する。

① 確固となる(信頼性の高い)、専門知識に基づくこと。

② データから近い方が良い。データに対して操作が入らない方が良い。

③ 統計的に記述できること。

(3) 整合性基準の構成の留意点

望ましい整合性基準の構成において、留意すべき点をあげる。

① 評価の対象は標本上に限局する。

② 整合性を診るときの突合せとなる対照は、固有知識か、データに基づいた統計的基本知識が望ましい。

③ 基準関数は整合性からの乖離に対して、単調性を有すること

とくに、下記の点について、注意する。

i 整合性の存在(有無)

① 悪条件、奇異となる現象に注目して、その生起領域を活用する。

② 基本情報である相関係数に対し、論理上の整合性に基づいて生起領域を生成する。

ii 整合性の程度

整合性基準関数は、上記にある(3)の②と③を満たさなければならない。Shieh(2001)が提示しているH関数は、抑制(suppression)の程度を表す指標であるが、データに基づく基本知識の乖離としての関数として表現されていないことに注意が要る。

(4) 整合性領域の構成の視点

i 悪条件、あるいは奇異な現象を生起する領域に着目する。それらの補集合の領域が整合性領域となる。

ii 回帰データの生成に着目し、固有知識を活用して、整合性を回帰モデルの構成要素とデータで定式化する。

iii 回帰データ(行列)の分割を通して、行方向である観測個体の層別、列方向である説明変数の選定において、回帰平方和間での整合性関係(劣加法性)を導入する。

本稿では、整合性領域の構成の1例として、視点iに沿う下記の2点をとりあげる。

・準共線性の現象の特定

・抑制(suppression)の現象

6. 回帰分析と整合性

本節では、回帰分析での整合性問題について、整合性の標的対象の明確化と分析後の解釈可能性を確保する視点に絞って、若干の整理を試みる。回帰分析の実施にあたっては、分析の目的が第1にあり、その上で、標本の抽出と観測変数の選定(回帰分析では、目的変数と説明変数)が実施される。所謂、観測過程に十分注意を払うとする前提において、回帰分析の実行がなされることになる。しかし、実地においては、往々にして、分析結果である推定回帰係数が必ずしも固有知識と合致することがないことに遭遇する。これらの原因の1つに、整合性の対象が明確でないこと、また推定方式が統計的性能の改良に重きを置かれている点があげられる。

(1) 整合性を診る標的対象

・推定結果と固有知識の乖離:推定回帰係数の解釈

・回帰方式に付随する前提・制約条件の成立(満足):加法性、独立性など

・関係構造の把握・予測問題における母集団の明確化:地域、観測個体集団など

(2) 解釈可能性の確保

・観測データの抽出が分析目的に合致しているか。

・観測データ上での整合性:半標本の活用

・基本情報と分析結果との整合性:相関

　　情報と推定回帰係数

以下、回帰分析の過程と整合性と他の性質について整理する。

（1）研究主題の明確化
　　統計的問題へ咀嚼：忠実性
（2）標本の抽出・データの観測
　　データの偏りとバラつき：信頼性
（3）回帰モデルの同定
　　前提条件の確認
（4）回帰式の推定
　・推定方式の統計的妥当性：母数などの統計性質の確保
　・推定結果と事前の知識：合致性、
　・観測変数間での解釈上の整合性
　・データ固有に内在する制約条件を満たす内的整合性
＊　一連の手順に対し、分析方針に沿う首尾一貫性
（5）推定（予測）回帰式の妥当性
　　外的妥当性、内的妥当性

図2には、望ましい推定回帰式の手順を要約する。

主題
↓　　統計的咀嚼
↓　　回帰標本の抽出
↓　　回帰変数の選定
回帰データ
↓
↓　基本統計量　　相関係数
↓
回帰分析（モデル）　回帰係数
↓　　推定方式
回帰結果　　回帰係数推定値
↓　　決定係数
↓
解釈　　固有知識
↓　　整合性確認
適切な推定回帰式

図2. 望ましい推定回帰式

7. 相関係数の活用

本節では、2個の説明変数からなる回帰分析に限定して、最小二乗推定値の整合性について考察する。とくに、回帰係数に着目し、最小二乗推定値間の符号、大きさをとりあげる。解釈上の整合性を確保する立場に沿って回帰推定を考察する。その際、観測データを得た事後の情報として、目的変数と説明変数に対応する観測特性間の相関情報に注目する。たとえば、整合性を課するものとして、説明変数間の相関係数の符号に基づく指標が考えられる。

指標（1）、(2)、(3）は、準共線性の現象、そして指標（4）は抑制（suppression）の生起領域の排反を意味する。

（1）① $r_{12} > 0 \rightarrow b_1 b_2 > 0$
　　② $r_{12} < 0 \rightarrow b_1 b_2 < 0$
（2）① $r_{12} > 0 \rightarrow b_{1\cdot(1,2)}\, b_{2\cdot(1,2)} > 0$
　　② $r_{12} < 0 \rightarrow b_{1\cdot(1,2)}\, b_{2\cdot(1,2)} < 0$
（3）① $r_{12} > 0 \rightarrow b_1 b_2 > 0$
　　　　$b_{1\cdot(1,2)}\, b_{2\cdot(1,2)} > 0$
　　② $r_{12} < 0 \rightarrow b_1 b_2 < 0$
　　　　$b_{1\cdot(1,2)}\, b_{2\cdot(1,2)} < 0$
（4）$R^2 < r_{y1}^{\,2} + r_{y2}^{\,2}$

相関データの形式のもとで、準共線性の現象の生起領域（(1)、(2)、(3)）は、相関係数の比γ（$= r_{y2} / r_{y1}$）とr_{12}を用いて、

$R_C = \{(r_{12}, \gamma) \mid r_{12} > 0, \gamma < 0, (1 - \gamma r_{12})(\gamma - r_{12}) < 0\} \cup \{(r_{12}, \gamma) \mid r_{12} > 0, \gamma < 0\} \cup \{(r_{12}, \gamma) \mid r_{12} < 0, \gamma > 0, (1 - \gamma r_{12})(\gamma - r_{12}) > 0\} \cup \{(r_{12}, \gamma) \mid r_{12} < 0, \gamma > 0\}$

となる。

同様に、suppressionの生起領域（4）は、

$R_S = \{(r_{12}, \gamma) \mid r_{12} > 0, (1+\gamma^2) r_{12} > 2\gamma\} \cup \{(r_{12}, \gamma) \mid r_{12} < 0, (1+\gamma^2) r_{12} < 2\gamma\}$

となる。

上記の2生起領域は、図3を参照。ここに、灰色部分は、suppressionの不整合領域を表す。このとき、$R_C \supset R_S$を得る（田中（2012））。

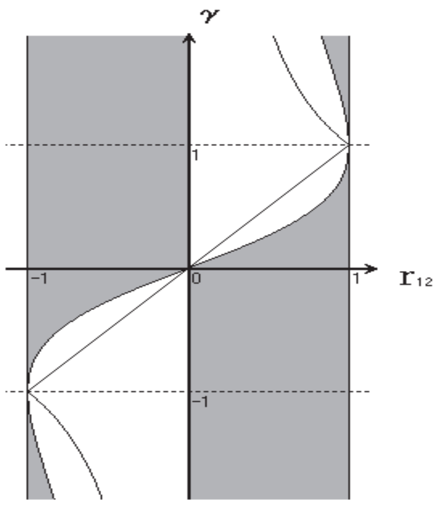

図 3. 生起領域 R_C、R_S

8. 数値例

本節では、最小二乗法に基づく推定回帰式に対し、整合性の点検の評価が必要であることについて、Schey（1993）のデータ集合（n=10、p =2）をとりあげて、考察・吟味する。Schey のデータは、8 本（A ～ H）のデータからなるデータ集合であり、本稿では、相関係数値と回帰係の最小二乗推定値の間で符号の意味で合致しない（不整合）データ（C、D、E、G）を取りあげて吟味する。各データの基本統計値と決定係数（寄与率）については、表 2 を参照する。

（C） 相関係数値と 2 つの単（重）回帰係数推定値は異符号

r_{12} =−0.258 < 0 であるが、単回帰係数値は、それぞれ b_1 = 3.160 > 0、b_2 = 0.150 > 0 であり、b_1b_2 > 0 となる。7 節の（1）の②に反する結果となる。また、重回帰推定値は、それぞれ $b_{1\cdot(1,2)}$ = 3.655 > 0、$b_{2\cdot(1,2)}$ = 0.300 > 0 であり、$b_{1\cdot(1,2)}\,b_{2\cdot(1,2)}$ > 0 となる。7 節の（2）の②に反する結果となる。次いで、決定係数では、R^2 = 0.964、$r_{y1}^2 + r_{y2}^2$ = 0.787 であり、$R^2 > r_{y1}^2 + r_{y2}^2$ となる。7 節の（4）に反する結果となる。

（D） 相関係数値と 2 つの単（重）回帰係数推定値は異符号

r_{12} = 0.707 > 0 であるが、単回帰係数値は、それぞれ b_1 =−0.413 < 0、b_2 = 0.364 > 0 であり、b_1b_2 < 0 となる。7 節の（1）の①に反する結果となる。また、重回帰推定値は、それぞれ $b_{1\cdot(1,2)}$ = -1.952 < 0、$b_{2\cdot(1,2)}$ = 0.995 > 0 であり、$b_{1\cdot(1,2)}\,b_{2\cdot(1,2)}$ < 0 となる。7 節の（2）の①に反する結果となる。次いで、決定係数では、R^2 = 0.862、$r_{y1}^2 + r_{y2}^2$ = 0.273 であり、$R^2 > r_{y1}^2 + r_{y2}^2$ となる。7 節の（4）に反する結果となる。

表 2. Schey（1993）のデータ

（C）

r_{y1}	r_{y2}	r_{12}	b_1	b_2	$b_{1\cdot(1,2)}$	$b_{2\cdot(1,2)}$	R^2	$r_{y1}^2 + r_{y2}^2$	γ
0.254	−0.258	3.160	0.150	3.655	0.300	0.964	0.787	0.299	0.850

（D）

r_{y1}	r_{y2}	r_{12}	b_1	b_2	$b_{1\cdot(1,2)}$	$b_{2\cdot(1,2)}$	R^2	$r_{y1}^2 + r_{y2}^2$	γ
−0.241	0.464	0.707	−0.413	0.364	−1.952	0.995	0.862	0.273	−1.931

（E）

r_{y1}	r_{y2}	r_{12}	b_1	b_2	$b_{1\cdot(1,2)}$	$b_{2\cdot(1,2)}$	R^2	$r_{y1}^2 + r_{y2}^2$	γ
0.853	−0.477	−0.866	0.784	−0.920	0.674	0.457	0.654	0.699	−0.731

（G）

r_{y1}	r_{y2}	r_{12}	b_1	b_2	$b_{1\cdot(1,2)}$	$b_{2\cdot(1,2)}$	R^2	$r_{y1}^2 + r_{y2}^2$	γ
0.454	−near0	−0.866	1.590	−0.834	2.037	1.699	0.826	0.206	(−)near0

（E）　相関係数値と２つの重回帰係数推定値は異符号

$r_{12}=-0.866<0$ であるが、単回帰係数値は、それぞれ $b_1=0.784>0$、$b_2=-0.920<0$ であり、$b_1b_2<0$ となる。7節の（1）の②に相応する結果となる。しかし、重回帰推定値は、それぞれ $b_{1\cdot(1、2)}=1.147>0$、$b_{2\cdot(1、2)}=0.674>0$ であり、$b_{1\cdot(1、2)}\,b_{2\cdot(1、2)}>0$ となる。7節の(2)の②に反する結果である。したがって、7節の(3)の②に反する結果となる。次いで、決定係数では、$R^2=0.457$、$r_{y1}^2+r_{y2}^2=0.654$ であり、$R^2<r_{y1}^2+r_{y2}^2$ となる。7節の（4）に相応する結果である。

（G）　相関係数値と２つの重回帰係数推定値は異符号で、抑制（suppression）が生起

$r_{12}=-0.866<0$ であるが、単回帰係数値は、それぞれ $b_1=1.590>0$、$b_2=-0.834<0$ であり、$b_1b_2<0$ となる。7節の（1）の②に相応する結果となる。しかし、重回帰推定値は、それぞれ $b_{1\cdot(1、2)}=2.037>0$、$b_{2\cdot(1、2)}=1.699>0$ であり、$b_{1\cdot(1、2)}\,b_{2\cdot(1、2)}>0$ となる。7節の(2)の②に反する結果である。したがって、7節の（3）の②に反する結果となる。次いで、決定係数では、$R^2=0.826$、$r_{y1}^2+r_{y2}^2=0.206$ であり、$R^2>r_{y1}^2+r_{y2}^2$ となる。7節の（4）に反するる結果である。

9. 若干の考察

本稿では、回帰分析の整合性評価の必要性を主張しているが、従来の用いられている方法と比較して、何故に要請されている利点があるかについて言及する。従来の方法との比較を通して、その利点を探る。以下の疑問を通して、整合性評価の有用性について整理する。

（1）従来（通常）に回帰分析との相違？
（2）推定回帰式の評価において、整合性基準は、従来（通常）の基準との相違？
（3）整合性基準の構成において、原因を特定化するのであれば、従来（通常）の点検方式との相違？
（4）標本上において、特定の原因である縛りを課しているが、制約つき最小二乗解との相違？

上記の疑問を考察することで、整合性評価の必要性について言及する。

（1）従来（通常）に回帰分析との相違？

往々にして、分析結果である推定回帰係数が必ずしも固有知識と合致しないことがある。原因の１つとして、従来の統計的評価基準が平均平方誤差や予測平均平方誤差を採用していること、すなわち期待的効果の最適化を図ることが主眼となり、回帰データそのものが評価対象となっていない。本稿での整合性評価の対象は回帰データであり、分析結果が基礎知識との合致することを意図する。すなわち、回帰データに基づく知識と合致することが要請されることになり、本稿では、基本統計値である相関係数値との関係が整合的であることを確保する。

（2）推定回帰式の評価において、整合性基準は、従来（通常）の基準との相違？

従来の基準が、推定量の統計的性能を測ることに集中するが、回帰データに近い水準での知識に基づいて、推定量間の関係から整合性を確保する。たとえば、本稿では、基本統計値である相関係数値と回帰推定値の関係において整合性を導入しているがこと、また回帰分析結果である決定係数に着目して、同時回帰と周辺回帰での決定係数間の大小関係（suppression：抑制）に対して整合性を導入することができる。

（3）整合性基準の構成において、原因を特定化するのであれば、従来（通常）の点検方式との相違？

（3）に対しては、（1）と（2）の疑問に対する回答と同様の、回帰データに密着する知識に基づく整合関係を原因とする。従来の点検方式は、モデルの構成要素をとりあげて、その関連知識に基づく仮説を原因とする。た

とえば、回帰係数の有意性検定、あるいは外れ値の検出・影響観測値の検出をあげることができる。

(4) 標本上において、特定の整合関係を縛りとして課しているが、制約つき最小二乗解との相違？

回帰標本において、整合性の領域内に存在するか否かは、分析結果に基づく観測個体が制約条件を成立するか否かを個々の観測個体ごと判定することである。この作業は、個々の観測個体の整合性の有無を確認する。一方、制約付き最小二乗解は個々の観測個体を対象とするのでなく、全体の観測個体集合を対象として、制約条件を満足することに狙いをおく。

10. おわりに

本稿では、回帰分析の適用に際して有用性の視点に基づいて、とくに、推定回帰式の整合性評価が必要であることを主張した。回帰分析での整合性を評価することは、実践的な要請に応えるものである。推定回帰式に対し、たとえば回帰係数の有意性検定を通しての従来の評価方法を適用する前に、整合性評価を実施することで、より有用な統計的知見を得ることになる。回帰診断の多様な方法を適用するとき、診断目的の対象となる前準備としての意味を有する。

参考文献

(1). Subrahmanyam,M. (1972). A property of simple least squares estimates. Sankhya, 355-356

(2). Suich,R.& Derringer,G.C. (1977). Is the regression equation adequate? one-criteria. Technometrics, 19, 213-216.

(3). Hamilton,D. (1987). Some times $R^2 > r_{y1}^2 + r_{y2}^2$, correlated variables are not always redundant.The American Statistician, 41, 2, 129-132.

(4). 田中浩光 (2003)。回帰分析における Suppression と準共線性、2003 年度統計関連学会連合大会講演報告集。

(5). 田中浩光 (2011)。適切な推定回帰式と整合性問題、2011 年度統計関連学会連合大会講演報告集。

(6). 田中浩光 (2012)。回帰標本上の整合性を有する推定回帰式について、2012 年度統計関連学会連合大会講演報告集。

(7). 田中浩光 (2014a)。整合性回帰における 1 つの定式化、第 36 回（通算 56 回）日本経営数学会研究大会報告要旨集。

(8). 田中浩光 (2014b)。回帰データと推定回帰式の整合性について、2013 年度統計関連学会連合大会講演報告集。

(9). 田中浩光 (2017a)。回帰分析における整合性問題について、2017 年度品質管理学会第 35 回中部支部研究会研究発表報告集。

(10). 田中浩光 (2017b)。回帰分析における整合性基準について、2017 年度統計関連学会連合大会講演報告集。