Statistics, EDA, Data mining

Stem-and-leaf

5 number summary
  box plot (batches)
Scatter plot

Percentile graph (empirical cdf)
  Q-Q plot

Magical thinking (Diaconis)

Summaries of location
  mean, median, trimmed mean, biweight

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

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 \(\text{diag}(X(X'X)^{-1}X')\)
orthogonality

Robust/resistant fitting
three groups
bisquare
M-estimate (IRLS)
Mallows rule/strategy
Residual analysis (pattern?)
data = fit + residual
ordinary
\[ r_i = y_i - x_i \beta \]
standardized
\[ \bar{r}_i = \frac{r_i}{s \sqrt{1-h_{ii}}} \]
cross-validation
\[ y_i - x_i \hat{\beta} - \hat{\epsilon} \]

Types
\[ r_i \text{ vs. } x_i \beta \]
\[ r_i \text{ vs. } x_{ij} \]
\[ r_i \text{ vs. new variables} \]
\[ r_i \text{ vs. } x_{ij} x_{ij} \]
\[ 1r_i \text{ vs. } x_i \beta \]
smoothed vs.

Partial residual plot

x-values
factors
Wavelets

Nonlinear least squares
ordinary robust/resistant

Sensitivity curve

\[(m+1)\{ T_{n+1}(y_1,...,y_m, y) - T_n(y_1,...,y_m) \}\]

Two-way array, \( y_{ij} \)

\[ y_{ij} = \mu + \alpha_i + \beta_j \text{ (additive)} \]
side conditions
residuals \( r_{ij} = y_{ij} - \mu - \alpha_i - \beta_j \)

Means \( \bar{y}_{..}, \bar{y}_{i.}, \bar{y}_{.j} \)

\[ y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) \]
ANOVA identity

\[ \sum_i \sum_j y_{ij}^2 \]

\[ = \sum_i \sum_j \overline{y}_{..}^2 + \sum_i \sum_j (\overline{y}_{..} - \overline{y}_{..})^2 \]

\[ + \sum_i \sum_j (\overline{y}_{.i} - \overline{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \overline{y}_{.i} - \overline{y}_{.j} + \overline{y}_{..})^2 \]

ANOVA Table

Residual analysis

\[ R_{ij} = y_{ij} - \overline{y}_{.i} - \overline{y}_{.j} + \overline{y}_{..} \]

Plot vs.
Fitted values
Row values, \( a_i \)
Column values, \( b_j \)
\[ \frac{a_i b_j}{m} \] diagnostic plot
Add resistant line
Look for
pattern(s)
outliers
surprises

$L_1$ approximation, robust
location $\sum_i |y_i - \Theta|$, median
linear $\sum_i |y_i - x_i' \beta|$

two-way array $\sum_{i,j} |y_{ij} - \mu - \alpha_i - \beta_j|$

Median polish
remove row and column medians
until each row/column has median 0

Can do by hand
resistant
missing values OK
M-estimate
\[ \sum_{i,j} \rho(y_{ij} - \mu - \alpha_i - \beta_j) \]

\{y_{ij}\} may be counts
contingency table
another approximation
\[ y_{ij} \sim \mu + \alpha_i \]

Tukey and Wilk