

Statistics 215a - 10/20/03 - D.R. Brillinger

C. Mallows and D. Pregibon "Some principles of data analysis"

DA ... usefully distinguished from ... formal statistical methods ... experimental design, hypothesis testing, estimation (including interval), likelihood, Bayesian

DA comprises ... exploration, ... graphical techniques, description and summarization

Modern theory - model specification (structural and stochastic parts) assumed and procedures are developed that aim at optimality relative to that specification.

assessing parameters in an assumed model

context of problem can be forgotten

many different fields can be attacked with similar methods

little to say about the exploratory phase, the searching for models

easy to fail to take the essential step back to the real problem

prudent to remain alert for new structure

Coen, Gomme and Kendall (1969) failed to allow for the possibility of serial correlation in the errors of their models

random walk better forecast than regression models

There is a continuing need for discussion of the principles that should guide work in the data analysis phase

teaching data analysis is difficult

organized principles

More often than not, data analysis is a series of dead ends, tedious searching for clues, and alternative explanations and transformations

Many good data analyses never see the light of day

much time spent getting the data into an acceptable form (data validation)

G. Polya (1957). *How To Solve It*.

Understand the problem.

interact with subject-oriented person

understand goals, constraints

can one get answers with available resources

Devise a plan.

rough-out a sequence of analysis steps

regression? classification? prediction?

Is n too small or large?

Redundancy?

Carry out the plan.

analyze the data

check each step

compute diagnostics

plot residuals

iterate (revise plan)

present results

state limitations of the methodology

Looking back.

examine solution and path to it

abstract for other problems

improvement of understanding of solution?

Can easily generate more numbers than started with

Can a description be sufficient? - perhaps

Daniel (1976) shows many cases in which a published analysis missed important features in the data.

Sensitivity analysis

e.g. drop-one-out

We need to choose a measure of the degree to which the description allows the raw data to be reconstructed.

What we need are ways of analyzing large data-sets that have a good chance of finding whatever structure may be present.

EDA: what seems to be going on?

CDA: what seems established beyond reasonable doubt?

*Lawyers aren't necessarily looking for 'truth'.
Data analysts should be.`*