Introduction to MATLAB

Violeta Ivanova, Ph.D.
Office for Educational Innovation & Technology

violeta@mit.edu
http://web.mit.edu/violeta/www
Introduction to MATLAB: Data Analysis and Statistics

Topics

- MATLAB Interface and Basics
- Calculus, Linear Algebra, ODEs
- Graphics and Visualization
- Basic Programming
- Programming Practice
- Statistics and Data Analysis
Resources

- **Class materials**
  - Previous sessions: InterfaceBasics, Graphics
  - This session: Statistics <.zip, .tar>

- **Mathematical Tools at MIT**
  http://web.mit.edu/ist/topics/math
MATLAB Help Browser

- MATLAB
  - Data Analysis
    - Preparing Data for Analysis
    - Data Fitting Using Linear Regression

- Curve Fitting Toolbox
  - Fitting Data

- Statistics Toolbox
  - Descriptive Statistics
  - Linear Models
  - Hypothesis Tests
  - Statistical Plots
MATLAB Data Analysis

Preparing Data
Correlation
Basic Fitting
Data Input / Output

- **Import Wizard** for data import
  
  File->Import Data ...

- File input with `load`
  
  ```matlab
  B = load('datain.txt')
  ```

- File output with `save`
  
  ```matlab
  save('dataout', 'A', '-ascii')
  ```
Missing Data

- Removing missing data
  - Removing NaN elements from vectors
    ```matlab
    >> x = x(~isnan(x))
    ```
  - Removing rows with NaN from matrices
    ```matlab
    >> X(any(isnan(X),2),:) = []
    ```

- Interpolating missing data
  ```matlab
  YI = interp1(X, Y, XI, 'method')
  ```
  Methods: 'spline', 'nearest', 'linear', ...
Correlation

- Definition
  Tendency of two variables to increase or decrease together.

- Measure
  Pearson product-moment coefficient

\[ \rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \]
Correlation Example

- Import Data: `cancersmoking.dat`
- Correlation coefficient & confidence interval

```matlab
>> [R, P] = corrcoef(X);
>> [i, j] = find(P < 0.05);
```

```
X =

18.2000  17.0500  6.1500
25.8200  19.8000  6.6100
18.2400  15.9800  6.9400
28.6000  22.0700  7.0600
31.1000  22.8300  7.2000
33.6000  24.5500  6.4500
40.4600  27.2700  7.0800
28.2700  23.5700  6.0700
20.1000  13.5800  6.6200
27.9100  22.8000  7.2700
26.1800  20.3000  7.0000
22.1200  16.5900  7.6900
21.8400  16.8400  7.4200
22.0000  17.7100  6.7100
```

```
>> [r,p]=corrcoef(X)

r =

    1.0000    0.6974   -0.0685
    0.6974    1.0000   -0.1516
   -0.0685   -0.1516    1.0000

p =

    1.0000    0.0000    0.6587
    0.0000    1.0000    0.3260
    0.6587    0.3260    1.0000
```
Data Statistics

**Figure Editor:** smokecancer.fig

Tools -> Data Statistics
Basic Fitting

Figure Editor: Tools -> Basic Fitting ...

![Graph showing the relationship between smoking and cancer incidence. The equation $y = 3x + 20$ is displayed, where $y$ represents cancer rate per 100,000 population and $x$ represents cigarettes per capita. The graph includes data points for different types of cancer.]
Statistics Toolbox

Probability Distributions
Descriptive Statistics
Linear & Nonlinear Models
Hypothesis Tests
Statistical Plots
Descriptive Statistics

- Central tendency
  \[
  \begin{align*}
  \text{mean}(X) & = \text{mean}(X) \\
  \text{geomean}(X) & = \text{geomean}(X) \\
  \text{median}(X) & = \text{median}(X) \\
  \text{mode}(X) & = \text{mode}(X)
  \end{align*}
  \]

- Dispersion
  \[
  \begin{align*}
  \text{std}(X) & = \text{std}(X) \\
  \text{var}(X) & = \text{var}(X)
  \end{align*}
  \]
Probability Distributions

- Probability density functions
  \[
  \text{>> } Y = \text{exppdf}(X, \text{mu})
  \]
  \[
  \text{>> } Y = \text{normpdf}(X, \text{mu}, \text{sigma})
  \]

- Cumulative density functions
  \[
  \text{>> } Y = \text{expcdf}(X, \text{mu})
  \]
  \[
  \text{>> } Y = \text{normcdf}(X, \text{mu}, \text{sigma})
  \]

- Parameter estimation
  \[
  \text{>> } m = \text{expfit}(\text{data})
  \]
  \[
  \text{>> } [m, s] = \text{normfit}(\text{data})
  \]
Statistical Plots

$$\text{>> bp = boxplot}(X, \text{group})$$
Polynomial Fitting Tool

>> polytool(X, Y)
Distribution Fitting Tool

>> dfittool
Linear Models

Definition:

\[ y = X\beta + \varepsilon \]

- \( y \): \( n \times 1 \) vector of observations
- \( X \): \( n \times p \) matrix of predictors
- \( \beta \): \( p \times 1 \) vector of parameters
- \( \varepsilon \): \( n \times 1 \) vector of random disturbances
Linear Regression

- Multiple linear regression
  
  ```matlab
  >> [B, Bint, R, Rint, stats] = regress(y, X)
  ```
  
  - \( B \): vector of regression coefficients
  - \( Bint \): matrix of 95% confidence intervals for \( B \)
  - \( R \): vector of residuals
  - \( Rint \): intervals for diagnosing outliers
  - \( stats \): vector containing \( R^2 \) statistic etc.

- Residuals plot
  
  ```matlab
  >> rcoplot(R, Rint)
  ```
Hypothesis Testing

- Definition: use of statistics to determine the probability that a given hypothesis is true.
  - Null hypothesis (observations are the result of pure chance) and alternative hypothesis.
  - Test statistic to assess truth of null hypothesis.
  - P-value: probability of test statistic to be significant if null hypothesis were true.
  - Comparison of P-value to acceptable $\alpha$-value.
Analysis of Variance (ANOVA)

- One-way ANOVA

\[
\text{>> } \text{anova1}(X, \text{group})
\]
Multiple Comparisons

```
>> [p, tbl, stats] = anovan(X, group)
>> [c, m] = multcompare(stats)
```
More Built-In Functions

- Two-way ANOVA
  \[ [P, \text{tbl}, \text{stats}] = \text{anova2}(X, \text{reps}) \]

- Other hypothesis tests
  \[ H = ttest(X) \]
  \[ H = \text{lillietest}(X) \]
Data Analysis Exercises

Exercise One: `dataanalysis.m`, `rfid.dat`, `barcode.dat`
- Correlation coefficient
- Hypothesis testing
- Statistical plots
- ANOVA

*Follow instructions in the m-file …*
Curve Fitting Toolbox

Curve Fitting Tool
Goodness of Fit
Analyzing a Fit
Fourier Series Fit
Curve Fitting Tool

>> cftool

[Graph showing data points and an exponential fit]

World population (in billions) vs. year
Exponential fit
Goodness of Fit Statistics

<table>
<thead>
<tr>
<th>Name</th>
<th>Data set</th>
<th>Type</th>
<th>SSE</th>
<th>R-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>exponential fit</td>
<td>Nbil vs. year</td>
<td>Exponential</td>
<td>0.1572972880</td>
<td>0.9973106593</td>
</tr>
<tr>
<td>gaussian fit</td>
<td>Nbil vs. year</td>
<td>Gaussian</td>
<td>0.0263778905</td>
<td>0.9995490123</td>
</tr>
</tbody>
</table>
Analyzing a Fit

- Evaluate fit at Xi
- Prediction or confidence bounds:
  - None
  - For function
  - For new observation
- Level: 95%
- 1st derivative at Xi
- 2nd derivative at Xi
- Integrate to Xi
  - Start from min(Xi)
  - Start from
- Plot results
- Plot data set: Nbil vs. year

Graph showing analysis of fit "exponential fit" for dataset 'Nbil vs. year':
- Exponential fit
- 95% prediction bounds
- Nbil vs. year
Fourier Series Fit

![Fourier Series Fit](image-url)
Data Analysis Exercises

Exercise Two: regression.m, worlddata.dat, star.txt
- Linear regression
- Polynomial fitting
- Probability density function fitting
- Goodness of Fit

Follow instructions in the m-file …