多変量解析特論

客員教授　藤越　康祝

各個体についていくつかの変数の組が観測されているとき、そのデータを解析することを多変量解析という。多変量解析においては、解析の目的に応じて、多くの変数のなかから少数個の有用な合成変数を抽出することが重要になる。このための手法として、正準相関分析、主成分分析、判別分析、多変量線形部分回帰法、などを取り上げる。また、分割表データから各カテゴリーを少数（１～３）次元空間に表示するための対応分析法、個体間の類似性データから各個体を少数次元に配置するための多次元尺度法、多変量データをいくつかのグループに分割するためのクラスター分析等についても講義する。

１．重回帰分析の例：複数個の変数 $x_1, x_2, \cdots, x_p$ に基づいて、ひとつの変数 y を推測する式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

は、y の $x_1, x_2, \cdots, x_p$ に対する重回帰式とよばれる。ここで $x_1, x_2, \cdots, x_p$ を説明変数（独立変数）、y を目的変数（従属変数）という。目的変数 y は計測がむずかしく、計測の容易な $x_1, x_2, \cdots, x_p$ から y を推測する場合や、$x_1, x_2, \cdots, x_p$ の値から y を予測する場合に用いられる。全世帯実質消費支出 y を勤労者世帯実質収入 x_1、消費者物価指数 x_2 から推測する式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

である。データは１９７３年１期から１９８０年２期までの期間で、下記の表に示す。

年	期	全世帯実質消費支出（ｙ）	勤労者世帯実質収入（ｘ１）	消費者物価指数（ｘ２）
1	1	145000.27	180863.27	67.37
1	2	147645.55	217986.30	70.83
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
7	2	161216.46	246369.33	137.13

これらから重回帰式を計算すると

$$\hat{y} = 86573.6 + 0.24667 x_1 + 136.52 x_2$$

となる。１年１期の説明変数 x_{11} と x_{21} は、$x_{11} = 180863.27, \quad x_{21} = 67.37$ であるから、回帰による推測値は、

$$\hat{y} = 86573.6 + 0.24667 \times 180863.27 + 136.52 \times 67.37 = 140384.31$$

を得る。１年２期の値 x_{12} と x_{22} を回帰式に代入して $\hat{y}_2$ を、以下同様にして $\hat{y}_3, \hat{y}_4, \cdots, \hat{y}_{30}$ を得る。

２．成績のデータ：複数個の変数 $x_1, x_2, \cdots, x_p$ に基づいて、互いに従属的な関係にある多次元的な情報をなるべく損わないように、しかもできる限り少ない次元に（２次元あるいは３次元）要約する手法を主成分分析法という。いま、国語 x_1、社会 x_2、数学 x_3、理科 x_4、音楽 x_5、美術 x_6、体育 x_7、技家 x_8、英語 x_9 の９科目について、166 人の成績があるとする。２３番目の学生の成績は、国語 60 点、社会 85 点、数学 88 点、理科 82 点、音楽 59 点、美術 78 点、体育 85 点、技家 47 点、英語 83 点である。このような１６６人の成績から９科目の主成分分析を行うと、第一主成分

$$y_1 = 0.388 x_1 + 0.339 x_2 + 0.377 x_3 + 0.337 x_4 + 0.349 x_5 + 0.227 x_6 + 0.172 x_7 + 0.298 x_8 + 0.472 x_9$$

を得る。主成分分析では変数 x_i の係数が大きければ大きいほど、その変数の y_1 への貢献度は高いと解釈する。このことから、第一主成分 y_1 は総合点の因子と考えられる。これは、全体の 66%の情報を持った重要な因子である。第二主成分 y_2 も得られるが、y_2 の一次式から体育だけの因子であり、これは総合特性値とは考えられない。