

## 1.6 節 「データ分析」 補足資料

(2024 年 8 月 30 日版)

### A.1.6.1 回帰分析

#### (1) 重回帰分析における説明変数の影響度の比較

重回帰分析では、複数の説明変数を用いて目的変数の予測をおこなう。このとき、どの説明変数が目的変数に大きな影響を与えているか、または、どの説明変数が予測の役に立っていないかを知りたいという場面は多いだろう。ここでは、目的変数に対する説明変数の影響度を比較する方法について説明する。

まず、例として 47 都道府県における窃盗犯認知件数 (人口千人あたり) を人口 (千人)、1 人あたり県民所得 (千円)、緯度、経度を用いて予測するモデルをみてみよう<sup>4)</sup>。このモデルにおける重回帰分析の結果は次のとおりである。

$$\begin{aligned} & \text{窃盗認知件数 (人口千人あたり)} \\ &= -1.0 + 0.00023 \times \text{人口 (千人)} \\ &\quad - 0.00012 \times 1 \text{ 人あたり県民所得 (千円)} \\ &\quad - 0.24 \times \text{緯度} + 0.16 \times \text{経度} \end{aligned}$$

この結果より、人口が多く、所得の低い地域のほうが窃盗が多く、また、南東の地域のほうが窃盗が多い傾向があることが読み取れる。ただし、窃盗犯認知件数とこのモデルを用いた予測値の散布図は図 A.1.2 であり、それほど当てはまりが良いわけではない (決定係数は 0.47 である)。

ここで、このモデルにおいて各変数が予測にどの程度の影響を与えているかを考える。人口の係数は緯度や経度の係数に比べて 0 に近いが、予測の役に立

---

<sup>4)</sup> 令和 2 年国勢調査 (総務省)、令和 3 年警察白書 (警察庁)、2019 年度県民経済計算 (内閣府)。

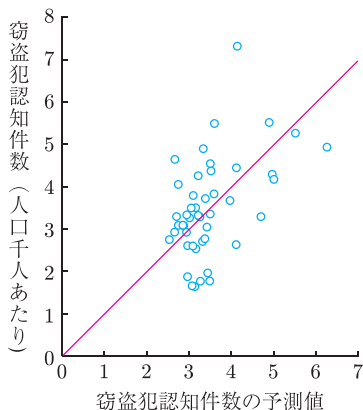


図 A.1.2 47 都道府県における窃盗犯認知件数（人口千人あたり）とその予測値

たないというわけではない。ここでは、人口の単位を千人としているが、もし十万人とすればどうなるだろうか。その場合、回帰係数を 100 倍の 0.023 とすることで、人口の単位が千人の場合とまったく同じ予測値が得られる。つまり、回帰係数の大きさは説明変数の大きさの影響を受けるので、スケールの異なる説明変数を用いている場合、その回帰係数の大きさの比較に意味はない。

説明変数の予測に対する影響を調べる際には、目的変数、説明変数ともに標準化（1.4 節を参照）をおこない、標準化したデータを用いて重回帰分析をおこなう必要がある。先ほどのデータについて、各変数を標準化して重回帰分析をおこなった結果が次のとおりである（各変数は標準化された値である）。

$$\begin{aligned} \text{窃盗認知件数} = & 0.56 \times \text{人口} - 0.051 \times 1 \text{ 人あたり県民所得} \\ & - 0.54 \times \text{緯度} + 0.51 \times \text{経度} \end{aligned}$$

このように、各変数を標準化した際に得られる回帰係数のことを**標準化偏回帰係数** (standardized partial regression coefficient) という。この結果から、人口は緯度や経度と同程度に予測に影響を与えるが、県民所得は予測への影響が小さいことがわかる。

### A.1.6.2 多変量データ解析

ここでは、ページの都合で本書では紹介できなかった多変量データ解析の手法のうち、**主成分分析** (principal component analysis) と**アソシエーション分析** (association analysis) について紹介する。主成分分析は多変量データのデータ間の関連を可視化する手法であり、アソシエーション分析は購買履歴データ (POS データ) について商品間の関連を調べる手法である。

#### (1) 主成分分析

クラスター分析ではデータをクラスターに分類できるが、多変量データは一般に (4 次元以上の場合) そのままでは可視化することができないので、適切に分類できているかどうかを把握することが難しい。そこで、多変量データをできるだけその相対的な位置関係を保ったまま散布図 (2 次元) で表現する方法について説明する。

多変量データを 2 次元で表す最も直観的な方法は、複数の変数のうちすべての 2 つのペアについて散布図を表示する方法であろう。すべての変数のペアの散布図を行列のように表示したものを**散布図行列** (scatter plot matrix) という。

図 A.1.3 の右のグラフは 4 つの点  $((1.1, 1.1, 1.1), (1, 0, 1), (0, 1, 1), (1, 1, 0))$  をそれぞれ赤、青、オレンジ、緑で 3 次元プロットしたものである。左の図はこのデータの散布図行列であるが、どの図を見ても同じような散布図となっており、4 つの点の関係を把握しづらい。これを別の方向からみるとどうなるだろうか。

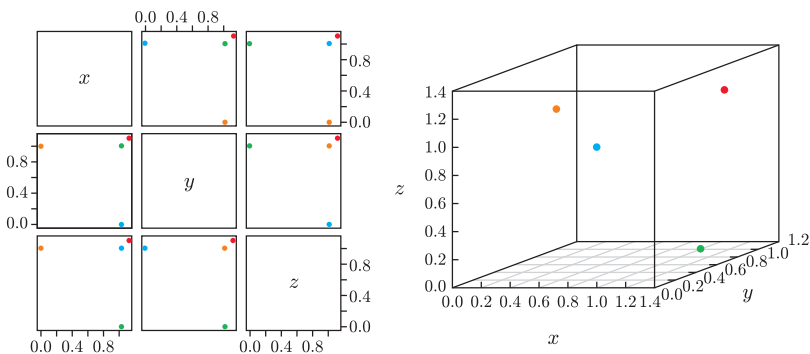


図 A.1.3 散布図行列

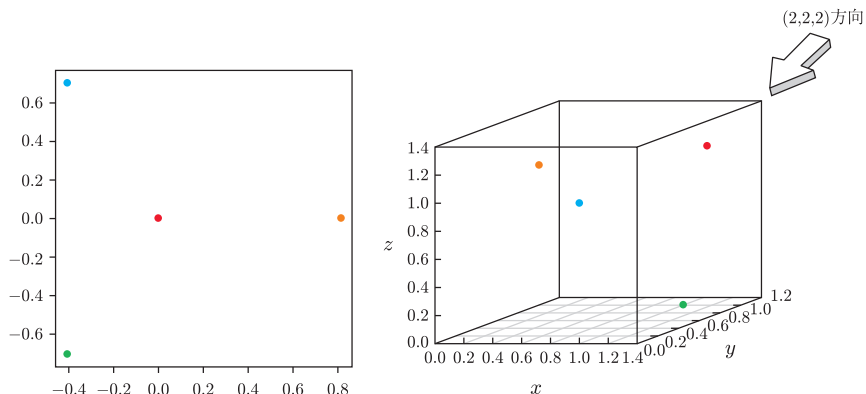


図 A.1.4 主成分分析の例

図 A.1.4 の右図のように、 $(2, 2, 2)$  方向から原点に向かってこれらの点を見てみると左図のようになり、4つの点の関係性を把握しやすくなっている。このように、元のデータの位置関係、関連性をできるだけ保持できる方向からデータを見たときの散布図を得る方法が主成分分析である<sup>5)</sup>。このとき、元のデータの情報をどの程度の割合で保持できているかを示す指標として**累積寄与率** (cumulative contribution ratio) がある。図 A.1.4 の左図の累積寄与率は 0.83 であり、元の 3 次元データの多くの情報を 2 次元で保持できていることがわかる。

## (2) アソシエーション分析

購買履歴データ (POS データ) を分析する場合、どの商品がよく売れているか、またはどの商品が同時購入されているか、などに興味があるだろう。このような情報を抽出する方法としてアソシエーション分析が使われる。アソシエーション分析では、ある商品を購入することが別の商品の購入にどのような影響を与えるかを調べる。商品 A と B の関係を調べるとき、商品 A が商品 B に与える影響を「 $A \Rightarrow B$ 」と表し、A を条件部、B を帰結部という。

アソシエーション分析では**支持度** (support)、**確信度** (confidence)、**リフト値** (lift) という指標が用いられる。A を条件部、B を帰結部としたときの支持度、

<sup>5)</sup> 厳密には 2 次元に限らず、データの情報をできるだけ保持したまま低次元にデータを圧縮する方法である。

確信度、リフト値は次のように定義されている。

- 支持度  
全購入者に対して、商品 A と B の両方を購入した人の割合
- 確信度  
商品 A を購入した人のうち、商品 B も購入している人の割合
- リフト値  
$$\frac{\text{商品 A を購入した人のうち、商品 B も購入している人の割合}}{\text{商品 B を購入している人の割合}}$$

リフト値は商品 A を購入することで、商品 B を購入する割合が何倍になるかを示す指標であり、アソシエーション分析で重要視される指標である。これらの指標を調べることで、商品間の関連を調べることが可能となる。

たとえば、「ピザ ⇒ コーラ」の支持度が 0.02、確信度が 0.3、リフト値が 10 であるとする。ピザとコーラの両方を同時に購入する人の割合が 0.02、ピザを購入した人だけで見たときにコーラを購入する人の割合が 0.3、この 0.3 という数字が、全体でコーラを購入する人の割合の 10 倍であることを意味している。さまざまな商品の関連を調べる際、確信度やリフト値について降順にソートすることで関連の強い商品を調べることができる。

### 課題学習

**A.1.6-1** 図 A.1.5 は 2020 年の「小売物価統計調査」<sup>6)</sup>に基づく 47 都道府県の食料費、住居費、光熱・水道費、家具・家事用品代、被服及び履物代、保険医療費、交通通信費、教育費、教養娯楽代、諸雑費(全国平均を 100 とした指数)のデータ(全国の都道府県庁所在市以外の 91 市において、スーパーを中心に代表的な店舗 500 店舗を対象として調査されたもの)を用い、階層型クラスタリング、および主成分分析をおこなった結果である(主成分分析の散布図は階層型クラスタリングの結果に基づき色分けしている)。この図からわかることについて検討せよ。

<sup>6)</sup> 2020 年「小売物価統計調査(構造編)結果」(総務省統計局)(<https://www.e-stat.go.jp/stat-search/file-download?statInfId=000032118470&fileKind=0>)

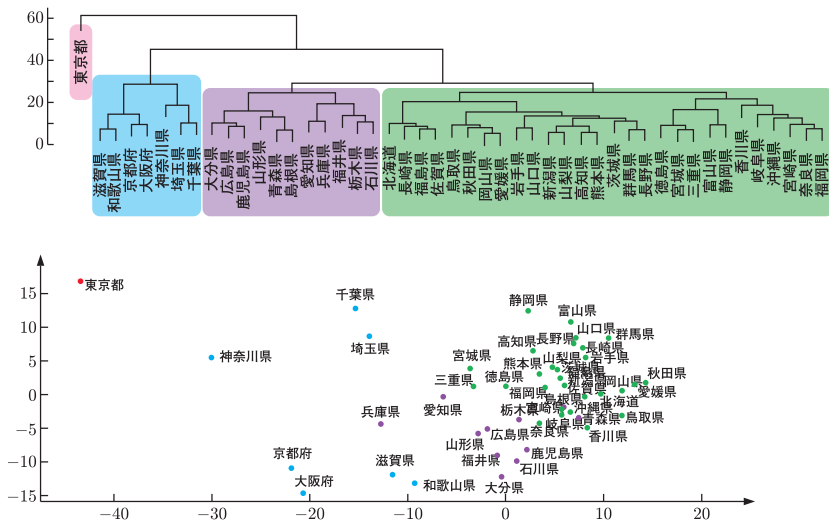


図 A.1.5 47 都道府県の各家庭の出費に対する階層型クラスターリングと主成分分析の結果