CATDAP 機能強化プラン
proposal of A super CATDAP

http://hdl.handle.net/10787/00034178

2021.1.17

石黒真木夫@統計数理研究所名誉教授

0. 概要

CATDAP（CATegorical Data Analysis Program) は、柱と坂元によって開発された、カテゴリカルデータ解析プログラムである (Katsura,K. and Sakamoto,Y.(1980);CATDAP, A categorical data analysis program; Computer Science Monographs, 14, The Institute of Statistical Mathematics, Tokyo)。

離散目的変数の分布の変動、離散・連続値変数あるいはその組み合わせで説明する、極めて使いやすいプログラムであった。

このプログラムは、中野と嵯峨によって、 R の関数としてパッケージ化され、広く使われるものとなっている。

SuperCATDAP は、CATDAP に以下の機能強化を施そうとするものである。

1. 他のソフトによる解析結果と比較可能な形の AIC を出力するようにする。
2. 連続値目的変数をとれるようにする。
3. 欠測値を許容するようにする。
4. 方向データ（circular data) を扱えるようにする。
5. グラフィカルな結果表示が出力されるようにする。

1. 分割表解析一般論

N 個の M 次元データ $\{(v_{i1}, v_{i2}, \dots, v_{iM}) \mid i = 1,2, \dots, N\}$ が得られたものとする。

ここで v_{ik} は $1 \le v_{ik} \le C_k$ の範囲の整数値をとるものとする$(i = 1,2, \dots, N,\ \ k = 1,2, \dots, M)$。

v_{i1} $(i = 1,2, \dots, N)$ が確率

このとき、j_k を $1 \le j_k \le C_k$ の範囲の整数としてM 変数関数

$$N_{1,2,\dots,M}(j_1, j_2, \dots, j_M) = \text{「}v_{i1}, = j_1\ v_{i2} = j_2, \dots, v_{iM} = j_M \text{ となるデータの数」}$$

を定義する。データをこの形に集約したものを分割表という。

v_{i1} $(i = 1,2, \dots, N)$ が値 j_1 を確率 $P_{1|2,3,\dots M}(j_1 | j_2, \dots, j_M)$ でとる確率変数の独立な実現値なら、$P_{1|2,3,\dots,M}(j_1 | j_2, \dots, j_M)$ を

$$\hat{P}_{1|2,3,\dots,M}(j_1 | j_2, \dots, j_M) = \frac{N_{1,2,3,\dots,M}(j_1, j_2, \dots, j_M)}{N_{2,3,\dots,M}(j_2, \dots, j_M)}$$

で推定できる。ここで

$$N_{2,3,\dots,M}(j_2, \dots, j_M) = \text{「}v_{i2} = j_2, \dots, v_{iM} = j_M \text{ となるデータの数」}$$

である。一般に

$$N_{k_1 k_2 \dots k_m}(j_{k_1}, j_{k_2} \dots, j_{k_m}) = \text{「}v_{ik_1} = j_{k_1}, \dots, v_{ik_m} = j_{k_m} \text{ となるデータの数」}$$

という表記法を使うことにすると、

$$\hat{P}_{1|k_1 k_2 \dots k_m}(j_1 | j_{k_1}, j_{k_2} \dots, j_{k_m}) = \frac{N_{1k_1 k_2 \dots k_m}(j_1, j_{k_1}, j_{k_2} \dots, j_{k_m})}{N_{k_1 k_2 \dots k_m}(j_{k_1}, j_{k_2} \dots, j_{k_m})}$$

が定義される。

2. オリジナル CATDAP

この場合、v_{i1} $(i = 1,2, \dots, N)$ の分布を説明するのに、たとえば、$\hat{P}_{1|2\dots M}(j_1 | j_2, \dots, j_M)$ と $\hat{P}_{1|34}(j_1 | j_3, j_4)$ のどちらがより適しているか、という問題が生ずる。 $\hat{P}_{1|23\dots M}(j_1 | j_2, \dots, j_M)$ の方が$\hat{P}_{1|34}(j_1 | j_3, j_4)$ より偏りが少ない推定を与えるが、はるかに多いパラメータを含み、それらの推定誤差の影響が大きいので、どこで妥協すべきかという問題が生ずるのである。

CATDAP は赤池情報量 AIC でモデルを評価する。モデル $\hat{P}_{1|k_1 k_2 \dots k_m}(j_1 | j_{k_1}, j_{k_2} \dots, j_{k_m})$ の AIC は

$$\begin{aligned}\mathrm{AIC}_{1|k_1 k_2 \dots k_m} &= -2 \sum_{j_1=1}^{C_1} \sum_{j_{k_1}=1}^{C_{k_1}} \sum_{j_{k_2}=1}^{C_{k_2}} \cdots \sum_{j_{k_m}=1}^{C_{k_m}} N_{1k_1 k_2 \dots k_m}(j_1, j_{k_1}, j_{k_2} \dots, j_{k_m}) \log \hat{P}_{1|k_1 k_2 \dots k_m}(j_1 | j_{k_1}, j_{k_2} \dots, j_{k_m}) \\ &+ 2(C_1 - 1) \prod_{j=1}^{m} C_{k_j}\end{aligned}$$

で計算される。説明変数なしのモデルの AIC は

$$\mathrm{AIC}_1 = -2 \sum_{j_1=1}^{C_1} N_1(j_1) \log \hat{P}_1(j_1) + 2(C_1 - 1)$$

もし、

$$\mathrm{AIC}_{1|k_1 k_2 \dots k_m} > \mathrm{AIC}_1$$

なら、$\mathrm{v}_{ik_1}, \mathrm{v}_{\mathrm{ik}_2}, \dots, \mathrm{v}_{\mathrm{ik}_\mathrm{m}}$ の値を知っても v_{i1} の値に関する情報は得られないと考えるべきである。そう考えて オリジナルCATDAP ではモデル $\hat{P}_{1|k_1 k_2 \dots k_m}(j_1 | j_{k_1}, j_{k_2} \dots, j_{k_m})$ を評価する数値として

$$\mathrm{AIC}_{1|k_1 k_2 \dots k_m} - \mathrm{AIC}_1$$

が出力されるようになっている。

3. オリジナルCATDAP での連続値の扱い

オリジナル CATDAP はデータ $\{(v_{i1}, v_{i2}, \ldots, v_{iM}) \mid i = 1,2, \ldots, N\}$ の v_{i1} が離散値であるとして作られている。しかし、$v_{i2}, \ldots, v_{iM}$ は離散値と連続値が混在したものであって構わない。説明のため、たとえば、2番目の変数が実数値 r_{i2} で $\{(v_{i1}, r_{i2}, \ldots, v_{iM}) \mid i = 1,2, \ldots, N\}$ という場合を想定しよう。この場合、適当な単調増大実数列 $\{b_i \mid i = 0,2, \ldots C_2\}$ を用意して

$$v_{i2} = k \qquad (if \;\; b_{k-1} < r_{i2} \leq b_k)$$

で、離散化することによって分割表解析の形に持ち込める。オリジナル CATDAP のモデル探索には、詳細は避けるが、$\{b_i \mid i = 0,2, \ldots C_2\}$ の選択も含まれている。v_{i1} 以外の他の説明変数候補も同様。

この機能によって、CATDAP は判別分析を目的とする使い方ができることになる。

- 坂元・石黒・北川(1983). 情報量統計学. 共立出版.
- 坂元 慶行(1985). カテゴリカルデータのモデル分析、共立出版.
- 坂元(2001). 質的データのデータマイニング|最適なクロス表の自動探索 CATDAP(1)(2),ESTRELA,No.91(pp.82-85) No.92(pp.84-87).1

4. AsuperCATDAP に付加された機能

1. AIC 出力

   モデル $P\hat{}\_(1|k\_1\ k\_2 \ldots k\_m)\ (\ j\_1\ |j\_(k\_1\ ), j\_(k\_2\ ) \ldots, j\_(k\_m\ )\ )$ を評価する数値として
   AIC_(1|k_1 k_2…k_m)−AIC_1
   だけでなく、AIC_1 も出力する。

2. 連続値目的変数
   連続値変数を目的変数として選ぶことを許す。その場合、適当な単調増大実数列$\{b_i\ \mid i$=0,2,…$C\_2\ \}$ を用意して

   $$v_{i1} = k \qquad (if \;\; b_{k-1} < r_{i2} \leq b_k)$$

   で、離散化することになる。この離散化は天下りで固定化し最適化の対象としない。

3. 欠測値の扱い
   実数値データの離散化機能を利用して「極端に大きなr_i2」で欠測を表現して、解析対象に含める。

4. 方向データの扱い
   変数が方向データ r_{i2} で周期性をもつ場合、適当な単調増大実数列$\{b_i \mid i = 0,2, \ldots C_2\}$ を用意して

   $$v_{i2} = \begin{cases} k & (if \;\; b_{k-1} < r_{i2} \leq b_k) \\ 1 & (if \;\; b_k < r_{i2}) \end{cases}$$

   と、離散化することによって分割表解析の形に持ち込む。

5. 条件付き確率のグラフィカル出力

5. FORTRAN版 CATDAP

オリジナルCATDAP は統計数理研究所発行の Computer Science Monographs の一冊として公開されている(Katsura,K. and Sakamoto,Y.(1980);CATDAP, A categorical data analysis program; Computer Science Monographs, 14, The Institute of Statistical Mathematics, Tokyo)。

本稿で紹介した機能強化を実装した版が AsuperCATDAP.F である。コードは GFORTRAN であり、

gfortran -ffixed-line-length-72 AsuperCATDAP.F

でコンパイルできる。データの与え方に関しては

CATDAP manual http://hdl.handle.net/10787/3821

を参照されたい。

6. R Package 'catdap'

https://CRAN.R-project.org/package=catdap

Package 'catdap'
March 12, 2020

Version1.3.5
TitleCategorical Data Analysis Program Package
AuthorThe Institute of Statistical Mathematics
Maintainer: Masami Saga <msaga at mtb.biglobe.ne.jp>
Depends R (>= 3.2.0)
Suggests utils, datasets, methods
Imports graphics, grDevices
Description Categorical data analysis by AIC. The methodology is described in
Sakamoto (1992) <ISBN 978-0-7923-1429-5>.
License GPL (>= 2)
MailingList Please send bug reports to ismrp@jasp.ism.ac.jp.
NeedsCompilation yes
Repository CRAN
Date/Publication 2020-03-12 11:00:02 UTC

このパッケージはAsuperCATDAP に付加された機能のうち、「 1.AIC_1 出力」,「2.連続値目的変数の扱い」「3.欠測値の扱い」が既に実装されている。