平均への回帰を考慮した テストスコア変化の分析について

斉　藤　慎　一

1.　はじめに

本稿では、標準化されたテストを 2 回受験した場合のスコアの変化をどのように分析するべきかについて、平均への回帰（regression toward the mean=RTM）（回帰効果［regression effect］とも言う）に焦点をあてて論じていく。事前—事後デザイン（pretest—posttest designs）の研究から得られたデータを分析する際、平均への回帰の影響を調整した場合としない場合で分析結果が異なってくることが知られている（e.g., Bonate, 2000; Campbell & Kenny, 1999）。そこで、本稿ではある標準化されたテストを 2 回受験した場合を例として、反復測定されたデータの分析によく用いられる統計手法の分析結果を比較することを通じて、平均への回帰を調整した場合としない場合で具体的にどのように結果が異なるのかを検討することを目的とする。さらに、1970 年代に提唱された、平均への回帰（回帰効果）の大きさを推定する方法についても論じていく。

2.　平均への回帰

2–1　平均への回帰とは

平均への回帰（回帰効果）とは、1 回目に極端な値が観測された場合、2 回目には 1 回目よりも平均に近い値が観測されることが多いという現象を指す（時間的順序は逆でも同じことが言える）。多くの研究者により、同一人物に繰り返し同じもの（例えば，テストのスコアやコレステロール値など）を測定した場合、2 回のスコアが完全に相関している場合でない限り平

均への回帰が生じる（e.g., Bonate, 2000; Campbell & Kenny, 1999; Iwasaki & Kawada, 2007）と指摘されてきた。従って、反復測定データを分析する際には平均への回帰を考慮に入れなくてはならない（e.g., Bonate, 2000; Campbell & Kenny, 1999; 岩崎・河田 2007）。

ただし、厳密に言うと Rogosa, Brandt, and Zimowski（1982; Rogosa, 1995）によって、平均への回帰は 1 回目の測定値と変化得点（=2 回目と 1 回目の測定値の差）との間に負の相関がある時にのみ生じるもので、どのような場合にも必ず生じるものではないことが証明されている。しかし、現実には、1 回目の測定値と変化得点は負の相関を示すことが多いため、平均への回帰は事実上普遍的な現象といえるかもしれない（Allison, 1999）(1)。

この平均への回帰は Galton（1886）により報告されて以来、これまで医学や疫学などの分野では盛んに検討されてきたが（徳永 2001）、例えば社会心理学（Nielsen, Karpatschof, & Kreiner, 2007; Yu & Chen, 2015）やコミュニケーション研究（Hansen, & Pedersen, 2014）などの分野では必ずしも十分に考慮されてきたわけではないといわれる。

2–2　平均への回帰が生じる理由

標準化されたテストを例にすると、平均への回帰が生じる理由として一般に次のように説明される。受験者がテストを受けた時たまたまいつもより体調が悪かったとか、いつも以上に緊張して実力を出し切れなかったため（つまり運が悪かったため）実際の実力より低いスコアになってしまった。逆に、たまたまテスト内容の一部が自分に馴染みのあるテーマであった、あるいは当て推量の結果が通常より多く正解であったなどの理由（つまり運が良かったおかげ）で実力以上のスコアを得られた。しかし、1 回目は上に挙げたような理由でたまたま運が良かったとか、逆に運が悪かった場合、2 回目も同じくらいに運が良いとか悪いということは可能性として低い。従って、1 回目のスコアがかなり良かった（あるいは悪かった）人は、2 回目のスコアは 1 回目より悪くなる（あるいは良くなる）傾向がある（要するに「平均

へ回帰」する）。

2–3　平均への回帰の調整法

同じテストを 2 回受験した場合のスコアの変化について、2 つ以上の群間に有意差があるかどうかを検討する場合、2 要因分散分析（対応あり×対応なし）や、2 回目のスコアから 1 回目のスコアを引いた変化得点（ないしは差得点）を従属変数とした *t* 検定（2 群間の場合）や 1 要因分散分析（ANOVA）を用いることが少なくない。しかし、多くの専門家が指摘するように、2 要因分散分析や変化得点を用いた分散分析は平均への回帰の影響を考慮していない（e.g., Bonate, 2000; Campbell & Kenny, 1999）[(2)]。平均への回帰の影響を調整する方法はこれまでいくつか提起されてきたが、データ収集後の事後的調整法として最もよく用いられる手法の一つは、1 回目の測定値を共変量として用いた共分散分析（ANCOVA）であろう（Bonate, 2000; Dimitrov & Rumrill, 2003）。また、共分散分析ほど知られていないようだが、それ以外に残差得点分析（Residual［change］score analysis）と呼ばれる、1 回目の測定値から回帰分析によって予測した 2 回目の値と実際の 2 回目の測定値との差（つまり「残差」（Residual））を従属変数として分散分析を行う方法も検討されてきた（e.g, Campbell & Kenny, 1999; MacKinnon, 2008）[(3)]。ただし、この方法には反対の立場を表明する研究者も存在する（e.g., Maxwell, Delaney, & Manheimer, 1985; Forbes & Carlin, 2005）。本稿では、これらの分析方法を比較検討する。

2–4　平均への回帰の大きさを推定する方法

次に、平均への回帰の大きさを予測する方法について述べる。社会心理学やコミュニケーション研究の分野ではあまり知られていないようだが、1970 年代に Davis（1976）や Gardner & Heady（1974）によって平均への回帰を予測する方法が提唱されている（近年の適用例として、例えば Barnett et al., 2004; Gmel et al., 2007; Linden, 2013 などがある）。

ここでは、ある標準化されたテストを 2 回受験した場合について考える。初回受験時のスコアを y_1 とし再受験時のスコアを y_2 とする。また、いずれのテストも母平均μ、母標準偏差σの正規分布に従うと仮定する。ρは 2 回のテストスコアの相関係数である。ある値（ここでは k とする）をカットオフ値に設定し、k を $z = |k - \mu| / \sigma$で標準化すると、カットオフ値 k_1 未満の者の 1 回目の受験時のスコアの平均値の予測値は

$$E(y_1 \mid y_1 < k_1) = \mu - C\sigma \qquad (1\text{–}1)$$

同様に、カットオフ値 k_2 を超えた者の 1 回目の受験時の平均値の予測値は

$$E(y_1 \mid y_1 > k_2) = \mu + C\sigma \qquad (1\text{–}2)$$

ここで、$C = \varphi(z) / [1 - \Phi(z)]$, $\varphi(z)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right]$, $\Phi(z) = \int_{-\infty}^{z}\varphi(x)dx$ である。

また、これらの受験者の再受験時のスコアの平均値の予測値は

$$E(y_2 \mid y_1 < k_1)=\mu - \rho C\sigma \quad \text{ないしは } E(y_2 \mid y_1 > k_2) = \mu + \rho C\sigma \qquad (2)$$

（1）から（2）を引いたものが平均への回帰の幅となる.

$$\text{RTM effect} = (\mu \pm C\sigma) - (\mu \pm \rho C\sigma) = C\sigma\,(1 - \rho) \qquad (3)$$

一般には母平均μおよび母標準偏差σは未知なので、実際に利用する際にはそれぞれの推定値$\hat{\mu}$と$\hat{\sigma}$を用いる。

3. 方法

3–1 分析に用いるデータの説明

本稿では、ある大学の1年生300人が入学直後の4月にTOEFLやTOEICのような標準化されたテストを受験し、その同じ学生たちが1年後に再度同じテストを受験した場合のスコアの変化を例にして考えていく(4)。その際、本稿では一般的に馴染みが深い0〜100の間でスコアを表示するテストとして論じていく。また、今回は1回目のスコアが60点未満の学生62人（20.7%）（ここでは便宜上「成績下位群」と呼ぶ）には他の80%の学生（ここでは便宜上「一般群」と呼ぶ）とは別に週一回の補習授業を行ったと仮定しておく。

最初に、図1〜図2および表1で今回使用するデータの基本情報を示しておく。図1は2回のテストスコアをヒストグラムで比較したもので、図2は散布図である。図2の点線（Y=X）は1回目のスコアと2回目のスコアが同じであった場合である。

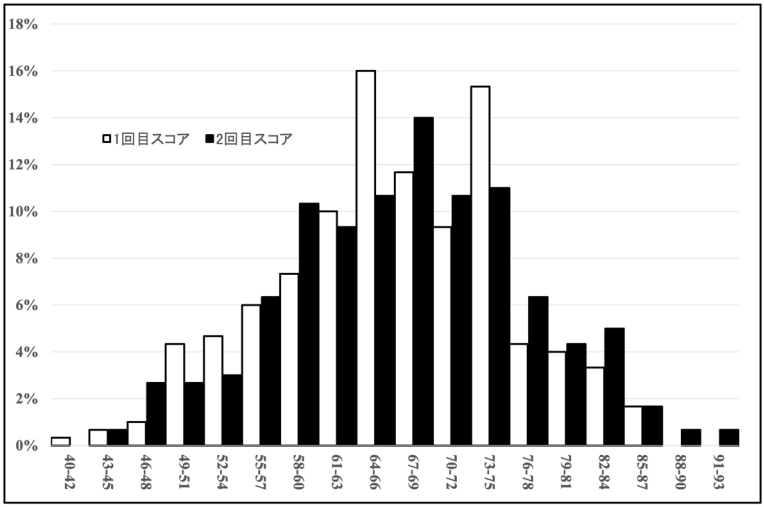

図1　使用するデータのヒストグラム

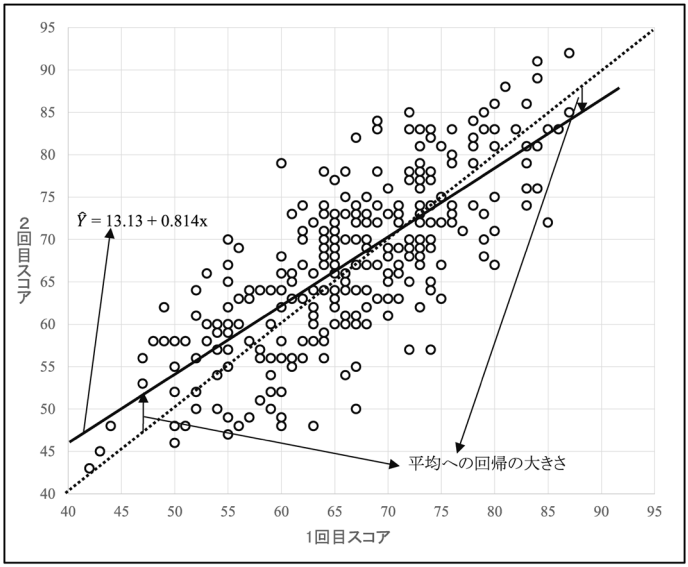

図 2　1 回目のスコアと 2 回目のスコアの散布図

表 1　本稿で使用したデータの平均値、標準偏差、標準誤差および *t* 検定の結果

	M	*SD*	*SE*
1 回目スコア（n=300）	66.38	8.98	0.518
2 回目スコア（n=300）	67.13	9.56	0.552
対応のある *t* 検定	$t(299)=-2.04$, $p=.042$, $d=0.12$; $r=.76$		

このデータに対して対応のある *t* 検定を行うと $t(299)=-2.04$, $p=.042$ で有意差が見られるが、効果量は $d=0.12$ と Cohen（1992）の基準からすると極めて小さい（表 1）。

本題に入る前に、今回使用するデータに平均への回帰が見られることを確認しておく。まず、2 回目のスコアから 1 回目のスコアを引いた変化得点（ないしは差得点）を算出した。図 3 は 1 回目のスコアと変化得点の散布図であるが、二つのスコアの間には負の相関が見られ、1 回目のスコアが高めの人は変化得点がマイナス（つまり 2 回目のスコアの方が低い）になる傾向

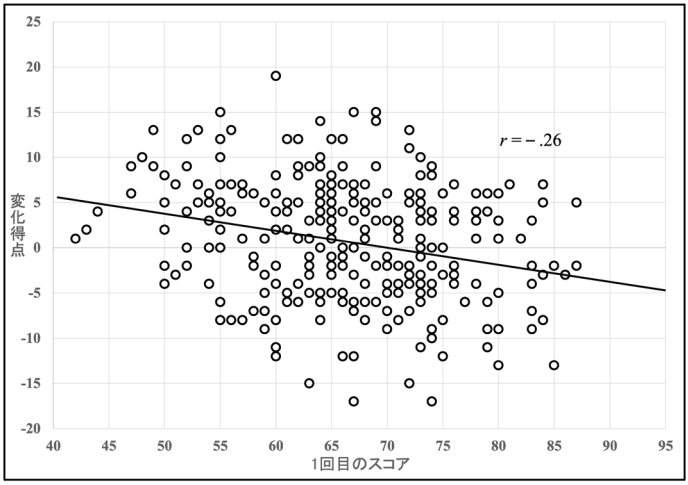

図 3　1 回目のスコアとスコアの変化量の関係

があるが、1 回目のスコアが低めの人は，変化得点がプラス（つまり 2 回目のスコアの方が高い）になる傾向にあり、平均への回帰が生じていることがみてとれる。

4.　分析結果

まず、先に述べた Davis（1976）や Gardner & Heady（1974）による平均への回帰の予測式を用いて検討してみる。今回は、カットオフ値 k_1 を 60 点とし初回受験時のスコアが下位約 20％の学生を選んだ。60 点を標準化すると z=0.710 であるから、エクセルの NORM.S.DIST 機能を使ってφ(z) およびΦ(z) を計算すると、φ(z)=0.31 およびΦ(z)=0.76 となる（図 4 も参照のこと）。従って、$C = \varphi(z)/[\,1 - \Phi(z)\,] = 1.29$ で、成績下位群の初回受験時のスコアの平均値の予測値は

$$E(y_1 \mid y_1 < k_1) = \hat{\mu} - C\hat{\sigma} = 66.38 - 1.29 \times 8.98 = 54.80$$

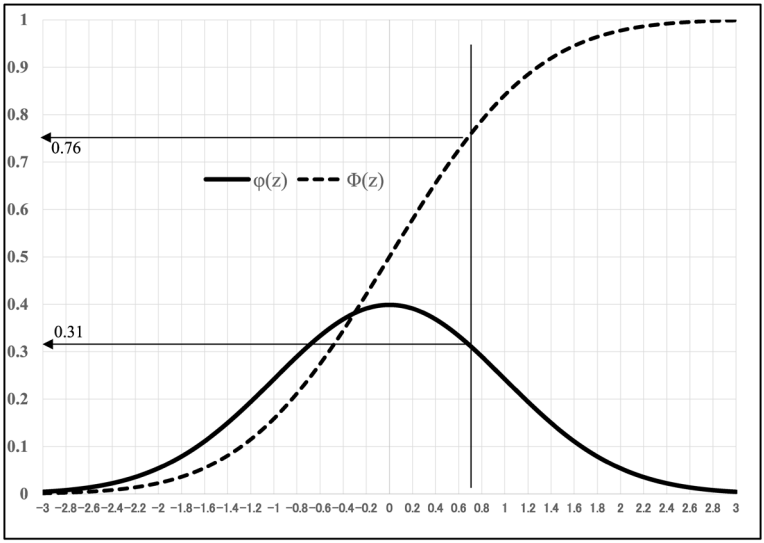

図 4　標準正規分布の確率密度関数および累積分布関数

表 2　成績下位群における 1 回目スコアと 2 回目スコアの観測値および平均への回帰の予測値の比較

	観測値の平均（1）	予測値の平均（2）	（1）と（2）の差
	下位約 20％の受験者（1 回目スコア＜60）		
1 回目のスコア	53.39	54.80	−1.41
2 回目のスコア	56.53	57.58	−1.05
回帰効果	3.14	2.78	0.36

同様に、成績下位群の 2 回目の受験時スコアの平均値の予測値は

$$E(y_2 \mid y_1 < k_1) = \hat{\mu} - \rho C\hat{\sigma} = 66.38 - 0.76 \times 1.29 \times 8.98 = 57.58$$

となり、$E(y_2 \mid y_1 < k_1) - E(y_1 \mid y_1 < k_1) = 57.58 - 54.80 = 2.78$ となる。同様に、（3）式を用いて計算しても $C\hat{\sigma}(1-\rho) = 1.29 \times 8.98 \times (1 - 0.76) = 2.78$ である。要するに、教育効果が全くない場合でも、平均への回帰によって、下位約

20％の受験生のスコアは 2.78 ほど伸びることが予想される。表 2 に示すとおり、実際の観測値では、成績下位群の初回受験時のスコアの平均が 53.39 で、再受験時の平均スコアは 56.53 なので、これらの受験者は平均 3.14 スコアを伸ばしたことになるが、上記の予測式を元に考えると、そのうちの約 89％（2.78÷3.14）の部分は平均への回帰によるものと考えることができる。

次に、同じデータに対して、成績別 2 群（対応なし要因）×2 回のテストスコア（対応あり要因）の 2 要因分散分析を行ったところ、交互作用が有意であったため（$F[1, 298]=11.31$, $p=.001$, $\eta_p^2= .037$）、群別に単純主効果の検定を行った。図 5 に示したように成績下位群ではスコアの変化量に有意差が見られ、効果量も中程度の大きさであるが（$p<.001$, $d=.52$）、一般群ではスコアの変化に有意差は見られない。従って、この結果だけを見ると、成績下位約 20% の学生に対する補習が効果を上げているように見える。しかし、この分析は平均への回帰を考慮にいれていない。

さらに、同じデータに対して、変化得点を従属変数とした 1 要因分散分析、1 回目のスコアを共変量に組み込んだ共分散分析、および残差得点を従

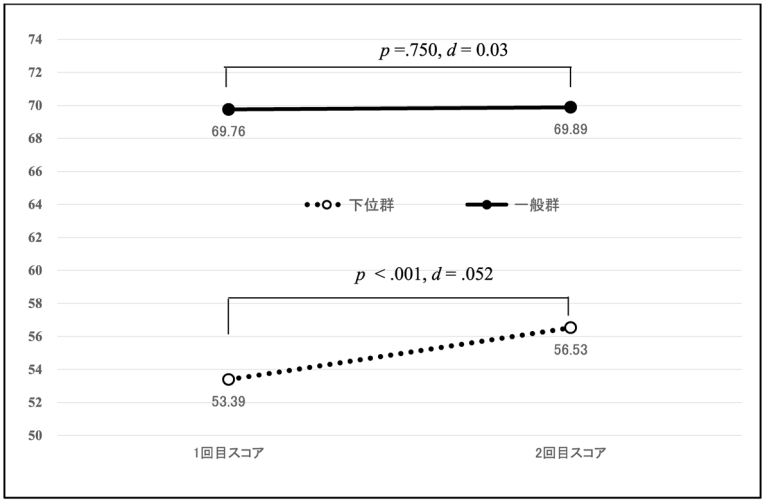

図 5　2 要因分散分析の結果

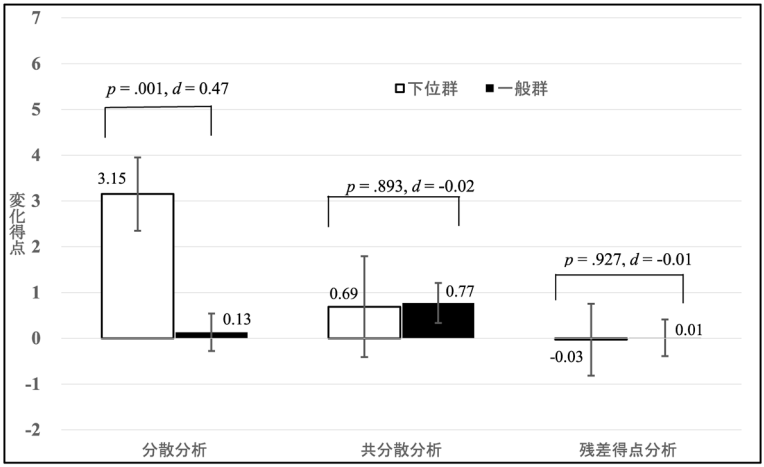

図 6　3 つの分析方法の結果比較

属変数とした分散分析の 3 つの分析方法を用いて分析を行い、結果の比較を行ってみる。

まず、共分散分析を行うための前提条件として、「回帰係数の等質性の仮定」を満たしているかどうか（つまり群ごとの回帰直線の傾きが等しい）を確認しておく。変化得点を従属変数とし、成績別 2 群を独立変数、1 回目のスコアを共変量、1 回目のスコアと成績別 2 群の交互作用をモデルに組み込んだ分析を行った結果、交互作用は有意でなかった（F[1, 296]=2.003, p=.158, η_p^2=.007）。従って、回帰係数の等質性の仮定は満たされていると判断できる。

図 6 に 3 つの方法で分析した結果の要点をまとめておく。まず変化得点を従属変数、成績別 2 群を独立変数にした 1 要因分散分析を行った結果, F(1, 298)=11.31, p=.001, η^2=.037 と群の主効果が有意であり、一般群（M = 0.13, 95%CI[−0.62〜0.93], SE=0.41）にはスコアの変化は見られないが、成績下位群（M=3.15, 95%CI[1.57〜4.72], SE=0.80）では 3.15 スコアを伸ばしていることになる(5)。

一方、1回目のスコアを共変量に組み込んだ共分散分析の結果を見ると、1回目のスコアの主効果は有意であるが（F[1, 297]=10.19, p=.002, η^2=.033）、成績別2群の主効果は有意ではなかった（F［1, 297］=0.04, p=.95, η^2=.000）。変化得点の推定値を見ると、一般群（M=0.77, 95%CI［−0.13〜1.65］, SE=0.45）、成績下位群（M=0.69, 95%CI［−1.48〜2.85］, SE=1.10）ともに95%CIに0が含まれており、スコアの変化は見られない。

同様に、残差得点を従属変数とした1要因分散分析（要するに残差得点分析）でも、成績別2群の主効果は有意ではなかった（F[1, 298]=0.02, p=.966, η^2=.000）。変化得点の推定値を見ると、一般群（M=.01, 95%CI［−0.78〜0.80］, SE=0.40）、成績下位群（M=−.03, 95%CI［−1.58〜1.55］, SE=0.79）ともにスコアの変化は見られない。

従って、平均への回帰を調整している共分散分析および残差得点分析の結果、一般群および成績下位群のいずれの群にもスコアの伸びは認められないことになる。この結果は、先に述べた平均への回帰の予測式で得られた結果とも整合性がある。従って、平均への回帰を十分考慮に入れないで分析を行うと、誤った結論を導く恐れがある。

5. 考察

本稿では、大学生が標準化テストを2回受験した場合のスコアの変化について、平均への回帰に焦点をあてて論じてきた。その際、1回目のスコアが下位約20％の学生には補習を行った場合を仮定した。ここまで見てきたとおり、平均への回帰の影響を調整しない2要因分散分析や変化得点を従属変数とした1要因分散分析の結果と、平均への回帰の影響を調整した共分散分析および残差得点分析の結果には明確な違いがあることが明らかになった。また、Davis（1976）やGardner & Heady（1974）によって提唱された平均への回帰の大きさを推定する方法についても言及し、この予測式が（たとえ近似的とはいえ）平均への回帰の大きさをある程度予測できることも明らかにした(6)。

今回分析したデータは、あくまで架空の例なので、現実に必ずこのような結果になるとは限らないが、平均への回帰を考慮に入れないで分析を行うと、補習授業の結果、成績下位群の 2 回目スコアがある程度伸びたと誤った結論を導く恐れがある。実際、そのような誤った分析結果が報告されているケースもあると言う（岩崎・河田，2007）。事前—事後研究のデータ分析には十分な注意が必要である。

なお、テストスコアの変化を分析する場合、集団間でスコアの伸びの違いを分析する場合と個人レベルの変化を見る場合があるが、本稿では前者の問題のみ論じてきた。しかし、テストの受験者個々人にとっては、自分のスコアが伸びたのかどうかの方が問題となる。この個人レベルの変化については、「差の標準誤差」（Standard Error of Difference = SED or SE*diff*）について考える必要がある(7)。

最後に、本稿では共分散分析と残差得点分析の関係については十分に議論出来なかった。先述のとおり、残差得点分析の使用には慎重な態度を示す研究者もいるため、この点についてはさらに研究が必要である。

注

（1） 1 回目の測定値と変化得点の間の相関係数は以下の式で表せる（Campbell & Kenny, 1999, p. 88）。

$$\frac{rS_y - S_x}{\sqrt{\left(S_x^2 + S_y^2 - 2rS_xS_y\right)}}$$

この式の分母はプラスなので、1 回目の測定値と変化得点の間の相関係数がプラスになるには、分子もプラスの値でなければならない。$rSy-Sx>0$, すなわち $r>S_x/S_y$ の場合のみ二つの相関係数がプラスになり得る。言うまでもなく r は 1 以下なので、これを満たすには、2 回目の測定値の標準偏差が 1 回目よりかなり大きくなる必要がある。

（2） ただし、平均への回帰はどのような場合にも調整しなければならないわけではない。例えば、男女間や人種間でスコアを比較するような場合は、本稿で論じる場合と状況が異なる。この点については、Saito（2019）などを参照いただきたい。

（3） 残差得点分析とは $Y_i - \hat{Y}(=\beta_0+\beta_1X_i)$ で算出される残差を従属変数とする分析である。

（4） 本稿で使用するデータは、実際のテストデータを元に筆者が作成した架空のものである。スコアは素点ではなく、項目反応理論（Item response theory）に基

づいて等化（equating）を行ったものとする。等化することで、異なるテストフォームでも同じ尺度上で解釈することが可能となる。

（5） 2 要因分散分析の交互作用の結果は、変化得点を従属変数とした 1 要因分散分析の主効果の結果と一致する。実際、本文中で言及しているとおり、いずれも（$F[1, 298]=11.31$, $p=.001$, $\eta_p^2=.037$）となる。

（6） ただし、この予測式が使えるのはデータが正規分布している場合である。Linden（2013）の分析で、データが正規分布していない場合は、この予測式から得られる結果と実際の回帰効果との間にはかなりのズレが生じることがわかっている。

（7） 通常どのようなテストにも「測定の標準誤差」（Standard Error of Measurement = SEM）と呼ばれる誤差が存在する（Lord & Novick, 1968; Dudek, 1979）が、テストを 1 回のみ受験した場合のスコアの解釈とは多少異なり、同一人物が同一のテストを異なる時期に 2 回受けた場合のテストスコアの差について検討するには「差の標準誤差」（Standard Error of Difference = SED or SE*diff*）と呼ばれる誤差を知る必要がある。

引用文献

Allison, P. D.（1990）. Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J.（2005）. Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*, 215–220.

Bonate, P. L.（2000）. *Analysis of pretest–posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.

Campbell, D. T., & Kenny, D. A.（1999）. *A primer on regression artifacts*. New York, NY: Guilford.

Cohen, J.（1992）. A power primer. *Psychological Bulletin*, *112*（1）, 155–159.

Gmel, G., Wicki, M., Rehm, J., & Heeb, J-L.（2007）. Estimating regression to the mean and true effects of an intervention in a four-wave panel study. *Addiction*, *103*, 32–41.

Davis, C. E.（1976）. The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology*, *104*（5）, 493–498.

Dimitrov. D. M., & Rumrill, P. D. Jr.（2003）. Pretest–posttest designs and measurement of change. *Works* 20（2）, 159–165.

Dudek, Frank J.（1979）The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, 86（2）, 335–337.

Forbes, A. B., & Carlin, J. B.（2005）. “Residualized change” analysis is not equivalent to analysis of covariance. *Journal of Clinical Epidemiology*, *58*, 540–541.

Galton, F.（1886）Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15. 246–263.

Gardner, M. J., & Heady, J. A.（1974）. Some effects of within-person variability in epidemiological studies. *Journal of Chronic Diseases*, *26*, 781–795.

Hansen, K. M. & Pedersen, R. T.（2014）. Campaigns matter: How voters become knowledgeable and efficacious during election campaigns. *Political Communica-*

tion, *31*, 303–324.
岩崎学・河田祐一（2007）．処置前後研究における平均への回帰とその周辺．日本統計学会誌 *36*(2), 131–145.
Lord, F. M., and Novick, M. R.（1968）*Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.
Linden, A.（2013）. Assessing regression to the mean effects in health care initiatives. *Medical Research Methodology*, *13*, 119.
MacKinnon, D. P.（2008）. *Introduction to Statistical Mediation Analysis*. Routledge: New York.
Maxwell, S. E., Delaney, H. D., & Manheimer, J. M.（1985）. ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics*, *10*, 197–209.
Nielsen, T., Karpatschof, B., & Kreiner, S.（2007）. Regression to the mean effect: When to be concerned and how to correct for it. *Nordic Psychology*, *59*(3), 231–250.
Rogosa, D., Brandt, D., & Zimowski, M.（1982）. A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748.
Rogosa, D.（1995）. Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman（Ed.）, *The analysis of change*（p.3–66）. Mahwah, New Jersey: Lawrence Erlbaum.
Saito, S.（2019）. "*Do Whites' scores change more than Blacks' in the vocabulary test?: Inappropriate Use of Control for Initial Score*". Poster presented at the 31st Association for Psychological Science Annual Convention, May 23–26, Washington, D.C.
徳永章二（2001）．統計解析，特に回帰解析をめぐる 3 つの話題―平均への回帰，Multilevel models, ポアソン回帰―，日本救急医学会雑誌，*12*(7), 333–342.
Yu, R. & Chen, L.（2015）. The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, *5*, Article 1574.

心理コミュニケーション学科コミュニケーション専攻

Key words

Regression toward the mean, ANOVA, ANCOVA, residual change score analysis

キーワード

平均への回帰、ANOVA、ANCOVA、残差得点分析