COMP 190-088: Systems Performance Analysis

# Analyzing Measured Data

**Jasleen Kaur**

Department of Computer Science
The University of North Carolina at Chapel Hill

Spring 2005



## Analysis of Measured Data

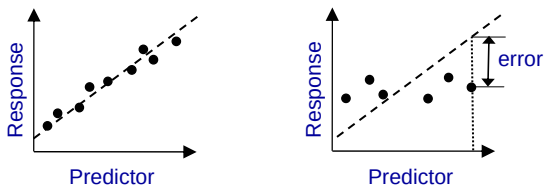
- Summarizing the Average
- Summarizing the Variability
- Comparing Systems Using Sample Data
- **Simple Linear Regression Models**

# Linear Regression Models

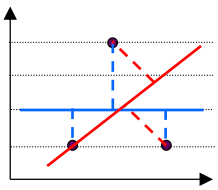
- Need to find relation between different measured quantities
- Regression Models:
  - Allow prediction of a random variable as a function of other variables
    - Response variable
    - Predictor variables
  - Assume that predictor variables are quantitative
- Linear Regression Models:
  - Response variable is modeled as a linear function of predictors
    - Limit focus to one predictor variable
  - Least-square fitting of straight lines to data



# What's a Good Regression Model?




- Basic Idea:
  - Model line should be closer to more observations
- Measure the vertical distance between observations and the line
  - Modeling error in predicting response for a given predictor value
- Positive and negative errors should cancel out
  - Too many lines satisfy that!
- Magnitude of errors should be small



***Use the least-square fitting line***

## Least-square Fitting Line

- Find the line that minimizes the sum of squares of the errors
- Linear model: $\hat{y} = b_0 + b_1 x$
- Given n observation points: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Error = $e_i = y_i - \hat{y}_i$
    $= y_i - b_0 - b_1 x_i$
- Best linear model is given by the $(b_0, b_1)$ that minimize: $\sum_{i=1}^{n} e_i^2$
  - Given the constraint that the mean error is zero: $\sum_{i=1}^{n} e_i = 0$



## Estimation of Model Parameters

- If mean error = 0: $\bar{y} - b_0 - b_1 \bar{x} = 0$
- Sum of Squared Errors (SSE):

$$SSE = \sum_{i=1}^{n} \left[ (y_i - \bar{y})^2 - 2b_1 (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 (x_i - \bar{x})^2 \right]$$
$$= \sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

- Differentiating with respect to $b_1$ and equating to 0 yields:

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

## Allocation of Variation in Response Variable

- Purpose of any model is to predict the response with minimum variability in predicted values
- One source of variability explained by regression model
  - What if we use the mean response as the predicted value for all x?
  - Error variance without regression = $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$
    = variance of observed y
- Total Sum of Squares (SST) = $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}y_i^2 - n\bar{y}^2 = SSE + SSR$
- Two sources of variability in response variable:
  - Variability in yi as a function of xi (SSR); explained by regression
  - Variability not explained by regression due to modeling errors (SSE)
- Coefficient of Determination = $R^2$ = SSR/SST
  - Measures the goodness of regression
  - Square of correlation-coefficient of (x,y)



## Standard Deviation of Errors

- Variance of errors = $s_e^2$ = SSE/n-2
  - (n-2) degrees of freedom, since errors are computed after calculating two regression parameters from the data
- Degrees of freedom for other sum of squares:
  - $\sum_{i=1}^{n}y_i^2$ : n
    - Obtained from n independent observations
  - $\sum_{i=1}^{n}\bar{y}^2$ : 1
    - Can be computed from the sample mean
  - SST : (n-1)
    - Can be computed only after mean has been computed
  - SSR : 1
    - Computed from SST and SSE
    - Degrees of freedom add up just like sum of squares do

## Confidence Intervals for Regression Parameters

- Coefficients $b_0$ and $b_1$ are estimates from a single sample
  - Using these, only probabilistic statements can be made about the true parameters $B_0$ and $B_1$ of the population
- Mean values of $b_0$ and $b_1$
  - As computed earlier

$$s_{b_0} = s_e\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}\right]^{1/2}$$

- Standard deviations:
  - $\bar{x}$ : sample mean of x
  - $s_e$ : standard deviation of errors

$$s_{b_1} = \frac{s_e}{\left[\sum x^2 - n\bar{x}^2\right]^{1/2}}$$

- 100(1-a)% confidence intervals for $b_0$ & $b_1$:

$$b_0 \mp t_{[1-\alpha/2;n-2]}s_{b_0}$$
$$b_1 \mp t_{[1-\alpha/2;n-2]}s_{b_1}$$



## Confidence Intervals for Predictions

- Regression enables prediction of response for any value of predictor using: $\hat{y}_p = b_0 + b_1 x_p$
- Above formula gives only the mean value of the response, based on the sample
- Standard deviation of predicted mean of a large number of observations:

$$s_{\hat{y}_p} = s_e\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}\right]^{1/2}$$

- Confidence interval can be computed using a t-variate with (n-2) degrees of freedom

$$\hat{y}_p \mp t_{[1-\alpha/2;n-2]}s_{\hat{y}_p}$$

## Regression Assumptions

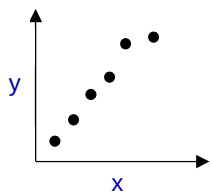
- Derivation of regression parameters uses several assumptions:
  - True relationship between response y, and predictor x, is linear
  - Predictor variable is non-stochastic and is measured without error
  - Model errors are statistically independent
  - Errors are normally distributed with zero mean and constant standard deviation
- Need to verify assumptions before conclusions of regression can be used

***Some assumptions can be verified visually***

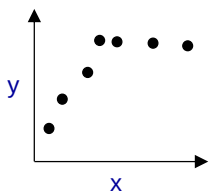


## Verifying Linear Relationship

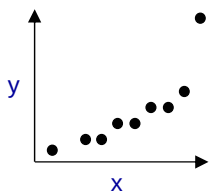
- Inspect the scatter plot of y versus x
  - Reject assumption if non-linear relationship seen in the plot



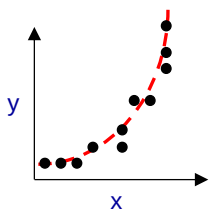

Linear




Multilinear



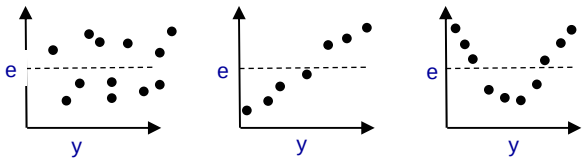

Outlier



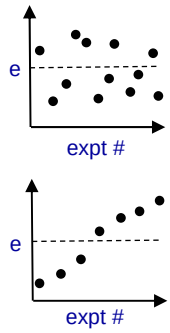

Nonlinear

# Verifying Independence of Errors

- Independence of errors and predicted response
  - Look for trends in scatter plot of $e_i$ versus $\hat{y}_i$




- Independence of errors across successive experiments
  - Look for trends in plot of $e_i$ versus experiment number
  - Trend indicates influence of factors that varied across experiments
    - eg, incorrect initialization

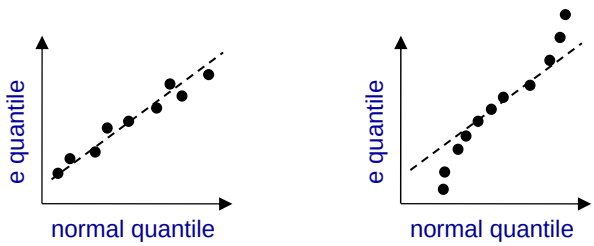



***Dependence can be proved, but independence can not!***



# Verifying Normal Distribution of Errors

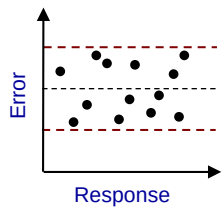
- Inspect the quantile-quantile plot of errors vs. normal distribution
  - If plot is linear, assumption is satisfied



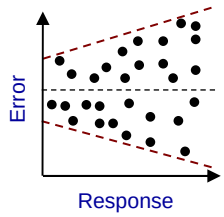

Normal Non-normal

## Verifying Constant Standard Deviation of Errors

- Inspect the scatter plot of errors versus response
  - If spread in one part looks different from others, assumption is not valid




No trend in spread




Increasing spread



## Course Outline

- Selection of metrics
- Performance Evaluation Methodologies
- Workload selection
- Measurements tools
- Analysis and **visualization** of measured data
- System Modeling
- Simulations
- Case studies
- Distributed monitoring infrastructures
- PA in the Research and Industrial communities