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
Regression Analysis

- **Simple Statistics (PT5 in Text)**
- **Curve Fitting** –
  - 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
- **Least Squares Regression** – desire a curve (equation) to follow the general trend of the data (Model).
- **Interpolation** –
  - 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)
   \[
   \bar{y} = \frac{\sum y_i}{n}
   \]

2. **Standard Deviation**: measure of the spread of the data
   \[
   s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}
   \]

---

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**
\[
Y = a_0 + a_1x + \epsilon
\]

With 7 pts we can fit up to a 6th order polynomial

---

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**
\[
Y = a_0 + a_1x + \epsilon
\]
Least Squares Regression

\[ y = a_0 + a_1 x + e \]
\[ e = y - (a_0 + a_1 x) \] - Error

Residual between measured \( y \) and the calculated \( y \) with the linear models

“Sum of the Squares” of the residual

\[ S_y = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_{i, measured} - y_{i, model}]^2 \]

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

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

\[ S_y = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_{i, measured} - y_{i, model}]^2 \]

\[ S_y = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - (a_0 + a_1 x_i)]^2 \] → Best Fit – Minimize \( S_y \)

Least Squares Fit of a Straight Line

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

1. Minimize \( S_y \) - Differentiate \( S_y \) w.r.t. each coefficient \( a_j \)
2. Set each equation equal to zero
3. Solve \( n \) equations for \( a_j \) unknowns

\[ \frac{\partial S_y}{\partial a_0} = \frac{\partial}{\partial a_0} \left( \sum_{i=1}^{N} [y_i - (a_0 + a_1 x_i)]^2 \right) = \sum_{i=1}^{N} \frac{\partial}{\partial a_0} [y_i - (a_0 + a_1 x_i)]^2 \]
Linear Regression – Error Quantification

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

Standard Deviation – most common measure of data spread

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

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

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

The greater the spread about the mean, the greater $S_y$

For Linear Regression

$S_e = \sum_{i=1}^{N} [y_i - (a_0 + a_1x_i)]^2$

Standard Error of the Estimate: the error for a predicted value of $y$

Corresponding to a particular value of $x$.

$S_{y|x} = \left[ \frac{S_e}{N-2} \right]^{1/2}$

Divide by (N-2) because we have lost 2 degrees of freedom from data derived estimates of $a_0$ 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 – $r^2$

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

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

Residual relative to the regression line: $S_r = \sum (y_i - (a_0 + a_1 x_i))^2$

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

Correlation Coefficient – $r$, often given in commercial packages to express “goodness of fit”

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

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

Exponential

\[ y = ae^{bx} \]

Power

\[ y = ax^b \]

Saturation-growth Rate

\[ y = a \frac{x}{b + x} \]

ln y

log y

1/y

Non-Linear Relationships - Linearization

1. Exponential Model \( y = ae^{bx} \)

2. Power Model \( y = ax^b \)

3. Saturation Growth Rate \( y = a \frac{x}{b + x} \)
Polynomial Regression

- Extend Linear Regression to higher order
- Consider a 2nd order polynomial:
  \[ y = a_0 + a_1 x + a_2 x^2 + e \]
  \[ e_i = y_i - \left( a_0 + a_1 x_i + a_2 x_i^2 \right) \quad \text{Residual} \]
- Follow the same procedure as linear regression
  \[ S_y = \sum_{i=1}^{N} \left( y_i - \left( a_0 + a_1 x_i + a_2 x_i^2 \right) \right)^2 \]
- Take derivatives of \( S_y \) w.r.t. \( a_0, a_1, a_2 \) and set equal to zero

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

\[
\begin{bmatrix}
  \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum y_i \\
  \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum x_i y_i \\
  \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \sum x_i^2 y_i
\end{bmatrix} \begin{bmatrix}
  a_0 \\
  a_1 \\
  a_2
\end{bmatrix} = \begin{bmatrix}
  \sum y_i \\
  \sum x_i y_i \\
  \sum x_i^2 y_i
\end{bmatrix}
\]
Polynomial Regression – Error Quantification

For 2nd order Polynomial Regression:

\[ S_e = \sum_{j=1}^{N} \left[ y_j - \left( a_0 + a_1 x_j + a_2 x_j^2 \right) \right]^2 \]

Standard Error of the Estimate:

\[ S_{y,x} = \left( \frac{S_e}{N-3} \right)^{1/2} \]

Divide by (N-3) because we have lost 3 degrees of freedom from data derived estimates of \(a_0, a_1, a_2\)

Polynomial Regression – Error Quantification

For an Mth order Polynomial Regression:

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

\[ S_e = \sum_{j=1}^{N} \left[ y_j - \left( a_0 + a_1 x_j + a_2 x_j^2 + \cdots + a_m x_j^m \right) \right]^2 \]

\[ \begin{bmatrix}
\sum x_i \\
\sum x_i^2 \\
\vdots \\
\sum x_i^m
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
a_2 \\
a_m
\end{bmatrix} = \begin{bmatrix}
\sum y_i \\
\sum x_i y_i \\
\vdots \\
\sum x_i^m y_i
\end{bmatrix} \text{ (m+1 equations, m+1 unknowns)}
\]

Standard Error of the Estimate:

\[ S_{y,x} = \left( \frac{S_e}{N-(m+1)} \right)^{1/2} \]

Polynomial Regression – Matlab Example

built in function
Multiple Linear Regression

• Extend Linear Regression such that \( y \) is a function of 2 or more variables

\[
y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + e
\]

\( y \) is a linear function of \( x_1 \) and \( x_2 \)

**Result is a “Regression Plane”**

• Example: Heat Transfer Nusselt Number correlation:

\[
Nu = c \cdot Re^{\alpha_1} Pr^{\alpha_2}
\]

\[
\log Nu = \log c + m \log Re + n \log Pr
\]

\[
\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{x}_1 + \hat{\alpha}_2 \hat{x}_2
\]

Multiple Linear Regression

• Procedure is the same

• Consider a 2 variables \( x_1 \) and \( x_2 \):

\[
S_e = \sum_{i=1}^{N} [y_i - (\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})]^2
\]

• Take derivatives of \( S_e \) w.r.t. \( \alpha_0, \alpha_1, \alpha_2 \) and set equal to zero

\[
\frac{\partial S_e}{\partial \alpha_0} = \sum_{i=1}^{N} [y_i - (\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})] = 0
\]

\[
\frac{\partial S_e}{\partial \alpha_1} = \sum_{i=1}^{N} [y_i - (\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})] x_{i1} = 0
\]

\[
\frac{\partial S_e}{\partial \alpha_2} = \sum_{i=1}^{N} [y_i - (\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})] x_{i2} = 0
\]

• Rearrange equations and use \( \sum a_i = ma_i \)

\[
\begin{align*}
\sum y_i &= m \alpha_0 + \alpha_1 \sum x_{i1} + \alpha_2 \sum x_{i2} \\
\sum x_{i}y_i &= \alpha_0 \sum x_{i} + \alpha_1 \sum x_{i}^2 + \alpha_2 \sum x_{i} x_{i2} \\
\sum x_{i}x_{i2} &= \alpha_0 \sum x_{i} x_{i1} + \alpha_1 \sum x_{i} x_{i1} + \alpha_2 \sum x_{i} x_{i1}^2
\end{align*}
\]

3 equations, 3 unknowns
### Multiple Linear Regression

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

\[
\begin{bmatrix}
\sum x_1 & \sum x_1 y_1 & \sum x_1 y_2 \\
\sum x_2 & \sum x_2 y_1 & \sum x_2 y_2 \\
\sum x_3 & \sum x_3 y_1 & \sum x_3 y_2
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_3
\end{bmatrix}
= 
\begin{bmatrix}
\sum x_1 y_1 \\
\sum x_2 y_1 \\
\sum x_3 y_1
\end{bmatrix}
\]

**Standard Error of the Estimate:**

\[
S_{/\hat{y}} = \left( \frac{S}{N-3} \right)^{1/2}
\]

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

\[
S_{/\hat{y}} = \left( \frac{S}{N-(m+1)} \right)^{1/2}
\]

### General Least Squares Method

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

\[
y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + \varepsilon
\]

- Where \( z_m \)'s represent functions for each type of model.

<table>
<thead>
<tr>
<th>Order</th>
<th>Function 1</th>
<th>Function 2</th>
<th>Function 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>( x )</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2nd</td>
<td>( x ) ( x^2 )</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>( x ) ( x^2 ) ( x^3 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi</td>
<td>( x_j ) ( x_2 )</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### General Least Squares Method

- In matrix form the general model is:

\[
[\mathbf{y}] = [\mathbf{Z}] \mathbf{a} + \mathbf{\varepsilon}
\]

\[
\begin{bmatrix}
z_{01} & z_{02} & \ldots & z_{0m} \\
z_{11} & z_{12} & \ldots & z_{1m} \\
z_{21} & z_{22} & \ldots & z_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
z_{n1} & z_{n2} & \ldots & z_{nm}
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_3 \\
\vdots \\
a_m
\end{bmatrix}
= 
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n
\end{bmatrix}
\]

\( m = \# \text{ variables} \)
\( n = \# \text{ data points} \)

Calculated values of the \( z \) functions at measured locations of the independent variable.
General Least Squares Method

- Typically \( n > m + 1 \) \( \rightarrow \) so \( z \) is NOT square

\[
\begin{pmatrix}
 y_1 \\
 y_2 \\
 \vdots \\
 y_n
\end{pmatrix} =
\begin{pmatrix}
 a_0 \\
 a_1 \\
 \vdots \\
 a_m
\end{pmatrix} +
\begin{pmatrix}
 \epsilon_1 \\
 \epsilon_2 \\
 \vdots \\
 \epsilon_n
\end{pmatrix}
\]

General sum of the squares is:

\[
S_e = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{m} \beta_j x_{ij}]^2
\]

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

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

\[
[Z'Z][A] = [Z'Y]
\]

- Solve using a matrix inversion technique (LU decomposition) for the matrix \( [A] \)

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

Non-linear Regression

- For models that DO NOT fit the general least squares equation

- Ex,

\[
y = a_0 (1 - e^{-ax}) + e
\]

- Solve using an iterative method that starts by using a Taylor series to express the non-linear equation in an approximate linear form.