Statistics 215a – 10/1/03 – D. R. Brillinger

Two-way arrays.

two-way table
two-way layout
two-factor array
contingency tables

New data type/structure (for course)

Rectangular display

rows, columns, responses

different factors/classifications
vary separately

response for each combination of levels of the factors

y_{ij} $i=1,\ldots,I;\ j=1,\ldots,J$

$n = I \times J$ observations (cells)

y numerical

Factors may be labels, ordered, numerical

Interested in relation between response and rows and columns

wish summary highlighting relation between response and each factor

Example – area burned in wildfires by month and year

Question – prediction?

the data

row is month, column is year (92-02)

$I = 12, J = 10$

(months have differing numbers of days)

boxplots for rows, columns

Conceptualization.

Response

$$\approx \text{summary} + \text{row effect} + \text{column effect}$$

$$y_{ij} \approx \mu + \alpha_i + \beta_j$$

Separate contribution for each factor

Additive dependence

(May need to transform. Later)

Old β is now $\theta = (\mu, \alpha_i, \beta_j)$

Paradigm.

data = fit + residual

Fitting.

OLS

$$\min_\theta \sum_{i,j} (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

overparametrized

side conditions

$$\sum_i \alpha_i \, , \, \sum_j \beta_j = 0$$

normal equations

$m = \bar{y}$, $a_i = (\bar{y}_{i.} - \bar{y})$, $b_j = (\bar{y}_{.j} - \bar{y})$

ANOVA identity

$\sum_i \sum_j y_{ij}^2$

$= \sum_i \sum_j (\bar{y})^2 + \sum_i \sum_j (\bar{y}_{i.} - \bar{y})^2 + \sum_i \sum_j (\bar{y}_{.j} - \bar{y})^2$

from orthogonality relations

ANOVA TABLE

Source	SS	DF
mean	$\sum_i \sum_j (\bar{y})^2$	1
rows	$\sum_i \sum_j (\bar{y}_{i.} - \bar{y})^2$	(I-1)
columns	$\sum_i \sum_j (\bar{y}_{.j} - \bar{y})^2$	(J-1)
residual	$\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$	(I-1)(J-1)

total $\sum_i \sum_j y_{ij}^2$ $n = IJ$

Wildfire example.

Plot effects a_i, b_j (parallel boxplots)

ANOVA TABLE

Source	SS	DF
mean	14.912	1
rows	17.270	11
columns	3.720	9
residual	10.642	99
total	46.544	120

twoway(trim=0), aov()

Response may be summary of a batch

Finding patterns difficult with large tables

If classical test rejects, what next?

EDA can suggest