mwp-116

判別分析概要　【 評価版 】

Stata における判別分析に関する機能は `discrim` 系の各種サブコマンドによって提供されます。個々のサブコマンドの機能、用法については別の whitepaper に譲るとして、本 whitepaper では判別分析全般に関する概念や用例について説明を行います。

1. 判別分析

判別分析 (discriminant analysis) という手法はグループ間の違いを記述し、それを用いて帰属のわからない観測データをグループに割り振る（分類する）機能を提供します。この技術は医療診断やマーケットリサーチ等、幅広い分野への適用が考えられます。

クラスタ分析（`[MV]` **cluster** (*mwp-110*) 参照）と似た機能を持つものと言えますが、グループへの帰属関係が既知のデータが利用できない場合にはクラスタ分析を用いることになります。これに対し、グループへの帰属関係が既知のデータが存在する状態で未分類のデータが与えられたときに、判別分析は利用されます。この場合、グループとの対応が既知のデータを用いてグループ間の違いがモデル化され、それを用いて帰属関係が未知のデータに対する分類が行われます。この前段を記述的判別分析 (descriptive discriminant analysis)、後段を予測的判別分析 (predictive discriminant analysis) と呼ぶことがあります。

Stata では次のような判別分析手法がサポートされています。

コマンド	機能
`discrim knn`	k 近傍法判別分析
`discrim lda`	線形判別分析
`discrim logistic`	ロジスティック判別分析
`discrim qda`	2 次判別分析

2. 判別分析の用例

▷ Example 1

[MV] **discrim** の Example 1 には Example データセット `lawnmower2.dta` を用いた用例が紹介されています。

```
. use http://www.stata-press.com/data/r16/lawnmower2.dta [*1]
(Johnson and Wichern (2007) Table 11.1)
```

このデータセット中にはある都市で標本抽出された乗用芝刈り機 (riding-lawnmower) 所有者 12 人と非所有者 12 人に関するデータが記録されています。次に示すのは 24 件の中から抽出した 8 件の観測データです。

	owner	income	lotsize
1.	owner	90.0	18.4
2.	owner	115.5	16.8
3.	owner	94.8	21.6
4.	owner	91.5	20.8
13.	nonowner	105.0	19.6
14.	nonowner	82.8	20.8
15.	nonowner	94.8	17.2
16.	nonowner	73.2	20.4

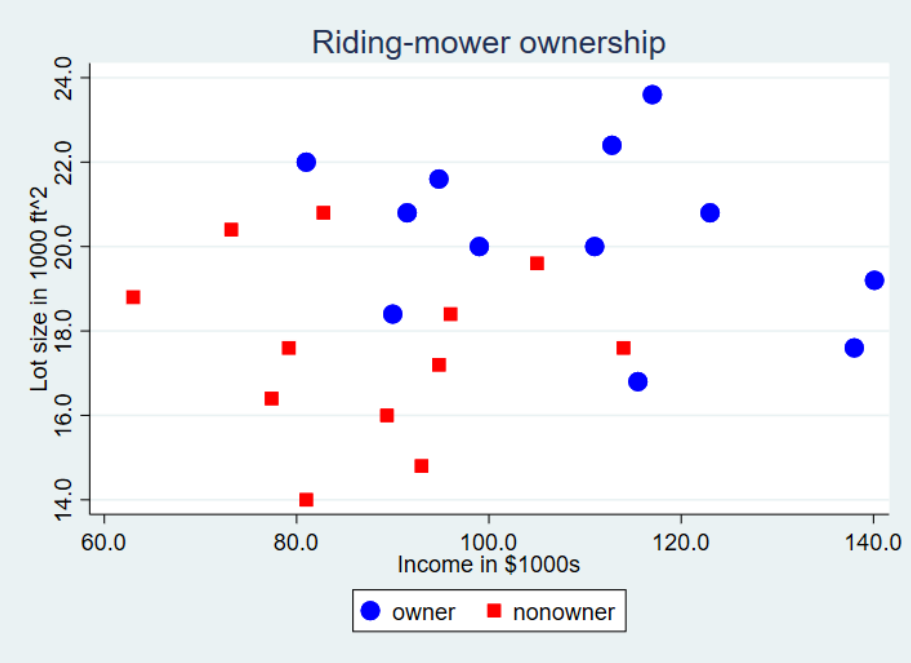
`owner` という変数は指標変数（ダミー変数）であり、乗用芝刈り機所有者の場合に 1、非所有者の場合に 0 という値を取ります [*2]。一方、変数 `income` は年収を$1,000 単位で、`lotsize` は敷地面積を 1,000 平方フィート単位で表現したものです。芝刈り機のメーカとしては `income` と `lotsize` が所有者/非所有者を区分する上で有力な変数であるかどうかを知りたいわけです。もしそれが確認できればマーケティング活動の対象を絞り込むことが可能になります。

[*1] メニュー操作： File ▷ Example Datasets ▷ Stata 16 manual datasets と操作、Multivariate Statistics Reference Manual [MV] の discrim の項よりダウンロードする。

[*2] 文字列変数のように見えるのは値ラベルが設定されていることによるものです。

次に示すのは芝刈り機所有者と非所有者に区分した形での散布図です。

```
. twoway (scatter lotsize income if owner == 1, mcolor(blue) msize(large)
> msymbol(circle)) (scatter lotsize income if owner == 0, mcolor(red)
> msymbol(square)), title("Riding-mower ownership")
> legend(order(1 "owner" 2 "nonowner"))
```
[*3]



このグラフからすると所有者と非所有者の分離はある程度できていますが、重複もあり、それほど明確とは言えません。予測的判別分析の機能を使用すると、この区分の能力を定量的に示すことができます。ここでは線形判別分析、すなわち `discrim lda` コマンドを実行することにします。

- Statistics ▷ Multivariate analysis ▷ Discriminant analysis ▷ Linear (LDA) と操作
- Model タブ: Type of discriminant analysis: Linear
  Variables: `lotsize income`
  Group variable: `owner`

[*3] メニュー操作： Graphics ▷ Twoway graph (scatter, line, etc.)

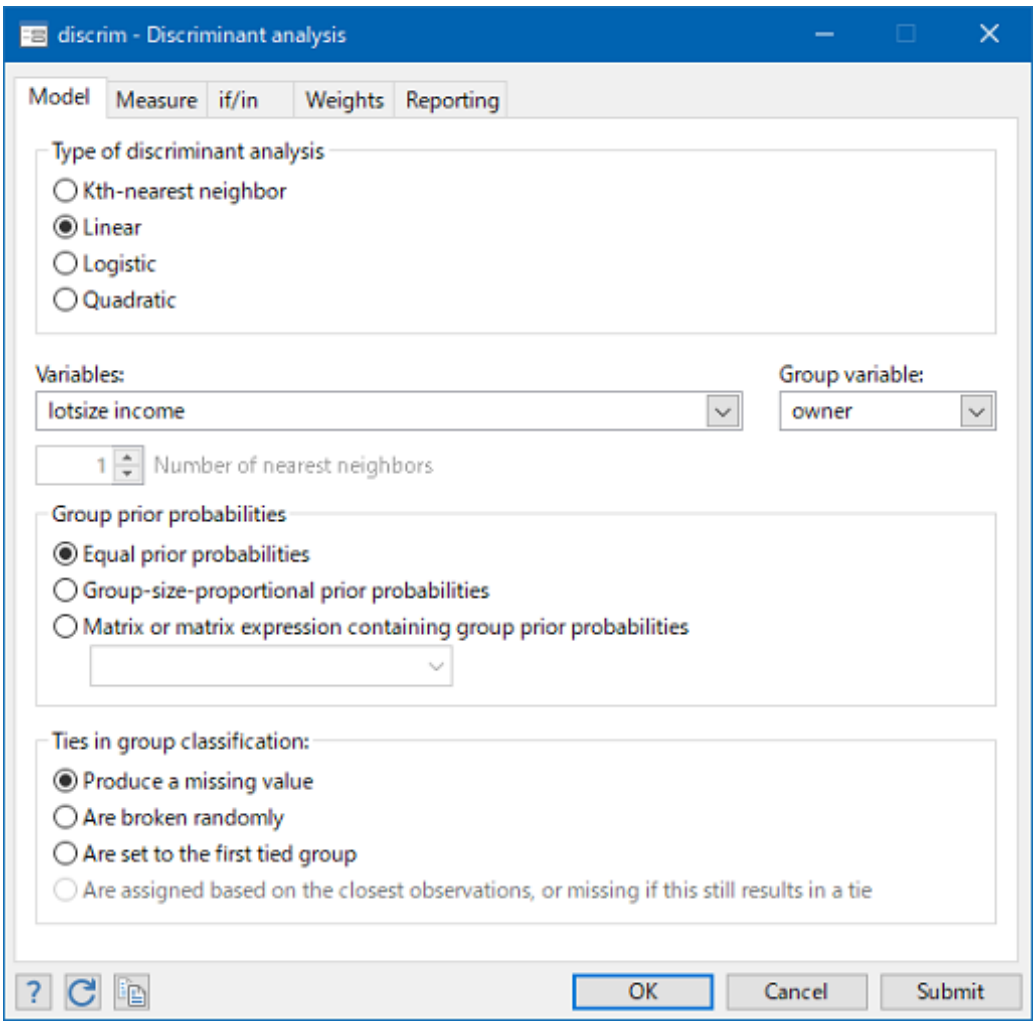


図 1 `discrim lda` ダイアログ - Model タブ

```
. discrim lda lotsize income, group(owner)

Linear discriminant analysis
Resubstitution classification summary

        ------------
        | Key      |
        |----------|
        | Number   |
        | Percent  |
        ------------
```

True owner	Classified nonowner	owner	Total
nonowner	10	2	12
	83.33	16.67	100.00
owner	1	11	12
	8.33	91.67	100.00
Total	11	13	24
	45.83	54.17	100.00
Priors	0.5000	0.5000	

`discrim` 系のコマンドから出力されるテーブルは分類表 (classification table) あるいは混同行列 (confusion matrix) と呼ばれます。ヘッダ部に再代入 (resubstitution) という言葉が表示されていますが、それは判別モデルの推定に用いられたのと同じ観測データを使って分類が行われているからです。表の対角要素は正しく分類されたデータの件数とパーセンテージを、非対角要素は正しく分類できなかったデータの件数とパーセンテージを示しています。この例では 1 人の所有者と 2 人の非所有者が誤って分類されたことになります。

評価版では割愛しています。

▷ Example 2

評価版では割愛しています。

補足１ – 2 次判別分析とロジスティック判別分析

評価版では割愛しています。

■