Stat 215a D.R. Brillinger 12/1/04

Statistics, EDA, Data mining

Stem-and-leaf

5 number summary
boxplot (batches)
Scatter plot

Percentile graph (empirical cdf)
Q-Q plot

Magical thinking (Diaconnis)

Summaries of location
mean, median, trimmed mean, biweight

Spread vs. level plot
transformations
log, Box-Cox, logit, $\sqrt{\ }$

Smoothing scatter plot
polynomial
bin smoother
running mean
running line

kernel
running median

regression spline
cubic smoothing spline
locally-weighted running line
super smoother (cross validation)
thin-plate spline

Future of data analysis

Linear fitting
  OLS
  residuals
  WLS, NLS
  multiple predictors
  leverage ($diag(X(X'X)^{-}X')$)
  orthogonality
Robust/resistant fitting
  three groups
  bisquare
  M-estimate (IRLS)
  Mallows rule/strategy

Residual analysis (pattern?)

data = fit + residual

ordinary

$$r_i = y_i - x_i' b$$

standardized

$$r_i / s\sqrt{1-h_{ii}}$$

cross-validation

$$y_i - x_i' b_{-i}$$

Types

r_i vs. $x_i' b$

r_i vs. x_{ij}

r_i vs. new variables

r_i vs. x_{iJ} x_{iJ}

$|r_i|$ vs. $x_i' b$

smothed vs.

Partial residual plot

x-values

factors