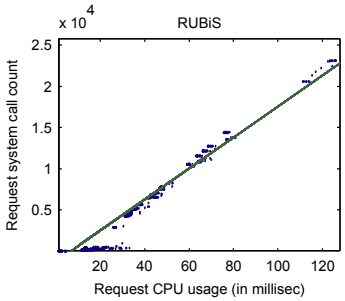
# Regression

**Kai Shen**

**Dept. of Computer Science, University of Rochester**

## Regression

- What is regression?
  - predict a target random variable as a function of one or more input variables
- Linear regression
  - linear functional relationship
  - $y = b_0 + b_1 x$






## Regression Error

- Least-square fitting
  - $y_1, \ldots, y_n$ are observed samples for target variable while $y_1^*, \ldots, y_n^*$ are corresponding predictions
  - minimize mean square error
    - $((y_1^*-y_1)^2 + \ldots + (y_n^*-y_n)^2) / N$
- Problems with mean square errors
  - doesn't give a true sense of modeling difficulty
    - coefficient of determination: $1 - E_{model} / E_{mean}$
    - F-test – must also consider degrees of freedom
      - trivial to find a regression function with two parameters that fit perfectly with two sample observations
  - too much weight for outliers - mean absolute error



## Non-linear Regression

- Curvilinear
  - nonlinear relationships that can be converted to linear forms with variable transformations
- Exponential
  - $y = b\, a^x$
  - $\ln y = \ln b + x \ln a$
- Powerlaw
  - $y = b\, x^a$
  - $\ln y = \ln b + a \ln x$
  - Powerlaw relationships in practice

## Powerlaw Fitting

- Loglog-transformations and then linear fitting
- Problems
  - Heavy weight on tails after transformations
  - Enlarged errors on non-tails



## Other Prediction Models

- Nearest matching
- Decision trees
- Bayesian networks



## Other Issues

- Outliers
- Confidence beyond measured range
- Intuitive explanation of derived models
- Correlation vs. Dependency



## Disclaimer

- Most materials in these slides were developed from the book "The Art of Computer Systems Performance Analysis", R. Jain, 1991, Wiley.