Computer Systems Performance Analysis and Benchmarking (37-235)

Analytic Modeling

Simulation

Measurements / Benchmarking

Lecture by:
Michela Taufer

Assignments/Projects:
Christian Kurmann

Textbook:
Raj Jain, “The Art of Computer Systems Performance Analysis”, 1991 Wiley & Sons, New York

Topic of Today:

- **Clustering Algorithm**
- **Confidence Intervals**
- **Linear Regression**

Confidence Intervals

Sample vs. population

- Sample is the random subset taken as an estimate. I can vary depending on size.
- A sample has statistics like sample mean or sample standard deviation.
- Population is the distribution as it really is. It is fixed but might be unknown.
- A population has parameters like population mean or population standard deviation.
- Statistics are estimates of parameters

Confidence Interval for a Mean

Definitions:

$$P\{c_1 \le \mu \le c_2\} = 1 - \alpha$$

- α significance level (0.05 or 0.10)
- 1-α confidence level (0.95 or 0.90)

Central Limit Theorem:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Calculation of the confidence Interval

$$\left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

x = sample mean

s = sample std.dev.

n = number of samples (n>30)

z’s = standard variate read from table

Calculation of the confidence Interval

$$\left(\bar{x} - t_{\left[1-\frac{\alpha}{2}, n-1\right]}\frac{s}{\sqrt{n}}, \bar{x} + t_{\left[1-\frac{\alpha}{2}, n-1\right]}\frac{s}{\sqrt{n}}\right)$$

x = sample mean

s = sample std.dev.

n = number of samples (n>30)

t’s = t variate read from Table

Testing for zero mean

- Calculate confidence interval
- Is zero part of it?

Comparing two Alternatives

- paired observations/value
- just compute difference
- subject to zero mean test

Comparing two Alternatives

- Unpaired observations
- A procedure called the t-test, see book on page 210.

Hint: Visualize the Confidence Intervals

Sample size for a simple mean:

- Formulas can be inverted to tell us the minimum sample size

Minimum sample size for telling alternatives are different.

- Same formula, just two times the slack.

Simple Linear Regression Model

Good - Bad Model

- measured in terms of residual

Criterion of the least squares

Model and Error:

$$\hat{y} = b_0 + b_1 x$$
$$\hat{y}_i = b_0 + b_1 x_i$$
$$e_i = \hat{y}_i - y_i$$

Sum of Squared Errors:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0$$

Computing the Parameters

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

Derivation:

Simple substitute and differentiate-> book.

Quality of linear regression

SSE: Sum of Squared Errors (see. above)

SST: Sum of Squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Coefficient of Determination:

$$R^2 = \frac{SST - SSE}{SST}$$

- 1 for perfect regression - 0 for a bad one

Hint: Again visualize and check against errors on the plotted graph.

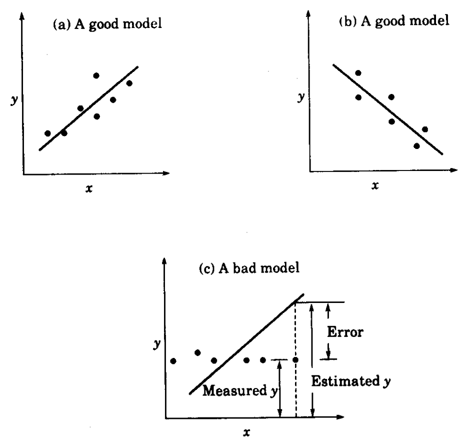

FIGURE 14.1 Good and bad regression models.