線形回帰分析における誤差項の非正規性および不均一性が分析結果に及ぼす影響
――仮想データによるシミュレーション――[1)]

Influence of non-normality and heterogeneity in the error term on the results of linear regression analysis: A computer simulation.

池田　龍也

Tatsuya IKEDA

要　旨

心理学研究において，しばしば最小二乗法による線形回帰分析が用いられる。この分析を実施するためには，4 つの仮定を満たす必要がある。本稿では，正規性の仮定または均一性の仮定を逸脱したとき，分析結果が受ける影響を調べた。具体的には，仮想のデータセットを用い，パラメータやサンプルサイズの変化が分析結果に及ぼす影響をシミュレーションした。その結果，次の 3 点が示された。(a) 回帰係数と決定係数の点推定値は，誤差分布の形状に影響を受ける，(b) 誤差分布が正規分布から逸脱していたり，極端な形状であったりした場合は回帰係数および決定係数を正しく推定できない可能性がある，(c) 均一性の仮定に違背した場合，サンプルサイズを増やしても正しい回帰係数が推定できない可能性がある，であった。したがって誤差分布の形状は回帰係数と決定係数の双方に影響するため，分析の仮定を充足する必要性が改めて確認された。

キーワード：最小二乗法による線形回帰，正規性の仮定，均一性の仮定，心理統計法，シミュレーション

1.　はじめに

心理学の研究において，しばしば重回帰分析やロジスティック回帰分析が用いられる。しかし吉田・村井 (2021) が指摘している通り，必ずしも十分な理解に基づいて使用されているわけではない。吉田・村井 (2021) の指摘によれば，多くの研究において研究目的・研究デザイン・結果の解釈に齟齬がある。吉田・村井 (2021) の指摘によって，回帰分析の実施に関する議論が活性化し，2021 年 7 月 2 日に日本心理学会が意見交換の場を開くと，多くの研究者が参加した。このことは心理統計法に対する関心の高さを反映していると考えられ，より適切かつ妥当な研究成果を生み出そうとする研究者の多さが窺われる。

取得したデータを適切に処理することは，妥当な研究成果を生み出すことに繋がる。このため研究者は自身の用いる分析手法について，理解を深めておくことが求められる。さらに，心理学がヒトを含めた生物を対象とする以上，取得したデータを適切に処理することは研究参加者に対する研究倫理としても要請される事柄であろう。心理統計において用いられる様々な分析には，それぞれに仮定がある。当然ながら，仮定を満たさない分析は，そうでない場合よりも妥当な結果を得られる見込みが乏しい。よって研究デザインや結果の解釈以前に，使用する分析手法の仮定を理解しておく必要がある。もっとも，このことは学部教育の段階で教育されるであろう。

それでは仮定を逸脱した場合に，どのように結果が歪曲されるのであろうか。統計解析ソフトウェアの普及によって，コンピュータさえあれば誰でも容易に分析できる状況になりつつある。コンピュータの画面上を数回クリックするだけで，複雑な分析であっても比較的容易に実施できる。その反面，本来は適用不可能な分析や無意味な検定を施してしまうこともある。本稿では最小二乗法による線形回帰における正規性の仮定および均一性の仮定に着目し，この仮定を満たさないデータへ線形回帰を当てはめたとき分析結果にどのような影響があるか，シミュレーションを用いて検討する。

2.　線形回帰分析の仮定

最小二乗法における線形回帰分析の仮定は 4 つある。それは (a) 独立性，(b) 正規性，(c) 均一性，そして (d) 線形性であり，このうち (a) 〜 (c) は誤差項に関する仮定である (佐和, 1979)。以下，本稿で着目する (b) 正規性および (c) 均一性について述べる。

正規性の仮定とは，回帰分析における誤差項が正規分布に従うことを意味する。ただし

誤差は直接観測できず，真値との和として観測値が得られる。そのため実際に誤差項が正規分布に従っているか否か，直接調べることはできない。誤差のこうした性質から，観測値から推測される誤差を残差と呼ぶ。本稿では後述の通り誤差が既知であるため，引き続き誤差と呼ぶこととする。

均一性の仮定とは，誤差項の分散が説明変数の値に左右されず均一であることを意味する。例えば説明変数の値が大きいときは誤差項の分散が小さいものの，説明変数の値が小さいと誤差項の分散が大きい場合，均一性の仮定を満たしていないことになる。こうした現象は，異なる母集団からデータを収集した際にもみられる。

4 つの仮定のうち，臨床心理学や異常心理学において懸念されるのが正規性の仮定であろう。たとえば一般群と呼ばれる比較的健康度の良好な集団を調査対象とした場合，抑うつ状態や解離傾性などの精神病理指標の観測値は正規分布には見えない形状となる。そのためこれらを目的変数とする回帰分析を実施したとき，誤差項が正規性の仮定を満たすか否か，判断が困難なことがある。

正規性の仮定を満たすか否かを判断する材料となるのが，正規 QQ プロットである。正規 QQ プロットは横軸に正規分布に従う場合の期待値，縦軸に観測値をとったグラフである。よって検討対象のデータが正規分布に従っていれば，正の相関がみられる散布図のごとく対角線上にプロットが整列する。一方で正規分布に従わない場合，対角線上からプロットが逸脱する。もちろん竹内 (1978) が指摘している通り，誤差項の非正規性はサンプルサイズが十分であれば大きな問題にはなりづらい。しかし竹内 (1978) は主に回帰係数の *t* 検定，すなわち回帰係数の有意性に関する指摘であり，回帰係数や決定係数への影響を確認する必要がある。

3.　使用機材

本稿の分析には，統計解析ソフト R for mac (ver. 3.6.2) (R Core Team, 2019) および統合開発環境 RStudio (ver. 1.1.463) を用いた。さらに“psych” (ver. 1.9.12) (Revelle, 2019) を用いて記述統計量を計算した。ヒストグラムや散布図などの図の作成に“ggplot2” (ver. 3.3.5) (Wickham, 2016)，“jtools” (ver. 2.1.0) (Long, 2020)，“patchwork” (ver. 1.1.1) (Pedersen, 2020) を用い，“devEMF” (Johnson, 2020) (ver. 3.8) によりグラフの拡張メタファイルを出力した。R スクリプトコードは，筆者の ResearchGate にて公開している [2)]。

4.　仮想データの生成と分析結果

仮想データ生成の手続き　まず (1) 式に示す単回帰式を想定した。B0 は切片であり，0 と指定した。次に B1 は回帰係数を表し，B1 = 1.00 と指定した。X は説明変数であり，一様分布から生成した。このときのシード値は 123 であり，範囲は 1.00 から 5.00，サンプルサイズは 50 であった。目的変数の真値 Y_{true} は，B1 と X の積に B0 を足した値であった。

$$Y_{true} = B1X + B0 \qquad (1)式$$

続いて，(2) 式に従って目的変数の観測値 Y_{obs} を計算した。E は誤差項を意味し，E の従う確率分布によって 3 種類作成された。それぞれ正規分布に従う E_normal (M = 0, SD = 1)，ガンマ分布に従う E_gamma (shape = 3, scale = 1)，そして指数分布にしたがう E_exp (λ = 1) であった。最後に目的変数の真値 Y と誤差項の値の和から，目的変数の観測値を作成した。誤差項の従う分布によって，Y_normal，Y_gamma，および Y_poison の 3 つが算出された。以上の手続きによって得られたデータの概要を Table 1 に示す。

$$Y_{obs} = Y_{true} + E \qquad (2)式$$

分析結果　3 種類の目的変数について，それぞれ単回帰分析を実施した。結果を Fig. 1 および Table 2 に示す。まず誤差項が正規分布に従う場合，B0 = 0.12, B1 = 0.97, R^2 = 0.60 であった。次にガンマ分布に従う誤差項の場合，各係数は B0 = 2.80, B1 = 1.00, R^2 = 0.38 であった。最後に誤差項が指数分布に従った場合，B0 = 0.68, B1 = 1.14, R^2 = 0.63 と推定された。

Table 1. 各データの概要 (N = 50)

	M	*SD*	*Me*	*Min*	*Max*	range	skew	kurtosis
Y	3.08	1.18	3.01	1.10	4.98	3.88	0.02	-1.31
Y_normal	3.11	1.47	3.09	-0.80	6.39	7.18	-0.17	-0.18
Y_gamma	5.89	1.90	6.11	1.83	11.88	10.05	0.49	0.44
Y_exp	4.18	1.67	3.97	1.27	8.80	7.53	0.75	0.34
X	3.08	1.18	3.01	1.10	4.98	3.88	0.02	-1.31
E_normal	0.03	0.93	-0.07	-1.97	2.17	4.14	0.16	-0.52
E_gamma	2.81	1.48	2.57	0.65	7.05	6.40	1.05	0.73
E_exp	1.10	1.02	0.91	0.00	4.37	4.36	1.26	1.27

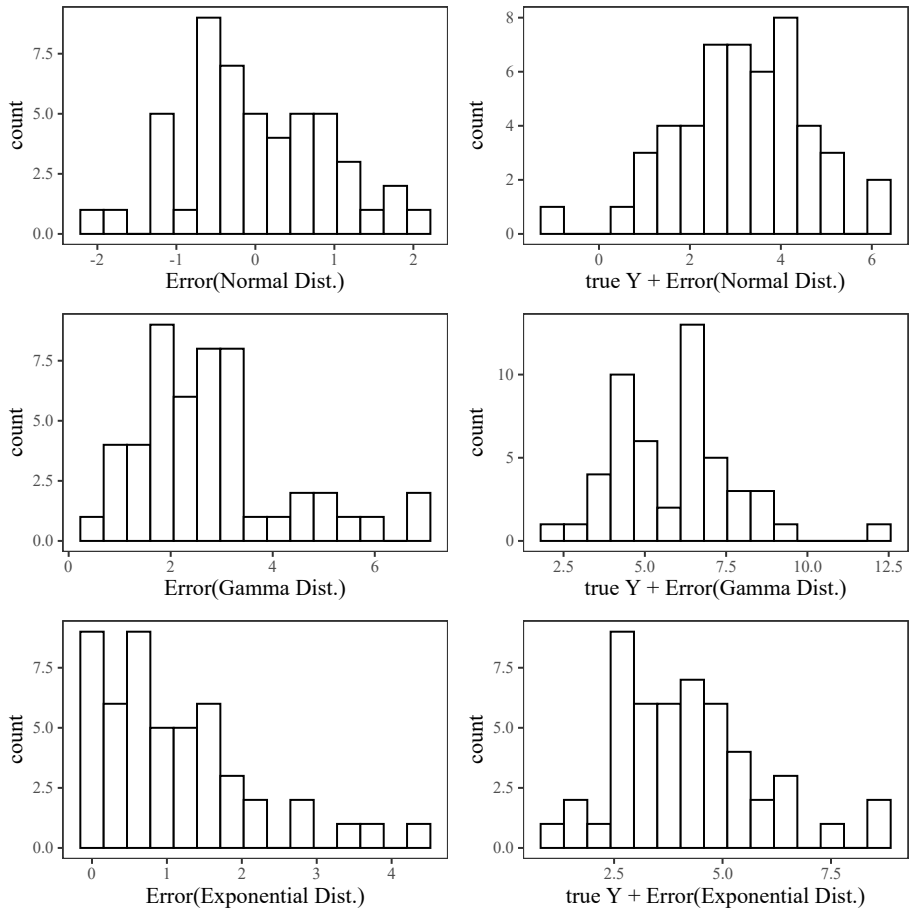

Fig. 1. 各誤差項と目的変数の観測値のヒストグラム。左側が誤差項，右側が目的変数の観測値である。“Dist.” はDistribution であり，“Eerror”に続く括弧内は誤差項が従う確率分布を示す。

正規分布に従う誤差項の場合，B0 と B1 のいずれもが真値に近かった。一方で誤差項がガンマ分布に従うケースでは，B1 は真値と一致したもののB0 が真値から大きく逸脱していた。また指数分布に従う誤差項を含めた分析結果は，B0 が真値からやや逸脱することが示された。

5.　シミュレーションによる検討

誤差分布の形状の変化が係数の点推定に与える影響　先述の通り，指数分布に従う誤差

Table 2. 異なる誤差分布による
回帰分析結果の相違

目的変数の誤差項	$B0$	$B1$	$R^{2\ 1)}$
誤差なし	0.00	1.00	1.00
正規分布	0.12	0.97	0.60
ガンマ分布	2.80	1.00	0.38
指数分布	0.68	1.14	0.63

[1)] 決定係数は自由度調整済み

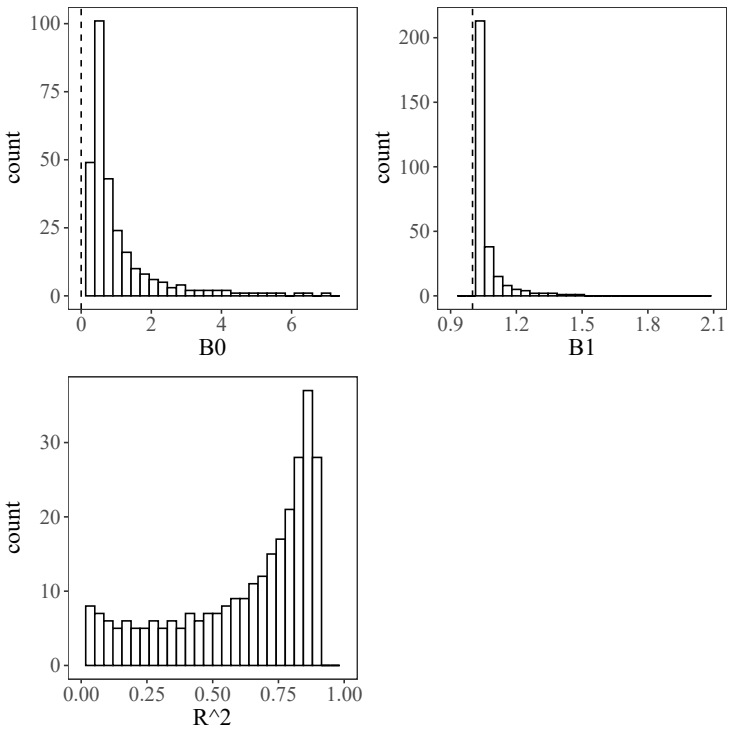

Fig. 2. 指数分布のパラメータλの変化による切片（左上），傾き（右上），決定係数（左下）。左上および右上の図中における破線は，B0 および B1 の真値を表す。

項はガンマ分布ほど真値から乖離していなかった。そこで指数分布のパラメータ λ を 0.1 から 3 まで 0.01 刻みで増加させ，292 の誤差項を得た。シード値は先述の通り 123 であった。これらの誤差項と Y の和を目的変数とし，単回帰分析を 292 回実施した。回帰係数と決定係数のヒストグラムを Fig. 2 に，λの変化にともなう回帰係数と決定係数の推移を Fig. 3 に示す。

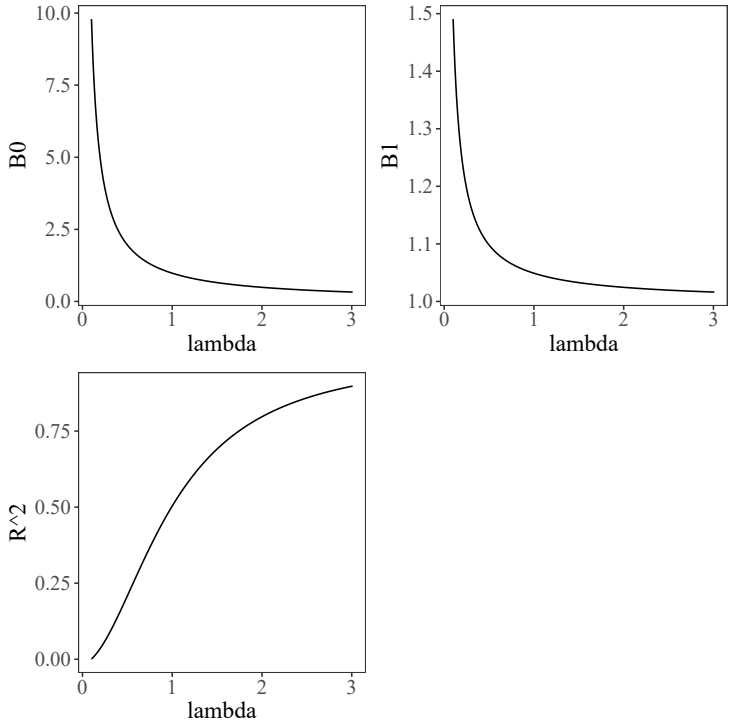

Fig. 3. 指数分布のパラメータ λ の変化が切片（左上），傾き（右上），決定係数（左下）に及ぼす影響のシミュレーション。

指数分布の形状に依存せずに回帰係数および決定係数を計算できるなら，Fig.3 では回帰係数の真値 B0 = 0.00, B1 = 1.00 付近を水平に推移するはずである。しかしシミュレーションの結果はバスタブ状に推移しており，λ の値が小さいと B0 と B1 が過大に推定された。ただし実際のバスタブの形状とは異なり，再び回帰係数の推定値が上昇することはなかった。逆に，決定係数の推移は右肩上がりの形状であり，λ の値が小さいと R^2 が極めて小さく推定されていた。したがって，回帰係数も決定係数も指数分布の形状に依存したと解釈できる。

サンプルサイズの変化が正規性の仮定を逸脱する回帰分析に及ぼす影響　サンプルサイズの影響を調べるために，サンプルサイズを 50 から 999 まで 1 ずつ増加させ，目的変数と説明変数を生成した。誤差項が従う分布は指数分布とした。パラメータは回帰係数の真値からの逸脱が大きかった $\lambda = 0.1$ とし，$N = 50$ から $N = 999$ まで 950 回の回帰分析を実施した。各分析で得られた回帰係数および決定係数のヒストグラムを Fig. 4 に示す。さらに

サンプルサイズの変化がB0，B1，R^2に及ぼす影響をFig. 5に示した。

サプルサイズの増加によって回帰係数および決定係数の推定精度が向上するのであれば，Fig.5ではサンプルサイズの増加にともなって回帰係数の真値B0＝0.00, B1＝1.00付近に収束するはずである。たしかにシミュレーションの結果は振幅が小さくなっていた。しかしB0は過大に推定され，N=999時点でB0＝10.74であった。B1については真値付近を推移する様子が認められたものの振幅が大きく，N＝999のときB1＝0.87であった。決定係数については負の値が散見され，N=999であってもR^2=0.009と極めて低い値であった。以上のように，サンプルサイズが小さいよりも大きいほうが，B1の推定精度が高まる傾向にあったものの，B0やR^2は過大または過小に推定された。したがって誤差分布の形状によっては，サンプルサイズを増やしたとしても，回帰係数および決定係数の推定精度は一概に向上しないと考えられる。つまりサンプルサイズを単純に増やしただけでは，誤差項の

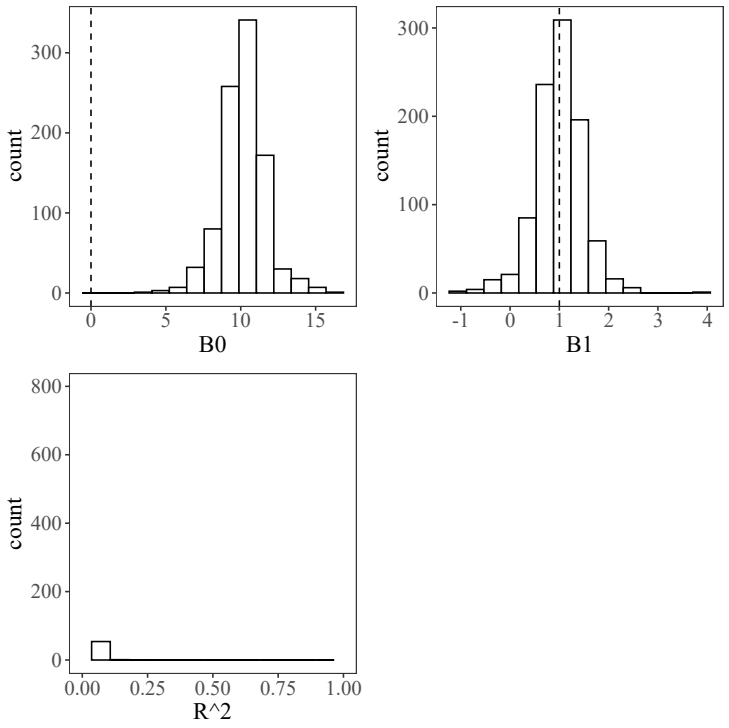

Fig. 4. λ＝0.1をパラメータとする指数分布に従う誤差項と切片（左上），傾き（右上），決定係数（左下）。左上および右上の図中における破線は，B0およびB1の真値を表す。

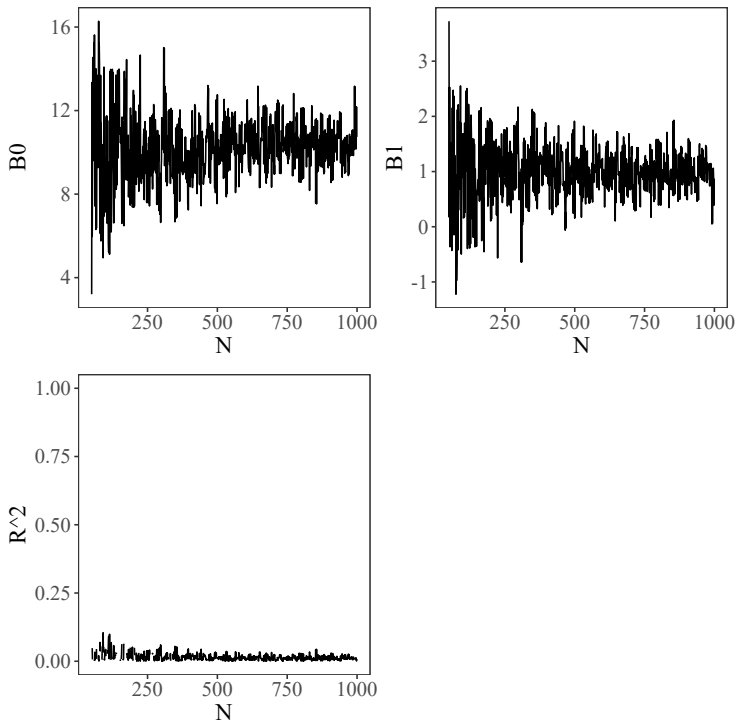

Fig. 5. 誤差項が正規性の仮定から逸脱したとき，サンプルサイズの変化が切片（左上），傾き（右上），決定係数（左下）の推定に及ぼす影響。

非正規性への対処として十分でないことが示唆された。

6.　不均一分散の影響

仮想データによる検討　ここまでの検討は誤差項が正規分布に従っていないものの，同一の確率分布から生成されていた。つまり正規性の仮定は違背していたものの，均一性の仮定は満たす分析であった。そこで異なる母集団からのサンプリングを想定し，誤差項の不均一性の影響を調べた。新たに作成した誤差項はすべて正規分布から生成されたものの，Table 3 に従って X の値によって *M* および *SD* を変化させた。サンプルサイズは $N = 50$ とし，乱数生成のシード値は，引き続き 123 であった。この誤差項と目的変数の真値 Y を足し合わせ，目的変数の観測値 Y_unbalance を作成した (Fig. 6)。Shapiro-Wilk の正規性検定

Table 3. 不均一なパラメータ

	正規分布のパラメータ	
Xの値	*M*	*SD*
X ≦ 2	0.00	2.00
2 < X ≦ 3	1.00	3.00
3 < X ≦ 4	2.00	4.00
4 < X	3.00	1.00

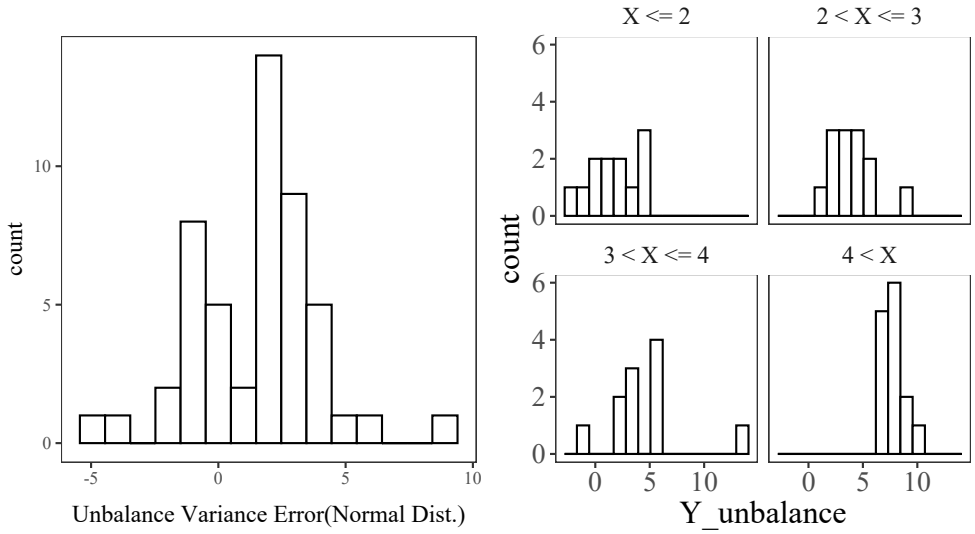

Fig. 6. 不均一分散に基づく誤差項をおよびXの値で層別化したYのヒストグラム。

の結果，Y_unbalance の正規性は否定されなかった (W=0.99, p=0.99)。Y_unbalance を目的変数，X を説明変数とする単回帰分析を実施したところ，B0 = -1.13，B1 = 1.86，R^2 = 0.48 であった。B0 は真値の 0 から負の方向へ逸脱し，B1 は正の方向へ逸脱した。

シミュレーションによる検討　サンプルサイズの影響を調べるために，サンプルサイズを 50 から 999 まで 1 ずつ増加して目的変数と説明変数を生成した。誤差項の従う正規分布のパラメータは，先述の通りであった。そして N = 50 から N = 999 まで 950 通りの回帰分析を実施した。各分析で得られた回帰係数と決定係数をヒストグラムに示す (Fig. 7)。さらにサンプルサイズの増加が各係数の点推定値に及ぼす影響を Fig. 8 に示した。

B0 はサンプルサイズが増加するにしたがい，およそ-2.00〜-1.00 の間に収束する様子が観察された。ただし B0 の真値は 0.00 であることから，負の方向に逸脱していたと評価できる。B1 はサンプルサイズの増加に伴い，おおむね 1.80〜2.00 の間に収束した。しかし B1

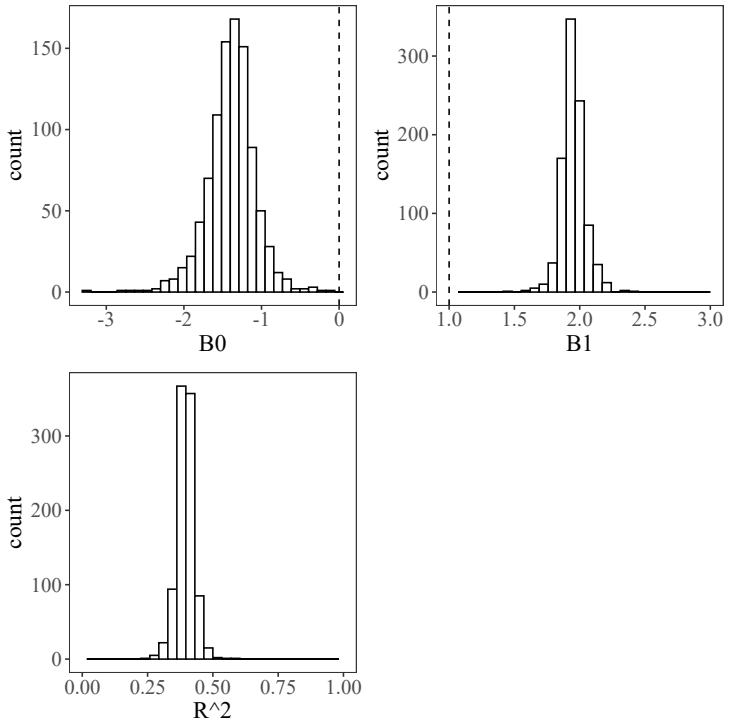

Fig. 7. 回帰分析の誤差項が不均一分散であったときの切片（左上），傾き（右上），決定係数（左下）。左上および右上の図中における破線は，B0 および B1 の真値を表す。

の真値は 1.00 であり，過大に推定されていた。今回のシミュレーションでは真値と類似した値に収束することはなく，サンプルサイズが増加しても B1 を正しく推定できないことが窺われた。

以上を踏まえると，均一性の仮定を違背したデータを分析するとき，サンプルサイズを増加させても正しい回帰係数を得られる見込みが乏しいと考えられる。本稿では正規分布のパラメータ M が異なっていたものの，正規性の検定では Y_unbalance の正規性が否定されていない。よって正規性の仮定は，少なくとも正規性検定のうえでは満たされていたと考えられる。しかしながら回帰係数の推定結果は真値と乖離したため，実際の分析では分散不均一性検定や誤差項の分散を可視化することで仮定を満たすか確認することは特に重要であろう。

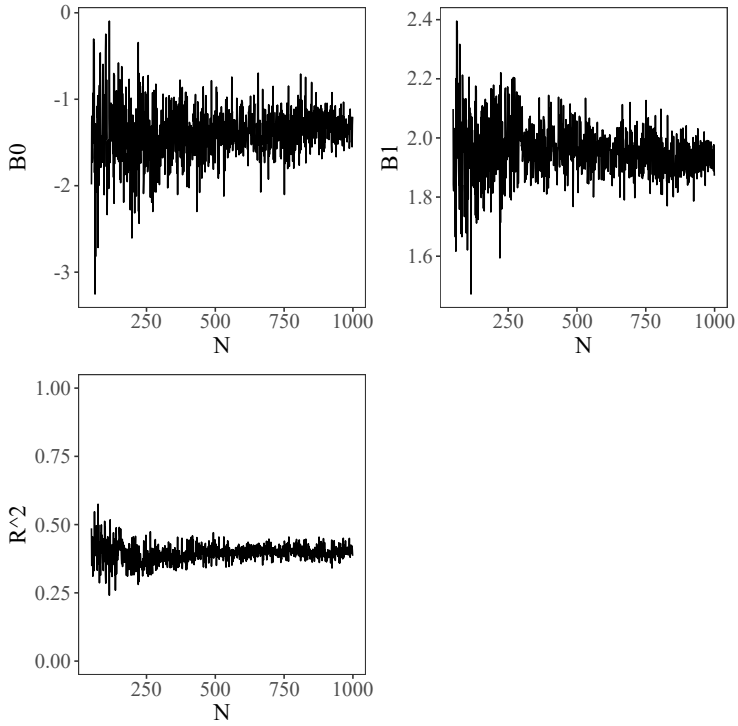

Fig. 8. 単回帰分析の誤差項が不均一分散であったとき，サンプルサイズの変化が切片（左上），傾き（右上），決定係数（左下）に及ぼす影響。

7.　おわりに

最小二乗法による線形回帰は，心理学の研究においてしばしば用いられる分析手法である。この分析には 4 つの仮定が存在し，実施にあたっては仮定を満たさなければならない。本稿では 4 つの仮定のうち正規性と均一性に着目し，それぞれを満たさないデータは分析結果にどのような影響をもたらすのかを，仮想データを用いたシミュレーションに基づいて調べた。その結果，次の 3 点が示された。(a) 回帰係数と決定係数の点推定値は，誤差分布の形状から影響を受ける，(b) 誤差分布が正規分布から逸脱していたり，極端な形状であったりした場合は回帰係数および決定係数を正しく推定できない可能性がある，(c) 均一性の仮定に違背した場合，サンプルサイズを増やしても正しい回帰係数が推定できない可能性がある，であった。

最小二乗法による線形回帰の分析結果は，誤差分布の形状に影響された。たとえ同じ種

類の確率分布であっても，そのパラメータによって点推定値が異なった。さらに誤差分布の形状によっては，たとえサンプルサイズを増やしても回帰係数を正しく推定できない可能性が示された。くわえて一見すると正規分布のようであり，正規性検定において正規性が否定されなかったとしても，均一性の仮定を違背すると回帰係数が過大あるいは過小に推定された。この結果はサンプルサイズを増やしても同様であった。

したがって分析の前に，必要な仮定を満たしているか確認するべきである。さらに分析後も回帰係数のみに注目するのではなく，決定係数自体が小さすぎないか，残差の正規性が満たされていそうか，誤差項の分散が不均一でないか，などを可視化・検定すべきであろう。誤差項や残差は研究者の関心の対象外であることが多い。そのため注意を払われなかったり，確認されなかったりすることもある。しかし本稿で示された知見によれば，誤差項は推定に大きく影響する。したがって研究成果の正確性を保持するために，これらの確認は必要性が大きいと考えられる。

本稿のシミュレーションは，全ての確率分布を網羅したものではない。一部の確率分布の限られたパラメータに基づいたシミュレーションに過ぎない。しかしながら本稿で示された通り，限定的な条件下であっても，誤差分布の形状が分析結果に及ぼす影響は無視できない。そのためデータを取得したら回帰分析のまえに各変数を可視化したり，分析後には残差の正規性や均一性を確認したりする必要があると言えよう。

引用文献

Johnson, P. (2020). devEMF: EMF Graphics Output Device. R package version 3.8. <URL: https://CRAN.R-project.org/package=devEMF>

Long, J.A. (2020). jtools: Analysis and Presentation of Social ScientificData_. R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.

Pedersen, T. L. (2020). patchwork: The Composer of Plots. R package version 1.1.1. <URL: https://CRAN.R-project.org/package=patchwork>.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <URL: https://www.R-project.org/>.

Revelle, W. (2019) psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois, USA. <URL: https://CRAN.R-project.org/package=psych Version =

1.9.12.>.

佐和隆光 (1979). 回帰分析　朝倉書店.

竹内　啓 (1978). 誤差分布の非正規性の処理　オペレーションズ・リサーチ, **23** (5), 305-310.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

吉田寿夫・村井潤一郎 (2021). 心理学的研究における重回帰分析の適用に関わる諸問題　心理学研究, **92** (3), 178-187.

註

[1] 本研究の一部は，日本学術振興会科学研究補助金 (19K19628 / 研究代表者:池田龍也) の助成を受けた。また，本研究は一般社団法人日本心理学会の研究倫理を遵守して遂行された。

[2] https://www.researchgate.net/publication/359641888_R_Script_Code_Used_in_Seisen_Univ_Bulletin_2022