Ⅳ-26　　予測誤差を考慮した重回帰モデルの変数選択について

北海道大学　学生員　○岩立　忠夫
北海道大学　正　員　山形　耕一
北海道大学　正　員　佐藤　馨一

1．はじめに

重回帰モデルは、交通計画の諸分野でも多く用いられており、その変数選択の方法については、RSS法やPSS法などいろいろな研究がなされている。しかし、従来までの研究では、現象をどの程度説明できるかという点に対する配慮が主であり、将来予測に際しての説明変数のもつ不確実性については、考慮されていない。本研究では、説明変数の将来予測値のもつ不確実性に注目し、この不確実性の影響を含めた上で、目的変数の将来予測値の分散を最小にするように、重回帰モデルにおける変数選択を行なう方法を考察した。

2．予測誤差分析と発生重回帰モデル決定への応用

予測誤差分析は、数学的モデルによる予測値に含まれる誤差の原因を考慮して、それらの原因の特性から、その予測値に含まれる誤差の大きさを、予測時点で事前に推定するものである。重回帰モデルを用いた場合、目的変数の将来予測値に含まれる誤差は、大別すると、①モデルによって現象を説明しきれないことによる誤差、②説明変数の将来予測値の不確実性に基づく誤差、の2つがある。従って、信頼性の高い目的変数の予測値を求めるためには、予測誤差の原因となる①，②の誤差について小さくすることが必要となる。ところが、①の誤差は、主に残差の影響によるものであり、説明変数の数を多くすることにより小さくなるが、他方、②の誤差は、説明変数の数が多くなるとその不確実性が増し、一般には大きくなる。このように、予測誤差には、説明変数の数について相反する2種類の誤差が含まれている。本研究では、この2種類の誤差の和としての予測誤差が、最小となるようなモデルを最良の予測モデルと考え、発生交通量予測モデルに応用した。

3．予測誤差分析の方法と発生重回帰モデル決定の手順

予測誤差分析の手法として、確率分布形，モーメント，あるいは、信頼限界を求めるなど、いくつかあげられるが、本研究では、モーメント計算法により予測誤差分析を行なった。線形回帰モデルとして、(1)式を考える。

$$y = \hat{a}_0 + \sum_{i=1}^{p} \hat{a}_i x_i + \varepsilon \quad \cdots\cdots \quad (1)$$

ここで、y：目的変数，x：説明変数，$\hat{a}_0, \hat{a}_i$：パラメーターの推定値，ε：誤差項

(1)のモデルにより、将来予測値を求める場合には、(2)式のようになる。

$$\hat{y}^*_\nu = \hat{a}_0 + \sum_{i=1}^{p} \hat{a}_i \hat{x}^*_{\nu i} + \varepsilon \quad \cdots\cdots \quad (2)$$

ここで、$\hat{x}^*_{\nu i}$：i番目説明変数の第ν番目ゾーンの将来推定値

この(2)式より、第ν番目ゾーンの将来予測値の分散：$V[\hat{y}^*_\nu]$を求める。計算式誘導に際して、説明変数間が独立であり、パラメーターと説明変数の推定値が独立であり、パラメーターおよび説明変数の推定値と誤差項εが独立であると仮定すると、第ν番目ゾーンの将来予測値の分散：$V[\hat{y}^*_\nu]$は、(3)式で与えられる。

$$V[\hat{y}^*_\nu] = V[\hat{a}_0] + 2\sum_{i=1}^{p} \hat{x}^*_{\nu i} Cov[\hat{a}_0, \hat{a}_i] + \sum_{i=1}^{p}\sum_{j=1}^{p} \hat{x}^*_{\nu i}\hat{x}^*_{\nu j} Cov[\hat{a}_i, \hat{a}_j] + \sum_{i=1}^{p} \hat{a}_i^2 V[\hat{x}^*_{\nu i}] + \sum_{i=1}^{p} V[\hat{a}_i] V[\hat{x}^*_{\nu i}] + V[\varepsilon] \quad \cdots\cdots \quad (3)$$

(3)式で、第1項、第2項、第3項は、パラメーター推定値の誤差が目的変数の将来予測値に対して及ぼす影響を表わす。第4項は、説明変数の予測の不確実性による影響である。第5項は、パラメーター推定値の誤差と説明変数の予測の不確実性との相乗したものが及ぼす影響であるが、テーラー展開による近似式で表わしたときには、2次の項で出現するものであり、本研究の応用例においても実際にその値の占める割合が小さいことがわかったので省略できると考えられる。また、最後の項は、重回帰モデル不一致による誤差の影響である。(3)式の中で、$V[\hat{a}_i]$，$Cov[\hat{a}_0, \hat{a}_i]$，$Cov[\hat{a}_i, \hat{a}_j]$，および、$V[\varepsilon]$については、回帰分析の際に求めることができる。(3)式から得ら

れるのは、ゾーンごとの目的変数の予測誤差であり、これらを全ゾーンについて総合的に評価して、変数を選択する必要がある。このために、たとえば、予測誤差の総和$\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]$を用いることができる。しかしながら、(3)式では、モデル作成段階で、事前に、説明変数の各ゾーンごとの将来値を算出する必要があるが、ゾーンごとの将来値を求めるのは難しくもあり、また煩雑である。そこで、平均的なゾーンにおける将来予測値のみによって、全体の分散を近似する方法を考えた。(3)式において、説明変数のゾーン別推定値が独立であると仮定すると、

$$\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]/n = V[\hat{a}_0] + 2\sum_{i=1}^{p} \bar{\hat{x}}_i^* Cov[\hat{a}_0, \hat{a}_i] + \sum_{i=1}^{p}\sum_{j=1}^{p} \bar{\hat{x}}_i^* \bar{\hat{x}}_j^* Cov[\hat{a}_i, \hat{a}_j] + \sum_{i=1}^{p} \hat{a}_i^2 V[\bar{\hat{x}}_i^*] + V[\varepsilon] \quad \cdots\cdots \quad (4)$$

ここで、$\bar{\hat{x}}_i^*$：i番目説明変数の将来推定値の平均値 $\left(\bar{\hat{x}}_i^* = \sum_{\nu=1}^{n} \hat{x}_{i\nu}^*/n\right)$

(4)式では、各説明変数に対して全ゾーンの総和としての値を予測すればよいことになり、実用的である。

本研究における予測誤差を考慮した重回帰モデルの決定手順は、以下のようなものである。まず、モデルとしての現象説明性を確かめるために、ステップワイズ法を適用して、説明力のある妥当な変数の候補を選択する。その後、それらの変数の候補に、予測の不確実性を導入して、予測誤差分析を行ない、説明変数を定める。

4．発生重回帰モデル決定への応用例

全目的発生交通量予測モデルに予測誤差分析を応用した例を以下に示す。なお、データとして、昭和47年度道央都市圏パーソントリップ調査（札幌市他3市3町,82ゾーン）の結果を用いた。

表－1　全目的発生交通量予測モデルにおける予測誤差分析例

ステップ	ステップワイズ法により作成されたモデル式	$\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]$	$\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]/n$	C.B.Dゾーンの$V[\hat{y}_\nu^*]$	RSS	R^2（寄与率）	$RSS/\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]$
1	$y = 3.2343x_1 + 19438.7844$	2.6687×10^{10}	3.0390×10^{8}	0.1596×10^{10}	2.3563×10^{10}	88.02%	88.29%
2	$y = 2.9841x_1 + 1.2204x_2 + 380.0488$	0.7913×10^{10}	0.8035×10^{8}	0.0885×10^{10}	0.5557×10^{10}	97.17%	70.23%
3	$y = 3.2244x_1 + 4.9454x_2 - 8.1311x_3 - 285.7479$	1.4393×10^{10}	1.2914×10^{8}	0.0901×10^{10}	0.2381×10^{10}	98.79%	16.54%
4	$y = 3.5471x_1 + 4.7067x_2 - 7.3353x_3 - 2.5456x_4 + 66.3514$	1.2289×10^{10}	1.1740×10^{8}	0.1057×10^{10}	0.1877×10^{10}	99.05%	15.27%

ただし、x_1：事業所従業員数，x_2：夜間人口，x_3：就業人口，x_4：第2次従業人口

$V[\hat{x}_{i\nu}^*] = (\hat{x}_{i\nu}^* \times w_i/100)^2$ で与える．

各説明変数の将来予測値の不確実性は、予測値の相対誤差：wで表わし、$w = 5$%の水準にとった。wは、事前に正確な値を求めることは難しく、経験的にどの程度の値をとるかが推測できるにすぎない。本研究における$w = 5$%は、比較的小さい値であり、このことは、将来時点における説明変数の不確実性を小さいものとみたことを意味するが、その場合でも、上例のごとく、選択される変数は少なくなり、モデル簡略化の方向が示される。

全目的発生交通量推定のための重回帰モデルとして、RSS基準によるステップワイズ法では、ステップ4の$y = 3.5471x_1 + 4.7067x_2 - 7.3353x_3 - 2.5456x_4 + 66.3514$が妥当なモデルとなるが、本研究では、予測誤差が最小値をとったステップ2において作成された、$y = 2.9841x_1 + 1.2204x_2 + 380.0488$というモデルが妥当であるという結論に至った。このモデルは、現象説明性を示すと考えられる指標である、RSS，R^2（寄与率）についても、ある程度の水準が保たれていると思われる。

5．おわりに

本研究では、説明変数の将来値算出の不確実性を重回帰モデルの設定に導入することを試み、発生交通量予測モデルに適用した。そして、説明変数の不確実性による誤差が、モデルが複雑になった段階では、全体の予測誤差に対して大きな割合を占めており、説明変数の不確実性の導入により、モデルの簡略化の方向が必要であることがわかった。このような予測誤差分析は、説明変数の選択方法として、有効な方法であることが確認できた。今後の課題としては、予測誤差を示す指標として、本研究で用いた$\sum_{\nu=1}^{n} V[\hat{y}_\nu^*]/n$というすべてのゾーンを等しい重みで総合化したもの以外に、重回帰モデルの適用目的に対応して、より合理的な指標をみつけ出す必要があると思われる。また、本研究では、説明変数の相対誤差を一律5%としたが、説明変数ごとに予測の難しさを配慮して、相対誤差の水準を定めていくことが必要である。

本論文の作成にあたって、御指導をいただいた北海道大学五十嵐日出夫教授に、感謝の意を表します。