論文引用に影響を与える要因：負の二項重回帰による検討

芳鐘冬樹† 辻慶太 小野寺夏生
筑波大学図書館情報メディア研究科
† fuyuki@slis.tsukuba.ac.jp

抄録
論文の被引用数を論文の計量書誌学的要因から説明するモデルとして，負の 2 項重回帰モデルと線形回帰モデルを設定し，説明要因の有効性を調べるとともに，両モデルの当てはまりのよさを比較した。自然科学系 6 分野のデータを対象に分析した結果，どちらのモデルにおいても Price 指数が全分野で有効であること，負の 2 項重回帰モデルが線形回帰モデルよりも観測値に対する当てはまりがよいことが示された。

1.はじめに

研究評価を行う際の参考データとして，論文の被引用数に基づく指標がよく用いられる。被引用数は論文の影響度を示す統計的測度としては適切であると考えられるが，これを研究評価に用いるには，その特性をよく知り，十分な注意を払う必要がある [1)]。たとえば，論文が掲載された雑誌のインパクトファクターを論文の評価指標に転用する例がしばしば見られるが，これは極めて不適切な使用例である [2)]。

更に，単純に論文間の被引用数を比較することには問題がある。論文の被引用数は，分野，掲載誌の発行国，論文の種類（原著論文，短報，レビュー等），使用言語等，多くの要因に影響されるからである。

論文の被引用数と様々な要因との関係を検討した研究は多数報告されている。しかし，それらの結論は必ずしも一致していない。その理由の 1 つは，これらの研究の多くがある単一の要因に着目しているか，複数の要因をそれぞれ独立に見ていることにある。従って，ある要因が被引用数と相関があっても，それが他の要因の交絡因子である可能性を否定できない。

重回帰分析は，個々の説明変数による効果を分離できるので，考え得る多くの要因の中のどれが被引用数に対する説明力が高いかを推察することができる。この方法を用いた研究はいくつかあり [3-10)]，それぞれ興味深い結果が得られている。しかし，これらの研究も，特定の分野を対象にしている，研究目的が被引用数に影響を与える要因の解明ではないなどの点から，汎用的な結論を出しているとは言えない。

筆者らは，①複数の分野とそれらの中の複数の雑誌について，②同一年に発表された原著論文を採り，③比較的長期間(6 年間)の被引用数を目的変数とし，④想定される多くの潜在要因を説明変数として，重回帰分析を進めている。これによって，被引用数に対する各要因の寄与を検討し，論文の被引用数に対するより精密なベースラインを与えることを目標としている。6 つの分野について，9 つの要因を考慮した中間的な結果を昨年報告した [11)]。Price 指数，参考文献数，論文長が被引用数に対する有効な説明要因であり，分野を越えてある程度一貫性のある予測モデルが得られる可能性を示した。

本研究は，これに以下の 2 つの改善を行ったものである：①説明変数に被引用論文著者の過去の業績（論文数，被引用数，活動期間）を加える，②非線形回帰の 1 つである負の 2 項重回帰(NBR)モデルを採用する。目的変数である被引用数は，非負の値をとる整数であり，その分布が極めて歪度が大きいことから，線形回帰(LR)モデルより NBR モデルが適していると考えられ，いくつかの先例 [7-9)]がある。

2.データ

本研究では物性物理学，無機・核化学，電気・電子工学，生化学及び分子生物学，生理学，消化器病医学の 6 分野から各 4 誌（すべて英文誌）を選び，それぞれから 2000 年発表の原著論文 50-60 件（計 1,395 件）を無作為抽出して，調査対象論文とした。手順としてはまず 2006 年 7 月に，Thomson の Web of Science から，24 誌の 2000 年における原著論文(articles)のデータ

表 1　説明変数

	変数名	変数略号	変数の説明
第1著者の属性	第1著者発表実績	Publ_Prim	2000年までの第1著者の論文数
	第1著者活動期間	Age_Prim	第1著者の最初の論文発表から2000年までの期間
	第1著者引用実績	Impact_Prim	Publ_Primの全論文の2000年までの被引用数合計（自己引用を除く）
著者集合属性	著者数	Authors	対象論文の著者数
	所属機関数	Insts	Authorsの所属機関数（同一機関内の異なる部門は1機関とする）
	所属国数	Countries	Instsが属する国の数
	*最高著者発表実績	Publ_Max	2000年までの最高論文数共著者の論文数
	*最長活動期間	Age_Max	2000年までの活動期間最長の共著者の活動期間
	*最高著者引用実績	Impact_Max	2000年までに最高の引用を得た共著者の被引用数合計（自己引用を除く）
論文の属性	論文長	Length	対象論文のページ数を紙面係数で補正
	図数	Figures	対象論文中の図の数
	表数	Tables	対象論文中の表の数
	数式数	Eqs	対象論文中の数式の数
	参考文献数	Refs	対象論文中の参考文献数
	Price指数	Price	Refsのうち最近5年以内のもの（1996以降）の比率
	雑誌	1A～4A	掲載誌を示すダミー変数（0か1）

をダウンロードした後，各誌から 50-60 論文を無作為抽出して調査用標本とした（2 ページ以下の論文，会議プロシーディング論文は除外した)。これらの標本論文の書誌データと，その参考文献及び引用文献のデータを，Thomson から購入し，これから後述の説明変数及び目的変数のデータを取得した。図の数，表の数，数式の数は，別途原論文に当たって取得した。

3.方法

NBR モデルに基づく回帰分析では，個体 i に対する目的変数 y_i（論文 i の被引用数）が負の 2 項分布：

$$\Pr(y_i = k) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{k}$$

に従うとし，その期待値 μ_i を次式で予測する：

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

β_0，β_1，β_2，…，β_p 及び θ の推定値は、$\{X_{i1}, X_{i2}, \cdots, X_{ip}; y_i\}$ の入力データに基づいて算出される（θ は i によらないとする)。

本研究では，以下の目的変数，説明変数を取り上げ，先述の 6 分野について非線形回帰分析（NBR）と線形回帰分析（LR）を行った：

[目的変数]

自己引用も含む被引用数 TC_Total とした[1]。LR においては ln(TC_Total+1)に変換した。

[説明変数]

表 1 のものを用いた。「第 1 著者発表実績」「第 1 著者引用実績」「最高著者発表実績」「最高著者引用実績」は，完全計数法による[2]。本研究では基本的に 6 分野各々4 つの雑誌を調査に用いたが，被引用論文著者の最高過去業績（表 1 中「*」を付した「最高著者発表実績」「最長活動期間」「最高著者引用実績」）についてはその算出作業量が膨大なため，6 分野各々1 雑誌のみを対象とした。即ち，本研究では 6 分野それぞれにおいて，(1) 1 誌だけを用い，最高過去業績の 3 変数も含める（ダミー変数「雑誌」は含めない）回帰分析と，(2) 4 誌すべてを用い，最高過去業績は含めない回帰分析の 2 種類を行った。

統計解析ソフトには R を用い，NBR については glm.nb 関数，LR については lm 関数を用いた。両者共に step 関数を用いてステップワイズ法で説明変数の選択を行った。

4.結果

4.1 回帰分析結果の概要

分野ごとの回帰分析結果の概要として，標本数 n，および NBR，LR 各々の AIC（赤池情報量規準）と LR の Rc^2（自由度調整済み決定係数）を表 2 に示した。AIC と Rc^2 は共にモデルのよさを評価する尺度（AIC は小さいほど評価が高く Rc^2 は大きいほど評価が高い）であるが，

[1] 自己引用を除いても結果に大きな差異はなかった。

[2] 調節計数法を用いても結果に大きな差異はなかった。

表 2　分野ごとの回帰分析結果

			NBR				LR			
分野	対象	n	AIC	相対残差平方和		残差平方和	Rc^2	AIC	相対残差平方和		残差平方和
物性物理学	4誌	230	1499.0	**181.2**	**(0.79)**	24074.4	0.262	-50.5	**468.5**	**(2.04)**	27366.6
	1誌	55	408.2	**34.1**	**(0.62)**	11453.9	0.231	-7.2	**98.2**	**(1.79)**	12629.5
無機･核化学	4誌	227	1466.9	**142.1**	**(0.63)**	18625.9	0.320	-107.5	**332.0**	**(1.46)**	20156.8
	1誌	54	354.4	**19.3**	**(0.36)**	2696.3	0.326	-30.4	**36.5**	**(0.68)**	3044.8
電気･電子	4誌	229	1201.8	**290.8**	**(1.27)**	13940.7	0.271	-86.4	**1068.6**	**(4.67)**	15097.5
工学	1誌	59	368.0	**39.9**	**(0.68)**	3950.5	0.306	-26.9	**80.4**	**(1.36)**	3746.2
生化学及び	4誌	240	1798.6	**104.6**	**(0.44)**	59498.3	0.475	-194.7	**172.8**	**(0.72)**	60364.3
分子生物学	1誌	60	488.7	**21.4**	**(0.36)**	18394.6	0.143	-64.8	**31.4**	**(0.52)**	19273.6
生理学	4誌	236	1600.9	**122.4**	**(0.52)**	28237.3	0.607	-198.1	**251.4**	**(1.07)**	29027.7
	1誌	58	444.7	**18.0**	**(0.31)**	8196.3	0.285	-61.1	**29.2**	**(0.50)**	8848.4
消化器病	4誌	233	1805.8	**161.5**	**(0.69)**	144454.1	0.497	-75.9	**385.3**	**(1.65)**	152235.8
医学	1誌	59	554.7	**34.3**	**(0.58)**	83866.5	0.192	-32.1	**93.0**	**(1.58)**	94743.4

どの分野で評価が高いか，また，4 誌すべてを用いて最高過去業績の 3 変数は含めない場合と，1 誌だけを用いて最高過去業績も含める場合のどちらが評価が高いかは, NBR と LR とで異なり，かつ同じ LR であっても AIC と Rc^2 とで異なり，一貫した傾向は認められなかった。

ステップワイズ法による変数選択において，6 分野のうち 2 分野以上で選ばれた変数を表 3 に示した。NBR，LR 共に，また，4 誌の場合も 1 誌の場合も共に，Price 指数が最もよく選択されており（特に 4 誌の場合は全分野で選択された), 分野を越えて被引用数に対して説明力を持つ変数であると確認できる。一方，被引用論文著者の過去業績（第 1 著者の業績と全共著者を考慮した最高業績）については，選択された分野数はすべて 2 以下に留まった。

4 誌すべてを用いる場合に関しては，掲載誌を示すダミー変数が多くの分野で選択された。掲載誌も，被引用数に対する有効な説明要因であることが示唆される。

4.2 NBR と LR の当てはまりのよさの比較

前述の Rc^2 は NBR に適用できず[3]，AIC も NBR と LR という異なるモデル間の比較には必ずしも適さない。そこで，モデルの当てはまりのよさについて NBR と LR を比べる尺度として, 相対残差平方和 SS を用いることとした。

$$SS = \sum \left(\frac{y_i - \mu_i}{\mu_i} \right)^2$$

相対残差平方和は，目的変数の観測値 y_i と予測値（期待値）μ_i のずれを示す尺度で，両者の差を期待値 μ_i で規格化した値の平方和である。NBR においても $\log(y)$ を目的変数とする LR においても, 残差は μ_i に比例して大きくなりやすいため，より合理的に比較できるよう，μ_i で規格化して残差の均等化を図った。表 2 に各々の相対残差平方和（括弧内は n で規格化した相対残差平方平均）を示した。参考までに，規格化していない残差平方和も併せて示した。

相対残差平方和と相対残差平方平均は，どの分野においても，また対象が 4 誌の場合も 1 誌の場合も，NBR の方が LR よりも値が小さい，即ち当てはまりがよいことが示された[4]。

次に，期待値ではなく，分布全体の当てはまりを見るために，被引用数ごとの論文数の観測値と予測値の分布を調べた。NBR に関しては，各論文について被引用数 $\Pr(y_i=k)$ の確率分布を計算し，それを積算することで予測値の分布を求めた。図 1 は，物性物理学の 4 誌を例に，被引用数ごとの分布を示したものである。図から，特に低被引用数において, NBR が LR よりも当てはまりがよいことを見て取れる。

5. まとめ

[3] NBR に適用できる Pseudo R^2 という尺度はあるが，一般的でなく，また LR の決定係数と定義が異なるため比較に適さない。

[4] 残差平方和の比較でも，1 つの例外（電気電子工学の 1 誌）を除き，NBR の方が値が小さい。

負の 2 項重回帰モデルにおいても，線形回帰モデルと同様，Price 指数が被引用数の説明に有効であることが確認できた。本研究で説明変数として新たに加えた，著者の過去業績については，その有効性を明確に示すことはできなかった。負の 2 項重回帰モデルと線形回帰モデルの比較では，前者の方が観測値に対する当てはまりがよいことが示された。

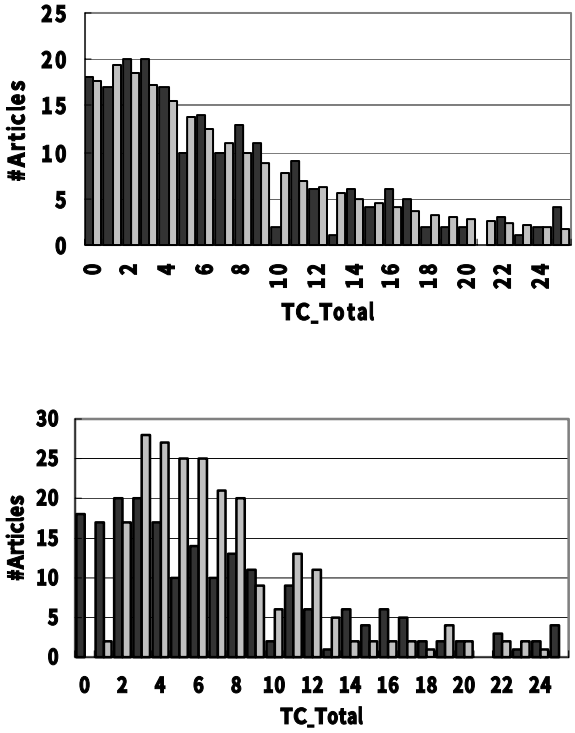

図 1　被引用数ごとの分布（上：NBR, 下：LR）（黒：観測値，グレー：予測値）

謝辞

本研究は，科学研究費補助金（基盤研究 (B)）により行っているものである。

引用文献

(1) 根岸正光；山崎茂明. 研究評価. 丸善, 2001.

(2) Seglen, P. O. The skewness of science. *J. Am. Soc. Inf. Sci.* 43(9), 628-638, 1992

(3) Snizek, W. E. et al. Textual and non-textual characteristics of scientific papers. Neglected science indicators. *Scientometrics.* 20(1), 25-35, 1991.

(4) Peters, H. P. F.; van Raan, A. F. J. On determinants of citation scores: A case study in chemical engineering. *J. Am. Soc. Inf. Sci.* 45(1), 39-49, 1994.

(5) Callaham, M. et al. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-revied journals. *JAMA.* 287(21), 2847-2850, 2002.

(6) Basu A.; Lewison, G. Going beyond journal classification for evaluation of research outputs. *Aslib Proc.* 57(3), 232-246, 2005.

(7) van Dalen, H. P.; Henkens, K. Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics.* 64(2), 209-233. 2005.

(8) Bornmann, L.; Daniel, H-D. Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. *J. Am. Soc. Inf. Sci Technol,* 58(8), 1100-1107, 2007.

(9) Bornmann, L.; Daniel, H-D. Selecting manuscripts for a high-impact journal through peer review: A citation analysis of Communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *J. Am. Soc. Inf. Sci Technol,* 59(11), 1841-1852, 2008.

(10) Davis, P. M. Author-choice open-access publishing in the biological and medical literature: A citation analysis. *J. Am. Soc. Inf. Sci. Technol.* 60(1), 3-8, 2009.

(11) 小野寺夏生ほか. 論文の被引用数に影響する要因に関する統計学的研究. 第 56 回日本図書館情報学会研究大会発表要綱, 41-44, 奈良，帝塚山大学，2008 年 11 月.

表 3　よく選択された説明変数，およびそれらが選択された分野数

NBR				LR			
4誌		1誌		4誌		1誌	
Price	6	Price	5	Price	6	Price	4
Refs	5	Length	4	Refs	5	Impact Max Full	2
Age Prim	2	Authors	3	Authors	3	Length	2
Authors	2	Refs	2	Figures	3	Age Max	2
Figures	2	Tables	2	Length	3	Authors	2
Length	2	Impact Max Full	2	Impact Prim Full	2	Eqs	2
Tables	2	Eqs	2	Insts	2		
				Tables	2		
				Countries	2		