Data Reduction

“It’s Poetic”

16.621

March 18, 2003

Introduction

- A primary goal of your efforts in this course will be to gather empirical data so as to prove (or disprove) your hypothesis
- Typically the data that you gather will not directly satisfy this goal
- Rather, it will be necessary to "reduce" the data, to put it into an appropriate form, so that you can draw valid conclusions
- In our discussion today we will examine some typical methods for processing empirical data
- Caution-garbage in/garbage out still applies

Hiawatha Designs an Experiment

by

Maurice G. Kendall

From *The American Statistician*
Vol. 13, No. 5, 1959, pp 23-24

Verses 1 through 6

Deyst’s 16.62X Project

- I have performed a very simple experiment
- The hypothesis was: my driving route distance, from West Garage to my driveway in Arlington, is eight miles
- On a number of trips I recorded the mileage, as indicated by the odometer of my automobile
- I now wish to reduce the data and draw some conclusions

Experimental Project (cont.)

- My experimental procedure was: at the exit from West Garage I zeroed my trip odometer and when I reached my driveway at home I recorded the odometer reading
- On each of ten trips I took the same route home

Error Sources

Random errors

Odometer readout resolution

Odometer mechanical variations

Route path variations

Tire slippage

Systematic errors

Bias in the odometer readings

Odometer scale factor error

Tire diameter decreases due to wear

Error Sources (cont.)

- The resolution I achieved in reading the odometer was within ±.025 miles
- The best knowledge I have about the other random errors is that they were all in the range of ±.10 miles
- I zeroed the odometer at the beginning of each trip so any bias in the measurements is small (i.e. about ±.005 miles)

Error Sources (cont.)

- I did a scale factor calibration by driving 28 miles, according to mileage markers on Interstate 95, and in both directions I recorded 27.425 miles on my odometer
- Thus, the scale factor is

$$S.F. = \frac{27.425}{28} = .980\,\frac{\text{odometer indicted miles}}{\text{actual miles}}$$

- And any error in the scale factor due to readout resolution is

$$e_{\mathrm{SF}} = \pm\frac{.025}{28\cdot\sqrt{2}} \cong \pm.0006$$

Recorded Data

Trip Number	Mileage Reading	S.F. Corrected Mileage reading
1	7.825	7.985
2	7.850	8.010
3	7.875	8.036
4	7.900	8.061
5	7.850	8.010
6	7.825	7.985
7	7.875	8.036
8	7.850	8.010
9	7.875	8.036
10	7.825	7.985

Mileage Data Analysis

- My system model is that the route distance is constant
- To minimize the effect of random errors take the sample mean (average) of the data to obtain an estimate

$$\hat{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = 8.015 \text{ miles}$$

Mileage Data Analysis (cont.)

- Variations of the individual measurements, about this estimate are

$$e_i = d_i - \hat{d}$$

- The sample mean of these variations is

$$\hat{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{d}) = 0$$

- So the estimate is <u>unbiased</u>

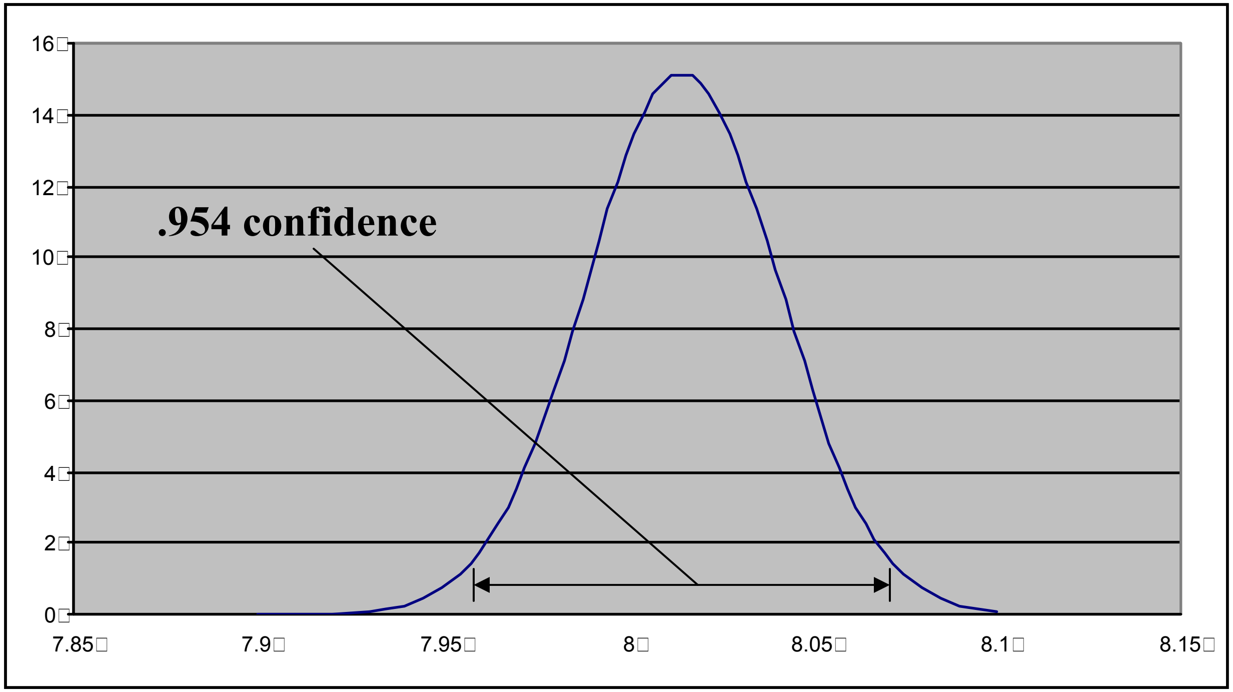
Mileage Data Analysis (cont.)

- Assuming the variations are statistically independent we can also compute the sample standard deviation of these variations as

$$\hat{\sigma} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(d_i - \hat{d})^2} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}{e_i}^2} = .026\,\text{miles}$$

Mileage Data Analysis (cont.)

- Since my experiment consisted of a number of independent trials it is reasonable to assume that the route distance, as determined by my measurements, is gaussian



Linear System Models

- The system model for my experiment assumed that the route distance is constant
- In many instances the system model is not constant but is a linear function
- Define a linear system model as

$$y = c_0 + c_1 x$$

where

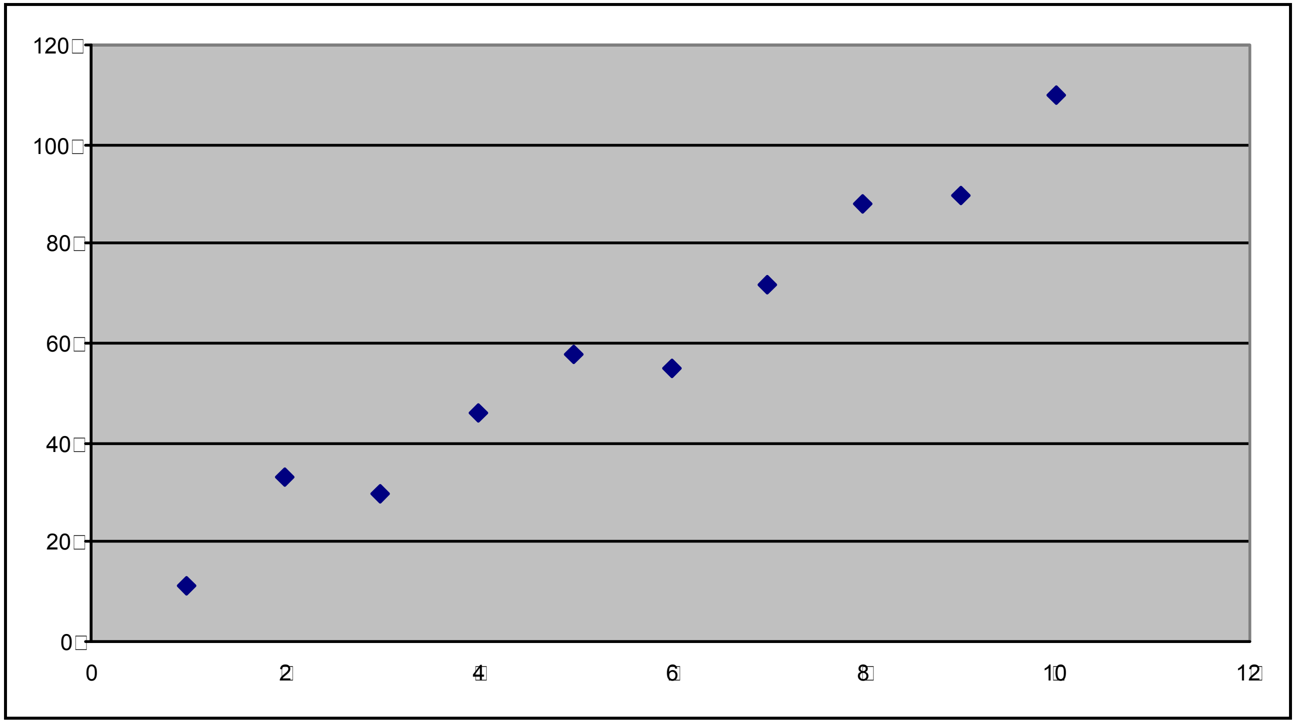
$$x \equiv \text{independent variable}$$

$$y \equiv \text{dependent variable}$$

Linear System Models (cont.)

Typically, for a number of values of the independent variable (*x*), the corresponding values of the dependent variable (*y*) are measured

measured values of dependent variable (y)



independent variable (x)

Straight Line Fit

- For a linear model, the object is to find the best straight line fit to the measured data
- We can characterize each measurement as

$$y_i = c_0 + c_1 x_i + e_i$$

where

$e_i =$ error or variation of the ith measuremer
from a straight line model

Straight Line Fit (cont.)

- To characterize the complete set of *n* measurements define the following arrays

$$\underline{y} = \begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \qquad \underline{c} = \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} \qquad \underline{e} = \begin{bmatrix} e_1 \\ . \\ . \\ e_n \end{bmatrix}$$

- So the measurement equation becomes

$$\underline{y} = X\underline{c} + \underline{e}$$

Straight Line Fit (cont.)

- Recall that we wish to find the best straight line fit to the measured data array $\underline{y}$
- A useful criterion for the best fit is to minimize the sum of the squared errors

$$\left\|\underline{e}\right\|^2 = \sum_{i=1}^{n} e_i^2 = \underline{e}^T \underline{e}$$

Straight Line Fit (cont.)

- And upon substitution from above

$$\begin{aligned}\|\underline{e}\|^2 &= (\underline{y} - X\underline{c})^T(\underline{y} - X\underline{c}) \\ &= \underline{y}^T\underline{y} - 2\underline{y}^T X\underline{c} + \underline{c}^T X^T X\underline{c}\end{aligned}$$

- Our goal is to find the array $\underline{c}$ so that the sum squared error is minimized

- First determine the gradient of the sum squared error with respect to $\underline{c}$

$$\frac{\partial\|\underline{e}\|^2}{\partial\underline{c}} = -2\underline{y}^T X + 2\underline{c}^T X^T X$$

Straight Line Fit (cont.)

- Setting the gradient to zero yields the optimum

$$\underline{\hat{c}} = (X^T X)^{-1} X^T \underline{y}$$

- Since the required inverse matrix is only 2×2 we can readily solve for the two elements of $\underline{\hat{c}}$

$$\hat{c}_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum (y_i x_i)}{n \sum x_i^2 - (\sum x_i)^2} \qquad \hat{c}_1 = \frac{n \sum (x_i y_i) - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

- These are the equations used in your calculator or computer to get a best straight line fit to data as

$$\hat{y}(x) = \hat{c}_0 + \hat{c}_1 x$$

Beam Deflection Example

- A cantilever beam deflects downward when a mass is attached to its free end. A beam model predicts that the deflection will be a linear function of the mass.
- A student places various masses on the end of the beam and records the deflections
- The masses are measured to within ± .11grams
- The error in reading the deflections is within ± .23 millimeters

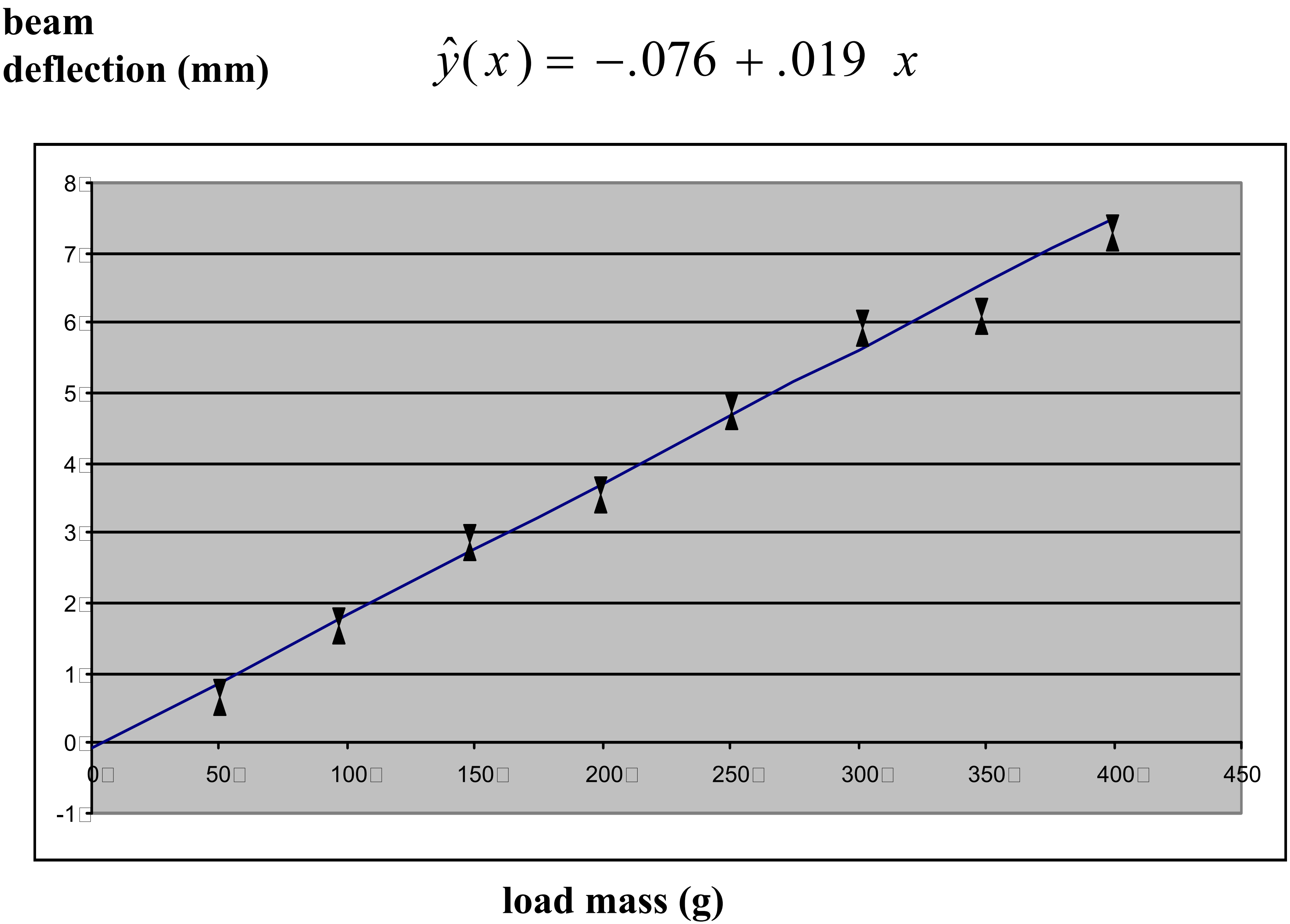
Excerpted from: Beckwith, T.G., Marangoni R.D.,and Lienhard V, J.H., *Mechanical Measurements,* Fifth Edition, Addison Wesley, Reading, MA, 1993, pp. 113-115

Beam Deflection Data

x value, load mass (gm)	y value, beam deflection (mm)
0	0
50.15	0.6
99.90	1.8
150.15	3.0
200.05	3.6
250.20	4.8
299.95	6.0
350.05	6.2
401.00	7.5

Straight Line Fit to Beam Data

$$\hat{y}(x) = -.076 + .019\ x$$



Hiawatha

Verses 7 through 12

Linear Fit Analysis

- Recall that the best fit to *y(x) is*

$$\hat{y}(x) = \hat{c}_0 + \hat{c}_1 x$$

- The variations or errors from the fit, at each measurement point, are then

$$e_i = y_i - \hat{y}(x_i) = y_i - (\hat{c}_0 + \hat{c}_1 x_i)$$

- So the array of measurement errors is

$$\underline{e} = \underline{y} - \underline{\hat{y}} = \underline{y} - X\underline{\hat{c}} = \underline{y} - X(X^T X)^{-1} X^T \underline{y}$$
$$= (I - X(X^T X)^{-1} X^T)\underline{y}$$

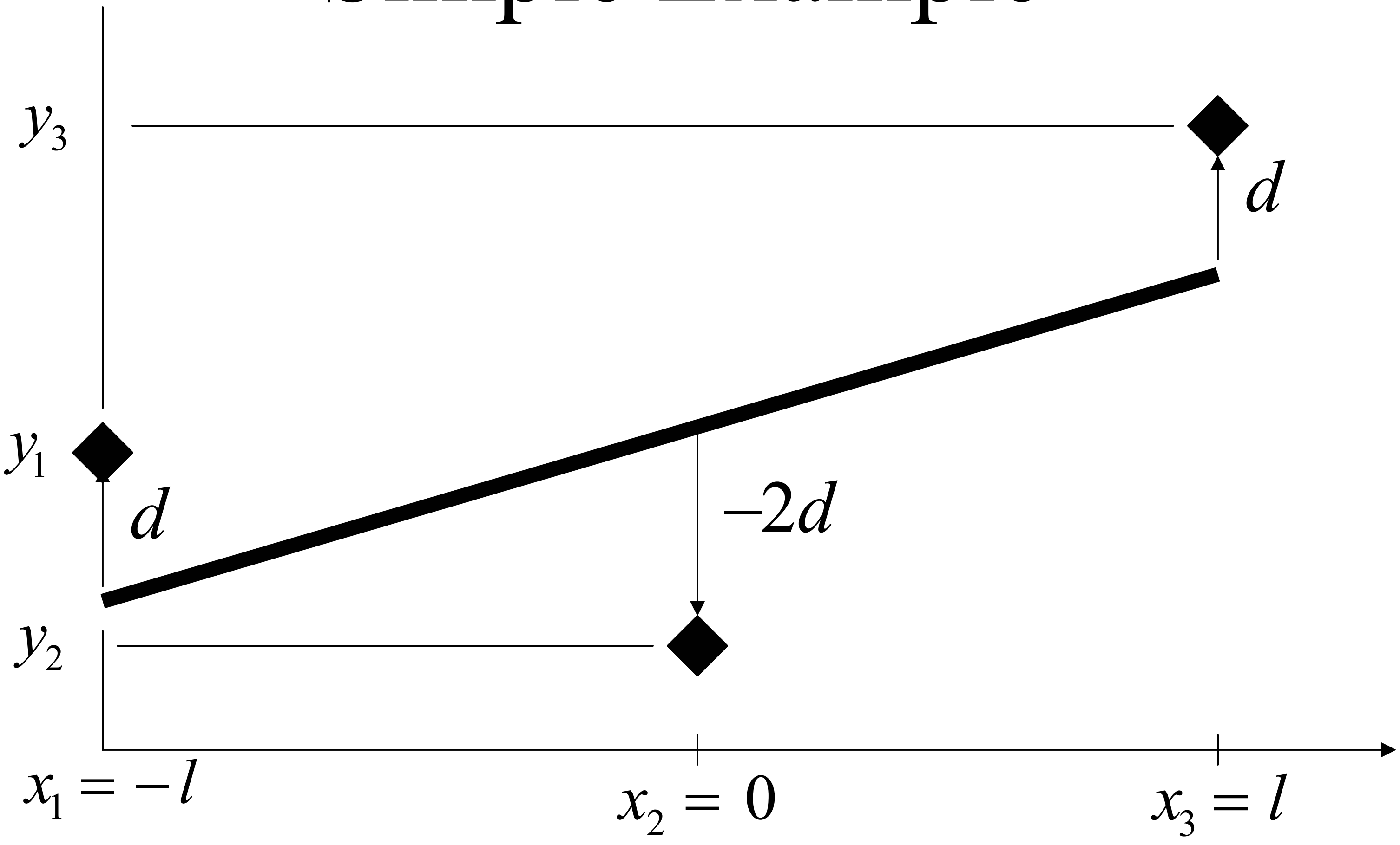
Straight Line Fit (cont.)

- A useful result is obtained by premultiplying both sides of this equation by the matrix X^T

$$X^T \underline{e} = \begin{bmatrix} \sum e_i \\ \sum x_i e_i \end{bmatrix} = (X^T - X^T X (X^T X)^{-1} X^T)\underline{y} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Thus, the sample mean and *x* weighted sample mean of the errors are both zero

Straight Line Fit
Simple Example



$$\sum e_i = d - 2d + d = 0$$

$$\sum x_i e_i = (-l)\cdot d + (0)\cdot(-2d) + l\cdot d = 0$$

Straight Line Fit (cont.)

- We can also derive an expression for the sample standard deviation, in terms of the measured data, by noting that

$$\hat{\sigma} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{(n-1)}\underline{e}^T\underline{e}} = \sqrt{\frac{1}{(n-1)}(\underline{y}-\underline{\hat{y}})^T(\underline{y}-\underline{\hat{y}})}$$

$$= \sqrt{\frac{1}{(n-1)}\underline{y}^T(I - X(X^T X)^{-1}X^T)\underline{y}}$$

Nonlinear System Models

- In many instances the system model will not be linear
- Often it is still possible to use a linear fit to analyze data
- For example, suppose the system is an electronic circuit, for which we measure the output voltage over time in response to an initial condition

Nonlinear System Models (cont.)

- The system model might be

$$v(t) = v(0)e^{-\alpha t}$$

where

$v(0)$ = initial condition voltage

$\alpha = 1/\tau =$ inverse time constant

- In this case the independent variable is time and the dependent (measured) variable is output voltage

Nonlinear System Models (cont.)

- To linearize take the natural log of both sides of this equation

$$\ln(v(t)) = \ln(v(0)) - \alpha t$$

- And we can obtain our previous linear equation

$$y = c_0 + c_1 x$$

by identifying

$$y \equiv \ln(v(t)) \quad c_0 \equiv \ln(v(0)) \quad c_1 \equiv -\alpha \quad x \equiv t$$

Nonlinear System Models (cont.)

- Thus the exponential system model is converted into a linear model
- The measured data is converted using the identities

$$x_i = t_i \quad \text{and} \quad y_{\mathrm{i}} = \ln(v_i)$$

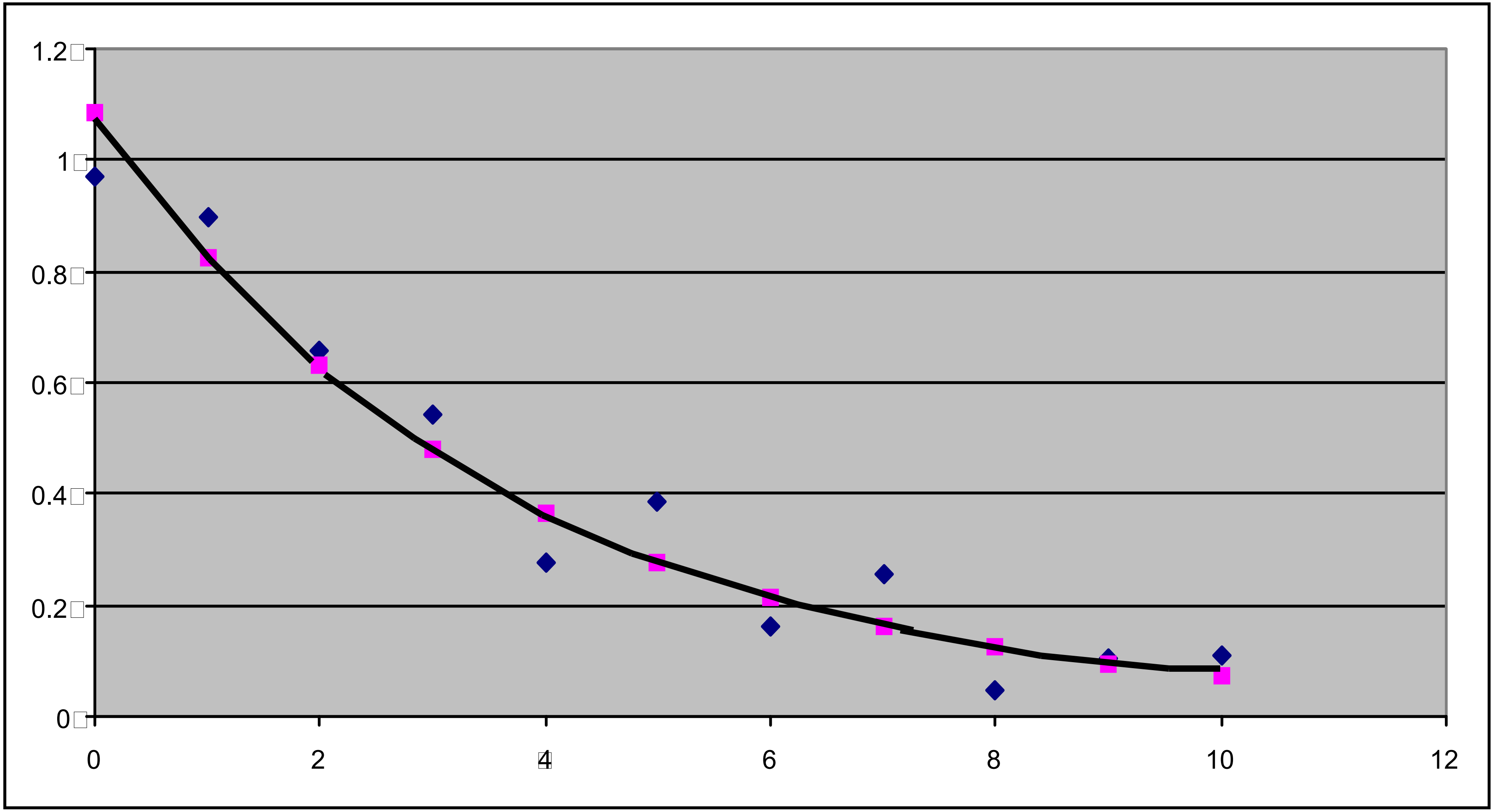
- These values are used to obtain $\underline{\hat{c}}$ as before

$$X = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \qquad \underline{\hat{c}} = (X^T X)^{-1} X^T \underline{y}$$

- The best exponential fit to the data is then

$$\hat{v}(t) = \exp(\hat{y}) = \exp(\hat{c}_0 + \hat{c}_1 t) = \exp(\hat{c}_0) \cdot \exp(\hat{c}_1 t)$$

Nonlinear System Model Example: Exponential Fit



Power Series Approximations

- Often, in cases where such a simple transformation is not available the data may be fit by a power series
- Suppose the dependent variable $y(x)$ can be approximated to sufficient accuracy by a finite power series in *x*, of degree m

$$y(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_m x^m$$

Power Series Approximations (cont.)

- Also, if we have *n* measurements of the dependent variable *y*, corresponding to *n* values of the independent variable *x*, then define the linear model as before so

$$\underline{y} = X\underline{c} + \underline{e}$$

- Where now

$$\underline{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_m \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}$$

Power Series Approximations (cont.)

- And our previous result can now be applied once again to get the best linear fit for $\underline{c}$ as

$$\underline{\hat{c}} = (X^T X)^{-1} X^T \underline{y}$$

- So the best linear fit is

$$\hat{y}(x) = \hat{c}_0 + \hat{c}_1 x + \hat{c}_2 x^2 + \cdots + \hat{c}_m x^m$$

- The solution is somewhat more difficult because the required inverse is $(m+1)\times(m+1)$, but for most situations the problem is still tractable

Fourier Series Approximations

- Sometimes the model may be periodic in nature and a truncated Fourier series can approximate the function
- If *y(x)* is an odd periodic function of *x*, with first harmonic wavelength *2L*, then a Fourier sine series approximation to *y(x) is*

$$y(x) \cong c_1 \sin(\pi x/L) + c_2 \sin(2\pi x/L) + \cdots + c_m \sin(m\pi x/L)$$

Fourier Series Approx. (cont.)

From: Beckwith, T.G., Marangoni R.D.,and Lienhard V, J.H., *Mechanical Measurements,* Fifth Edition, Addison Wesley, Reading, MA, 1993, p. 141

Figure 4.10 Plot of square-wave function: (a) plot of first three terms only (includes the fifth harmonic), (b) plot of the first five terms (includes the ninth harmonic), (c) plot of the first eight terms (includes the fifteenth harmonic)

Fourier Series Approx. (cont.)

- Thus, if n measurements y_i are taken at various values x_i of the independent variable, then the X matrix can be defined as

$$X = \begin{bmatrix} \sin(\pi x_1/L) & \sin(2\pi x_1/L) & \sin(3\pi x_1/L)\cdots\cdots & \sin(m\pi x_1/L) \\ \sin(\pi x_2/L) & \sin(2\pi x_2/L) & \sin(3\pi x_2/L)\cdots\cdots & \sin(m\pi x_2/L) \\ . & . & . & . \\ . & . & . & . \\ \sin(\pi x_n/L) & \sin(2\pi x_n/L) & \sin(3\pi x_n/L)\cdots\cdots & \sin(m\pi x_n/L) \end{bmatrix}$$

Fourier Series Approx. (cont.)

- And, as before, the array $\underline{\hat{c}}$ is obtained from

$$\underline{\hat{c}} = (X^T X)^{-1} X^T \underline{y}$$

- So the best linear fit for *y(x)* is

$$\hat{y}(x) \cong \hat{c}_1 \sin(\pi x / L) + \hat{c}_2 \sin(2\pi x / L) + \cdots + \hat{c}_m \sin(m\pi x / L)$$

Hiawatha

Verse 13