

情報量統計学的データ可視化ツール

石黒真木夫@統計思考院@統計数理研究所

2015. 2. 15

科学研究費によって製作したソフトウェア

モデル選択法による統計的推論へのデータ前処理組み込みに関する研究
平成23年度 ~ 平成26年度

挑戦的萌芽研究 課題番号 23650148

研究分担者

清水 悟	東京女子医科大学・医学部
種村 正美	統計数理研究所・名誉教授
三分一 史和	統計数理研究所・モデリング研究系

内容

現象解析の初期段階において多変数間の関係を見る道具として一般に広く使われているものとして相関係数行列がある。しかし、それには次の2つの大きな欠点がある。

- (1) カテゴリ変数が扱えない
- (2) 非線形関数関係が扱えない

本ソフトは、変数間に成り立つモデルの赤池情報量基準 (AIC) を考えることにより

- (1)' カテゴリ変数が扱える
- (2)' 非線形関数関係が扱える

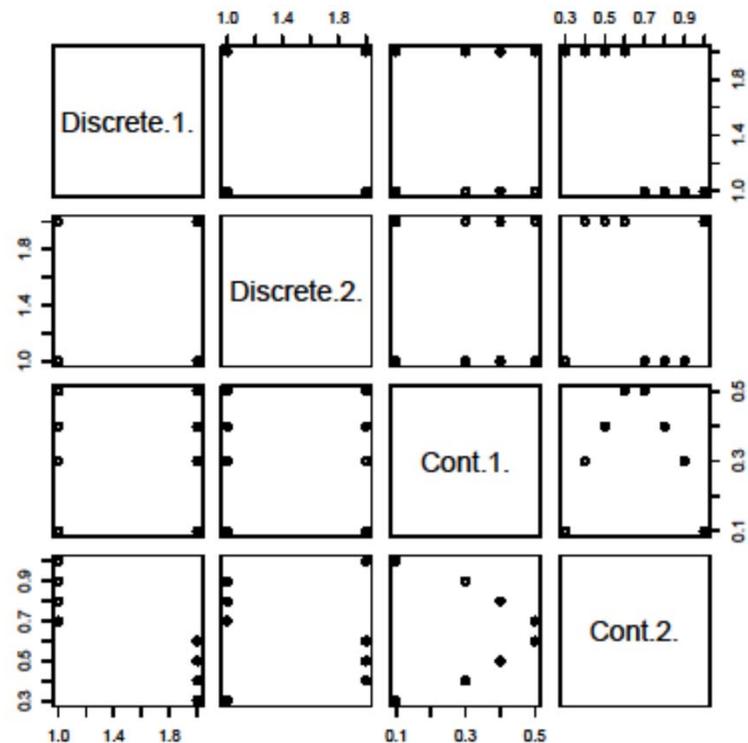
ようにした汎用ソフトである。

ファイル

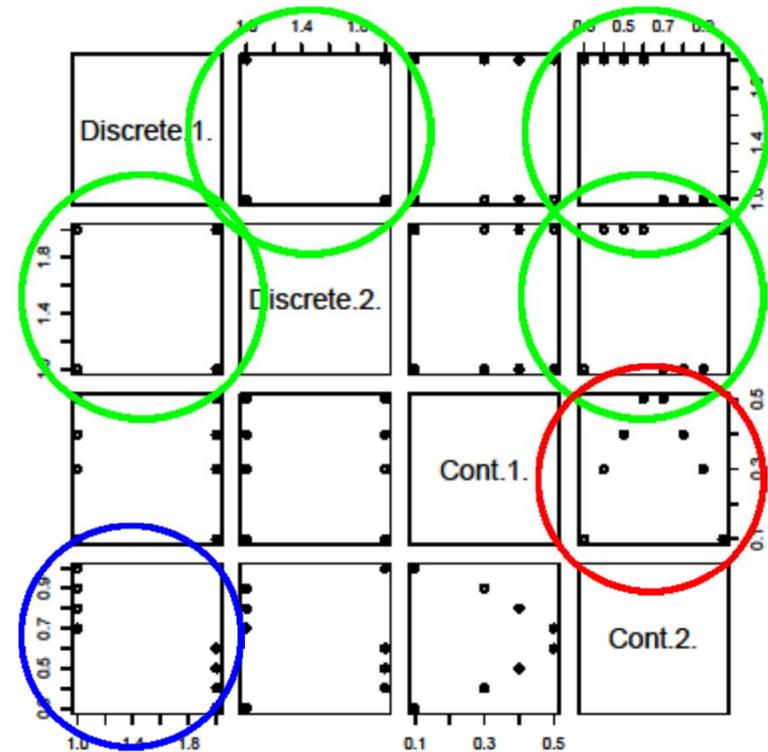
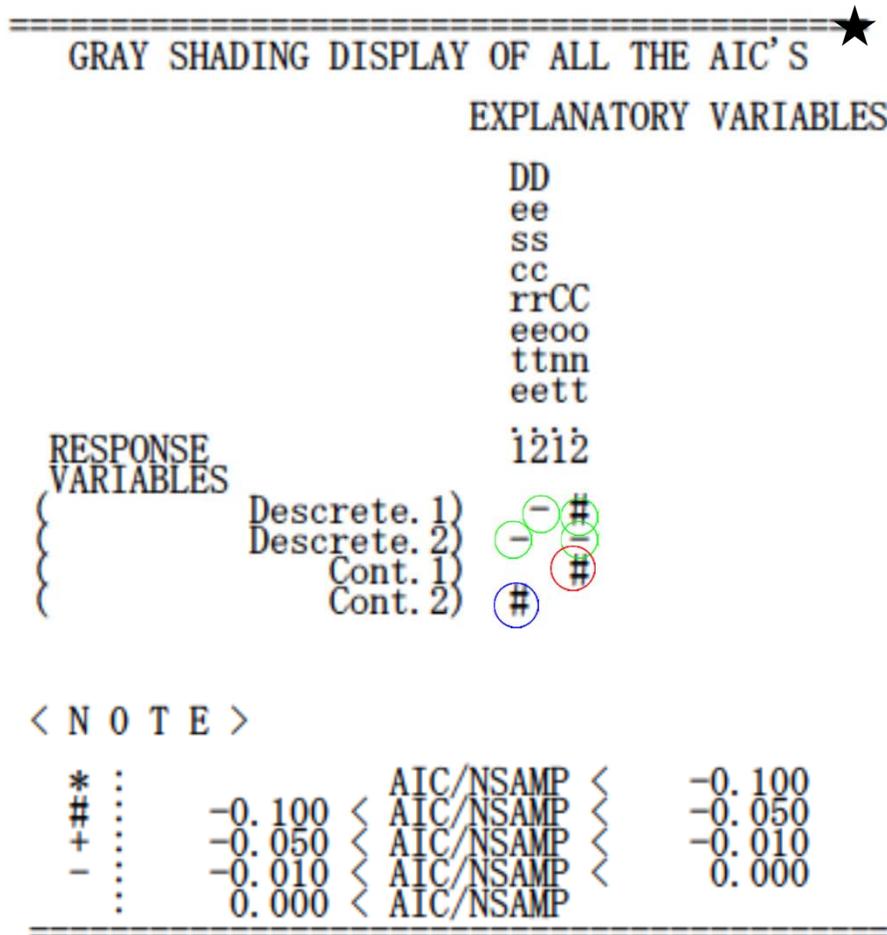
1. Visualize.infoStat_ver_1.2.R
2. AIC-based_Tool4DataVisualization.pdf
3. サンプルデータフォルダ
4. 出力結果見本フォルダ

離散連続混在データの散布図行列

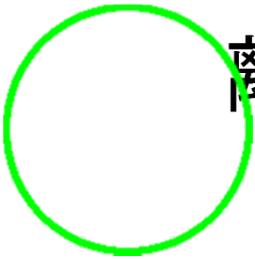
Discrete.1'b	Discrete.2'	Cont.1'	Cont.2'
1	2	0.1	1
1	1	0.3	0.9
1	1	0.4	0.8
1	1	0.5	0.7
2	2	0.5	0.6
2	2	0.4	0.5
2	2	0.3	0.4
2	1	0.1	0.3



散布行列要素の評価



★ CATDAP01(坂元他、1983)の図にならったものであるが、散布図行列の配置に合わせて目的変数と説明変数の配置を逆にしている。



離散目的変数・離散説明変数の場合

$MODEL0: P(I)$

$MODEL1: P(I | J)$

$AIC0 = MODEL0$ の AIC

$AIC1 = MODEL1$ の AIC

$AIC1 < AIC0$ なら変数 J は説明変数として有効
→ 変数 J と I の間に関連ありと考える。

連続目的変数・連続説明変数の場合

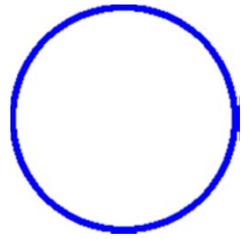
$$MODEL0: y \sim N(\mu, \sigma^2)$$

$$MODEL1: y \sim N\left(\mu + \sum_{m=1}^M x^m, \sigma^2\right)$$

$AIC0 = MODEL0$ の AIC

$AIC1 = MODEL1$ の AIC

$AIC1 < AIC0$ なら変数 x は説明変数として有効
→ 変数 x と y の間に関連ありと考える。



連続目的変数・離散説明変数の場合

$$MODEL0: y \sim N(\mu, \sigma^2)$$

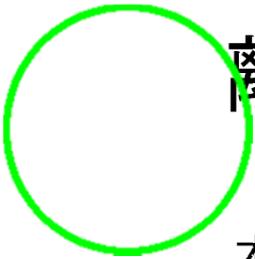
$$MODEL1: y \sim N(\mu_J, \sigma^2)$$

$AIC0 = MODEL0$ の AIC

$AIC1 = MODEL1$ の AIC

$AIC1 < AIC0$ なら変数 J は説明変数として有効

→ 変数 J と y の間に関連ありと考える。



離散目的変数・連続説明変数の場合

本来、logistic モデルのような「判別モデル」を利用すべきであるが、CA TDAP2の「標準装備」である、実数値データ離散化機能を利用して。。。

$MODEL0: P(I)$

$MODEL1: P(I | J(x))$

$AIC0 = MODEL0$ の AIC

$AIC1 = MODEL1$ の AIC

$AIC1 < AIC0$ なら変数 x は説明変数として有効

→ 変数 x と I の間に関連ありと考える。

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.