Statistics 215a – 10/2/03 – D. R. Brillinger

Residual analysis.

$r_{ij} = y_{ij} - m - a_i - b_j$

$= y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}$

Plot vs.

fitted values

row values, a_i

column values, b_j

the diagnostic plot

comparison value, $a_i b_j / m$

Look for

pattern(s)

outlier(s)

surprises

45^o line suggests log transform

Example. wildfire data

L1 approximation.

1. Location summary statistic

$$\min_\theta \sum_i |y_i - \theta|$$

minimized by: $y_{(m+1)}$ if $n = 2m+1$

any $(y_{(m)}, y_{(m+1)})$ if $n = 2m$

Proof. Perturb θ away

median()

2. Linear function

$$\min_\beta \sum_i |y_i - x_i^T\beta|$$

linear programming

l1fit()

3. Two-way array

i) least absolute residuals

$\min_{\mu,\alpha,\beta} \sum_{i,j} |y_{ij} - \mu - \alpha - \beta|$

l1fit()

diagnostic plot

residual, r_{ij} vs. $a_i b_j / m$

ii) median polish

operate iteratively removing and column medians until each row/column has median 0

(sweeping)

additive approximation

$y_{ij} = m + a_i + b_j + r_{ij}$

resistant

can carry out by hand

missing values OK

answer depends on whether begin with rows or columns

approximates L1 solution

if remove means get aov in one pass

$m = \bar{\bar{y}}_{..}$

$a_i = \bar{y}_{..} - \bar{y}_{i.}, \quad b_j = \bar{y}_{..} - \bar{y}_{.j}$

$r_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$

twoway()

diagnostic plot

residual, r_{ij} vs. $a_i b_j / m$

add resistant line

can suggest transformation to additivity

$$y_{ij} = m*(1+a_i/m)*(1+b_j/m) + r_{ij}$$

Example.

acres of wildfires

Other criteria

$$\min_{\mu,\alpha,\beta} \sum_{i,j} \rho(y_{ij} - \mu - \alpha - \beta)$$

biweight, trimmed mean

twoway(), medpolish(eda)

cp. OLS vs. resistant line

Statistics 215a – 10/6/03 – D. R. Brillinger

J. W. Tukey and M. B. Wilk (1966). Data analysis and statistics: an expository overview. *AFIPS Conference Proceedings* Vol. 29. Also in *The Collected Works of John W. Tukey* Vol. 3 and the *Statistics 215a Reader.*

Introduction

"The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer and recordable for posterity."

"Four major influences act on data analysis today (1966):

1. The formal theories of statistics.

2. Accelerating developments in computers and display devices.

3. The challenge, in many fields, of more and ever larger bodies of data.

4. The emphasis on quantification in an ever wider variety of disciplines."

“Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis.

Formal statistics has given almost no guidance to exposure; …”

Data analysis is like doing experiments

“Far too many people, … , have persisted in regarding statistics, even data analysis, as a branch of probability theory, … within modern mathematics. …

Statistical data analysis is much more appropriately associated with the sciences and with the experimental process in general.”

“The general purposes of conducting experiments and analyzing data match, point by point.

For experimentation, these purposes include

(1) more adequate description of experience and quantification of some areas of knowledge
(2) discovery or invention of new phenomena and relations
(3) confirmation, or labeling for change, of previous assumptions, expectations, and hypotheses
(4) generation of ideas for further useful experiments
(5) keeping the experimenter relatively occupied while he thinks.

Comparable objectives in data analysis are

(1) to achieve more specific description of what is loosely known or suspected

(2) to find unanticipated aspects in the data, and to suggest unthought-of-models for data’s summarization and exposure

(3) to employ the data to assess the (always incomplete) adequacy of a contemplated model

(4) to provide bith incentives and guidance for further analysis of the data

(5) to keep the investigator usefully stimulated while he absorbs the feeling of his data and what to do next.