Regression Analysis

Least-Squares Regression

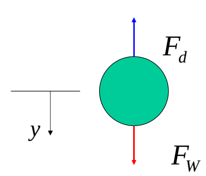
Ch. 17

Lecture Objectives

- To review some basic statistical definitions
- To understand why engineers use curve fitting so extensively
- To understand how and when to appropriately apply different curve fitting techniques

Recall Our Newton's 2nd Law Mathematical Model

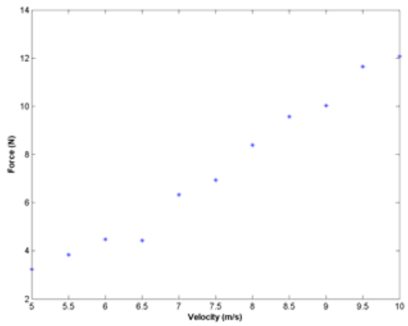
$$a = \frac{\sum F_y}{m} = \frac{F_d + F_W}{m}$$



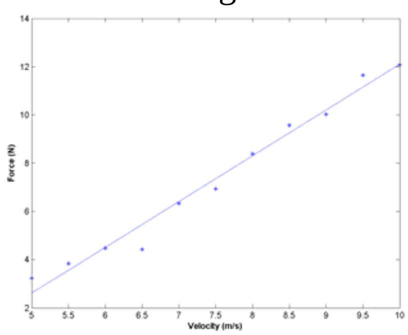
$$F_d = -\frac{1}{2}\rho v^2 C_d A$$

$$F_w = mg$$

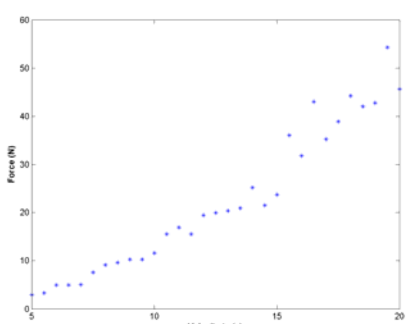
Drag Example



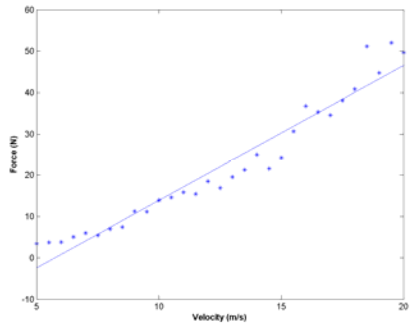
Linear Regression



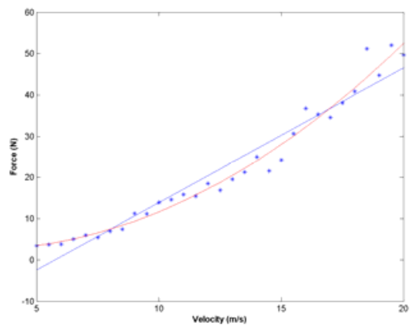
Larger Velocity Range



Larger Velocity Range
Linear Model



Larger Velocity Range
2^{nd} order polynomial



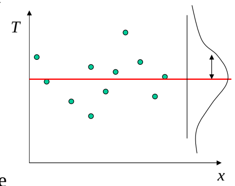
Regression Analysis

- <u>Simple Statistics (PT5 in Text)</u>
- <u>Curve Fitting</u> –
  - Data usually available at discrete points from measurements, tables of data, etc. We may want to
    1. Obtain estimates between data points
    2. Obtain a simplified version of a complicated function
- <u>Least Squares Regression</u> – desire a curve (equation) to follow the general trend of the data (Model).
- <u>Interpolation</u> –
  - Use with precise data
  - Curves that intersect all of the data points

Simple Statistics Background – PT5

1. Arithmetic Mean: (measure of the central tendency of the distribution)

$$\bar{T} = \frac{\sum_{i=1}^{n} T_i}{n}$$



2. Standard Deviation: measure of the spread of the data

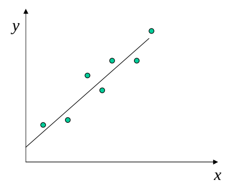
$$s_T = \sqrt{\frac{\sum_{i=1}^{n}\left(T_i - \bar{T}\right)}{n-1}}$$

Variance

$$s_T^2 = \frac{\sum_{i=1}^{n}\left(T_i - \bar{T}\right)}{n-1}$$

Least Squares Regression

- Used when there is error associated with data
- Desire a general trend of the data
- Least Squares Regression- method to determine the best fit of an equation to data.

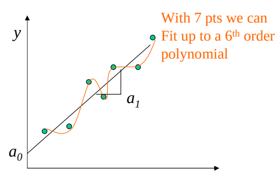


Straight Line – Simplest Approximation

$$y = a_0 + a_1 x + e$$

Least Squares Regression

- Used when there is error associated with data
- Desire a general trend of the data
- Least Squares Regression- method to determine the best fit of an equation to data.



Straight Line – Simplest Approximation

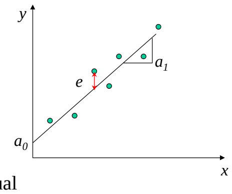
$$y = a_0 + a_1 x + e$$

Least Squares Regression

$$y = a_0 + a_1 x + e$$

$$e = y - (a_0 + a_1 x) \quad \text{- Error}$$

Residual between measured y and the calculated y with the linear models



"Sum of the Squares" of the residual

$$S_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_{i,measured} - y_{i,model} \right]^2$$

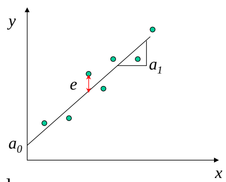
$$S_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i) \right]^2$$

Least Squares Regression – Best Fit

$$y = a_0 + a_1 x + e$$

$$e = y - (a_0 + a_1 x)$$

Residual between measured y and the calculated y with the linear models



"Sum of the Squares" of the residual

$$S_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_{i,measured} - y_{i,model} \right]^2$$

$$S_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i) \right]^2 \longleftarrow \text{Best Fit – Mininimize } S_r$$

Least Squares Fit of a Straight Line

$$S_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i) \right]^2$$

1. Minimize S_r - Differentiate S_r w.r.t. each coefficient a_i
2. Set each equation equal to zero
3. Solve n equations for a_n unknowns

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial}{\partial a_0}\left(\sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i) \right]^2 \right) = \sum_{i=1}^{N} \frac{\partial}{\partial a_0} \left[y_i - (a_0 + a_1 x_i) \right]^2$$

Linear Regression – Error Quantification

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$y = a_0 + a_1 x$$

- Method Produces “best” fit
- All other lines will result in larger S_r values
- How good is the linear regression method? Fit Statistics

Linear Regression – Error Quantification

Standard Deviation – most common measure of data spread

$$S_y = \left(\frac{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}{N-1} \right)^{1/2}$$

Total Sum of the Squares: $S_t = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$

$$S_y = \left(\frac{S_t}{N-1} \right)^{1/2}$$

The greater the spread about the mean, the greater S_y

Linear Regression – Error Quantification

For Linear Regression

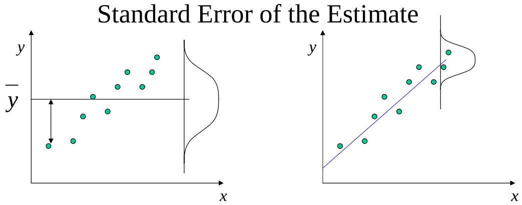
$$S_r = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i\right) \right]^2$$

Standard Error of the Estimate: the error for a predicted value of *y* Corresponding to a particular value of x.

$$S_{y/x} = \left(\frac{S_r}{N-2} \right)^{1/2}$$

Divide by (N-2) because we have lost 2 degrees of freedom from data derived estimates of a_o and a_1

Standard Error of the Estimate



- S_y is measure of the spread of the data about the mean
- $S_{y/x}$ is measure of the spread data about the regression line

We would like to quantify the "Goodness of Fit" and compare various fits.

Coefficient of Determination – r^2

$$r^2 = \frac{S_t - S_r}{S_t} = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Residual relative to the mean: $S_t = \sum \left(y_i - \bar{y}\right)^2$

Residual relative to the regression line: $S_r = \sum \left(y_i - \left(a_o + a_1 x_i\right)\right)^2$

Error Reduction due to describing the data with a straight line rather than the average: $S_t - S_r$

<u>Correlation Coefficient – r:</u> often given in commercial packages to express "goodness of fit"

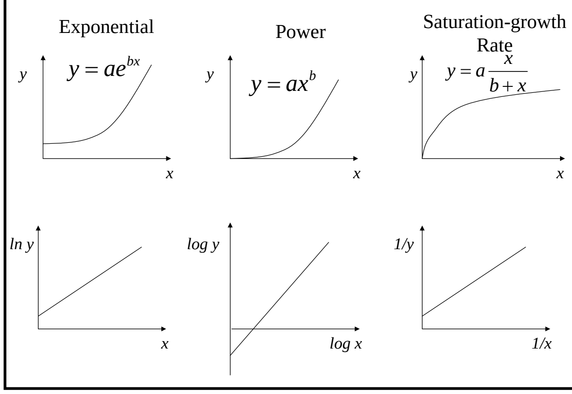
Perfect Fit: $S_r = 0$ $\quad r = 1$

NOTE: Always plot data & regression line to visually check goodness of fit.

Non-Linear Relationships - Linearization

- Some non-linear relationships can be transformed in to linear forms
- This way linear regression analysis can be used
- Transform back to get our fit

Linearization of Non-linear Relationships



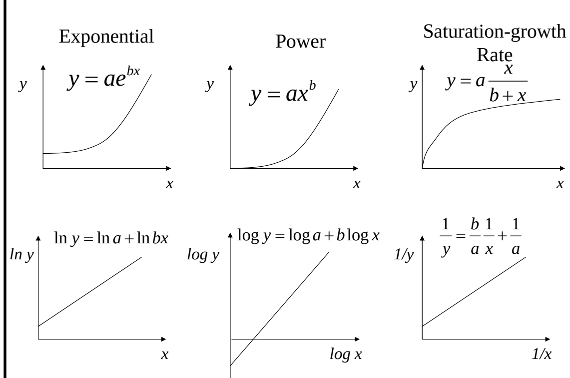
Non-Linear Relationships - Linearization

1. Exponential Model $y = ae^{bx}$

2. Power Model $y = ax^b$

3. Saturation Growth Rate $y = a\dfrac{x}{b+x}$

Linearization of Non-linear Relationships



Polynomial Regression

- Extend Linear Regression to higher order
- Consider a 2nd order polynomial:

$$y = a_0 + a_1 x + a_2 x^2 + e$$

$$e_i = y_i - \left(a_0 + a_1 x_i + a_1 x_i^2\right) \longrightarrow \text{Residual}$$

- Follow the same procedure as linear regression

$$S_r = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2\right)\right]^2$$

- Take derivatives of S_r w.r.t. a_o, a_1, a_2 and set equal to zero

Polynomial Regression

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2\right)\right] = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2\right)\right] x_i = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2\right)\right] x_i^2 = 0$$

- Rearrange equations and use $\sum a_o = n a_o$

$$\left.\begin{aligned} \sum y_i &= n a_0 + a_1 \sum x_i + a_2 \sum x_i^2 \\ \sum x_i y_i &= a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 \\ \sum x_i^2 y_i &= a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 \end{aligned}\right\} \text{3 equations, 3unknowns}$$

Polynomial Regression

Solve the 3x3 matrix using Gauss-Elimination, Gauss-Jordan, etc.

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{Bmatrix}$$

Polynomial Regression – Error Quantification

For 2nd order Polynomial Regression

$$S_r = \sum_{i=1}^{N}\left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2\right)\right]^2$$

Standard Error of the Estimate:

$$S_{y/x} = \left(\frac{S_r}{N-3}\right)^{1/2}$$

Divide by (N-3) because we have lost 3 degrees of freedom from data derived estimates of a_o , a_1 , a_2

Polynomial Regression – Error Quantification

For an Mth order Polynomial Regression

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e$$

$$S_r = \sum_{i=1}^{N}\left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x_i^m\right)\right]^2$$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{m+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^3 & \cdots & \sum x_i^{m+2} \\ . & . & . & & . \\ \sum x_i^m & \sum x_i^{m+1} & \sum x_i^{m+2} & \cdots & \sum x_i^{2m} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ a_m \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ . \\ \sum x_i^m y_i \end{Bmatrix}$$

m+1equations
m+1 unknowns

Standard Error of the Estimate: $S_{y/x} = \left(\dfrac{S_r}{N-(m+1)}\right)^{1/2}$

Polynomial Regression – Matlab Example built in function

Multiple Linear Regression

- Extend Linear Regression such that y is a function of 2 or more variables

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$

y is a linear function of x1 and x_2

Result is a "Regression Plane"

- Example: Heat Transfer Nusselt Number correlation:

$$Nu = c\,\mathrm{Re}^m\,\mathrm{Pr}^n$$

$$\underbrace{\log Nu}_{y} = \underbrace{\log c}_{a_0} + \underbrace{m \log \mathrm{Re}}_{a_1 x_1} + \underbrace{n \log \mathrm{Pr}}_{a_2 x_2}$$

Multiple Linear Regression

- Procedure is the same
- Consider a 2 variables x_1 and x_2:

$$S_r = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_{1i} + a_2 x_{2i} \right) \right]^2$$

- Take derivatives of S_r w.r.t. a_o, a_1, a_2 and set equal to zero

Multiple Linear Regression

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_{2i} \right) \right] = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_{2i} \right) \right] x_{1i} = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_{2i} \right) \right] x_{2i} = 0$$

- Rearrange equations and use $\sum a_o = n a_o$

$$\left.\begin{aligned} \sum y_i &= n a_0 + a_1 \sum x_{1i} + a_2 \sum x_{2i} \\ \sum x_{1i} y_i &= a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} x_{2i} \\ \sum x_{2i} y_i &= a_0 \sum x_{2i} + a_1 \sum x_{1i} x_{2i} + a_2 \sum x_{2i}^2 \end{aligned}\right\} \text{3 equations, 3unknowns}$$

Multiple Linear Regression

Solve the 3x3 matrix using Gauss-Elimination, Gauss-Jordan, etc.

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{Bmatrix}$$

Standard Error of the Estimate:

$$S_{y/x} = \left(\frac{S_r}{N-3} \right)^{1/2}$$

Standard Error of the Estimate for an M-dimensional problem:

$$S_{y/x} = \left(\frac{S_r}{N-(m+1)} \right)^{1/2}$$

General Least Squares Method

- Linear, polynomial & multiple linear regression models can be written in the following form:

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

- Where z_m's represent functions for each type of model

	z_0	z_1	z_2	z_3
1st order poly	1	x	-	-
2nd order poly	1	x	x^2	-
3rd order poly	1	x	x^2	x^3
2nd order multi-var	1	x_1	x_2	-

General Least Squares Method

- In matrix form the general model is:

$$\{Y\} = [Z]\{A\} + \{E\}$$

$$\begin{bmatrix} z_{01} & z_{11} & z_{21} & \cdots & z_{m1} \\ z_{02} & z_{12} & z_{22} & \cdots & z_{m2} \\ z_{03} & z_{13} & z_{23} & \cdots & z_{m3} \\ \cdot & \cdot & \cdot & & \cdot \\ z_{0n} & z_{1n} & z_{2n} & \cdots & z_{mn} \end{bmatrix}$$

m = # variables
n = # data points

Calculated values of the z functions at measured locations of the independent variable

General Least Squares Method

- Typically *n > m + 1* → so z is NOT square

$$\{Y\}=\begin{Bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{Bmatrix} \quad \{A\}=\begin{Bmatrix} a_0 \\ a_1 \\ . \\ a_m \end{Bmatrix} \quad \{E\}=\begin{Bmatrix} e_1 \\ e_2 \\ . \\ e_n \end{Bmatrix}$$

General sum of the squares is:

$$S_r = \sum_{i=1}^{N}\left[y_i - \sum_{j=0}^{m} a_j z_{ji} \right]^2$$

Again, minimize by taking partial derivatives with respect to The *a*'s and setting equal to zero. This results in the following equations

General Least Squares Method

- Normal Equations – which relate exactly to our previous specific examples

$$[Z]^T[Z]\{A\} = \{[Z]^T\{Y\}\}$$

- Solve using a matrix inversion technique (LU decomposition) for the matrix {A}

$$\{A\} = [[Z]^T[Z]]^{-1}\{[Z]^T\{Y\}\}$$

Non-linear Regression

- For models that DO NOT fit the general least squares equation
- Ex, $y = a_0\left(1-e^{-a_1 x}\right)+e$
- Solve using an iterative method that starts by using a Taylor series to express the non-linear equation in an approximate linear form.