1

カテゴリカルデータ解析プログラムCATDAP の構造と機能強化案

http://hdl.handle.net/10787/3887

石黒真木夫@統計思考院@統計数理研究所

2016.10. 22

2018.1.5　分割表解析と混合正規分布解析の関係を論じた記述を追加

2

概要

CATDAP(CATegorical Data Analysis Program) は、カテゴリカルデータの分布に影響を与える要因を探るためのプログラムとして開発されたものである。説明変数の候補としてカテゴリカル変数、連続変数いずれも与えることが可能であることから非常に使いやすい強力なソフトになっている。

このプログラムに簡単な機能強化をほどこすことによって、

1. 目的変数が連続値である場合への適用
2. 欠測値を含むデータの解析

が可能となる。また、わずかな改変を加えることで、混合正規分布解析プログラムが得られる。

3

CATDAPの理論的基礎：分割表モデルとその評価

離散値 1,2, … , c をとる確率変数 I と、離散値 1,2, … , d をとる変数 J の関係を確率関数モデル $P(I|J)$ で表す。データ

$$\{n_{ij}|I = i.J = j \text{ となるデータの数}, i = 1,2,..,c, j = 1,2,\dots,d\}$$

にもとづくこのモデルのパラメータの最尤推定値は

$$\hat{P}(i|j) = \frac{n_{ij}}{\sum_{i=1}^{c} n_{ij}}$$

で与えられ、このモデルの AIC は

$$\mathrm{AIC}_{\mathrm{I|J}} = \sum_{j=1}^{d}\{-2\sum_{i=1}^{c} n_{ij}\log\hat{P}(i|j) + 2(c-1)\}$$

となる。I の分布が J の値に依存しないとするモデル $P(I)$ のパラメータの推定値とAIC は

$$\hat{P}(i) = \frac{\sum_{j=1}^{d} n_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{d} n_{ij}}$$

$$AIC_I = -2\sum_{i=1}^{c}\sum_{j=1}^{d} n_{ij}\,\hat{P}(i) + 2(c-1)$$

で与えられる。

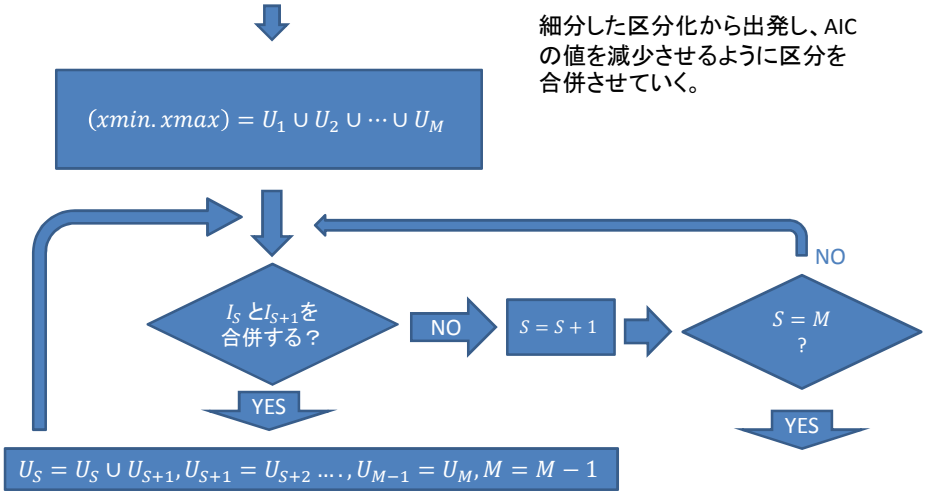
4

CATDAPを構成する技術要素

- 集計による分割表モデルの当てはめ
- AICの計算
- 変数増減法による説明変数の最適組み合わせ探索法
- 連続値説明変数の離散化

5

連続値変数離散化法



細分した区分化から出発し、AICの値を減少させるように区分を合併させていく。

6

連続値目的変数解析への適応

目的変数 y の確率密度関数を、一定区間巾 s の区間$I_1, , I_2, , \ldots, I_c$ で定義されたヒストグラムモデルで表現するモデルを採用する。このモデルのパラメータの最尤推定値は

$$\hat{H}(y|j) = \frac{\hat{P}(i|j)}{s} \qquad (y \in I_i \text{ の場合})$$

となり、AIC は

$$\mathrm{AIC_{I|J}} + 2\mathrm{n}\log \mathrm{s}$$

となる。ここで

$$n = \sum_{j=1}^{d}\sum_{i=1}^{c} n_{ij}$$

である。同様に y の確率密度関数が説明変数に依存しないとするモデルの場合は

$$\hat{H}(y) = \frac{\hat{P}(i)}{s} \qquad (y \in I_i \text{ の場合})$$

$$AIC_I + 2n\log s$$

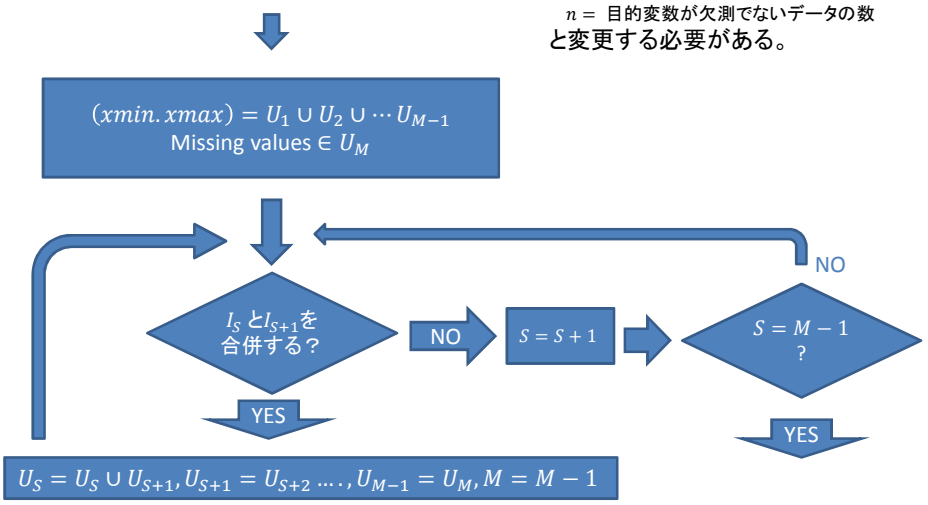
となる。

7

欠測値を含むデータへの適応

観測値がもともと離散変数である場合は、欠測は何らの問題も引き起こさない。
連続値観測が欠測を含む場合には、離散化手続きを以下のようにする。

さらに、前スライドの n の定義を
$n=$ 目的変数が欠測でないデータの数
と変更する必要がある。



8

欠測値を含むデータのスクリーニング

欠測値を含むデータを解析できるソフトにおいては、応答変数あるいは説明変数に欠測が含まれている場合、以下のような使い方が考えられる。

1. 全データを解析する。
    1. すべての変数を欠測と非欠測の2値変数に変換して解析する。
    2. 全データそのままで解析する
2. 応答変数に欠測がないデータのみを解析する。
    1. 説明変数を欠測と非欠測の2値変数に変換して解析する。
    2. 説明変数をそのままで解析する
3. 応答変数と説明変数ともに欠測がないデータのみを解析する。

「3」という使い方では。データの数が減ってしまう恐れがある。
「1.1」という使い方で, 欠測の間に何らかの関係があるか否か調べられる。
「2.1」という使い方で、欠測に偏りが有るか無いか調べられる。
「2.2」という使い方で、欠測の「影響」が調べられる。

前処理の段階でデータのスクリーニングすることによってこれらの使いわけが出来る。このスクリーニング機能を CATDAP に組み込むのは簡単である。

9

分割表解析と混合正規分布解析

	区分1	区分2	区分3
U_1	n_{11}	n_{12}	n_{13}
U_2	n_{21}	n_{22}	n_{23}
⋮	⋮	⋮	⋮
U_M	n_{M1}	n_{M2}	n_{M3}
合計	n_1	n_2	n_3

データ全体において、

$n_j =$ 値 j をとったデータの数

とする。データを互いに重なることのない M 個の部分集合に分割し、

$n_{ij} = U_i$ において値 j をとったデータの数

として、n_j を

$$n_j = \sum_i n_{ij}$$

の形に表現することの可否を問うのが分割表解析である。

	0 次和	1 次和	2 次和
U_1	s_{10}	s_{11}	s_{12}
U_2	s_{20}	s_{21}	s_{22}
⋮	⋮	⋮	⋮
U_M	s_{M0}	s_{M1}	s_{M2}
合計	s_0	s_1	s_2

データ $\{x_1, x_2, \dots, x_n\}$ が与えられたとき、

$$s_j = \sum_{i=1}^{n} x_i^j$$

とする。データを互いに重なることのない M 個の部分集合に分割し、

$$s_{ij} = \sum_{x_i \in U_i} x_i^j$$

として、s_j を

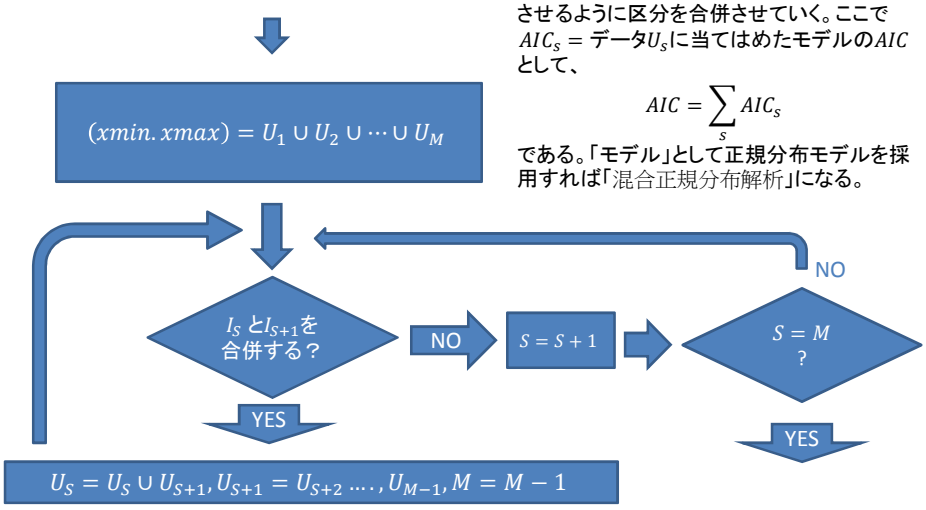
$$s_j = \sum_i s_{ij}$$

の形に表現することの可否を問うのが混合正規分布解析である。

この類似は CATDAP をほんの少し改変することによって混合正規分布解析機能を備えたプログラムが得られることを意味している。

10

混合正規分布解析



細分した区分化から出発し、AIC の値を減少させるように区分を合併させていく。ここで

$AIC_s =$ データU_sに当てはめたモデルのAIC

として、

$$AIC = \sum_s AIC_s$$

である。「モデル」として正規分布モデルを採用すれば「混合正規分布解析」になる。

11

感謝

CATDAP 製作者坂元慶行博士と桂康一氏に感謝。
問題意識を提供して下った小前和智氏に感謝。
ヒストグラムモデル（＝確率密度関数モデル）と確率関数モデルの関係に気付かせて下さった中村隆博士に感謝。

12

参考文献

- 坂元・石黒・北川（1983）.情報量統計学、共立出版
- Katsura,K. & Sakamoto,Y.(1980). CATDAP, A categorical data analysis program package, Computer Science Monographs, No.14, The Institute of Statistical Mathematics,Tokyo.
- 坂元慶行（1985）.カテゴリカルデータのモデル分析.共立出版.
- 情報量統計学的データ可視化ツール　An AIC-based Tool for Data Visualization　http://hdl.handle.net/10787/3614