タイトル	日本語教員養成課程における教案自動採点システムの評価モデルの再構築と比較検証
著者	歌代，崇史；UTASHIRO，Takafumi
引用	北海学園大学学園論集(197)：1-17
発行日	2025-07-25

日本語教員養成課程における
教案自動採点システムの評価モデルの
再構築と比較検証

歌　代　崇　史

1．はじめに

大学の日本語教員養成課程及び民間の日本語教師養成講座では，実践的な教授能力養成のため，実習もしくは模擬授業を実施しているが，教授経験の乏しい学生（以下，実習生）にとっては，想定される日本語学習者のレベル及び教科書の進度に合わせた授業を構成することが難しく，その準備に膨大な時間と労力を要する（平田 2008）。特に教案の作成と修正には多くの時間を要し，事前指導があっても独力で書くことは実習生にとって負荷の高い学習活動である。教案作成とその改善過程は実習生にとって多様な視点に気付く重要な機会であるものの，多くの養成課程では履修者が多く，担当教員１人が実習生の教案全てを添削し，個別に指導を行うのは教師としても非常に負荷が高く，時間的にも困難であることが多い。そのため，教案作りに不可欠な検討と改善の学習機会が実習生に不足する場合もある。

教員の負担を軽減し，実習生に対する学習機会の拡充を目的として，歌代（2021）ではティーチャートークを含めた教案全体の自動評価のための計算式を開発している。実習生が作成した教案を想定発話以外の要素も含めて分析し，教案としての適切さを自動推定するために有効な指標を探索し，その指標を基にした教案の適切さを推定するモデルを導出した。以下にその重回帰式を示す。

$$Y = 1.307 - 0.54X_1 + 0.024X_2 + 0.046X_3 + 1.658X_4$$

Y＝評価
X_1＝一文あたりの述語数
X_2＝括弧内文節数
X_3＝ターンの交代回数
X_4＝既習文型の割合
（R^2=.651，調整済み R^2=.630）

歌代（2024）では上記モデル（Initial Model）を使用した教案自動採点システム（Tetrater）を開

発し，グループ学習と組み合わせて実践環境に導入した。効果検証の結果，Tetrater が教案修正時に新たな視点を実習生に提供するだけでなく，グループ学習の支援にもつながる可能性が示唆された。一方で，Initial Model には変数として絶対数を用いたことに起因する問題があった。具体的には，「括弧内文節数」や「ターンの交代回数」は絶対数であり，これらの記述が増加すれば，それに伴って評価得点も高くなる傾向が見られた。これは，上記の重回帰式からも明らかである。つまり，記述量の多寡がそのまま評価に影響する構造となっており，教案の質的側面を適切に反映しているとは言いがたい。このようなモデルは，教案作成支援の目的には必ずしも合致しないと考えられるため，本稿では変数の見直しを通じてモデルの改良を試みる。

本研究の目的は，日本語教員養成課程における教案自動採点システムへの実装を想定し，絶対値を含まない変数によって構成される評価推定モデルを開発し，その予測性能を検証することである。

2．方　　法

ここでは，モデル構築に用いたデータ，変数およびデータの整理，モデル構築の手法，外れ値の検討について述べる。さらに，重回帰分析に加え，モデルの頑健性を高めるロバスト回帰，汎化性能検討のための交差検証についても説明する。

2.1. データの収集方法

本稿で利用したデータは歌代・須藤（2017）で収集したデータの一部である。歌代・須藤（2017）は教室内言語調整の能力を測定するため，ティーチャートーク・テストという記述式のテストを開発し，初級のはじめころの教室（『みんなの日本語』第 10 課終了）を想定した指導案を実習生に書かせ，その指導案を 3 名の日本語教師が 5 段階で採点した。3 名の評価の Cronbach's alpha は，$\alpha = .77$ で一定程度の一貫性が認められたため，3 名の平均値を評価とした。このようにして，学習者が記述した指導案とその評価のセットが 76 あった。データの概要を表 1 に示す。

2.2. 変数およびデータの整理

Initial Model の全変数の記述統計を表 2 に示す。Initial Model では「評価」を目的変数，その他の変数を説明変数の候補として，ステップワイズ法でモデル構築した。「既習文型の割合」や「一文あたりの述語数」などは比率化されているが，「教師の発話文数」や「ターンの交代回数」など

表 1　データの文字数の概要

最小	最大	合計	平均値	標準偏差
155	950	32,531	428.04	151.53

$n = 76$

表２　Initial Model の全変数の記述統計

	評価	教師の発話文数	既習語彙割合	既習文型割合	一文あたりの文節数	一文あたりの述語数	括弧の数	ターンの交代回数	総文節数	括弧内文節数
最小値	0	3	0.40	0.16	1.71	0.54	0	0	36	0
最大値	4.50	44	0.96	0.93	10.40	4.40	19	30	232	85
合計	———	1,230	———	———	———	———	479	549	7460	1906
平均値	2.28	16.18	0.67	0.51	4.21	1.44	6.30	7.22	98.16	25.08
標準偏差	1.16	8.25	0.11	0.16	1.47	0.57	4.10	5.41	34.24	20.20

$n=76$

は絶対数である。絶対数を変数に組み入れると記述量が単純に評価の高低につながってしまう。これが初期モデルの問題点であるとの認識から，比率化されていない変数を除外し，比率化した相対的な変数を組み入れた。具体的には，表２の「教師の発話文数」「括弧の数」「ターンの交代回数」「総文節数」「括弧内文節数」を取り除き，比率化した変数「ターンの交代回数／教師の発話文数」「括弧内文節数／教師の総文節数」を加えた。表３が目的変数「評価」および説明変数の候補とした全変数の記述統計である。表４はそれら全変数の相関をまとめたものである。表４から「一文あたりの述語数」と「一文あたりの文節数」の相関が非常に高いことが分かる（$r=.909$, $p<.01$）。これら二つの変数を同時に重回帰分析に投入すると，多重共線性が発生する可能性が極めて高いことから，「一文あたりの文節数」を説明変数の候補から除外した。

表３　全変数の記述統計

	評価	既習語彙割合	既習文型割合	一文あたりの文節数	一文あたりの述語数	ターンの交代回数／教師の発話文数	括弧内文節数／教師の総文節数
min	0	0.400	0.164	1.714	0.538	0	0
max	4.500	0.957	0.931	10.400	4.400	2.750	0.805
sum	———	———	———	———	———	———	———
mean	2.283	0.672	0.511	4.206	1.444	0.506	0.256
sd	1.161	0.107	0.162	1.471	0.566	0.453	0.192

$n=76$

表４　全変数の相関

		1	2	3	4	5	6	7
1	評価	—						
2	既習語彙割合	.450**	—					
3	既習文型割合	.573**	.569**	—				
4	一文あたりの文節数	−.607**	−.407**	−.631**	—			
5	一文あたりの述語数	−.646**	−.357**	−.624**	.909**	—		
6	ターンの交代回数／教師の発話文数	.269*	.359**	.392**	−.332**	−.333**	—	
7	括弧内文節数／教師の総文節数	.520**	.505**	.467**	−.425**	−.401**	.297**	—

$n=76$, $^{**}p<.01$, $^{*}p<.05$

また，多重共線性の発生を抑えるために，式（1）に示す方法で説明変数の中心化を行った。重回帰分析の前に説明変数のすべてのデータを中心化した。

$$X_i^f = X_i - \overline{X} \qquad (1)$$

X_i^f: 中心化後のデータ
X_i: 元のデータ
$\overline{X}$: 平均

2.3. 変数選択の方法

変数選択は R の step()関数と glmulti()関数を用いて行った。どちらも AIC が最小になる組み合わせを探索できるが，step()は変数を一つずつ追加したり，削除したりを繰り返すステップワイズ法を実施するのに対し，glmulti()は変数のすべての組み合わせを網羅的に探索することが可能である。Initial Model に交互作用項はなかったが，モデルの説明力を向上させることと変数同士の影響を考慮して，交互作用項をモデルに組み込むこととした。step()関数も glmulti()関数も交互作用項を考慮した探索が可能である。本稿では step()関数により導いたモデルを Step モデル，glmulti()関数により導いたモデルを Glmulti モデルと呼ぶ。さらに，外れ値の影響を低減するため 2 つのモデルに対し，ロバスト回帰を行い，汎化性能を高めた頑健なモデルを導出した。

2.4. 残差の検討方法

このように得られた二つのモデルの残差を検討するため，「残差のヒストグラム」，「残差 vs. 予測値のプロット」，「分位数分位数プロット」，「Scale-Location プロット」，「クックの距離」を描画し，Shapiro-Wilk 検定を行い，残差の状況を検討した。

2.5. 予測性能の評価

さらに，２つのモデルの予測性能を検討するため，10 分割交差検証（10-fold cross-validation）を行い，RMSE（Root Mean Squared Error：二乗平均平方根誤差）の平均，RMSE の標準偏差，RMSE の信頼区間，R^2，調整済み R^2 を比較した。RMSE は式（2）で導く。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

y_i: 観測値
$\hat{y}_i$: 予測値
n: データの個数

2.6. 外れ値の検討

モデルの構築に先立ち，データの予備的な精査を実施し，潜在的な外れ値を特定した。Step モデルと Glmulti モデルごとに，ステューデント化残差が ±3 を超えるものや，Cook の距離が高い観測値を影響の大きいデータ点として特定し，検討を進めた。

2.6.1. Step モデルの外れ値の検討

図 1 は Step モデルのステューデント化残差を示しており，点線が閾値の −3 である。±3 の閾値を超えるデータ点はインデックス 42（−3.313）のみであった。図 2 は Step モデルのクックの距離を示している。点線は 4/n の閾値（0.0526）を示しており，それを超えるデータ点は 4 つあった。インデックス 8（0.0536），35（0.3395），42（0.1097），74（0.3961）がそれである。インデックス 8 は閾値を超えているが，僅かであるため外れ値として除外しなかった。ステューデント化残差 ±3 以上のインデックスおよびクックの距離が 0.1 以上のインデックスを除外することとした。よって，除外インデックスは，35，42，74 である。

外れ値を除外後，再度データの予備的精査を実施した。図 3 がステューデント化残差，図 4 がクックの距離を表している。ステューデント化残差において ±3 以上のデータ点は無かった。クックの距離においては 4/n の閾値（0.0547）を超えるデータ点は 4 つあったが，いずれも大きく超えているということではなかったため（0.0569，0.0620，0.0641，0.0673），データから除外しなかった。

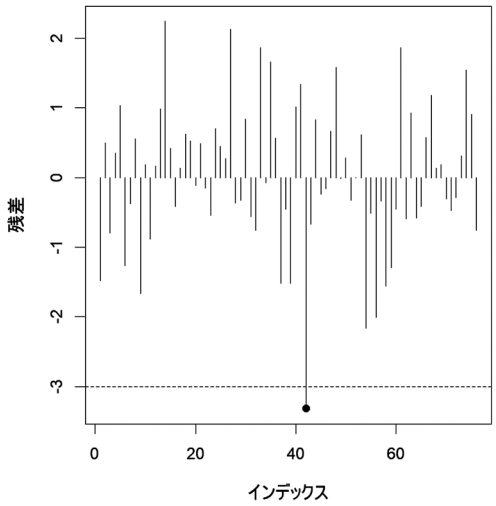

図 1　Step モデルのステューデント化残差①

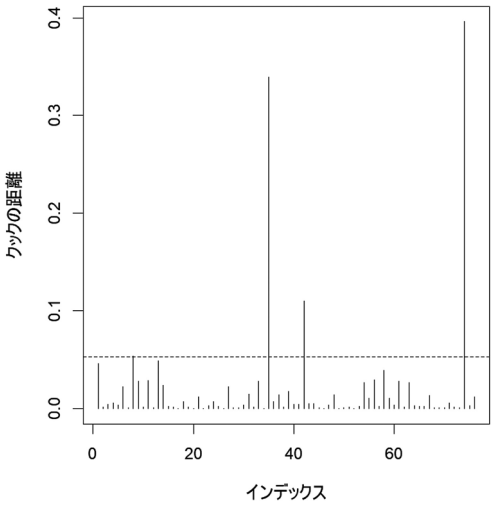

図 2　Step モデルのクックの距離①

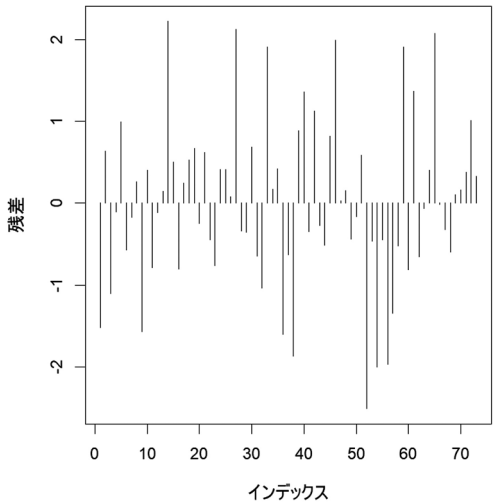

図 3　Step モデルのステューデント化残差②

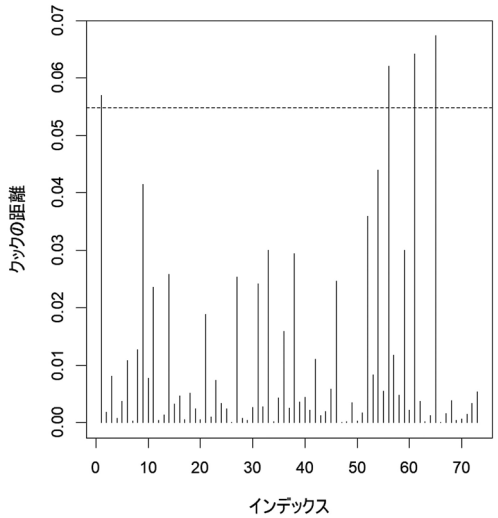

図 4　Step モデルのクックの距離②

2.6.2. Glmulti モデルの外れ値の検討

図 5 は Glmulti モデルのステューデント化残差を示しており，±3 の閾値を超えるデータ点はインデックス 42（−3.251）のみであった。図 6 は Glmulti モデルのクックの距離を示している。4/n の閾値（0.0526）を超えるデータ点は 3 つあった。インデックス 42（0.1097），58（0.0799），63（0.0860）がそれである。インデックス 58，63 は共に閾値を大きく超えているというわけではないため，データから除外しなかった。インデックス 42 がステューデント化残差 ±3 以上かつクックの距離が 0.1 以上であったため，外れ値として除外した。最終的に本稿では 10 分割交差検証を用いたモデル比較を行うが，その際，２つのモデル間で使用データが異なると比較の妥当性が損なわれる。予測性能の違いがデータの違いに起因することが無いよう，２つのモデル構築に使用されるデータは同一にする必要がある。そのため，Step モデルで外れ値として除外したインデックス 35，74 も Glmulti モデルで使用するデータから除外することとした。このように Glmulti モデルの除外インデックスは 35，42，74 となり，Step モデルのデータと同一である。

外れ値を除外後，再度データの予備的精査を実施した。図 7 がステューデント化残差，図 8 がクックの距離を表している。ステューデント化残差において ±3 以上の点は無かった。クックの距離においては 4/n の閾値（0.0547）を超えるデータ点は 4 つあった。その中のインデックス 3（0.0643），9（0.0666），56（0.0574）は僅かな超過であったためデータから除外しなかった。一方で，インデックス 61（0.1138）は 0.1 を超えていたため，外れ値として精査が必要となった。インデックス 61 を外れ値に含め，除外インデックスを 35，42，61，74 として Glmulti モデルを推定したところ，VIF（Variance Inflation Factor）が複数の変数で 10 を超える値となり，多重共線性が発生した。そのため，インデックス 61 はクックの距離が 0.1 を超えているが，そのままデー

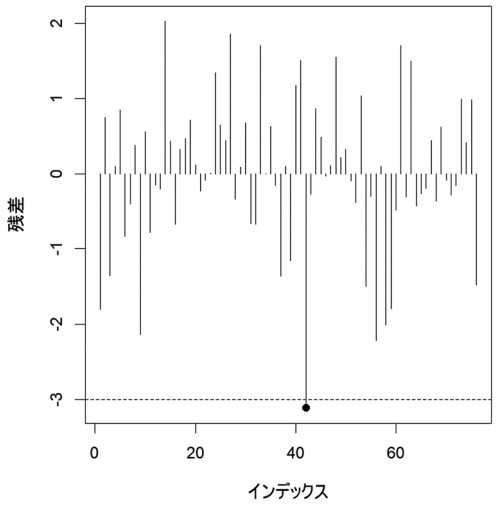

図 5　Glmulti モデルのステューデント化残差①

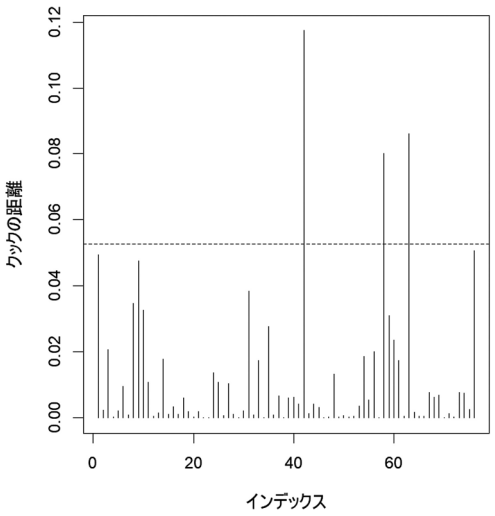

図 6　Glmulti モデルのクックの距離①

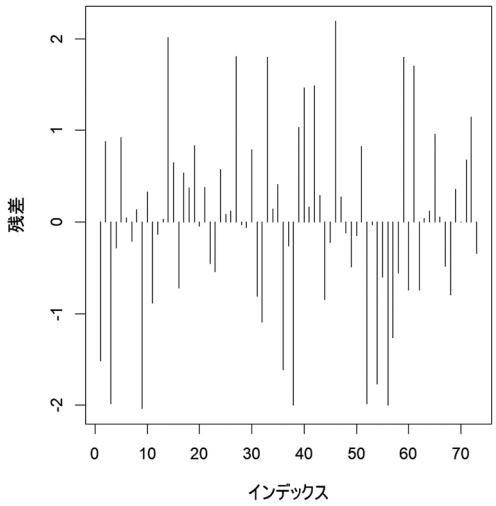

図 7　Glmulti モデルのステューデント化残差②

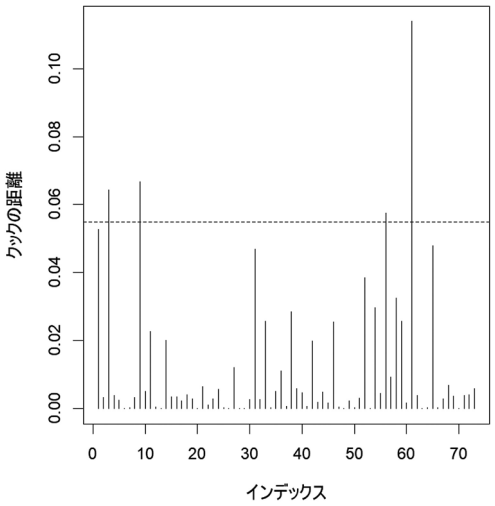

図 8　Glmulti モデルのクックの距離②

タに残すこととした。

2.7. 外れ値除外後の記述統計と相関

除外データを両モデルで共通化し，外れ値を除外済みの目的変数および説明変数の概要を表 5 の記述統計にまとめた。さらに，モデル構築に使用した変数の相関を表 6 にまとめた。「一文あたりの述語数」はいずれの変数に対しても有意な負の相関であった。

表 5　目的変数および説明変数の記述統計

	評価	既習語彙割合	既習文型割合	一文あたりの述語数	ターンの交代回数／教師の発話文数	括弧内文節数／教師の総文節数
min	0	0.213	0.467	0.538	0	0
max	4.500	0.931	0.957	4.400	2.750	0.805
mean	2.356	0.517	0.676	1.421	0.509	0.259
sd	1.123	0.155	0.104	0.539	0.458	0.190

n＝73

表 6　モデル構築に利用した変数の相関

		1	2	3	4	5	6
1	評価	—					
2	既習語彙割合	.433**	—				
3	既習文型割合	.560**	.558**	—			
4	一文あたりの述語数	−.637**	−.385**	−.607**	—		
5	ターンの交代回数／教師の発話文数	.267*	.390**	.396**	−.310**	—	
6	括弧内文節数／教師の総文節数	.523**	.510**	.424**	−.367**	.286*	—

n＝73，**p＜.01，*p＜.05

3．結　　果

step()関数と glmulti()関数による重回帰分析を実施し，Step モデルと Glmulti モデルを得た。さらに，汎化性能を高め，頑健なモデルを得るため 2 つのモデルに対し，ロバスト回帰を実施した。それぞれのモデルの概要を表 7 と表 8 に示す。各変数の「_c」は中心化済みの変数であることを示す。

表 7　Step モデルの概要

	B	*SE B*	*β*	t 値	p 値	偏 F 値	許容度	VIF
切片	2.547	0.087	2.592	29.172	0.000			
既習語彙割合_c	0.948	0.709	0.100	1.338	0.186	1.789	0.486	2.057
既習文型割合_c	2.349	0.771	0.347	3.048	0.003	9.287	0.180	5.568
一文あたりの述語数_c	−0.818	0.164	−0.452	−5.003	0.000	25.029	0.226	4.425
括弧内文節数／教師の総文節数_c	2.524	0.473	0.460	5.339	0.000	28.508	0.369	2.713
既習語彙割合_c：既習文型割合_c	−29.060	5.274	−0.471	−5.510	0.000	30.356	0.235	4.254
既習語彙割合_c：一文あたりの述語数_c	−8.035	1.784	−0.451	−4.505	0.000	20.291	0.152	6.561
一文あたりの述語数_c：(括弧内文節数／教師の総文節数)_c	4.179	1.050	0.427	3.981	0.000	15.850	0.350	2.857

n＝73，Durbin-Watson＝1.982，p＝0.450

表 8　Glmulti モデルの概要

	B	*SE B*	*β*	t 値	p 値	偏 F 値	許容度	VIF
切片	2.597	0.091	2.640	28.643	0.000			
既習文型割合_c	2.557	0.649	0.378	3.943	0.000	15.545	0.192	5.211
一文あたりの述語数_c	−0.718	0.159	−0.402	−4.501	0.000	20.259	0.220	4.546
括弧内文節数／教師の総文節数_c	2.778	0.417	0.511	6.656	0.000	44.301	0.435	2.298
ターンの交代数／教師の発話文数_c	0.626	0.259	0.279	2.420	0.018	5.858	0.194	5.155
既習文型割合_c：既習語彙割合_c	−31.063	5.453	−0.503	−5.696	0.000	32.450	0.215	4.645
一文あたりの述語数_c：既習語彙割合_c	−10.099	2.100	−0.565	−4.809	0.000	23.130	0.124	8.065
一文あたりの述語数_c：（括弧内文節数／教師の総文節数）_c	3.892	1.080	0.397	3.605	0.001	12.998	0.269	3.721
既習文型割合_c：（ターンの交代数／教師の発話文数）_c	−2.695	0.970	−0.192	−2.777	0.007	7.713	0.183	5.465

n=73，Durbin-Watson=2.070，p=0.592

3.1. モデルの導出

Step モデルでは主効果の変数が 4 つ，交互作用項が 3 つ選ばれた。「ターンの交代回数／教師の発話文数_c」は主効果でも交互作用項においても選択されなかった。Glmulti モデルでは主効果の変数が 4 つ，交互作用項が 4 つ選ばれ，Step モデルと比べると交互作用項が 1 つ多いという特徴が見られた。Glmulti モデルの主効果として「既習語彙割合_c」は選択されなかった。両モデルの主効果として「一文あたりの述語数_c」が選択され，符号は両モデルとも負であった。その他は同様の変数が選ばれ，符号と係数の大きさも同様の傾向が見られた。

変数の有意性に関して，Step モデルの「既習語彙割合_c」の p 値が 0.186 で有意ではなかったが，Glmulti モデルの変数はすべて有意（p< .05）であった。Step モデルの β の「既習語彙割合_c」を見ると 0.100 で，他の変数の β と比較して小さく，モデル内の相対的影響力が強くはないことがわかる。

多重共線性の指標となる VIF を見ると，値が 5.0 以上の変数が Step モデルでは 2 つ，Glmulti モデルでは 4 つあることが分かる。Step モデル，Glmulti モデル共に「既習語彙割合_c：一文あたりの述語数_c」において特に値が高め（Step モデル：6.561，Glmulti モデル：8.065）であった。多重共線性を判断する一般的な基準が VIF＞10 であることを踏まえ，交互作用項の「既習語彙割合_c：一文あたりの述語数_c」の値は若干高めだが，説明変数として残しても問題がないと判断した。

Durbin-Watson 統計量に基づいて評価すると，Step モデル（1.982）および Glmulti モデル（2.070）は，いずれも 2.0 に近い値を示しており，自己相関の問題はないと考えられる。さらに，p 値も両モデルとも 0.05 を超えており，統計的に有意な自己相関は確認されなかった。これら

表 9　モデル全体の統計値

	R^2	調整済み R^2	AIC	BIC	F 値	P 値	残差標準誤差
Step モデル	0.690	0.656	146.468	164.791	55.614	0.000	0.591
Glmulti モデル	0.698	0.661	156.815	177.429	55.320	0.000	0.626

表 10　各モデルの残差

	Min	1Q	Median	3Q	Max
Step モデル	−1.619	−0.365	−0.021	0.310	1.461
Glmulti モデル	−1.249	−0.351	0.005	0.345	1.350

から，両モデルは回帰分析の前提条件の一つである残差の独立性を満たしていると考えられる。

表 9 はモデル全体の統計値をまとめたものである。2 つのモデル共に $p < .01$ で有意であることが分かる。決定係数 R^2 は Step モデル（0.690）と Glmulti モデル（0.698）で，ほぼ同じ値であった。調整済み R^2 も Step モデル（0.656）と Glmulti モデル（0.661）で，ほぼ同じ値であった。AIC に関しては，Step モデル（146.468）の方が Glmulti モデル（156.815）よりも低く，Step モデルの方が赤池情報量基準においては優れていることがわかる。BIC に関しても同様で，Step モデル（164.791）の方が Glmulti モデル（177.429）よりも小さく，ベイズ情報量基準においても Step モデルの方が優れていることがわかる。残差標準誤差は，Step モデル（0.591）の方が Glmulti モデル（0.626）よりも小さく，データに対する当てはまりの良さは Step モデルの方が高いことを示している。

表 10 は各モデルの残差をまとめたものである。残差の中央値が Glmulti モデル（0.005）の方が Step モデル（−0.021）よりも 0 により近いことから，Glmulti モデルは Step モデルと比較して予測の偏りが少ない傾向があることが示唆される。残差の範囲（Max–Min）は Step モデルの方が大きく，特に最小値（Min）は −1.619 と低いことから，Glmulti モデルの方が極端な残差が少ない可能性が示唆される。ただし，第一四分位点（Q1）および第三四分位点（Q3）の範囲では両モデルの残差分布は類似しており，全体の残差分布に大きな差は認められない。

3.2. 残差の検討

Step モデルと Glmulti モデルの残差を評価するために，以下の図表を作成し，分析を行った。具体的には，残差のヒストグラム，Shapiro-Wilk 検定，残差 vs. 予測値のプロット，分位数–分位数プロット，Scale-Location プロットを用いて，残差の分布特性および等分散性の検証を行った。

3.2.1. 残差のヒストグラムおよび Shapiro-Wilk 検定

Step モデルおよび Glmulti モデルにおける残差の分布を，ヒストグラムとカーネル密度推定を

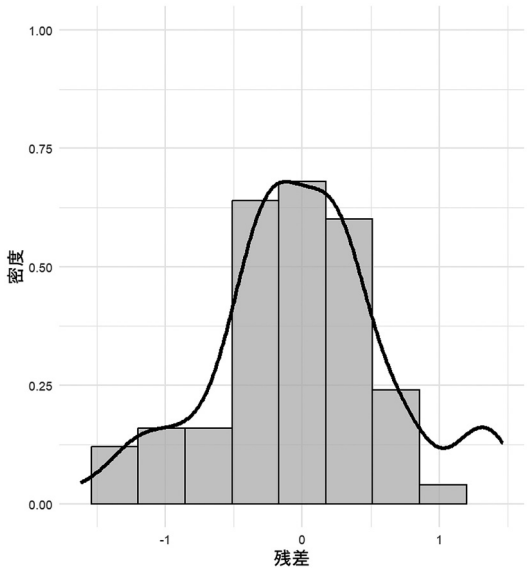

図 9　Step モデルの残差分布

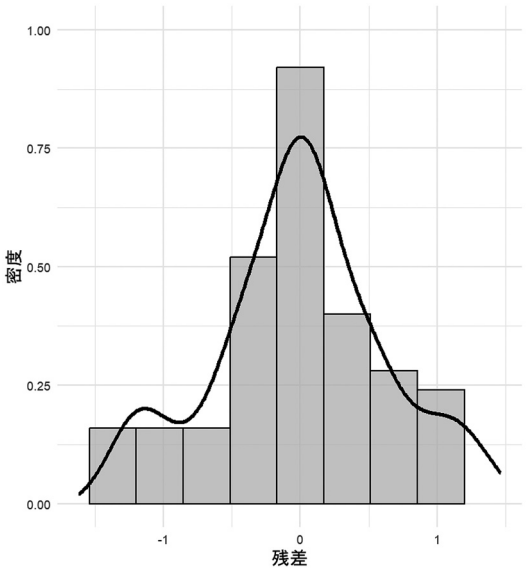

図 10　Glmulti モデルの残差分布

表 11　Shapiro-Wilk 検定の結果

	Step モデル	Glmulti モデル
W 値	0.979	0.974
p 値	0.276	0.130

用いて可視化したものが図 9 および図 10 である。Step モデルの残差分布は左右対称であるものの，Glmulti モデルと比較すると分散が大きく，ピークが低い傾向が見られる。一方，Glmulti モデルの残差分布は，中央に鋭いピークを持ち，左右対称な正規分布に近い形状を示している。両モデルとも残差の中央値は 0 付近に位置しており，大きなバイアスは認められなかった。特に，Glmulti モデルは残差が中心に集中していることから，より予測精度が高い可能性が示唆される。

残差が正規分布に従うかを検証するため，Shapiro-Wilk 検定を実施した。その結果を表 11 に示す。両モデルにおいて p 値はいずれも 0.05 を上回っており，残差が正規分布に従うという帰無仮説を棄却できなかった。したがって，両モデルの残差は正規性を満たしているとみなすことができる。また，W 値についても，Step モデル（0.979）および Glmulti モデル（0.974）はともに 1 に近く，正規性が高いことが示唆される。

3.2.2. 残差 vs. 予測値のプロット

残差のランダム性および等分散性を検証するために，両モデルの残差と予測値の散布図を作成し，その結果を図 11 および図 12 に示した。いずれのモデルにおいても，残差は予測値に対して

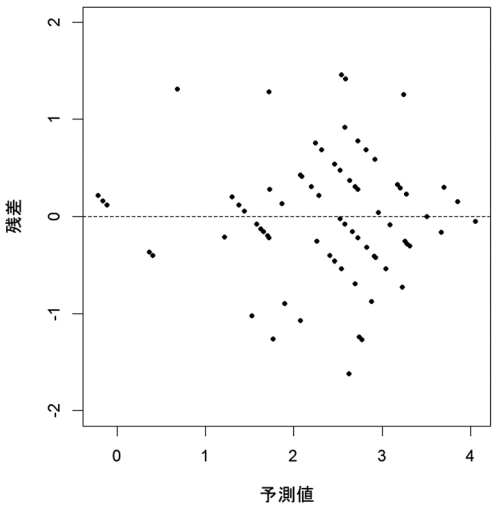

図 11　Step モデルの残差 vs 予測値

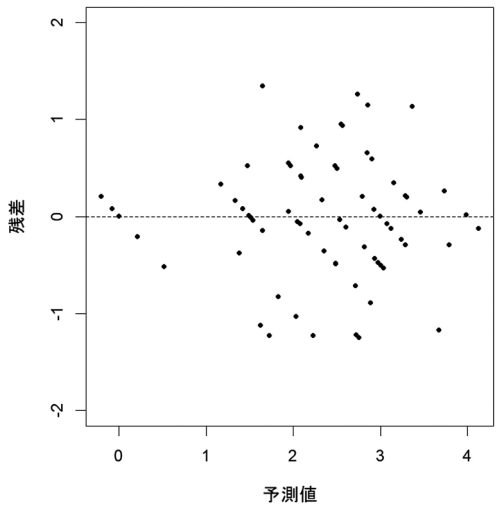

図 12　Glmulti モデルの残差 vs 予測値

ランダムに分布しており，線形性の仮定は概ね満たされていると判断された。等分散性の観点から比較すると，Step モデルでは予測値が大きい領域において残差がやや負に偏る傾向が見られたのに対し，Glmulti モデルでは予測値全体にわたって残差のばらつきが比較的均一に分布していた。外れ値の影響については，両モデルとも残差の大部分が ±2 の範囲内に収まっており，顕著な外れ値は認められなかった。これらの結果を踏まえると，両モデルは線形性，等分散性，外れ値の影響といった線形回帰の基本的仮定を概ね満たしていると言える。

3.2.3. 分位数–分位数プロット

残差の正規性を視覚的に確認するため，両モデルの分位数–分位数プロットを作成し，図 13 および図 14 に示した。分位数–分位数プロットでは，残差が理論上の正規分布に従う場合，プロットされた点は対角線上に並ぶとされる。Step モデルでは，中心付近の残差は対角線に沿って分布していたが，両端，特に右端において対角線からの逸脱が認められ，極端な値を含む可能性が示唆された。一方，Glmulti モデルにおいても中央付近の分布は良好であり，両端に若干の逸脱が見られたものの，その程度は Step モデルよりも小さかった。特に左端においてわずかに対角線から外れていたが，全体としては正規性が維持されていると判断される。また，両モデルとも残差の大部分が ±2 の範囲内に収まっており，顕著な外れ値は確認されなかった。これらの結果から，両モデルにおける残差の正規性に関して，重大な逸脱は認められなかった。

3.2.4. Scale-Location プロット

残差の均一分散性を視覚的に評価するため，両モデルの Scale-Location プロットを作成し，図 15 および図 16 に示した。Scale-Location プロットは，残差の絶対値の平方根（$\sqrt{|\text{残差}|}$）を縦

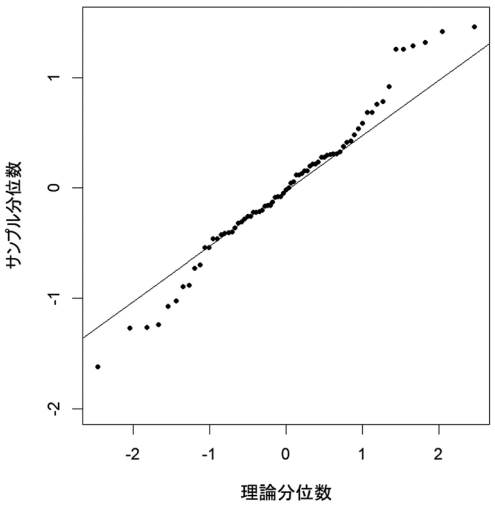

図 13　Step モデルの分位数–分位数プロット

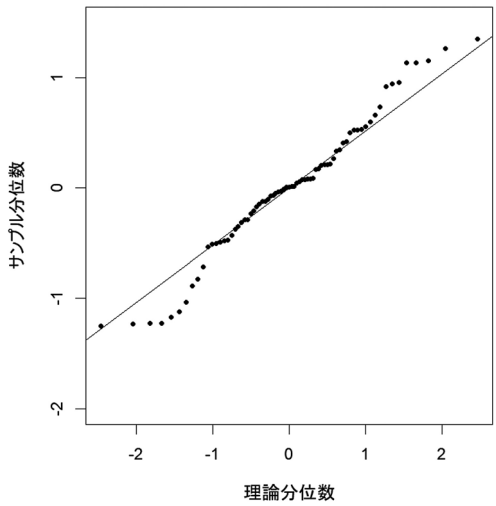

図 14　Glmulti モデルの分位数–分位数プロット

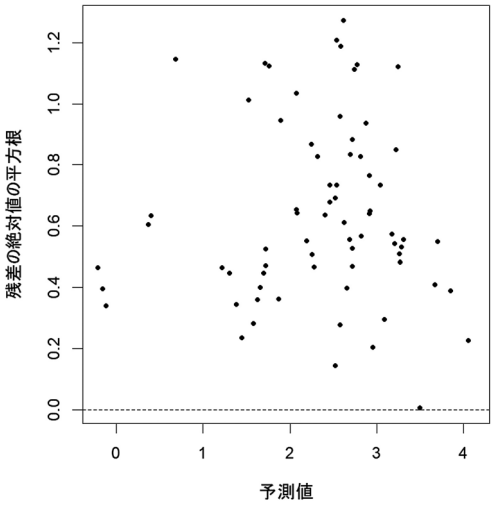

図 15　Step モデルの Scale-Location プロット

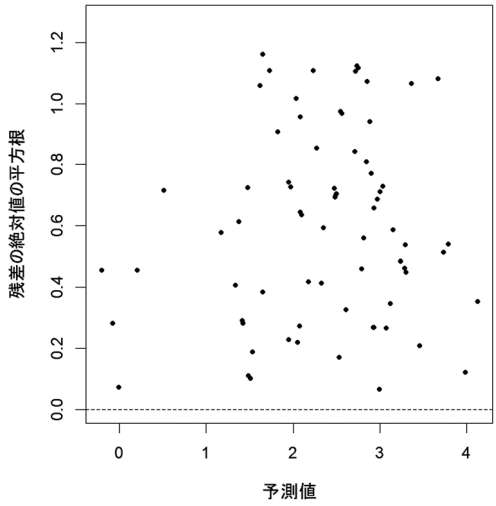

図 16　Glmulti モデルの Scale-Location プロット

軸，モデルの予測値を横軸に取り，予測値に対する残差の分散の変化を可視化する手法である。プロットされた点がランダムに分布し，特定の傾向を示さなければ，等分散性の仮定が満たされていると判断される。

Step モデルにおいては，残差のばらつきは予測値全体にわたりほぼ均一に分布しており，明確な右上がりの傾向や漏斗型のパターンは認められなかった。Glmulti モデルにおいても同様に，残差の分布に特定のパターンは見られなかった。以上の結果から，両モデルにおいて残差の均一分散性の仮定は満たされていると判断した。

3.3. 予測精度の比較：10 分割交差検証

Step モデルおよび Glmulti モデルの予測精度を評価するため，10 分割交差検証（10-fold cross-validation）を実施した。交差検証は，モデルが訓練データに過度に適合することを防ぎ，未知のデータに対する予測性能を評価する手法として広く用いられている。評価指標としては，予測誤差の大きさを示す RMSE（Root Mean Square Error）の平均値，予測の安定性を表す RMSE の標準偏差，精度のばらつきを示す RMSE の 95％信頼区間，決定係数（R^2），調整済み R^2 を用いた。これらの指標を基に，Step モデルと Glmulti モデルの予測性能を比較検討した。表 12 に交差検証の結果を示す。

予測誤差の大きさを表す RMSE の平均値においては，Glmulti モデル（0.671）の方が Step モデル（0.696）よりもわずかに小さかったが，その差は統計的に有意ではなかった（p=0.278）。また，RMSE の標準偏差についても，Glmulti モデルの方が小さく，予測の安定性において若干の優位性が認められたものの，有意差は確認されなかった（p=0.578）。さらに，RMSE の 95％信頼区間に基づく予測誤差のばらつきも，Glmulti モデル（0.599−0.743）の方が Step モデル（0.609−0.783）よりも狭く，より一貫した予測を示唆していたが，統計的な有意差は得られなかった（p=0.663）。予測力の指標である R^2 および調整済み R^2 についても，Glmulti モデルの R^2（0.670）および調整済み R^2（0.628）は，Step モデルの R^2（0.652）および調整済み R^2（0.614）をそれぞれ上回っていたが，これらの差も統計的に有意ではなかった（R^2：p=0.192，調整済み R^2：p=0.568）。以上の結果から，Glmulti モデルは全ての指標において Step モデルをわずかに上

表 12　10 分割交差検証の結果

比較項目	Step モデル	Glmulti モデル	有意差
平均 RMSE	0.696	0.671	なし $t(9)=1.155$，$p=.278$， 95% CI [−0.024，0.076]，paired t-test
RMSE の標準偏差	0.140	0.116	なし $F(9, 9)=1.467$，$p=.578$， 95% CI [0.364，5.904]，F-test
RMSE の信頼区間	(0.609,0.783)	(0.599,0.743)	なし $t(17.378)=0.443$，$p=.663$， 95% CI [−0.096，0.147]，Welch's t-test
R^2	0.652	0.670	なし $z=-1.305$，$p=.192$， 95% CI [−0.149，0.030]，Steiger's Z-test
調整済み R^2	0.614	0.628	なし $z=-0.571$，$p=.568$， 95% CI [−0.106，0.058]，Steiger's Z-test

回ったものの，いずれの指標においても統計的な有意差は認められず，両モデルの汎化性能に本質的な差はないと考えられる。

4．考　　察

step()関数およびglmulti()関数を用いて重回帰分析を行い，それぞれのモデルを構築した。さらに，モデルの頑健性を高めるために，両モデルに対してロバスト回帰を実施し，Stepモデルおよび Glmultiモデルを導出した。以下に2つのモデルを示す。

Stepモデル：

$$Y = 2.547 + 0.948X_1 + 2.349X_2 - 0.818X_3 + 2.524X_4 - 29.060X_1X_2 - 8.035X_1X_3 + 4.179X_3X_4$$

X_1：既習語彙割合_c
X_2：既習文型割合_c
X_3：一文あたりの述語数_c
X_4：（括弧内文節数／教師の総文節数）_c

Glmultiモデル：

$$Y = 2.597 + 2.557X_1 - 0.718X_2 + 2.778X_3 + 0.626X_4 - 31.063X_1X_5 - 10.099X_2X_5 + 3.892X_2X_3 - 2.695X_1X_4$$

X_1：既習文型割合_c
X_2：一文あたりの述語数_c
X_3：（括弧内文節数／教師の総文節数）_c
X_4：（ターンの交代数／教師の発話文数）_c
X_5：既習語彙割合_c

両モデルとも，モデル全体のp値はいずれも0.01未満であり，統計的に有意なモデルであると確認された。構成変数の観点では，Stepモデルにおける「既習語彙割合_c」のp値は統計的に有意ではなかったものの，使用される変数の数はGlmultiモデルより1つ少なく，Stepモデルはより簡潔な構造を有していた。一方，調整済みR^2は両モデルでほぼ同等であり，いずれも同程度の説明力を有していると考えられる。ただし，AICおよびBICの値はStepモデルの方が低く，情報量基準に基づく適合度においては，Stepモデルの方が良好である可能性が示唆された。残差の分布に関しては，Glmultiモデルの方が残差が0付近に集中しており，正規性がより保たれていた。これらの結果を総合すると，両モデルは訓練データに対する適合度において概ね同等の妥当性を有しているが，それぞれ異なる特性を持つと言える。すなわち，Stepモデルはモデルの簡

潔性および情報量基準の観点から優れており，Glmulti モデルは残差分布の正規性および予測精度においてより精緻である可能性がある。

未知のデータに対する予測性能を評価するために，10 分割交差検証を実施した。RMSE の 3 指標（平均値，標準偏差，95％信頼区間），R^2，調整済み R^2 のいずれの指標においても，Glmulti モデルの方がわずかに良好な結果を示したものの，統計的有意差は確認されなかった。James（2013）は「一般的には訓練データでどの程度機能するかはあまり重要ではない。むしろ，その訓練に使わなかったデータに対して，どの程度正確な予測ができるかについて関心がある」と述べており，予測性能評価の重要性を強調している。この指摘と，本研究の目的である自動採点モデルの開発という観点を踏まえると，交差検証の結果において有意差は見られなかったものの，すべての指標において Glmulti モデルが良好な値を示したことは注目に値する。特に，自動採点システムにおいては，予測の安定性や誤差の最小化が重要となるため，わずかな性能差であっても，ユーザーとしての実習生には大きな違いとして受け止められる可能性もある。したがって，本研究の目的との整合性から判断すると，Glmulti モデルの方が実践応用上優先されるべきモデルと言える。

一方で，Step モデルは使用変数が少なく，モデル構造が簡潔であるという利点を有している。この点は，モデルの解釈性や実装の簡便性という観点から重要であり，特に教育現場での応用を想定した場合には，指導や運用の容易さに貢献する可能性がある。つまり，予測精度よりも教育上の簡便性や解釈性を重視する場合には，Step モデルが選択肢となり得る。

5．結　　論

本稿では，日本語教員養成課程での活用を想定した教案自動採点システム（Tetrater）における評価の自動推定に用いる重回帰モデルの再構築を試みた。従来モデル（Initial Model）は，既習語彙数や文型数といった絶対値を用いており，記述量が多くなれば，それだけ得点が高くなる場合があるという問題点があった。そこで，変数の比率化を行い，記述量の影響をできるだけ抑制した上で，中心化を行い，R の step()関数および glmulti()関数を用いて重回帰分析を実施し，2 種類の改良モデルを構築した。さらに，両モデルに対してロバスト回帰を適用することで，外れ値の影響を抑えた頑健なモデルを導出した。その結果，両モデルはともに高い説明力を有しており，情報量基準（AIC・BIC）やモデルの簡潔性の観点では Step モデルがやや優れていることが分かった。また，10 分割交差検証の結果，統計的有意差は認められなかったが，RMSE の平均，調整済み R^2 などのすべての指標において Glmulti モデルはわずかに優れた予測性能を示した。このことから，自動採点システムにおいては誤差の最小化や安定した予測が求められるため，Glmulti モデルの方が実践教育上は適していると言えるだろう。一方，Step モデルは使用変数が少なく構造がより簡潔であり，教育現場での導入や説明の容易さを重視する場合には有用性が高いと言える。改善された 2 つのモデルは両者とも，十分な予測力，頑健性があり，それぞれの特

性を持ったモデルであることが分かった。

今後の課題としては，改善されたモデルをシステムに実装し，実際に日本語教員養成課程の実習生にシステムを使用してもらい，出力された評価の妥当性を検証する必要がある。また，システムに関する使用感や機械的な評価から受ける印象を尋ねる質問紙調査，インタビューなどを行い，教案作成における学習の認知，行動の変容を探りたい。さらに，システム使用後のグループ活動における発話内容にも着目し，自動採点と他者とのコミュニケーションが学習行動に及ぼす効果を明らかにしたい。

参考文献

平田歩（2008）「日本語教育課程における教育実習の試み」，『梅光学院大学論集』41，pp.54-59，梅光学院大学.

James, G., Witten, D., Hastie, T. J., Tibshirani, R. J.（2013）*An introduction to statistical learning with applications in R*. Boston: Springer.（落海浩・首藤信通（訳）R による統計的学習入門 朝倉書店）

歌代崇史（2021）「ティーチャートークの適切さの自動推定に向けた探究―日本語教員養成課程において―」，河西良治教授退職記念論文集刊行会（偏）『言語研究の扉を開く』．pp.14-28，開拓社.

歌代崇史（2024）「日本語教員養成のための教案自動採点システムは実践環境でどのように機能するのか？」，『日本教育工学会研究報告集』2024(4)，pp.221-228，日本教育工学会.

歌代崇史・須藤むつ子（2017）「教室内の言語調整の練習を支援するシステムの開発―実習生の意識と言語使用に注目した評価―」，『日本教育工学会論文誌』41(2)，pp.109-23，日本教育工学会.