*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 |

© Copyright Math 工房； 一部 © Copyright StataCorp LP (used with permission)

## 2. 判別分析の用例

### ▷ Example 1

評価版では割愛しています。

### ▷ Example 2

今度は年収 `income` と敷地面積 `lotsize` が芝刈り機の所有/非所有にどのような形で関わりを持つかに着目して分析を進めることにします。記述的判別分析に関連したツールを用いることによって、グループがどのように分離しているかを調べることができます。線形判別分析 (LDA) の場合、ベースとなるのは Fisher の線形判別関数 (linear discriminant functions) ですが、これについては **[MV] discrim lda** (*mwp-117*) と **[MV] discrim lda postestimation** (*mwp-118*) をご参照ください。Postestimation コマンドである `estat loadings` を用いると判別関数の係数値（負荷量 (loadings) とも呼ばれる）を確認することができます。

- Statistics ▷ Postestimation ▷ Discriminant analysis reports and graphs
  ▷ Canonical discriminant-function coefficients ▷ Launch と操作
- `estat` ダイアログ: Main タブ:
  Canonical coefficients (loadings)
  Display pooled within-group standardized canonical discriminant function coefficients: ✓
  Display unstandardized canonical discriminant function coefficients: ✓

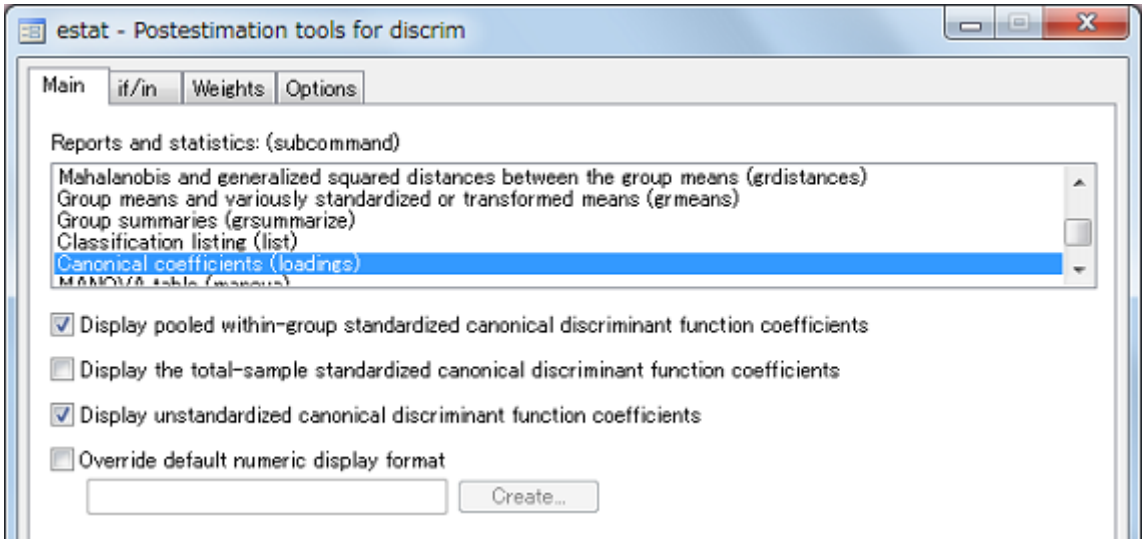



図 5 `estat loadings` ダイアログ

```
. estat loadings, standardized unstandardized

Canonical discriminant function coefficients

                 | function1
-----------------+-----------
         lotsize |  .3795228
          income |  .0484468
           _cons | -11.96094

Standardized canonical discriminant function coefficients

                 | function1
-----------------+-----------
         lotsize |  .7845512
          income |  .8058419
```

この操作では標準化した係数値と標準化していない係数値の双方を出力させています。後者は標準化していない変数に対する係数値です。これに対し前者はプール化された群内共分散 (pooled within-group covariance) を用いて標準化された変数に対する係数値で、判別関数に対する相対的な貢献度を評価することができます。

非標準化係数は芝刈り機の所有/非所有を区分する直線を規定するもので、この例では

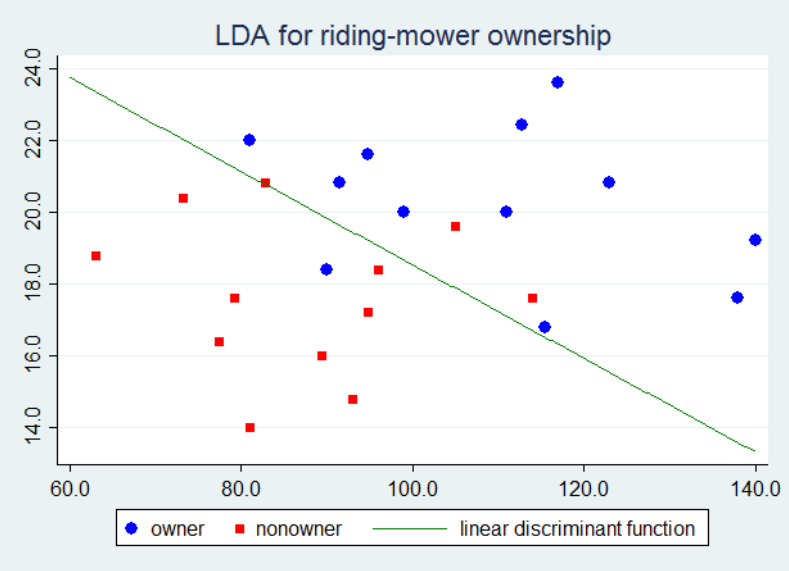
$$0 = 0.38 \cdot \texttt{lotsize} + 0.048 \cdot \texttt{income} - 11.96$$

すなわち

$$\texttt{lotsize} = -0.13 \cdot \texttt{income} + 31.52$$

で与えられることになります。これを散布図に重ねた形でプロットすると次のようになります。

```
. twoway (scatter lotsize income if owner == 1, mcolor(blue) msize(large)
> msymbol(circle)) (scatter lotsize income if owner == 0, mcolor(red)
> msymbol(square))
> (function -0.13*x + 31.52, range(60 140) lcolor(green) lwidth(medium)),
> title(LDA for riding-mower ownership) legend(order(1 "owner" 2 "nonowner"
> 3 "linear discriminant function") rows(1))
```




記述的判別分析に関連したツール全般については [MV] **discrim lda postestimation** (*mwp-118*) をご参照ください。 ◁

## 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。

■