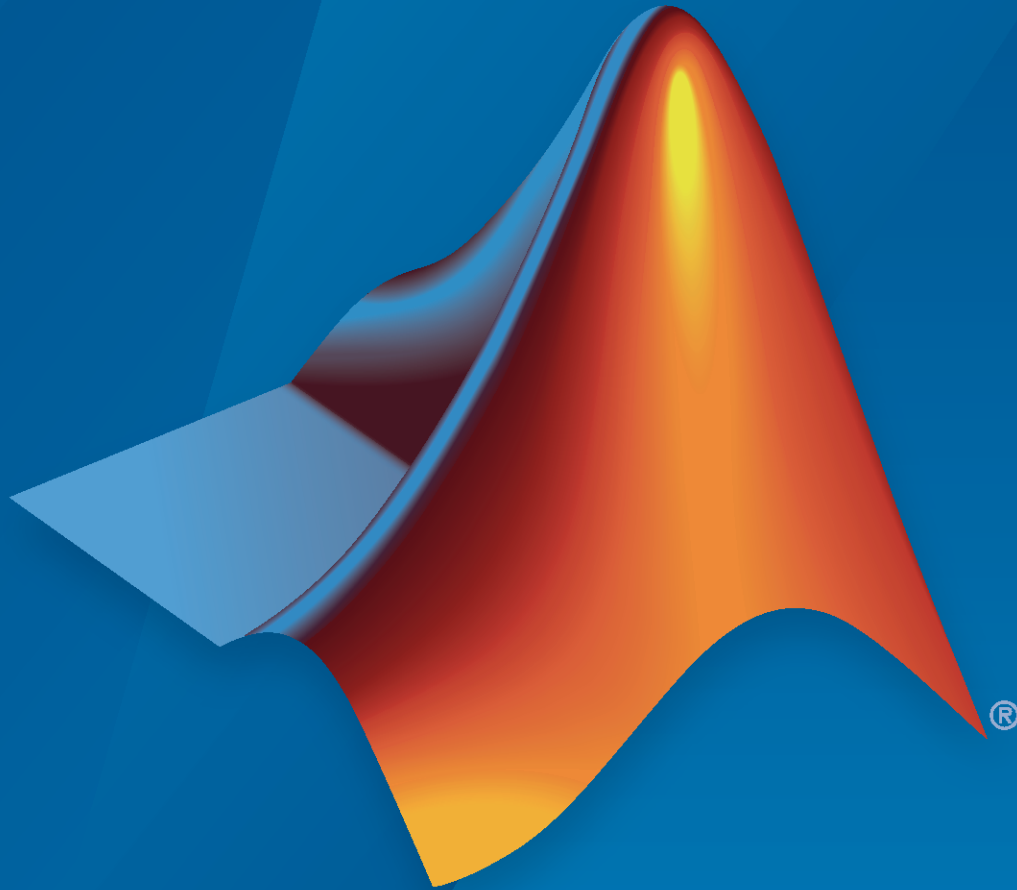


**MATLAB®**

Data Analysis



**MATLAB®**

R2022a



## How to Contact MathWorks



Latest news: [www.mathworks.com](http://www.mathworks.com)  
Sales and services: [www.mathworks.com/sales\\_and\\_services](http://www.mathworks.com/sales_and_services)  
User community: [www.mathworks.com/matlabcentral](http://www.mathworks.com/matlabcentral)  
Technical support: [www.mathworks.com/support/contact\\_us](http://www.mathworks.com/support/contact_us)



Phone: 508-647-7000



The MathWorks, Inc.  
1 Apple Hill Drive  
Natick, MA 01760-2098

*MATLAB® Data Analysis*

© COPYRIGHT 2005–2022 by The MathWorks, Inc.

The software described in this document is furnished under a license agreement. The software may be used or copied only under the terms of the license agreement. No part of this manual may be photocopied or reproduced in any form without prior written consent from The MathWorks, Inc.

FEDERAL ACQUISITION: This provision applies to all acquisitions of the Program and Documentation by, for, or through the federal government of the United States. By accepting delivery of the Program or Documentation, the government hereby agrees that this software or documentation qualifies as commercial computer software or commercial computer software documentation as such terms are used or defined in FAR 12.212, DFARS Part 227.72, and DFARS 252.227-7014. Accordingly, the terms and conditions of this Agreement and only those rights specified in this Agreement, shall pertain to and govern the use, modification, reproduction, release, performance, display, and disclosure of the Program and Documentation by the federal government (or other entity acquiring for or through the federal government) and shall supersede any conflicting contractual terms or conditions. If this License fails to meet the government's needs or is inconsistent in any respect with federal procurement law, the government agrees to return the Program and Documentation, unused, to The MathWorks, Inc.

### Trademarks

MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See [www.mathworks.com/trademarks](http://www.mathworks.com/trademarks) for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

### Patents

MathWorks products are protected by one or more U.S. patents. Please see [www.mathworks.com/patents](http://www.mathworks.com/patents) for more information.

## Revision History

September 2005	Online only	New for MATLAB Version 7.1 (Release 14SP3)
March 2006	Online only	Revised for MATLAB Version 7.2 (Release 2006a)
September 2006	Online only	Revised for MATLAB Version 7.3 (Release 2006b)
March 2007	Online only	Revised for MATLAB Version 7.4 (Release 2007a)
September 2007	Online only	Revised for MATLAB Version 7.5 (Release 2007b)
March 2008	Online only	Revised for MATLAB Version 7.6 (Release 2008a)
October 2008	Online only	Revised for MATLAB Version 7.7 (Release 2008b)
March 2009	Online only	Revised for MATLAB 7.8 (Release 2009a)
September 2009	Online only	Revised for MATLAB 7.9 (Release 2009b)
March 2010	Online only	Revised for MATLAB 7.10 (Release 2010a)
September 2010	Online only	Revised for MATLAB Version 7.11 (R2010b)
April 2011	Online only	Revised for MATLAB Version 7.12 (R2011a)
September 2011	Online only	Revised for MATLAB Version 7.13 (R2011b)
March 2012	Online only	Revised for MATLAB Version 7.14 (R2012a)
September 2012	Online only	Revised for MATLAB Version 8.0 (R2012b)
March 2013	Online only	Revised for MATLAB Version 8.1 (R2013a)
September 2013	Online only	Revised for MATLAB Version 8.2 (R2013b)
March 2014	Online only	Revised for MATLAB Version 8.3 (R2014a)
October 2014	Online only	Revised for MATLAB Version 8.4 (R2014b)
March 2015	Online only	Revised for MATLAB Version 8.5 (R2015a)
September 2015	Online only	Revised for MATLAB Version 8.6 (R2015b)
March 2016	Online only	Revised for MATLAB Version 9.0 (R2016a)
September 2016	Online only	Revised for MATLAB Version 9.1 (R2016b)
March 2017	Online only	Revised for MATLAB Version 9.2 (R2017a)
September 2017	Online only	Revised for MATLAB Version 9.3 (R2017b)
March 2018	Online only	Revised for MATLAB Version 9.4 (R2018a)
September 2018	Online only	Revised for MATLAB Version 9.5 (R2018b)
March 2019	Online only	Revised for MATLAB Version 9.6 (R2019a)
September 2019	Online only	Revised for MATLAB Version 9.7 (R2019b)
March 2020	Online only	Revised for MATLAB Version 9.8 (R2020a)
September 2020	Online only	Revised for MATLAB Version 9.9 (R2020b)
March 2021	Online only	Revised for MATLAB Version 9.10 (R2021a)
September 2021	Online only	Revised for MATLAB Version 9.11 (R2021b)
March 2022	Online only	Revised for MATLAB Version 9.12 (R2022a)



<b>1</b>	<b>Data Processing</b>
	<b>Importing and Exporting Data . . . . . 1-2</b>
	Importing Data into the Workspace . . . . . 1-2
	Exporting Data from the Workspace . . . . . 1-2
	<b>Plotting Data . . . . . 1-3</b>
	Introduction . . . . . 1-3
	Load and Plot Data from Text File . . . . . 1-3
	<b>Missing Data in MATLAB . . . . . 1-5</b>
	<b>Data Smoothing and Outlier Detection . . . . . 1-9</b>
	<b>Clean Messy Data and Locate Extrema Using Live Editor Tasks . . . . . 1-21</b>
	<b>Filter Data . . . . . 1-29</b>
	Filter Difference Equation . . . . . 1-29
	Moving-Average Filter of Traffic Data . . . . . 1-29
	Modify Amplitude of Data . . . . . 1-30
	<b>Smooth Data with Convolution . . . . . 1-33</b>
	<b>Detrending Data . . . . . 1-37</b>
	Introduction . . . . . 1-37
	Remove Linear Trends from Data . . . . . 1-37
	<b>Computing with Descriptive Statistics . . . . . 1-40</b>
	Functions for Calculating Descriptive Statistics . . . . . 1-40
	Example: Using MATLAB Data Statistics . . . . . 1-42
	Data Statistics . . . . . 1-42
	<b>Quantiles and Percentiles . . . . . 1-49</b>

<b>2</b>	<b>Regression Analysis</b>
	<b>Linear Correlation . . . . . 2-2</b>
	Introduction . . . . . 2-2
	Covariance . . . . . 2-2
	Correlation Coefficients . . . . . 2-3

<b>Linear Regression</b> .....	<b>2-5</b>
Introduction .....	2-5
Simple Linear Regression .....	2-5
Residuals and Goodness of Fit .....	2-9
Fitting Data with Curve Fitting Toolbox Functions .....	2-11
<b>Interactive Fitting</b> .....	<b>2-13</b>
Basic Fitting UI .....	2-13
Preparing for Basic Fitting .....	2-13
Opening the Basic Fitting UI .....	2-13
Example: Using Basic Fitting UI .....	2-14
<b>Programmatic Fitting</b> .....	<b>2-26</b>
MATLAB Functions for Polynomial Models .....	2-26
Linear Model with Nonpolynomial Terms .....	2-26
Multiple Regression .....	2-27
Programmatic Fitting .....	2-28

## Time Series Analysis

### 3

<b>What Are Time Series?</b> .....	<b>3-2</b>
<b>Time Series Objects and Collections</b> .....	<b>3-3</b>
Types of Time Series and Their Uses .....	3-3
Time Series Data Sample .....	3-3
Example: Time Series Objects and Methods .....	3-5
Time Series Constructor .....	3-12
Time Series Collection Constructor .....	3-12

# Data Processing

---

- “Importing and Exporting Data” on page 1-2
- “Plotting Data” on page 1-3
- “Missing Data in MATLAB” on page 1-5
- “Data Smoothing and Outlier Detection” on page 1-9
- “Clean Messy Data and Locate Extrema Using Live Editor Tasks” on page 1-21
- “Filter Data” on page 1-29
- “Smooth Data with Convolution” on page 1-33
- “Detrending Data” on page 1-37
- “Computing with Descriptive Statistics” on page 1-40
- “Quantiles and Percentiles” on page 1-49

## Importing and Exporting Data

In this section...
“Importing Data into the Workspace” on page 1-2
“Exporting Data from the Workspace” on page 1-2

### Importing Data into the Workspace

The first step in analyzing data is to import it into the MATLAB workspace. See “Supported File Formats for Import and Export” for information about importing data from specific file formats.

### Exporting Data from the Workspace

When you analyze your data, you might create new variables or modify imported variables. You can export variables from the MATLAB workspace to various file formats, both character-based and binary. You can, for example, create HDF and Microsoft® Excel® files containing your data. For details, see the documentation on “Supported File Formats for Import and Export”.



# Plotting Data

## In this section...

“Introduction” on page 1-3

“Load and Plot Data from Text File” on page 1-3

## Introduction

After you import data into the MATLAB workspace, it is a good idea to plot the data so that you can explore its features. An exploratory plot of your data enables you to identify discontinuities and potential outliers, as well as the regions of interest.

The MATLAB figure window displays plots. See “Types of MATLAB Plots” for a full description of the figure window. It also discusses the various interactive tools available for editing and customizing MATLAB graphics.

## Load and Plot Data from Text File

This example uses sample data in `count.dat`, a space-delimited text file. The file consists of three sets of hourly traffic counts, recorded at three different town intersections over a 24-hour period. Each data column in the file represents data for one intersection.

### Load the `count.dat` Data

Import data into the workspace using the `load` function.

```
load count.dat
```

Loading this data creates a 24-by-3 matrix called `count` in the MATLAB workspace.

Get the size of the data matrix.

```
[n,p] = size(count)
```

```
n = 24
```

```
p = 3
```

`n` represents the number of rows, and `p` represents the number of columns.

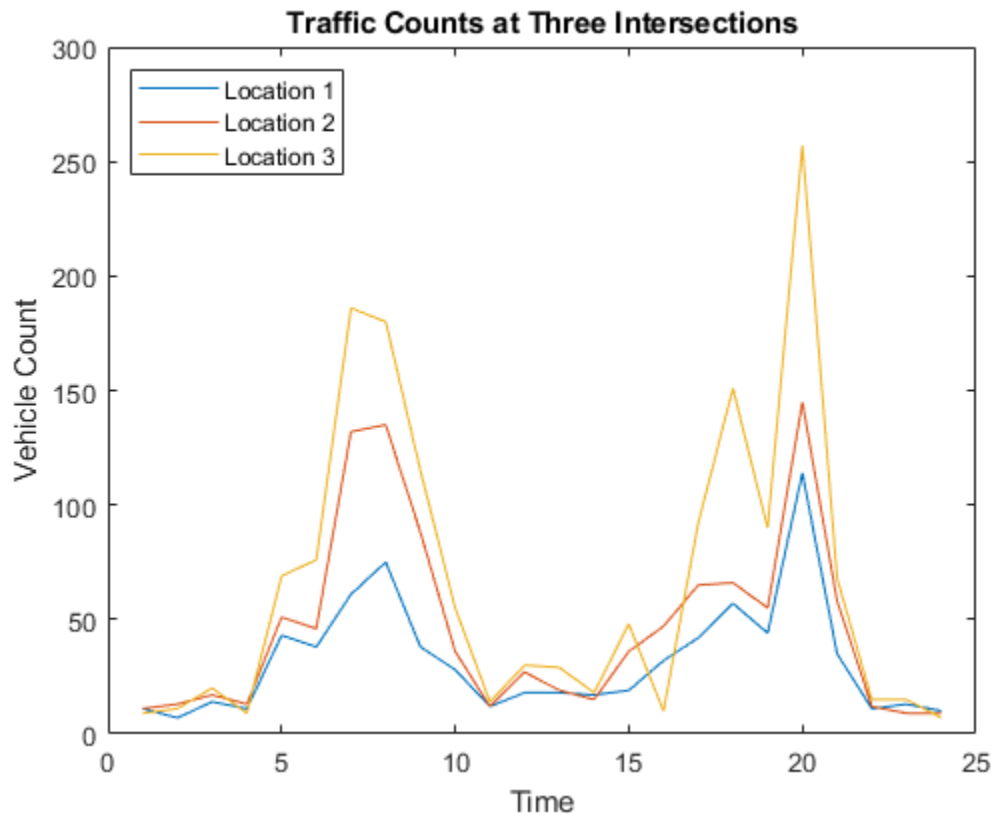
### Plot the `count.dat` Data

Create a time vector, `t`, containing integers from 1 to `n`.

```
t = 1:n;
```

Plot the data as a function of time, and annotate the plot.

```
plot(t,count),  
legend('Location 1','Location 2','Location 3','Location','NorthWest')  
xlabel('Time'), ylabel('Vehicle Count')  
title('Traffic Counts at Three Intersections')
```



### See Also

`load` | `plot` | `legend` | `xlabel` | `ylabel` | `title` | `size`

### More About

- "Types of MATLAB Plots"

## Missing Data in MATLAB

Working with missing data is a common task in data preprocessing. Although sometimes missing values signify a meaningful event in the data, they often represent unreliable or unusable data points. In either case, MATLAB® has many options for handling missing data.

### Create and Organize Missing Data

The form that missing values take in MATLAB depends on the data type. For example, numeric data types such as `double` use `NaN` (not a number) to represent missing values.

```
x = [NaN 1 2 3 4];
```

You can also use the missing value to represent missing numeric data or data of other types, such as `datetime`, `string`, and `categorical`. MATLAB automatically converts the missing value to the data's native type.

```
xDouble = [missing 1 2 3 4]
```

```
xDouble = 1x5
```

```
NaN    1    2    3    4
```

```
xDatetime = [missing datetime(2014,1:4,1)]
```

```
xDatetime = 1x5 datetime
```

```
NaT          01-Jan-2014    01-Feb-2014    01-Mar-2014    01-Apr-2014
```

```
xString = [missing "a" "b" "c" "d"]
```

```
xString = 1x5 string
```

```
<missing>    "a"    "b"    "c"    "d"
```

```
xCategorical = [missing categorical({'cat1' 'cat2' 'cat3' 'cat4'})]
```

```
xCategorical = 1x5 categorical
```

```
<undefined>    cat1    cat2    cat3    cat4
```

A data set might contain values that you want to treat as missing data, but are not standard MATLAB missing values in MATLAB such as `NaN`. You can use the `standardizeMissing` function to convert those values to the standard missing value for that data type. For example, treat 4 as a missing `double` value in addition to `NaN`.

```
xStandard = standardizeMissing(xDouble,[4 NaN])
```

```
xStandard = 1x5
```

```
NaN    1    2    3    NaN
```

Suppose you want to keep missing values as part of your data set but segregate them from the rest of the data. Several MATLAB functions enable you to control the placement of missing values before further processing. For example, use the `'MissingPlacement'` option with the `sort` function to move `NaNs` to the end of the data.

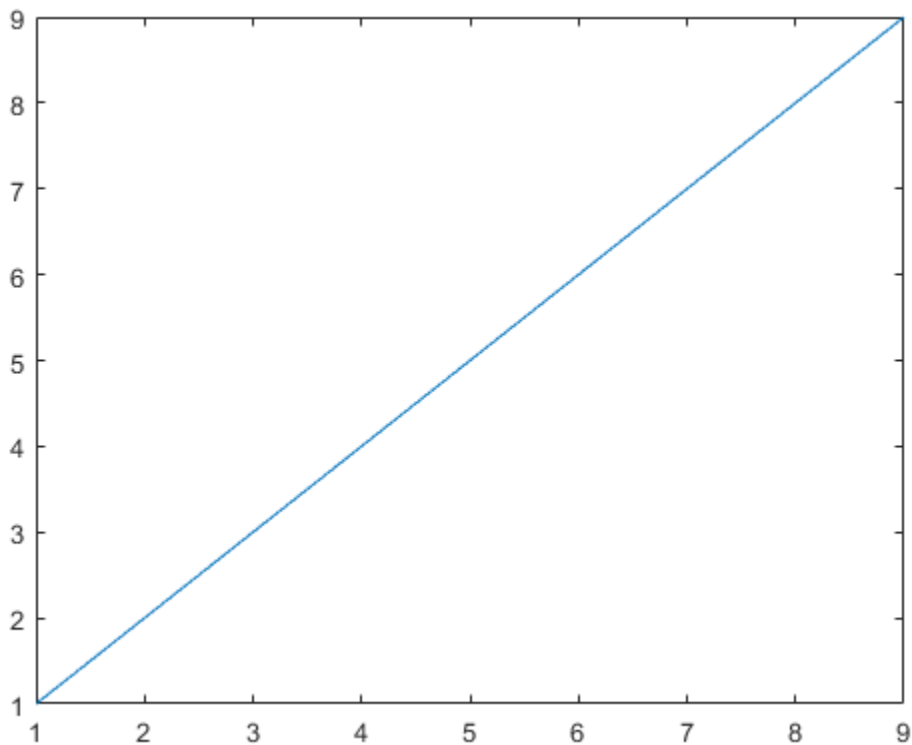
```
xSort = sort(xStandard, 'MissingPlacement', 'last')  
  
xSort = 1×5  
      1      2      3      NaN      NaN
```

### Find, Replace, and Ignore Missing Data

Even if you do not explicitly create missing values in MATLAB, they can appear when importing existing data or computing with the data. If you are not aware of missing values in your data, subsequent computation or analysis can be misleading.

For example, if you unknowingly plot a vector containing a NaN value, the NaN does not appear because the `plot` function ignores it and plots the remaining points normally.

```
nanData = [1:9 NaN];  
plot(1:10, nanData)
```



However, if you compute the average of the data, the result is NaN. In this case, it is more helpful to know in advance that the data contains a NaN, and then choose to ignore or remove it before computing the average.

```
meanData = mean(nanData)  
  
meanData = NaN
```

One way to find NaNs in data is by using the `isnan` function, which returns a logical array indicating the location of any NaN value.

```
TF = isnan(nanData)
```

```
TF = 1x10 logical array
```

```
0 0 0 0 0 0 0 0 0 1
```

Similarly, the `ismissing` function returns the location of missing values in data for multiple data types.

```
TFdouble = ismissing(xDouble)
```

```
TFdouble = 1x5 logical array
```

```
1 0 0 0 0
```

```
TFdatetime = ismissing(xDatetime)
```

```
TFdatetime = 1x5 logical array
```

```
1 0 0 0 0
```

Suppose you are working with a table or timetable made up of variables with multiple data types. You can find all of the missing values with one call to `ismissing`, regardless of their type.

```
xTable = table(xDouble',xDatetime',xString',xCategorical')
```

```
xTable=5x4 table
```

Var1	Var2	Var3	Var4
NaN	NaT	<missing>	<undefined>
1	01-Jan-2014	"a"	cat1
2	01-Feb-2014	"b"	cat2
3	01-Mar-2014	"c"	cat3
4	01-Apr-2014	"d"	cat4

```
TF = ismissing(xTable)
```

```
TF = 5x4 logical array
```

```
1 1 1 1
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
```

Missing values can represent unusable data for processing or analysis. Use `fillmissing` to replace missing values with another value, or use `rmmmissing` to remove missing values altogether.

```
xFill = fillmissing(xStandard, 'constant', 0)
```

```
xFill = 1×5  
      0     1     2     3     0
```

```
xRemove = rmmissing(xStandard)
```

```
xRemove = 1×3  
      1     2     3
```

Many MATLAB functions enable you to ignore missing values, without having to explicitly locate, fill, or remove them first. For example, if you compute the sum of a vector containing NaN values, the result is NaN. However, you can directly ignore NaNs in the sum by using the 'omitnan' option with the sum function.

```
sumNan = sum(xDouble)
```

```
sumNan = NaN
```

```
sumOmitnan = sum(xDouble, 'omitnan')
```

```
sumOmitnan = 10
```

## See Also

`ismissing` | `fillmissing` | `standardizeMissing` | `missing`

## Related Examples

- Clean Messy Data and Locate Extrema Using Live Editor Tasks on page 1-21
- “Clean Messy and Missing Data in Tables”

## Data Smoothing and Outlier Detection

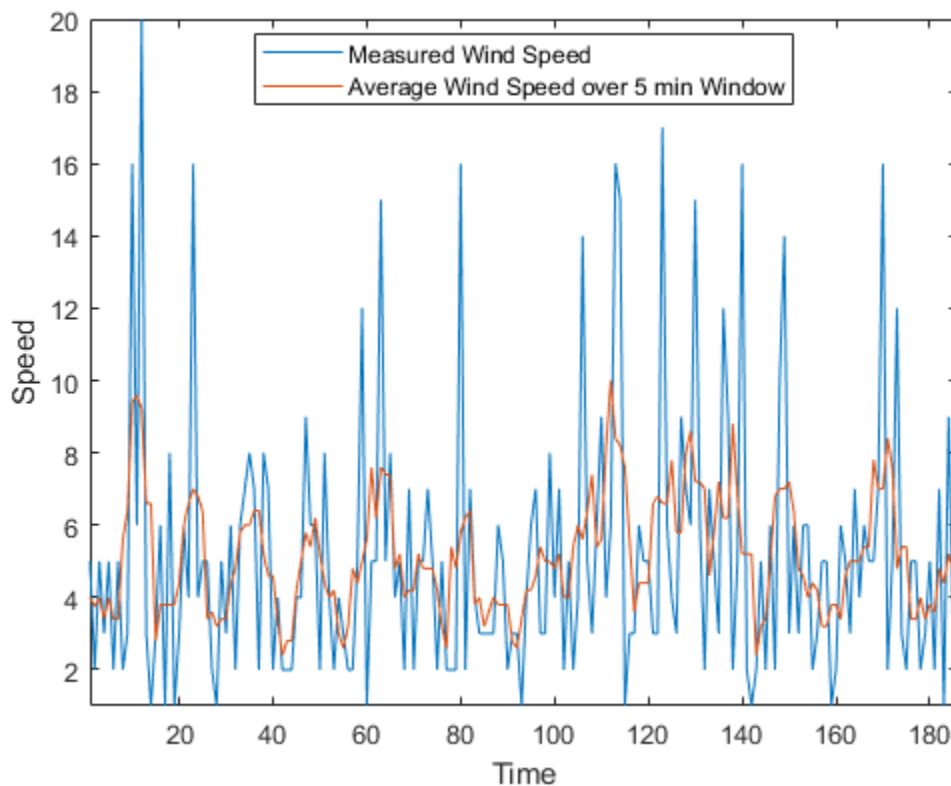
Data smoothing refers to techniques for eliminating unwanted noise or behaviors in data, while outlier detection identifies data points that are significantly different from the rest of the data.

### Moving Window Methods

Moving window methods are ways to process data in smaller batches at a time, typically in order to statistically represent a neighborhood of points in the data. The moving average is a common data smoothing technique that slides a window along the data, computing the mean of the points inside of each window. This can help to eliminate insignificant variations from one data point to the next.

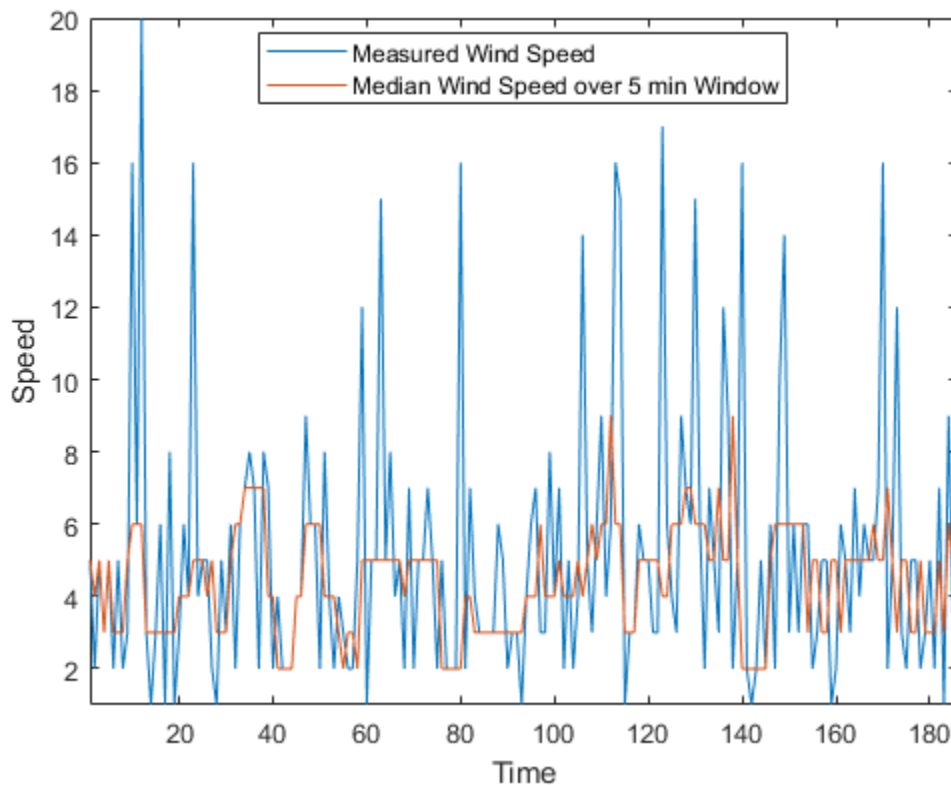
For example, consider wind speed measurements taken every minute for about 3 hours. Use the `movmean` function with a window size of 5 minutes to smooth out high-speed wind gusts.

```
load windData.mat
mins = 1:length(speed);
window = 5;
meanspeed = movmean(speed,window);
plot(mins,speed,mins,meanspeed)
axis tight
legend('Measured Wind Speed','Average Wind Speed over 5 min Window', ...
       'location','best')
xlabel('Time')
ylabel('Speed')
```



Similarly, you can compute the median wind speed over a sliding window using the `movmedian` function.

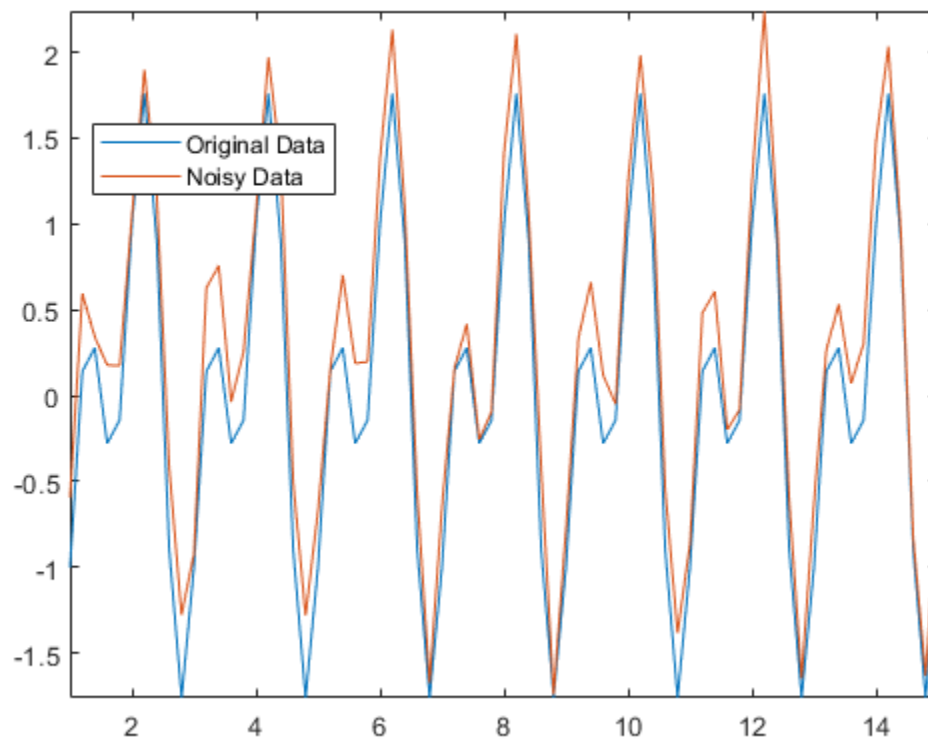
```
medianspeed = movmedian(speed>window);  
plot(mins,speed,mins,medianspeed)  
axis tight  
legend('Measured Wind Speed','Median Wind Speed over 5 min Window', ...  
      'location','best')  
xlabel('Time')  
ylabel('Speed')
```



Not all data is suitable for smoothing with a moving window method. For example, create a sinusoidal signal with injected random noise.

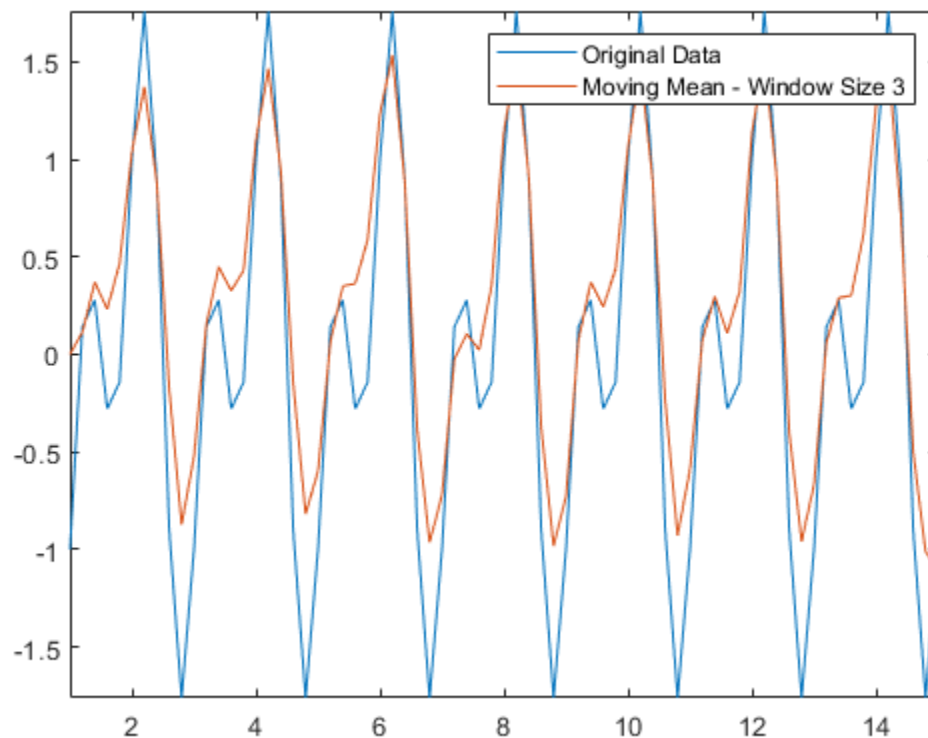
```
t = 1:0.2:15;  
A = sin(2*pi*t) + cos(2*pi*0.5*t);  
Anoise = A + 0.5*rand(1,length(t));  
plot(t,A,t,Anoise)  
axis tight  
legend('Original Data','Noisy Data','location','best')
```





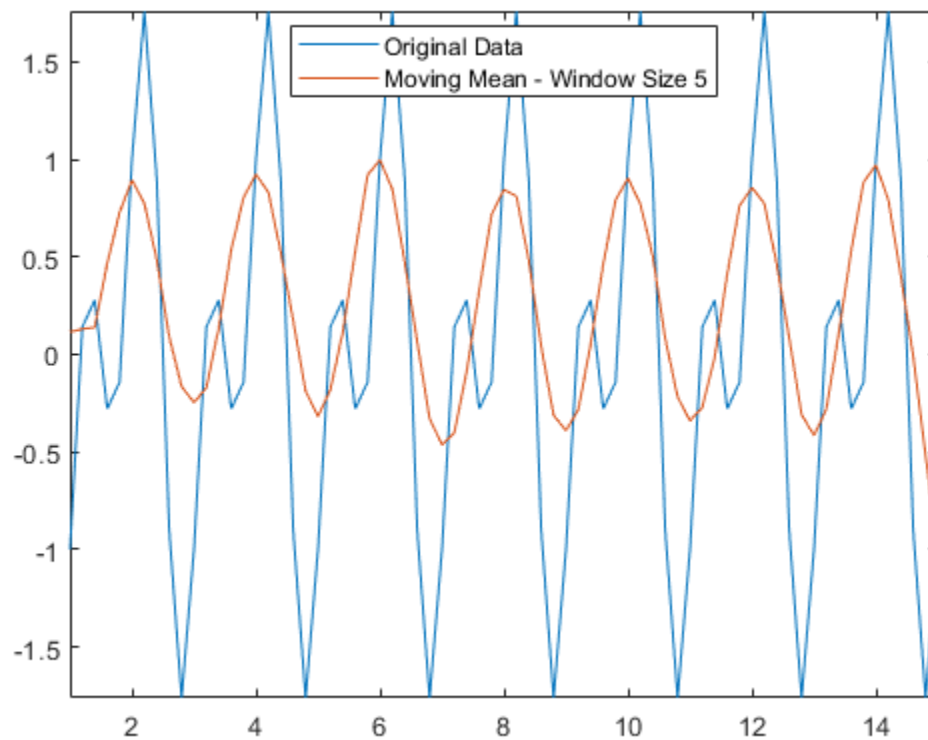
Use a moving mean with a window size of 3 to smooth the noisy data.

```
window = 3;  
Amean = movmean(Anoise,window);  
plot(t,A,t,Amean)  
axis tight  
legend('Original Data','Moving Mean - Window Size 3')
```



The moving mean achieves the general shape of the data, but doesn't capture the valleys (local minima) very accurately. Since the valley points are surrounded by two larger neighbors in each window, the mean is not a very good approximation to those points. If you make the window size larger, the mean eliminates the shorter peaks altogether. For this type of data, you might consider alternative smoothing techniques.

```
Amean = movmean(Anoise,5);  
plot(t,A,t,Amean)  
axis tight  
legend('Original Data','Moving Mean - Window Size 5', ...  
       'location','best')
```

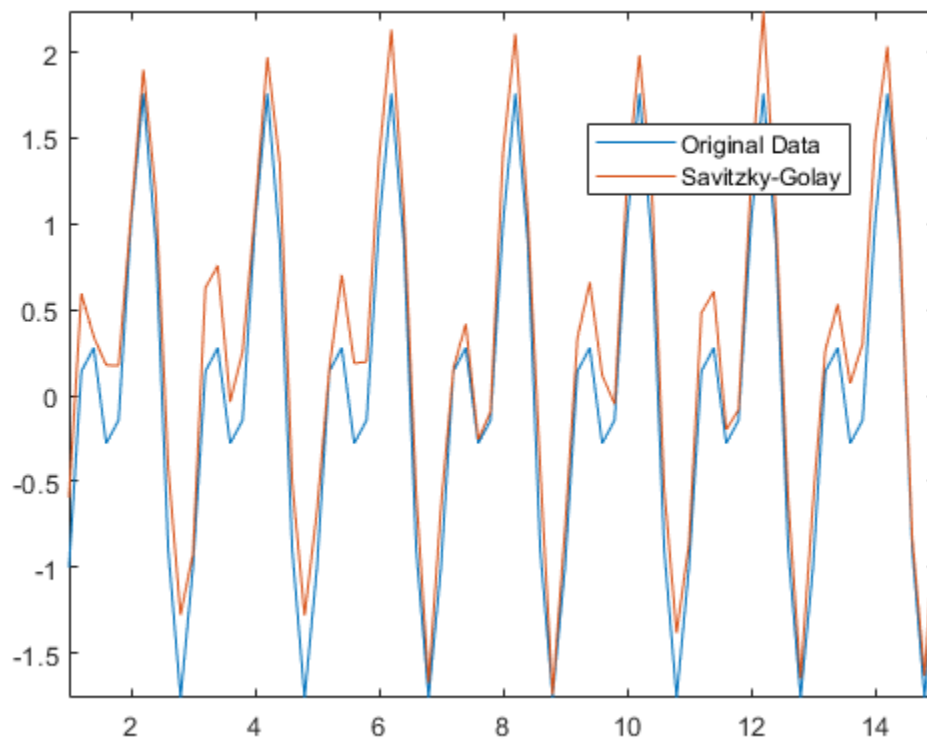


### Common Smoothing Methods

The `smoothdata` function provides several smoothing options such as the Savitzky-Golay method, which is a popular smoothing technique used in signal processing. By default, `smoothdata` chooses a best-guess window size for the method depending on the data.

Use the Savitzky-Golay method to smooth the noisy signal `Anoise`, and output the window size that it uses. This method provides a better valley approximation compared to `movmean`.

```
[Asgolay,window] = smoothdata(Anoise,'sgolay');  
plot(t,A,t,Asgolay)  
axis tight  
legend('Original Data','Savitzky-Golay','location','best')
```

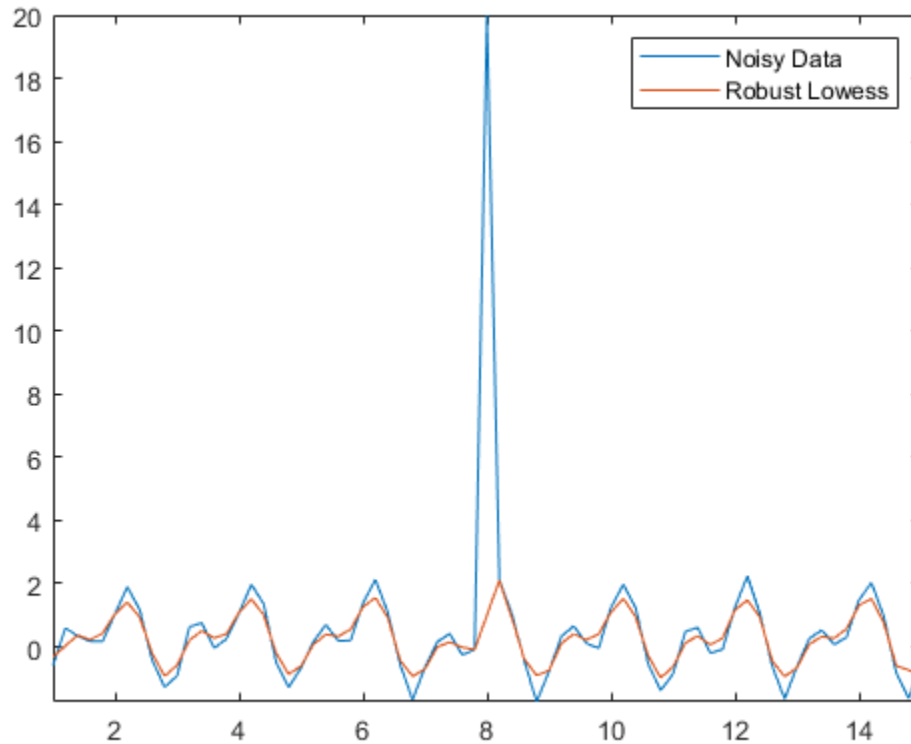


```
window
```

```
window = 3
```

The robust Lowess method is another smoothing method that is particularly helpful when outliers are present in the data in addition to noise. Inject an outlier into the noisy data, and use robust Lowess to smooth the data, which eliminates the outlier.

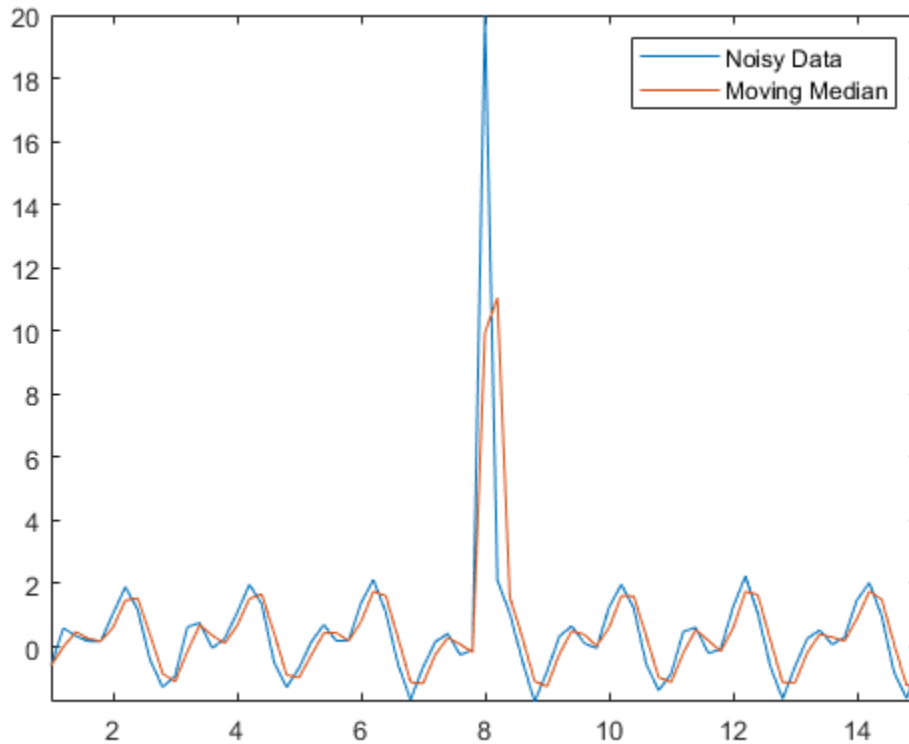
```
Anoise(36) = 20;  
Arlowess = smoothdata(Anoise,'rlowess',5);  
plot(t,Anoise,t,Arlowess)  
axis tight  
legend('Noisy Data','Robust Lowess')
```



### Detecting Outliers

Outliers in data can significantly skew data processing results and other computed quantities. For example, if you try to smooth data containing outliers with a moving median, you can get misleading peaks or valleys.

```
Amedian = smoothdata(Anoise, 'movmedian');  
plot(t, Anoise, t, Amedian)  
axis tight  
legend('Noisy Data', 'Moving Median')
```



The `isoutlier` function returns a logical 1 when an outlier is detected. Verify the index and value of the outlier in `Anoise`.

```
TF = isoutlier(Anoise);  
ind = find(TF)
```

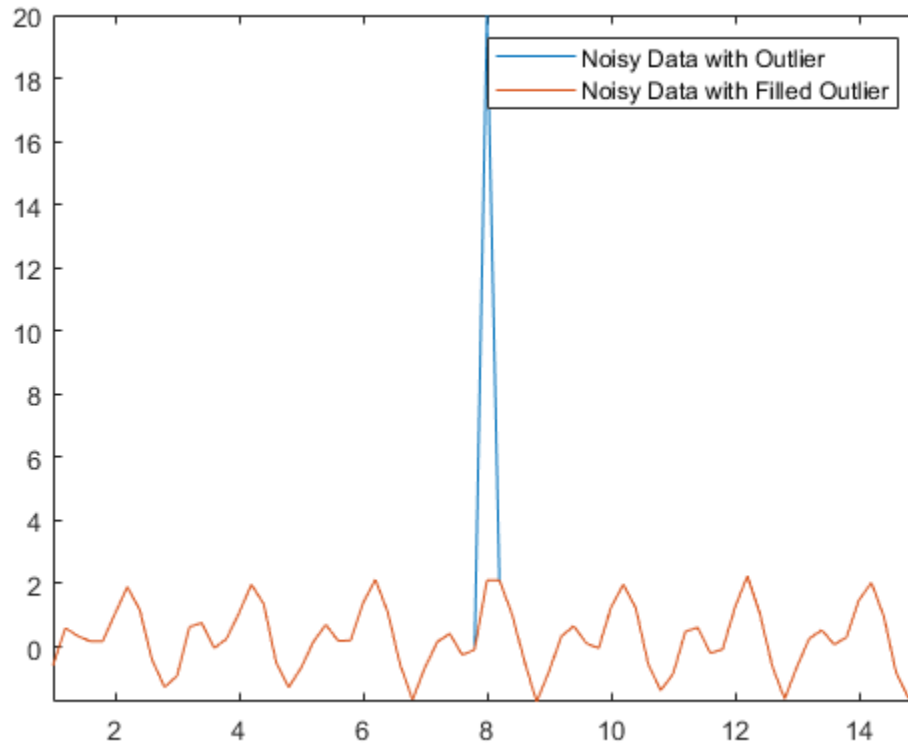
```
ind = 36
```

```
Aoutlier = Anoise(ind)
```

```
Aoutlier = 20
```

You can use the `filloutliers` function to replace outliers in your data by specifying a fill method. For example, fill the outlier in `Anoise` with the value of its neighbor immediately to the right.

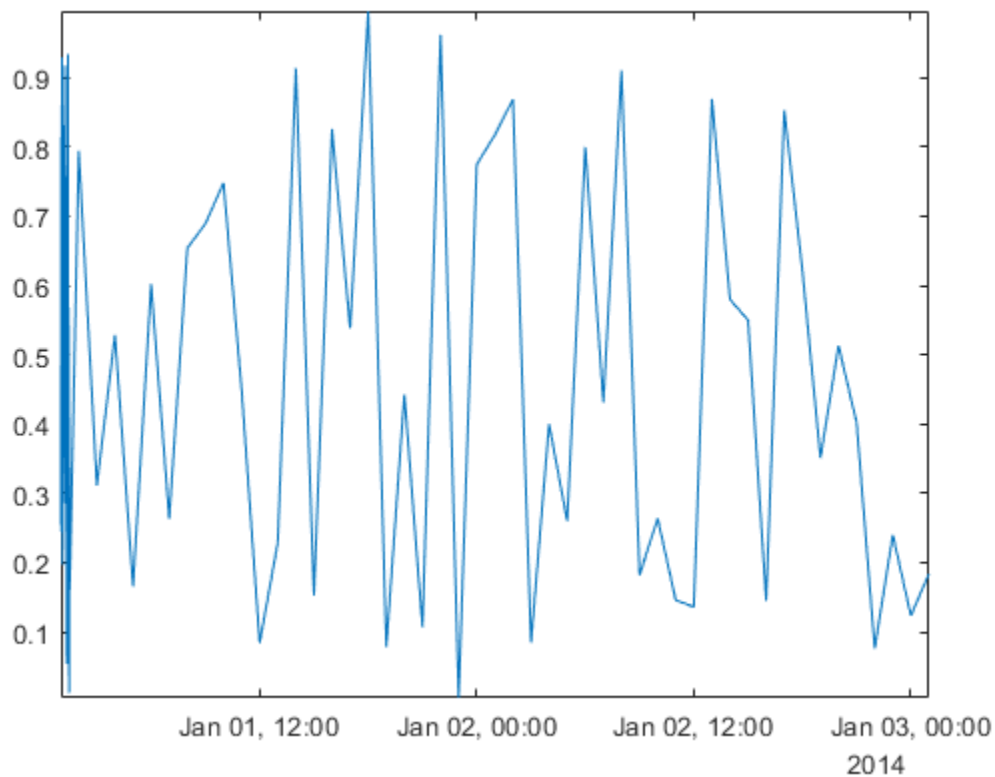
```
Afill = filloutliers(Anoise,'next');  
plot(t,Anoise,t,Afill)  
axis tight  
legend('Noisy Data with Outlier','Noisy Data with Filled Outlier')
```



### Nonuniform Data

Not all data consists of equally spaced points, which can affect methods for data processing. Create a `datetime` vector that contains irregular sampling times for the data in `Airreg`. The `time` vector represents samples taken every minute for the first 30 minutes, then hourly over two days.

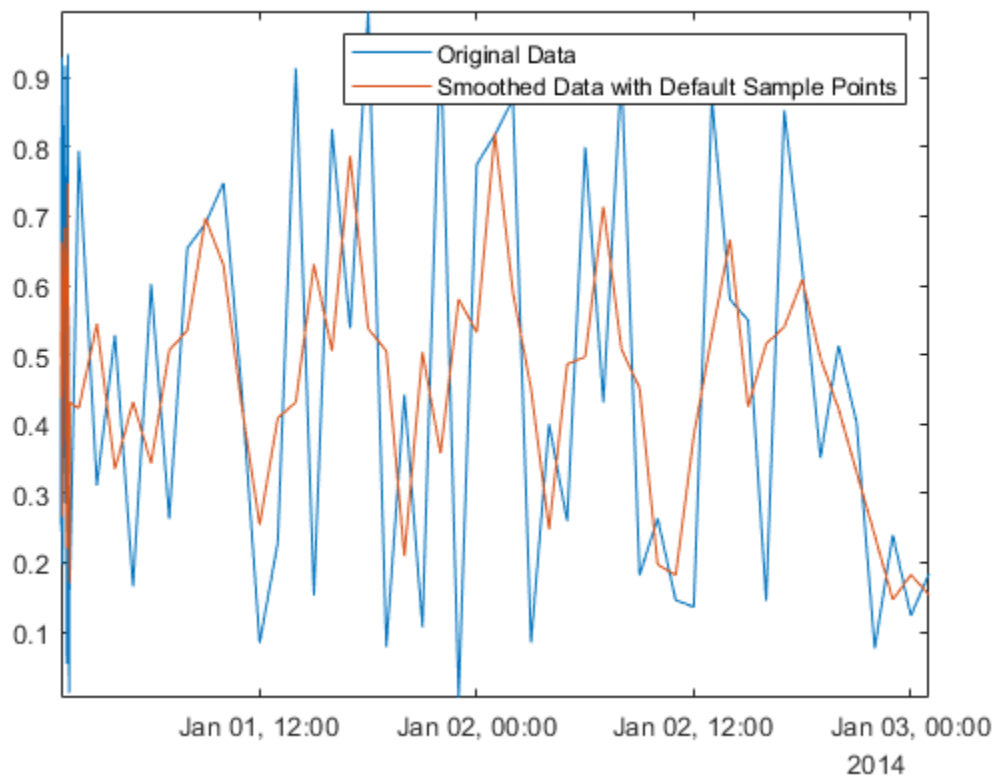
```
t0 = datetime(2014,1,1,1,1,1);  
timeminutes = sort(t0 + minutes(1:30));  
timehours = t0 + hours(1:48);  
time = [timeminutes timehours];  
Airreg = rand(1,length(time));  
plot(time,Airreg)  
axis tight
```



By default, `smoothdata` smooths with respect to equally spaced integers, in this case,  $1, 2, \dots, 78$ . Since integer time stamps do not coordinate with the sampling of the points in `Airreg`, the first half hour of data still appears noisy after smoothing.

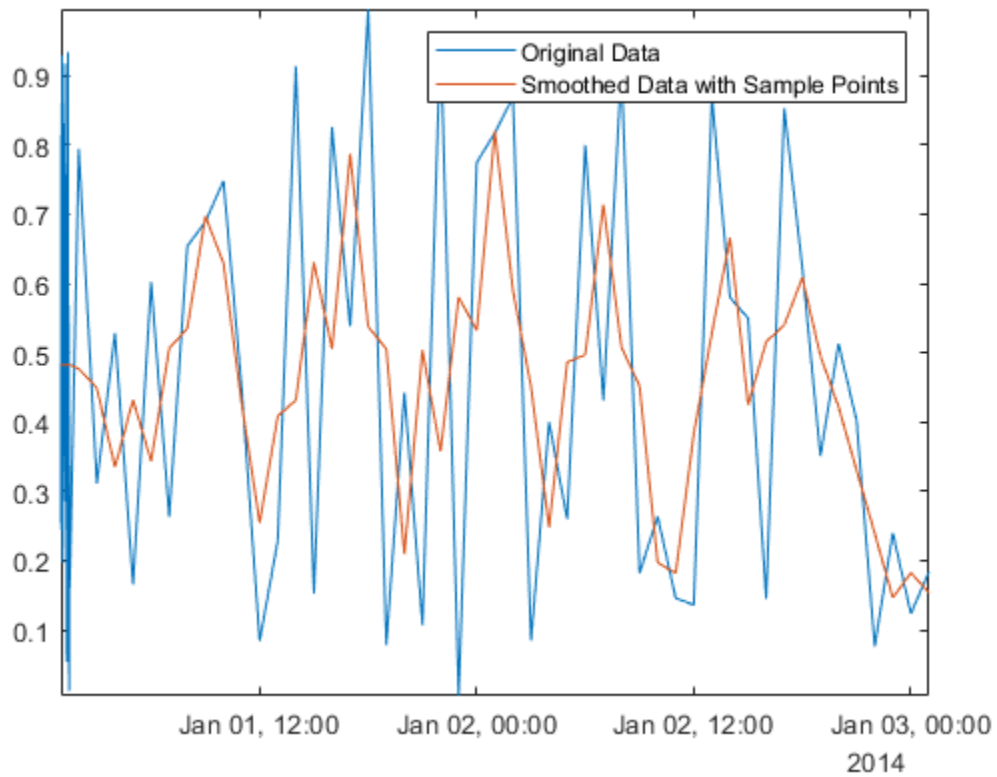
```
Adefault = smoothdata(Airreg,'movmean',3);  
plot(time,Airreg,time,Adefault)  
axis tight  
legend('Original Data','Smoothed Data with Default Sample Points')
```





Many data processing functions in MATLAB®, including `smoothdata`, `movmean`, and `filloutliers`, allow you to provide sample points, ensuring that data is processed relative to its sampling units and frequencies. To remove the high-frequency variation in the first half hour of data in `Airreg`, use the `'SamplePoints'` option with the time stamps in time.

```
Asamplepoints = smoothdata(Airreg,'movmean', ...  
    hours(3),'SamplePoints',time);  
plot(time,Airreg,time,Asamplepoints)  
axis tight  
legend('Original Data','Smoothed Data with Sample Points')
```



**See Also**

`smoothdata` | `isoutlier` | `filloutliers` | `movmean` | `movmedian`

**Related Examples**

- Clean Messy Data and Locate Extrema Using Live Editor Tasks on page 1-21
- “Filter Data” on page 1-29


## Clean Messy Data and Locate Extrema Using Live Editor Tasks

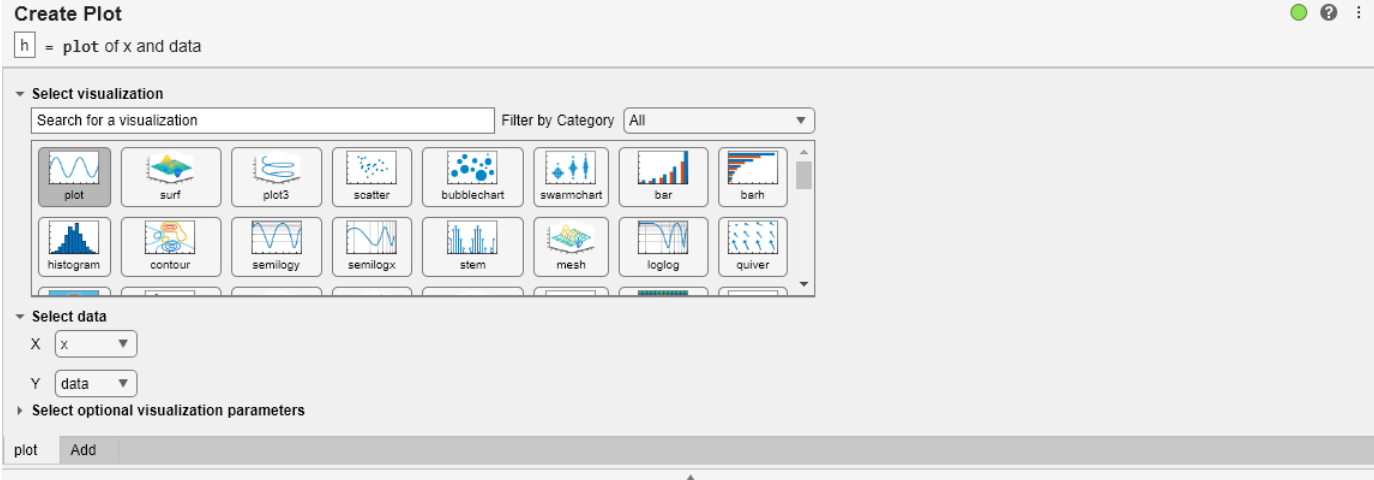
You can interactively preprocess data using sequences of Live Editor tasks, visualizing the data at each step. This example uses five tasks to clean noisy data with missing values and outliers in order to identify local minima and maxima. For more information on Live Editor tasks, see “Add Interactive Tasks to a Live Script”.

First, create and plot a vector of messy data, which contains four NaN values and five outliers.

```
x = 1:100;
data = cos(2*pi*0.05*x+2*pi*rand) + 0.5*randn(1,100);
data(20:20:80) = NaN;
data(10:20:90) = [-50 40 30 -45 35];
```

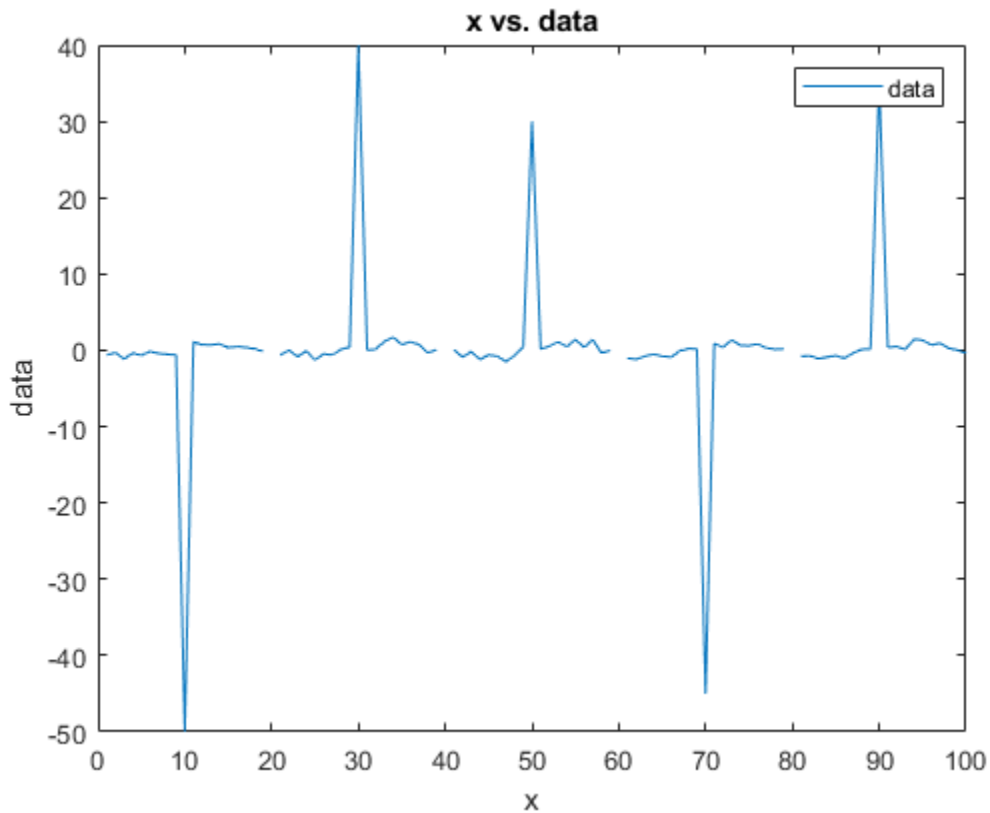
To plot the messy data, open the **Create Plot** task. Start by typing the keyword `plot` in a code block, and then click **Create Plot** when it appears in the menu. Select the plot type and input data to plot the data.

To see the code that this task generates, expand the task display by clicking  at the bottom of the task parameter area.




```
h = plot(x,data,"DisplayName","data");

% Add xlabel, ylabel, title, and legend
xlabel("x")
ylabel("data")
title("x vs. data")
legend
```



### Fill Missing Data

To replace NaN values in the data and visualize the results, open the **Clean Missing Data** task. Start by typing the keyword `missing` in a code block, and then click **Clean Missing Data** when it appears in the menu. Select the input data and the cleaning method to plot the filled data automatically.

To see the code that this task generates, expand the task display by clicking  at the bottom of the task parameter area.

**Clean Missing Data** ⊙ ? ⋮

`cleanedData` = Filled missing data in `data` using the linear interpolation method

---

▼ **Select data**

Input data:

X-axis:

▶ **Define optional missing value indicators**

▼ **Specify method**

Cleaning method:

Max gap to fill:

▼ **Display results**

Cleaned data  Filled missing entries

---

▲

```
[cleanedData,missingIndices] = fillmissing(data,"linear");
```

% Fill mis

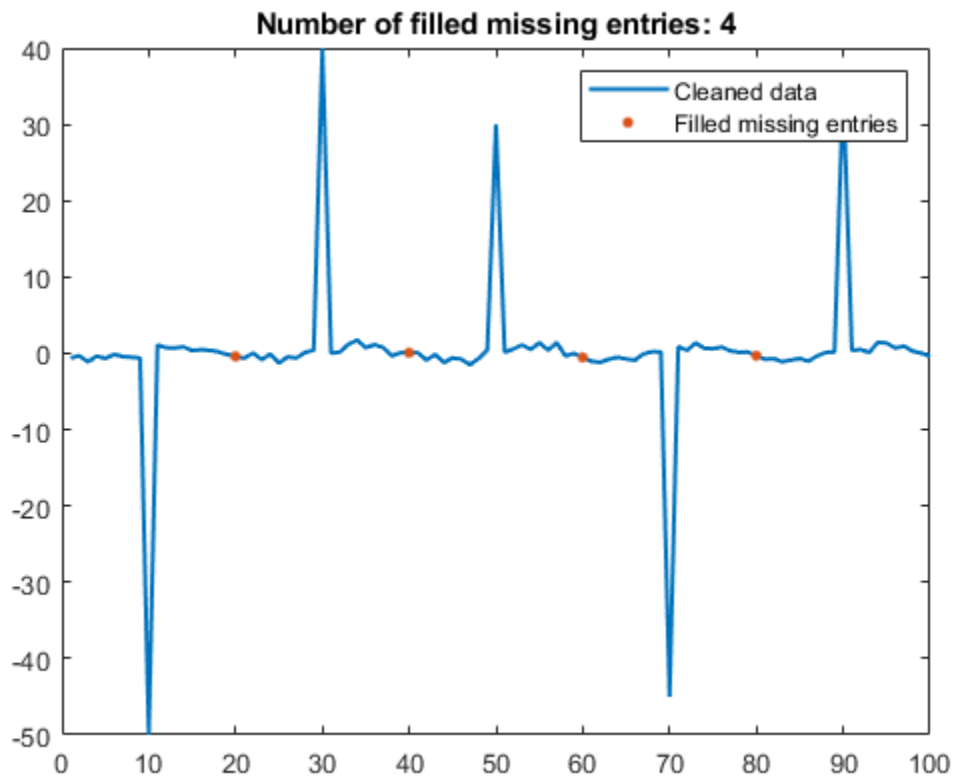
```

% Display results
clf
plot(cleanedData,"Color",[0 114 189]/255,"LineWidth",1.5,...
      "DisplayName","Cleaned data")
hold on

% Plot filled missing entries
plot(find(missingIndices),cleanedData(missingIndices),".","MarkerSize",12,...
      "Color",[217 83 25]/255,"DisplayName","Filled missing entries")
title("Number of filled missing entries: " + nnz(missingIndices))

hold off
legend

```



```
clear missingIndices
```

### Fill Outliers

You can now remove the outliers from the cleaned data in the previous task by using the **Clean Outlier Data** task. Type the keyword `outliers` in a new code block and click **Clean Outlier Data** to open the task. Select `cleanedData` as the input data. You can customize the methods for cleaning and detecting outliers and adjust the threshold to find more or fewer outliers.

To see the code that this task generates, expand the task display by clicking  at the bottom of the task parameter area.

**Clean Outlier Data**

`cleanedData2` = Filled outliers in `cleanedData` using the linear interpolation method

▼ Select data

Input data: `cleanedData`

X-axis: `default`

▼ Specify cleaning method

Cleaning method: `Fill outliers` `Linear interpolation`

▼ Define outliers

Detection method: `Median` Threshold factor: `3`

▼ Display results

Plot style: `Line`

Input data  Cleaned data  Outliers  Filled outliers  Outlier thresholds  Outlier center

% Fill out

```
[cleanedData2,outlierIndices,thresholdLow,thresholdHigh] = ...
    filloutliers(cleanedData,"linear");

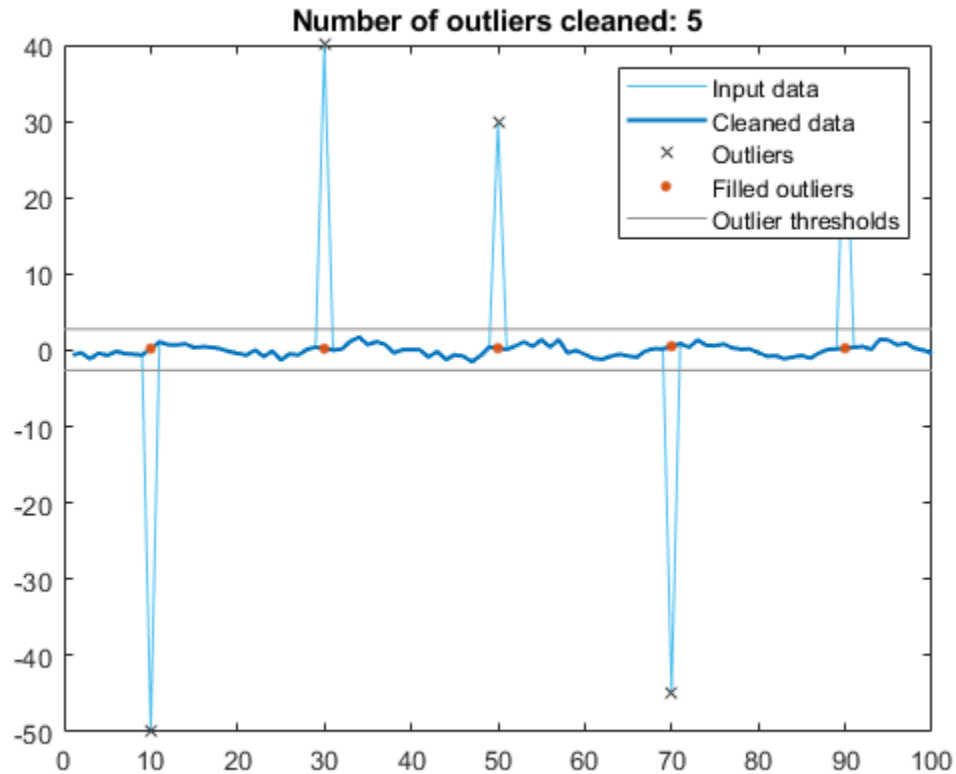
% Display results
clf
plot(cleanedData,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(cleanedData2,"Color",[0 114 189]/255,"LineWidth",1.5,...
    "DisplayName","Cleaned data")

% Plot outliers
plot(find(outlierIndices),cleanedData(outlierIndices),"x",...
    "Color",[64 64 64]/255,"DisplayName","Outliers")
title("Number of outliers cleaned: " + nnz(outlierIndices))

% Plot filled outliers
plot(find(outlierIndices),cleanedData2(outlierIndices),".","MarkerSize",12,...
    "Color",[217 83 25]/255,"DisplayName","Filled outliers")

% Plot outlier thresholds
plot([xlim missing xlim],[thresholdLow*[1 1] NaN thresholdHigh*[1 1]],...
    "Color",[145 145 145]/255,"DisplayName","Outlier thresholds")


hold off
legend
```



```
clear outlierIndices thresholdLow thresholdHigh
```

### Smooth Data

Next, smooth the cleaned data from the previous task by using the **Smooth Data** task. Type the keyword `smooth` and click the task when it appears. Select `cleanedData2`, the output from the previous task, as the input data. Select a smoothing method, and adjust the smoothing factor for more or less smoothing.

To see the code that this task generates, expand the task display by clicking  at the bottom of the task parameter area.

**Smooth Data** ● ? ⋮

`smoothedData` = Smoothed noisy data in `cleanedData2` using the Gaussian filter method

▼ Select data

Input data

X-axis

▼ Specify method and parameters

Smoothing method

Smoothing factor

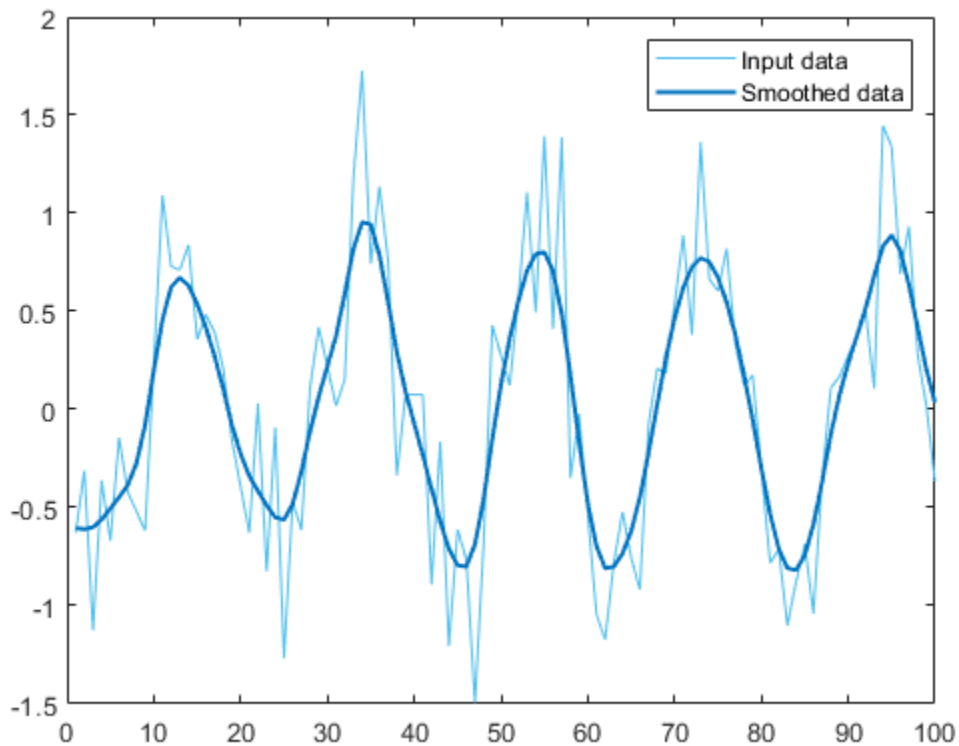
▼ Display results

Input data  Smoothed data

```
smoothedData = smoothdata(cleanedData2,"gaussian","SmoothingFactor",0.4);
```


% Smooth i

```
% Display results
clf
plot(cleanedData2,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(smoothedData,"Color",[0 114 189]/255,"LineWidth",1.5,...
     "DisplayName","Smoothed data")
hold off
legend
```



### Locate Extrema

Finally, start typing the keyword `extrema` and click `Find Local Extrema`. Use `smoothedData` as the input data and change the extrema type to find both the local maxima and local minima of the cleaned, smoothed data. You can adjust the local extrema parameters to find more or fewer maxima and minima.

To see the code that this task generates, expand the task display by clicking  at the bottom of the task parameter area.



**Find Local Extrema** ● ⓘ ⋮

`maxIndices`, `minIndices` = Local maxima and minima in `smoothedData`

▼ **Select data**

Input data: `smoothedData` ▼  
 X-axis: `default` ▼

▼ **Define local extrema**

Extrema type: `Maxima a...` ▼ Flat selection: `Center` ▼  
 Max. num. extrema: `100` Min. prominence: `0`  
 Min. separation: `0` Prominence window: `Centered` `100`

▼ **Display results**

Input data  Local maxima  Local minima

% Find loc

```

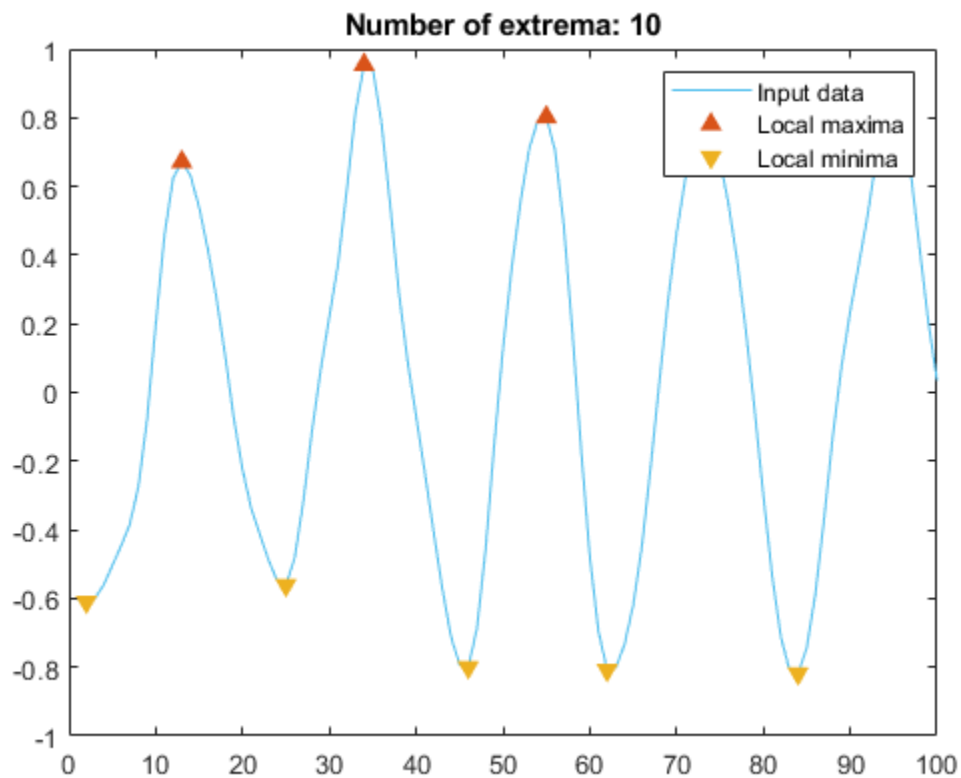
maxIndices = islocalmax(smoothedData);
minIndices = islocalmin(smoothedData);

% Display results
clf
plot(smoothedData,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on

% Plot local maxima
plot(find(maxIndices),smoothedData(maxIndices),"^","Color",[217 83 25]/255,...
     "MarkerFaceColor",[217 83 25]/255,"DisplayName","Local maxima")

% Plot local minima
plot(find(minIndices),smoothedData(minIndices),"v","Color",[237 177 32]/255,...
     "MarkerFaceColor",[237 177 32]/255,"DisplayName","Local minima")
title("Number of extrema: " + (nnz(maxIndices)+nnz(minIndices)))
hold off
legend

```



## See Also

### Live Editor Tasks

[Clean Missing Data](#) | [Clean Outlier Data](#) | [Find Change Points](#) | [Find Local Extrema](#) | [Smooth Data](#) | [Remove Trends](#)

### Functions

[ismissing](#) | [rmmissing](#) | [fillmissing](#) | [isoutlier](#) | [filloutliers](#) | [rmoutliers](#) | [ischange](#) | [islocalmin](#) | [islocalmax](#) | [smoothdata](#)

## Related Examples

- “Add Interactive Tasks to a Live Script”
- “Data Smoothing and Outlier Detection” on page 1-9
- “Missing Data in MATLAB” on page 1-5

## Filter Data

### Filter Difference Equation

Filters are data processing techniques that can smooth out high-frequency fluctuations in data or remove periodic trends of a specific frequency from data. In MATLAB, the `filter` function filters a vector of data  $x$  according to the following difference equation, which describes a tapped delay-line filter.

$$a(1)y(n) = b(1)x(n) + b(2)x(n-1) + \dots + b(N_b)x(n-N_b+1) \\ - a(2)y(n-1) - \dots - a(N_a)y(n-N_a+1)$$

In this equation,  $a$  and  $b$  are vectors of coefficients of the filter,  $N_a$  is the feedback filter order, and  $N_b$  is the feedforward filter order.  $n$  is the index of the current element of  $x$ . The output  $y(n)$  is a linear combination of the current and previous elements of  $x$  and  $y$ .

The `filter` function uses specified coefficient vectors  $a$  and  $b$  to filter the input data  $x$ . For more information on difference equations describing filters, see [1].

### Moving-Average Filter of Traffic Data

The `filter` function is one way to implement a moving-average filter, which is a common data smoothing technique.

The following difference equation describes a filter that averages time-dependent data with respect to the current hour and the three previous hours of data.

$$y(n) = \frac{1}{4}x(n) + \frac{1}{4}x(n-1) + \frac{1}{4}x(n-2) + \frac{1}{4}x(n-3)$$

Import data that describes traffic flow over time, and assign the first column of vehicle counts to the vector  $x$ .

```
load count.dat
x = count(:,1);
```

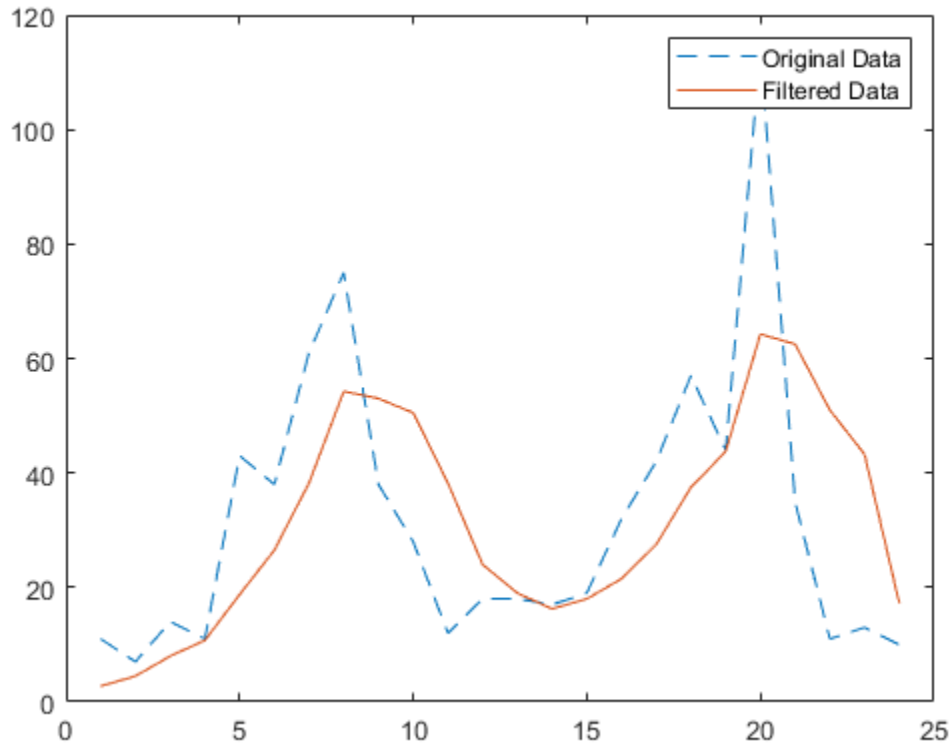
Create the filter coefficient vectors.

```
a = 1;
b = [1/4 1/4 1/4 1/4];
```

Compute the 4-hour moving average of the data, and plot both the original data and the filtered data.

```
y = filter(b,a,x);

t = 1:length(x);
plot(t,x,'--',t,y,'-')
legend('Original Data','Filtered Data')
```



## Modify Amplitude of Data

This example shows how to modify the amplitude of a vector of data by applying a transfer function.

In digital signal processing, filters are often represented by a transfer function. The Z-transform of the difference equation

$$a(1)y(n) = b(1)x(n) + b(2)x(n-1) + \dots + b(N_b)x(n-N_b+1) \\ - a(2)y(n-1) - \dots - a(N_a)y(n-N_a+1)$$

is the following transfer function.

$$Y(z) = H(z^{-1})X(z) = \frac{b(1) + b(2)z^{-1} + \dots + b(N_b)z^{-N_b+1}}{a(1) + a(2)z^{-1} + \dots + a(N_a)z^{-N_a+1}}X(z)$$

Use the transfer function

$$H(z^{-1}) = \frac{b(z^{-1})}{a(z^{-1})} = \frac{2 + 3z^{-1}}{1 + 0.2z^{-1}}$$

to modify the amplitude of the data in `count.dat`.

Load the data and assign the first column to the vector `x`.

```
load count.dat
x = count(:,1);
```

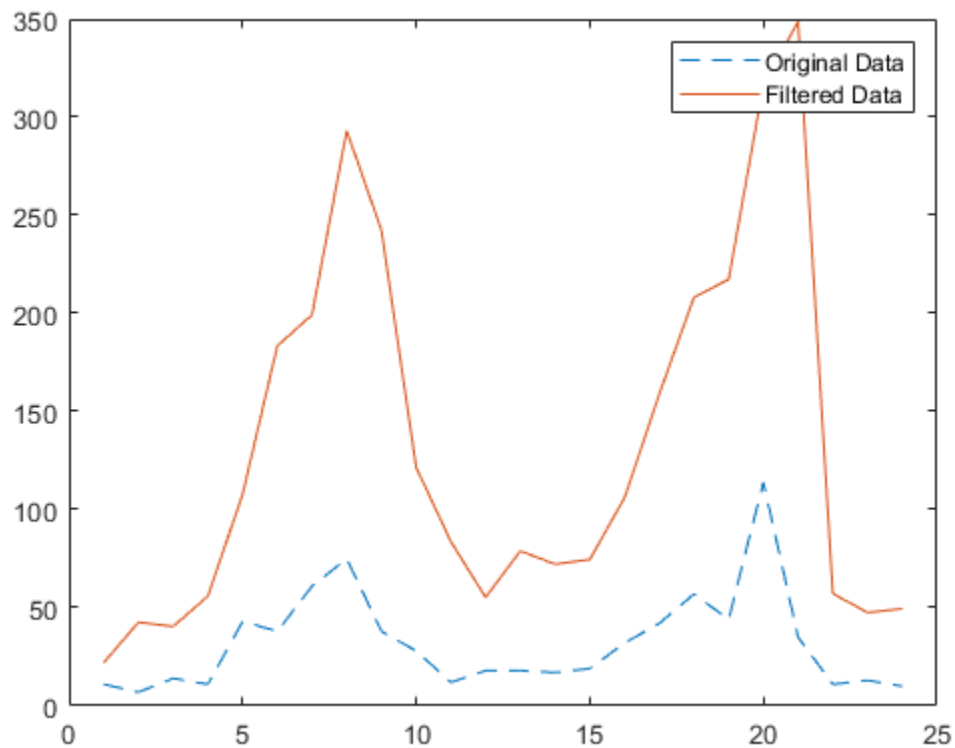
Create the filter coefficient vectors according to the transfer function  $H(z^{-1})$ .

```
a = [1 0.2];
b = [2 3];
```

Compute the filtered data, and plot both the original data and the filtered data. This filter primarily modifies the amplitude of the original data.

```
y = filter(b,a,x);

t = 1:length(x);
plot(t,x,'--',t,y,'-')
legend('Original Data','Filtered Data')
```



## References

- [1] Oppenheim, Alan V., Ronald W. Schaffer, and John R. Buck. *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.

## See Also

[filter](#) | [conv](#) | [filter2](#) | [smoothdata](#) | [movmean](#)

## **Related Examples**

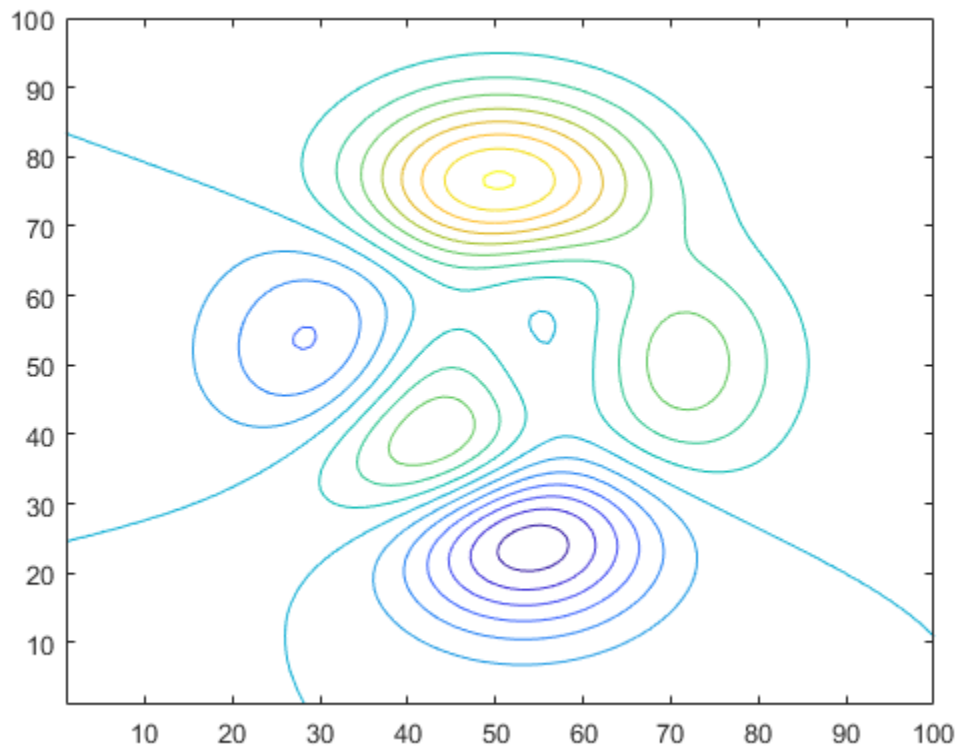
- “Smooth Data with Convolution” on page 1-33

## Smooth Data with Convolution

You can use convolution to smooth 2-D data that contains high-frequency components.

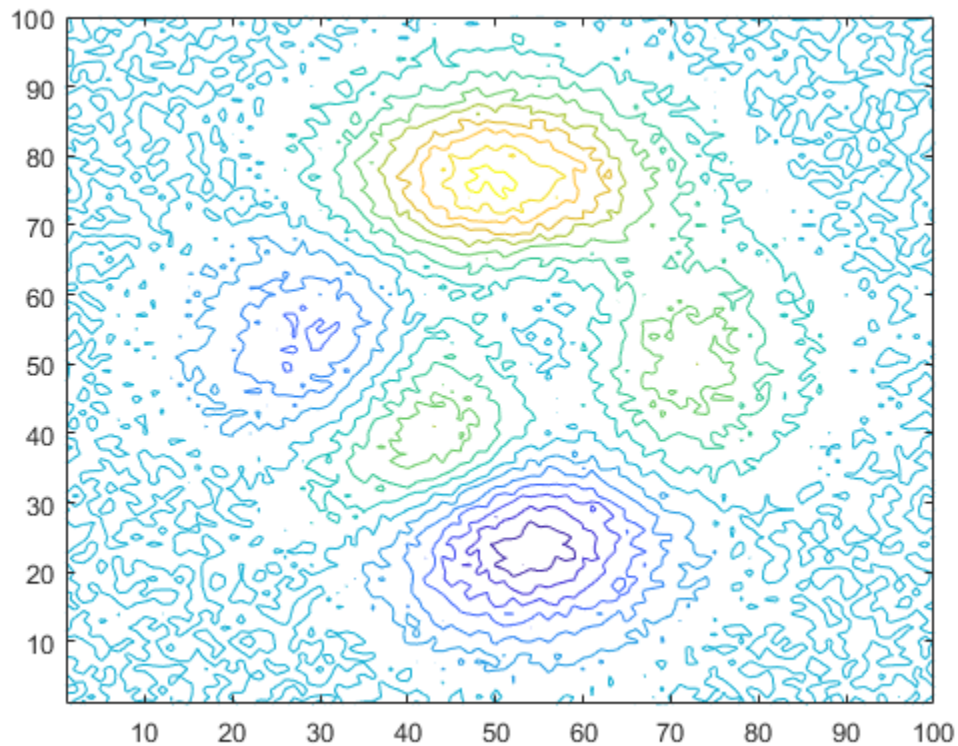
Create 2-D data using the `peaks` function, and plot the data at various contour levels.

```
Z = peaks(100);  
levels = -7:1:10;  
contour(Z, levels)
```



Inject random noise into the data and plot the noisy contours.

```
Znoise = Z + rand(100) - 0.5;  
contour(Znoise, levels)
```

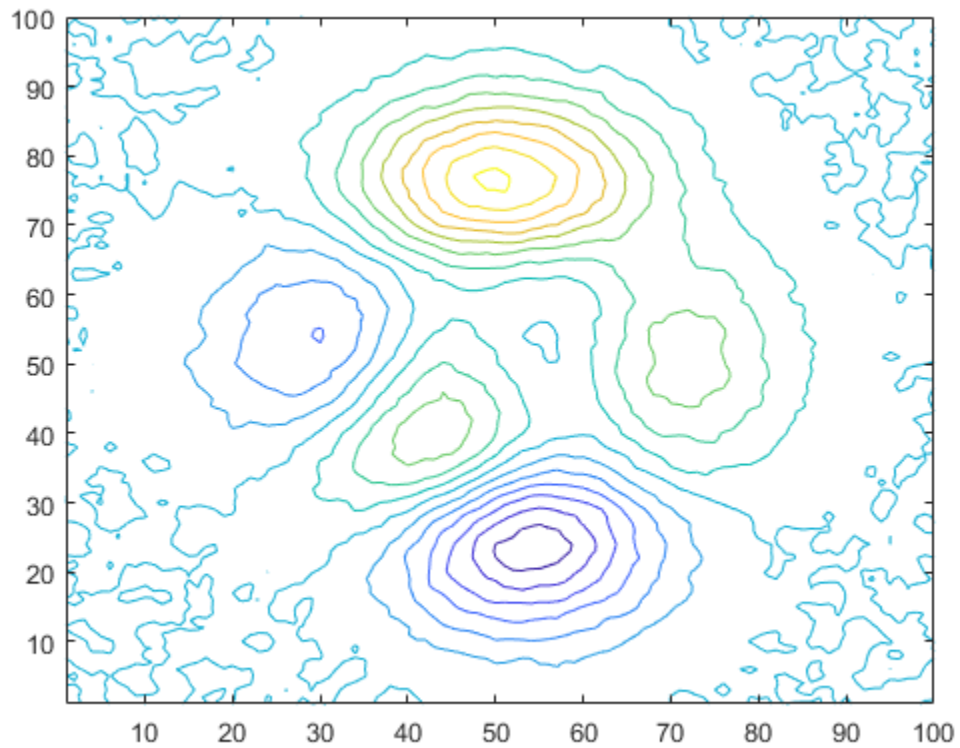


The `conv2` function in MATLAB® convolves 2-D data with a specified kernel whose elements define how to remove or enhance features of the original data. Kernels do not have to be the same size as the input data. Small-sized kernels can be sufficient to smooth data containing only a few frequency components. Larger sized kernels can provide more precision for tuning frequency response, resulting in smoother output.

Define a 3-by-3 kernel `K` and use `conv2` to smooth the noisy data in `Znoise`. Plot the smoothed contours. The `'same'` option in `conv2` makes the output the same size as the input.

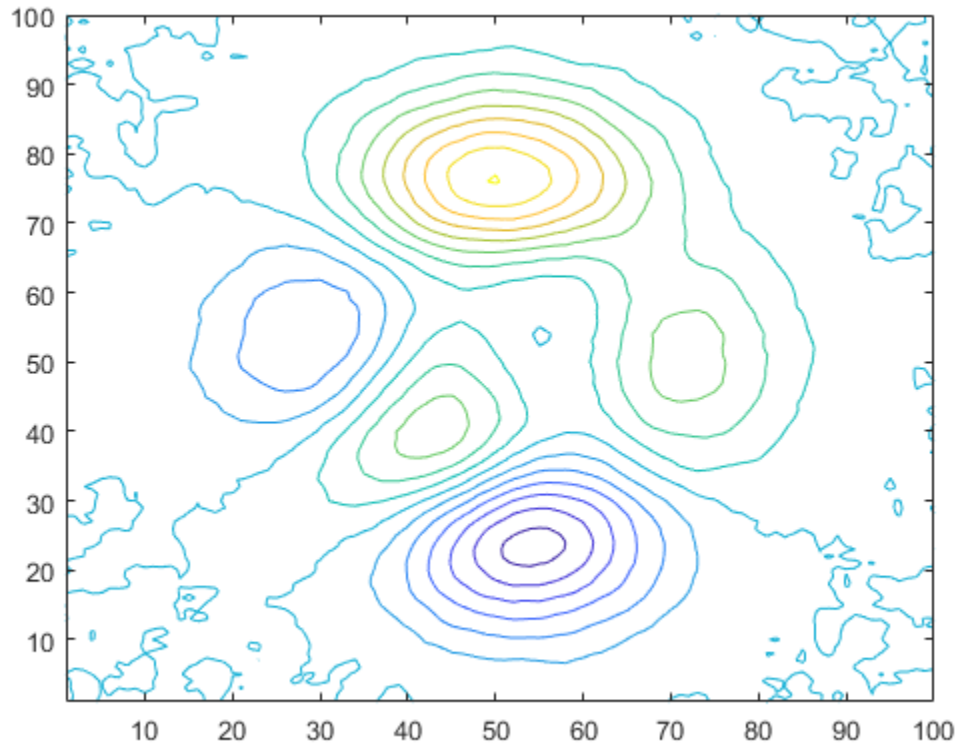
```
K = (1/9)*ones(3);  
Zsmooth1 = conv2(Znoise,K, 'same');  
contour(Zsmooth1, levels)
```





Smooth the noisy data with a 5-by-5 kernel, and plot the new contours.

```
K = (1/25)*ones(5);  
Zsmooth2 = conv2(Znoise,K,'same');  
contour(Zsmooth2,levels)
```



**See Also**

`conv2` | `conv` | `filter` | `smoothdata`

**Related Examples**

- “Filter Data” on page 1-29

## Detrending Data

### In this section...

“Introduction” on page 1-37

“Remove Linear Trends from Data” on page 1-37

### Introduction

The function `detrend` subtracts the mean or a best-fit line (in the least-squares sense) from your data. If your data contains several data columns, `detrend` treats each data column separately.

Removing a trend from the data enables you to focus your analysis on the fluctuations in the data about the trend. A linear trend typically indicates a systematic increase or decrease in the data. A systematic shift can result from sensor drift, for example. While trends can be meaningful, some types of analyses yield better insight once you remove trends.

Whether it makes sense to remove trend effects in the data often depends on the objectives of your analysis.

### Remove Linear Trends from Data

This example shows how to remove a linear trend from daily closing stock prices to emphasize the price fluctuations about the overall increase. If the data does have a trend, detrending it forces its mean to zero and reduces overall variation. The example simulates stock price fluctuations using a distribution taken from the `gallery` function.

Create a simulated data set and compute its mean. `sdata` represents the daily price changes of a stock.

```
rng(20)
t = 0:300;
dailyFluct = randn(size(t));
sdata = cumsum(dailyFluct) + 20 + t/100;
```

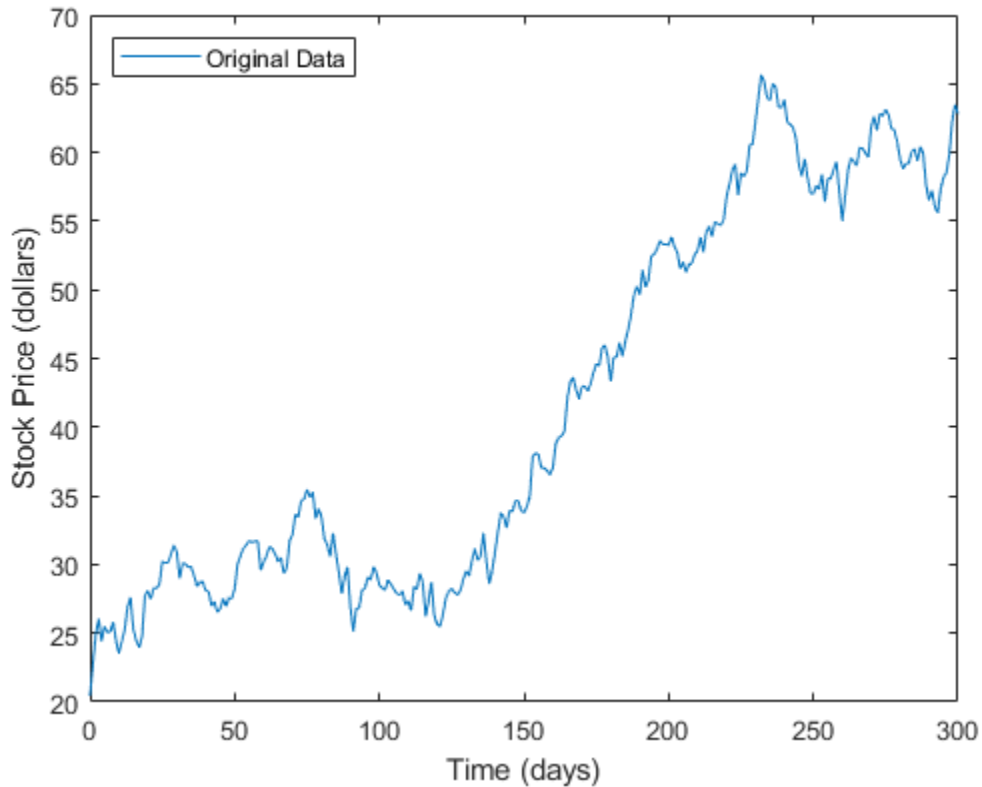
Find the average of the data.

```
mean(sdata)
```

```
ans = 41.5155
```

Plot and label the data. Notice the systematic increase in the stock prices that the data displays.

```
figure
plot(t,sdata);
legend('Original Data','Location','northwest');
xlabel('Time (days)');
ylabel('Stock Price (dollars)');
```



Apply `detrend`, which performs a linear fit to `sdata` and then removes the trend from it. Subtracting the output from the input yields the computed trend line.

```
detrend_sdata = detrend(sdata);
trend = sdata - detrend_sdata;
```

Find the average of the detrended data.

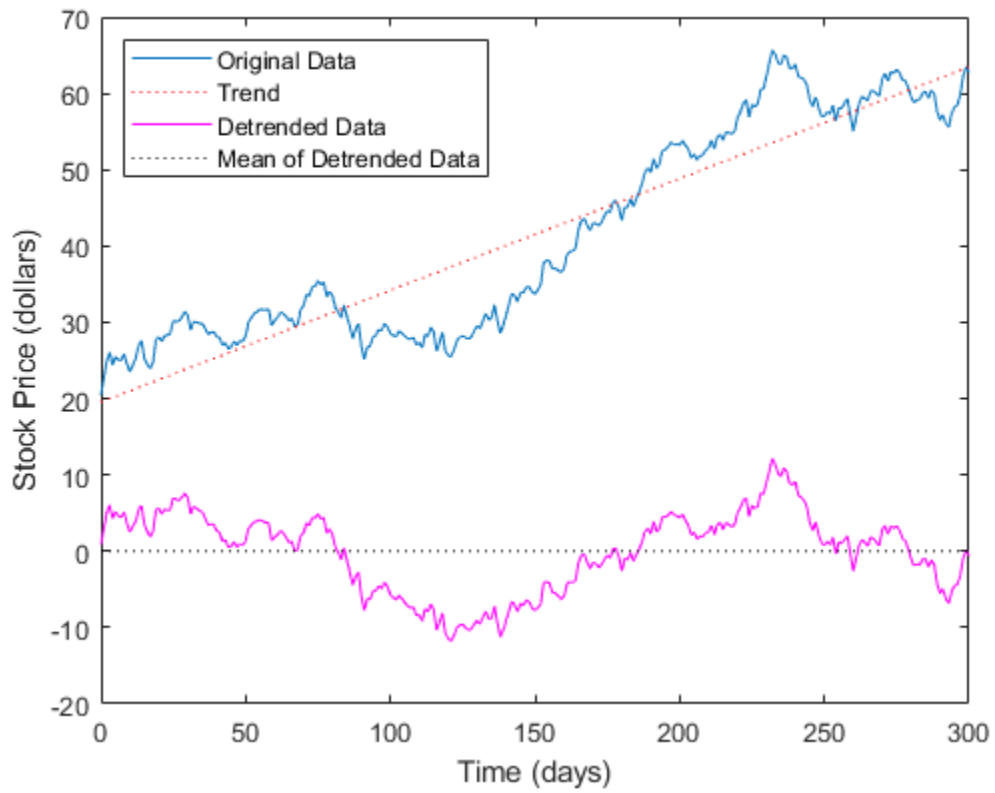
```
mean(detrend_sdata)
```

```
ans = -8.0497e-15
```

As expected, the detrended data has a mean very close to 0.

Display the results by adding the trend line, the detrended data, and its mean to the graph.

```
hold on
plot(t,trend,':r')
plot(t,detrend_sdata,'m')
plot(t,zeros(size(t)),':k')
legend('Original Data','Trend','Detrended Data',...
       'Mean of Detrended Data','Location','northwest')
xlabel('Time (days)');
ylabel('Stock Price (dollars)');
```

**See Also**

[detrrend](#) | [gallery](#) | [plot](#) | [cumsum](#)

## Computing with Descriptive Statistics

### In this section...

“Functions for Calculating Descriptive Statistics” on page 1-40

“Example: Using MATLAB Data Statistics” on page 1-42

“Data Statistics” on page 1-42

If you need more advanced statistics features, you might want to use the Statistics and Machine Learning Toolbox™ software.

### Functions for Calculating Descriptive Statistics

Use the following MATLAB functions to calculate the descriptive statistics for your data.

**Note** For matrix data, descriptive statistics for each column are calculated independently.

#### Statistics Function Summary

Function	Description
max	Maximum value
mean	Average or mean value
median	Median value
min	Smallest value
mode	Most frequent value
std	Standard deviation
var	Variance, which measures the spread or dispersion of the values

The following examples apply MATLAB functions to calculate descriptive statistics:

- “Example 1 — Calculating Maximum, Mean, and Standard Deviation” on page 1-40
- “Example 2 — Subtracting the Mean” on page 1-41

#### Example 1 — Calculating Maximum, Mean, and Standard Deviation

This example shows how to use MATLAB functions to calculate the maximum, mean, and standard deviation values for a 24-by-3 matrix called `count`. MATLAB computes these statistics independently for each column in the matrix.

```
% Load the sample data
load count.dat
% Find the maximum value in each column
mx = max(count)
% Calculate the mean of each column
mu = mean(count)
% Calculate the standard deviation of each column
sigma = std(count)
```

The results are

```

mx =
    114    145    257

mu =
    32.0000    46.5417    65.5833

sigma =
    25.3703    41.4057    68.0281

```

To get the row numbers where the maximum data values occur in each data column, specify a second output parameter `indx` to return the row index. For example:

```
[mx,indx] = max(count)
```

These results are

```

mx =
    114    145    257

indx =
    20    20    20

```

Here, the variable `mx` is a row vector that contains the maximum value in each of the three data columns. The variable `indx` contains the row indices in each column that correspond to the maximum values.

To find the minimum value in the entire `count` matrix, 24-by-3 matrix into a 72-by-1 column vector by using the syntax `count(:)`. Then, to find the minimum value in the single column, use the following syntax:

```

min(count(:))

ans =
     7

```

### Example 2 — Subtracting the Mean

Subtract the mean from each column of the matrix by using the following syntax:

```

% Get the size of the count matrix
[n,p] = size(count)
% Compute the mean of each column
mu = mean(count)
% Create a matrix of mean values by
% replicating the mu vector for n rows
MeanMat = repmat(mu,n,1)
% Subtract the column mean from each element
% in that column
x = count - MeanMat

```

---

**Note** Subtracting the mean from the data is also called *detrending*. For more information about removing the mean or the best-fit line from the data, see “Detrending Data” on page 1-37.

---

## Example: Using MATLAB Data Statistics

### Data Statistics

The Data Statistics dialog box helps you calculate and plot descriptive statistics with the data. This example shows how to use MATLAB Data Statistics to calculate and plot statistics for a 24-by-3 matrix, called `count`. The data represents how many vehicles passed by traffic counting stations on three streets.

This section contains the following topics:

- “Calculating and Plotting Descriptive Statistics” on page 1-42
- “Formatting Data Statistics on Plots” on page 1-45
- “Saving Statistics to the MATLAB Workspace” on page 1-46
- “Generating Code Files” on page 1-47

---

**Note** MATLAB Data Statistics is available for 2-D plots only.

---

### Calculating and Plotting Descriptive Statistics

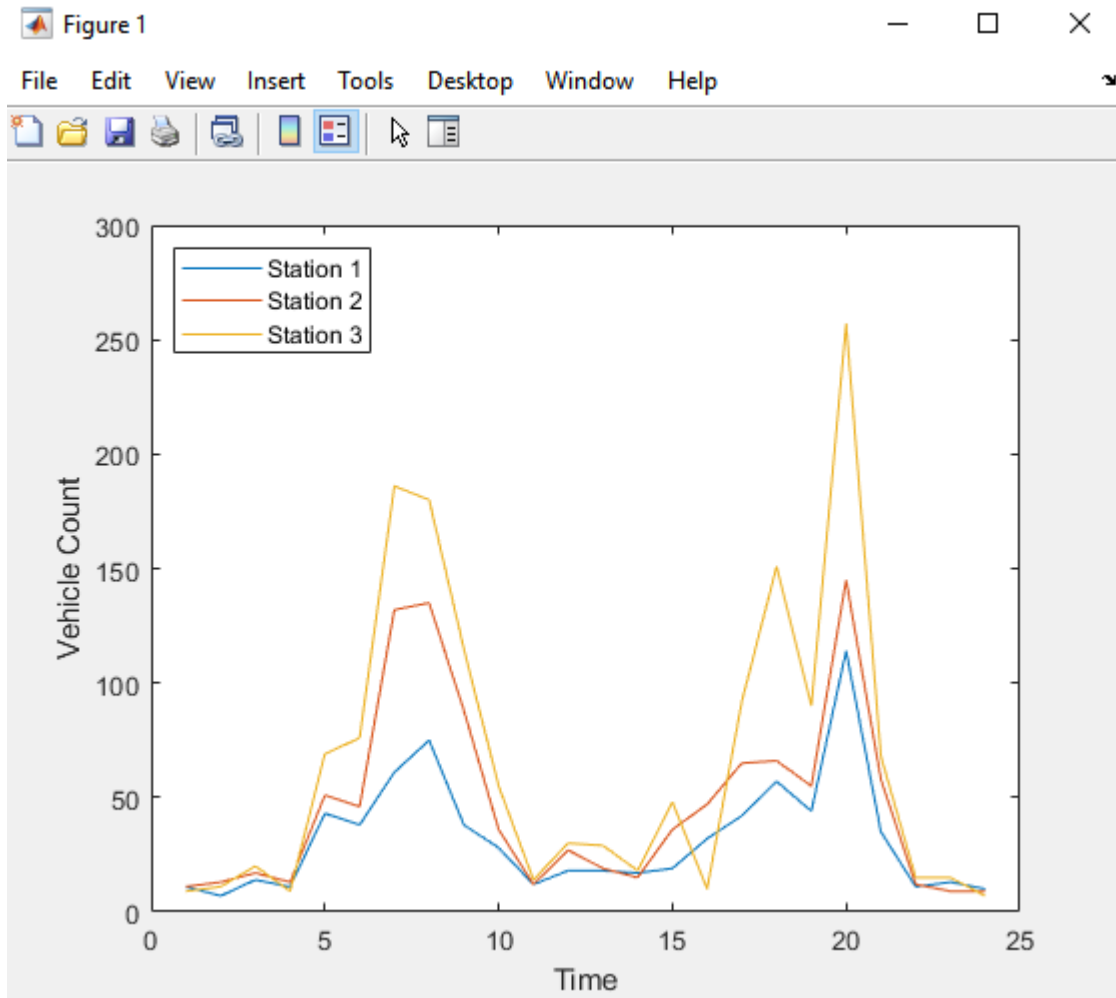
- 1 Load and plot the data:

```
load count.dat
[n,p] = size(count);

% Define the x-values
t = 1:n;

% Plot the data and annotate the graph
plot(t,count)
legend('Station 1','Station 2','Station 3','Location','northwest')
xlabel('Time')
ylabel('Vehicle Count')
```






---

**Note** The legend contains the name of each data set, as specified by the `legend` function: Station 1, Station 2, and Station 3. A *data set* refers to each column of data in the array you plotted. If you do not name the data sets, default names are assigned: `data1`, `data2`, and so on.

---

- 2 In the Figure window, select **Tools > Data Statistics**.

The Data Statistics dialog box opens and displays descriptive statistics for the X- and Y-data of the Station 1 data set.

---

**Note** The Data Statistics dialog box displays a *range*, which is the difference between the minimum and maximum values in the selected data set. The dialog box does not display the range on the plot.

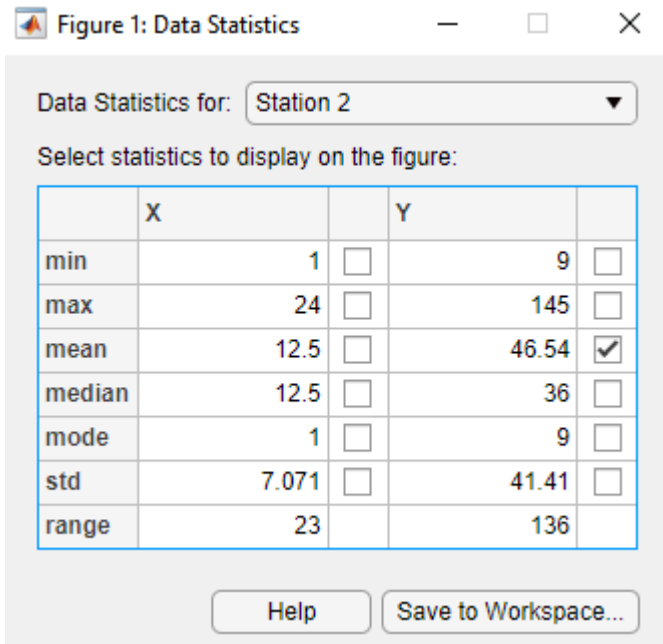
---

- 3 Select a different data set in the **Data Statistics for** list: Station 2.

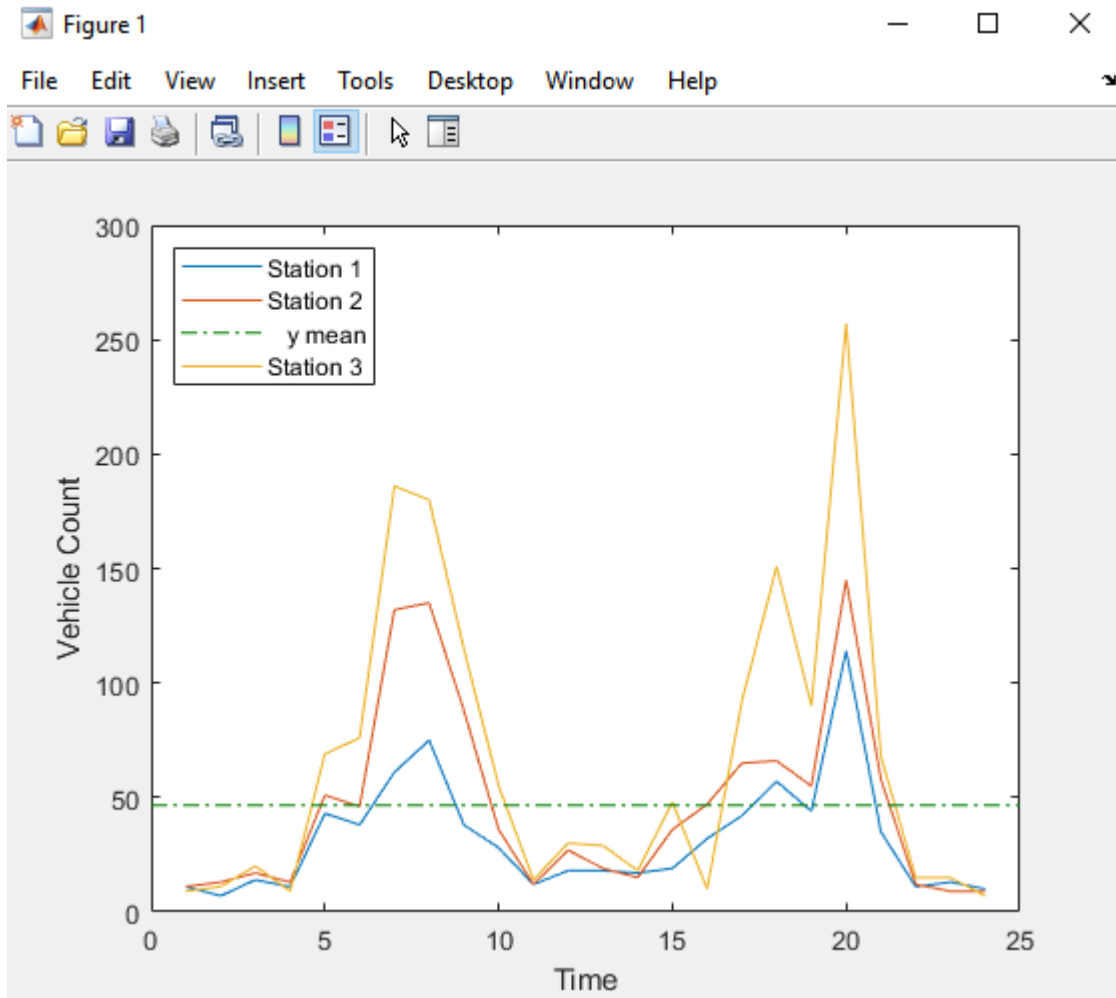
This displays the statistics for the X and Y data of the Station 2 data set.

- 4 Select the check box for each statistic you want to display on the plot, and then click **Save to Workspace**.

For example, to plot the mean of Station 2, select the **mean** check box in the **Y** column.



This plots a horizontal line to represent the mean of Station 2 and updates the legend to include this statistic.



### Formatting Data Statistics on Plots


The Data Statistics dialog box uses colors and line styles to distinguish statistics from the data on the plot. This portion of the example shows how to customize the display of descriptive statistics on a plot, such as the color, line width, line style, or marker.

---

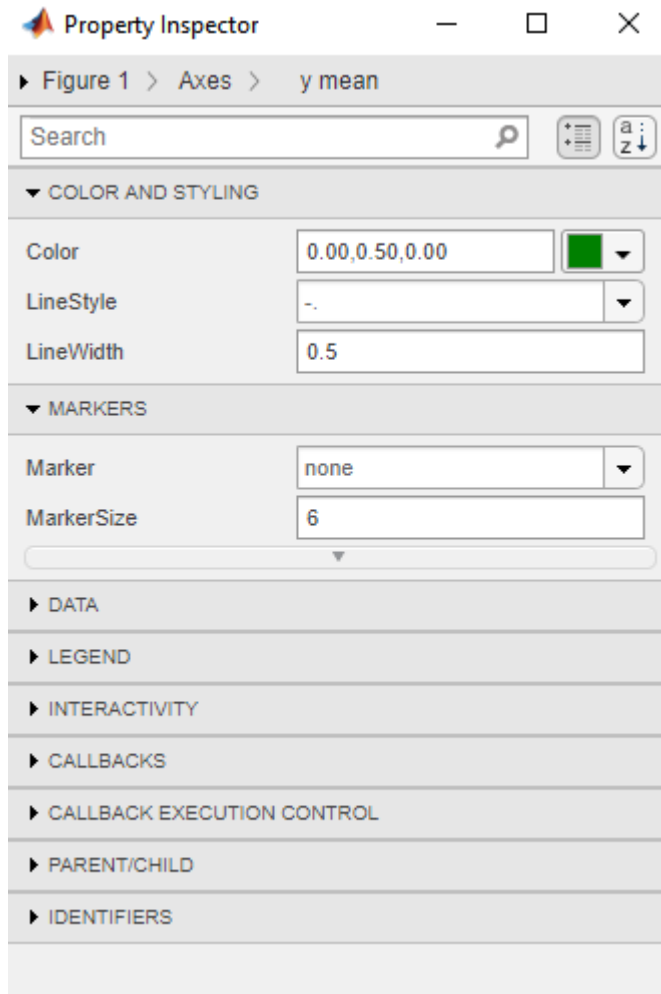
**Note** Do not edit display properties of statistics until you finish plotting all the statistics with the data. If you add or remove statistics after editing plot properties, the changes to plot properties are lost.

---

To modify the display of data statistics on a plot:

- 1 In the MATLAB Figure window, click the  (**Edit Plot**) button in the toolbar.  
This step enables plot editing.
- 2 Double-click the statistic on the plot for which you want to edit display properties. For example, double-click the horizontal line representing the mean of Station 2.

This step opens the Property Inspector, where you can modify the appearance of the line used to represent this statistic.



- 3 In the Property Inspector window, specify the line and marker styles, sizes, and colors.

---

**Tip** Alternatively, right-click the statistic on the plot, and select an option from the shortcut menu.

---

### Saving Statistics to the MATLAB Workspace

Perform these steps to save the statistics to the MATLAB workspace.

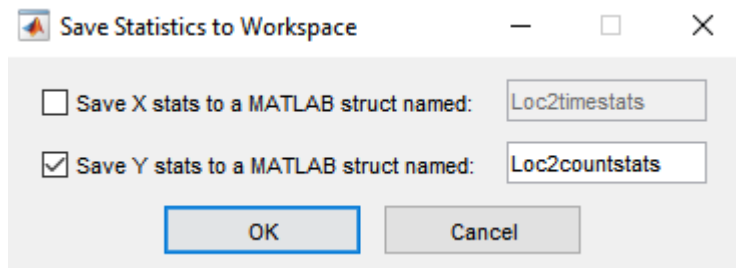
---

**Note** When your plot contains multiple data sets, save statistics for each data set individually. To display statistics for a different data set, select it from the **Data Statistics for** list in the Data Statistics dialog box.

---

- 1 In the Data Statistics dialog box, click the **Save to Workspace** button.
- 2 In the Save Statistics to Workspace dialog box, select options to save statistics for either X data, Y data, or both. Then, enter the corresponding variable names.

In this example, save only the Y data. Enter the variable name as `Loc2countstats`.



- 3 Click **OK**.

This step saves the descriptive statistics to a structure. The new variable is added to the MATLAB workspace.

To view the new structure variable, type the variable name at the MATLAB prompt:

```
Loc2countstats
```

```
Loc2countstats =
```

```
struct with fields:
```

```
    min: 9
    max: 145
   mean: 46.5417
  median: 36
    mode: 9
    std: 41.4057
   range: 136
```

### Generating Code Files

This portion of the example shows how to generate a file containing MATLAB code that reproduces the format of the plot and the plotted statistics with new data. Generating a code file is not available in MATLAB Online™.

- 1 In the Figure window, select **File > Generate Code**.

This step creates a function code file and displays it in the MATLAB Editor.

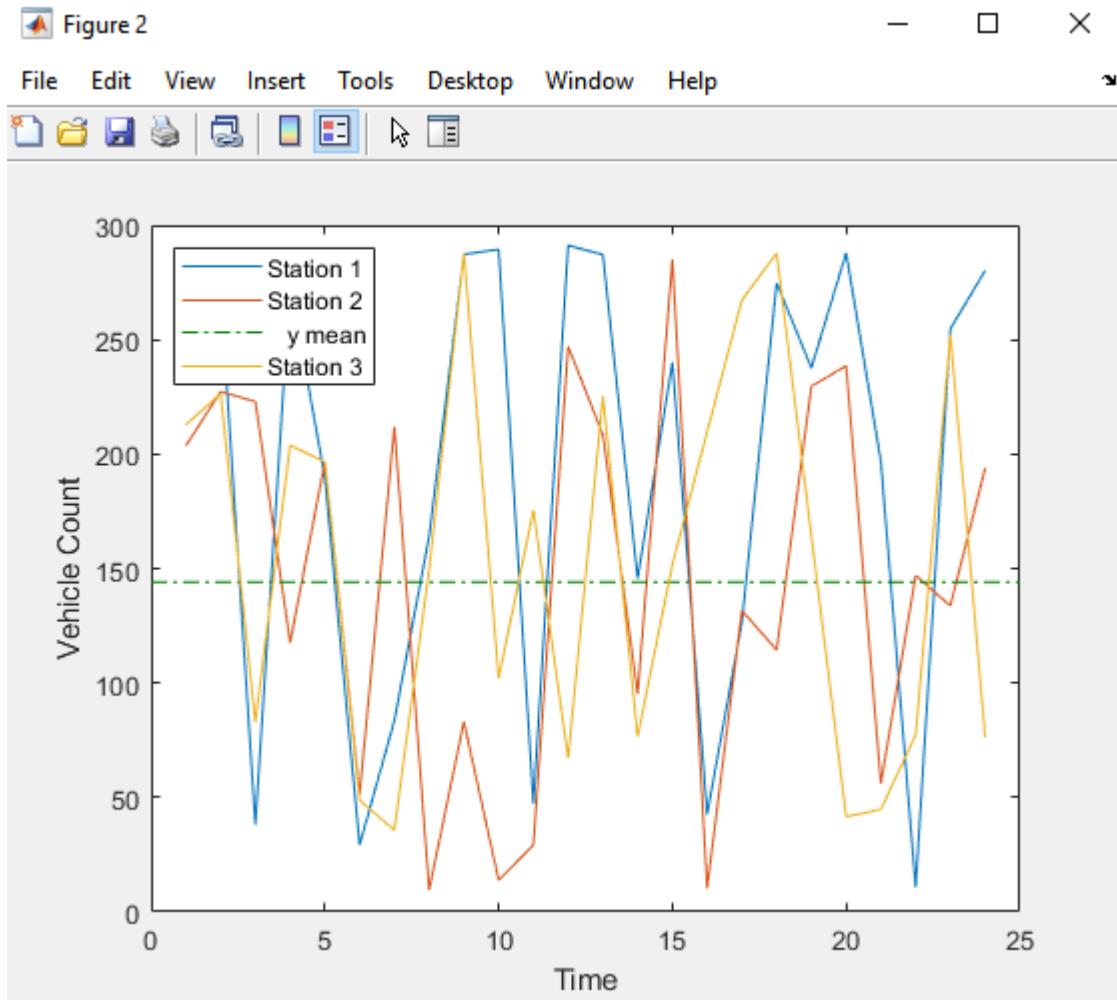
- 2 Change the name of the function on the first line of the file from `createfigure` to something more specific, like `countplot`. Save the file to your current folder with the file name `countplot.m`.

- 3 Generate some new, random count data:

```
rng('default')
randcount = 300*rand(24,3);
```

- 4 Reproduce the plot with the new data and the recomputed statistics:

```
countplot(t,randcount)
```



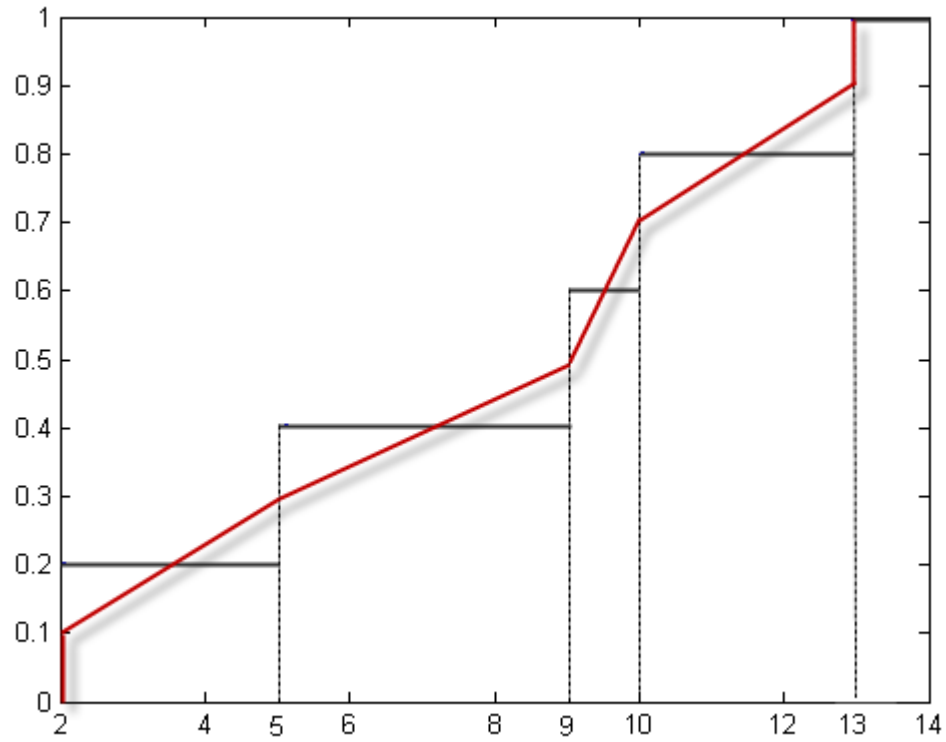
## Quantiles and Percentiles

This example explains how MATLAB functions `quantile` and `prctile` compute quantiles and percentiles.

The `prctile` function calculates the percentiles in a similar way to how `quantile` calculates quantiles. These steps in the computation of quantiles are also true for percentiles, given the fact that, for the same data sample, the quantile at the value  $Q$  is the same as the percentile at the value  $P = 100*Q$ .

- 1 `quantile` initially assigns the sorted values in  $X$  to the  $(0.5/n)$ ,  $(1.5/n)$ , ...,  $([n - 0.5]/n)$  quantiles. For example:
  - For a data vector of six elements such as  $\{6, 3, 2, 10, 8, 1\}$ , the sorted elements  $\{1, 2, 3, 6, 8, 10\}$  respectively correspond to the  $(0.5/6)$ ,  $(1.5/6)$ ,  $(2.5/6)$ ,  $(3.5/6)$ ,  $(4.5/6)$ , and  $(5.5/6)$  quantiles.
  - For a data vector of five elements such as  $\{2, 10, 5, 9, 13\}$ , the sorted elements  $\{2, 5, 9, 10, 13\}$  respectively correspond to the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles.

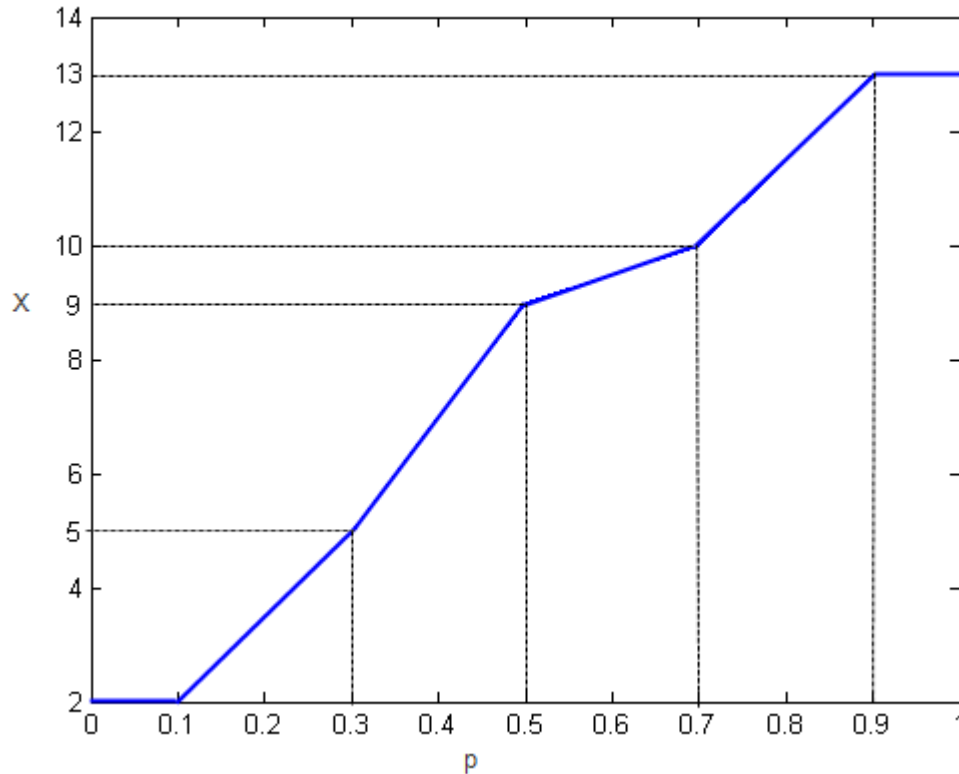
This figure illustrates this approach for data vector  $X = \{2, 10, 5, 9, 13\}$ . The first observation corresponds to the cumulative probability  $1/5 = 0.2$ , the second observation corresponds to the cumulative probability  $2/5 = 0.4$ , and so on. The step function in this figure shows these cumulative probabilities. `quantile` instead places the observations in midpoints, such that the first corresponds to  $0.5/5 = 0.1$ , the second corresponds to  $1.5/5 = 0.3$ , and so on, and then connects these midpoints. The red lines in the figure connect the midpoints.



**Assigning Observations to Quantiles**

By switching the axes, as in the next figure, you can see the values of the variable  $X$  that correspond to the  $p$  quantiles.





### Quantiles of $X$

- 2 quantile finds any quantiles between the data values using linear interpolation.

*Linear interpolation* uses linear polynomials to approximate a function  $f(x)$  and construct new data points within the range of a known set of data points. Algebraically, given the data points  $(x_1, y_1)$  and  $(x_2, y_2)$ , where  $y_1 = f(x_1)$  and  $y_2 = f(x_2)$ , linear interpolation finds  $y = f(x)$  for a given  $x$  between  $x_1$  and  $x_2$  as

$$y = f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1).$$

Similarly, if the  $1.5/n$  quantile is  $y_{1.5/n}$  and the  $2.5/n$  quantile is  $y_{2.5/n}$ , then linear interpolation finds the  $2.3/n$  quantile  $y_{2.3/n}$  as

$$y_{\frac{2.3}{n}} = y_{\frac{1.5}{n}} + \frac{\left(\frac{2.3}{n} - \frac{1.5}{n}\right)}{\left(\frac{2.5}{n} - \frac{1.5}{n}\right)}\left(y_{\frac{2.5}{n}} - y_{\frac{1.5}{n}}\right).$$

- 3 quantile assigns the minimum and maximum values of  $X$  to the quantiles for probabilities less than  $(0.5/n)$  and greater than  $([n-0.5]/n)$ , respectively.

### References

- [1] Langford, E. "Quartiles in Elementary Statistics", *Journal of Statistics Education*. Vol. 14, No. 3, 2006.

**See Also**

quantile | prctile | median

# Regression Analysis

---

- “Linear Correlation” on page 2-2
- “Linear Regression” on page 2-5
- “Interactive Fitting” on page 2-13
- “Programmatic Fitting” on page 2-26

## Linear Correlation

### In this section...

“Introduction” on page 2-2

“Covariance” on page 2-2

“Correlation Coefficients” on page 2-3

### Introduction

*Correlation* quantifies the strength of a linear relationship between two variables. When there is no correlation between two variables, then there is no tendency for the values of the variables to increase or decrease in tandem. Two variables that are uncorrelated are not necessarily independent, however, because they might have a nonlinear relationship.

You can use linear correlation to investigate whether a linear relationship exists between variables without having to assume or fit a specific model to your data. Two variables that have a small or no linear correlation might have a strong nonlinear relationship. However, calculating linear correlation before fitting a model is a useful way to identify variables that have a simple relationship. Another way to explore how variables are related is to make scatter plots of your data.

*Covariance* quantifies the strength of a linear relationship between two variables in units relative to their variances. Correlations are standardized covariances, giving a dimensionless quantity that measures the degree of a linear relationship, separate from the scale of either variable.

The following MATLAB functions compute sample correlation coefficients and covariance. These sample coefficients are estimates of the true covariance and correlation coefficients of the population from which the data sample is drawn.

Function	Description
<code>corrcoef</code>	Correlation coefficient matrix
<code>cov</code>	Covariance matrix
<code>xcorr</code>	Cross-correlation sequence of a random process (includes autocorrelation)

### Covariance

Use the MATLAB `cov` function to calculate the sample covariance matrix for a data matrix (where each column represents a separate quantity).

The sample covariance matrix has the following properties:

- $\text{cov}(X)$  is symmetric.
- $\text{diag}(\text{cov}(X))$  is a vector of variances for each data column. The variances represent a measure of the spread or dispersion of data in the corresponding column. (The `var` function calculates variance.)
- $\text{sqrt}(\text{diag}(\text{cov}(X)))$  is a vector of standard deviations. (The `std` function calculates standard deviation.)
- The off-diagonal elements of the covariance matrix represent the covariances between the individual data columns.

Here,  $X$  can be a vector or a matrix. For an  $m$ -by- $n$  matrix, the covariance matrix is  $n$ -by- $n$ .

For an example of calculating the covariance, load the sample data in `count.dat` that contains a 24-by-3 matrix:

```
load count.dat
```

Calculate the covariance matrix for this data:

```
cov(count)
```

MATLAB responds with the following result:

```
ans =
  1.0e+003 *
    0.6437    0.9802    1.6567
    0.9802    1.7144    2.6908
    1.6567    2.6908    4.6278
```

The covariance matrix for this data has the following form:

$$\begin{bmatrix} s^2_{11} & s^2_{12} & s^2_{13} \\ s^2_{21} & s^2_{22} & s^2_{23} \\ s^2_{31} & s^2_{32} & s^2_{33} \end{bmatrix}$$

$$s^2_{ij} = s^2_{ji}$$

Here,  $s^2_{ij}$  is the sample covariance between column  $i$  and column  $j$  of the data. Because the `count` matrix contains three columns, the covariance matrix is 3-by-3.

---

**Note** In the special case when a vector is the argument of `cov`, the function returns the variance.

---

## Correlation Coefficients

The function `corrcoef` produces a matrix of sample correlation coefficients for a data matrix (where each column represents a separate quantity). The correlation coefficients range from -1 to 1, where

- Values close to 1 indicate that there is a positive linear relationship between the data columns.
- Values close to -1 indicate that one column of data has a negative linear relationship to another column of data (*anticorrelation*).
- Values close to or equal to 0 suggest there is no linear relationship between the data columns.

For an  $m$ -by- $n$  matrix, the correlation-coefficient matrix is  $n$ -by- $n$ . The arrangement of the elements in the correlation coefficient matrix corresponds to the location of the elements in the covariance matrix, as described in “Covariance” on page 2-2.

For an example of calculating correlation coefficients, load the sample data in `count.dat` that contains a 24-by-3 matrix:

```
load count.dat
```

Type the following syntax to calculate the correlation coefficients:

```
corrcoef(count)
```

This results in the following 3-by-3 matrix of correlation coefficients:

```
ans =  
    1.0000    0.9331    0.9599  
    0.9331    1.0000    0.9553  
    0.9599    0.9553    1.0000
```

Because all correlation coefficients are close to 1, there is a strong positive correlation between each pair of data columns in the `count` matrix.

# Linear Regression

## In this section...

“Introduction” on page 2-5

“Simple Linear Regression” on page 2-5

“Residuals and Goodness of Fit” on page 2-9

“Fitting Data with Curve Fitting Toolbox Functions” on page 2-11

## Introduction

A data *model* explicitly describes a relationship between predictor and response variables. Linear regression fits a data model that is linear in the model coefficients. The most common type of linear regression is a least-squares fit, which can fit both lines and polynomials, among other linear models.

Before you model the relationship between pairs of quantities, it is a good idea to perform correlation analysis to establish if a linear relationship exists between these quantities. Be aware that variables can have nonlinear relationships, which correlation analysis cannot detect. For more information, see “Linear Correlation” on page 2-2.

The MATLAB Basic Fitting UI helps you to fit your data, so you can calculate model coefficients and plot the model on top of the data. For an example, see “Example: Using Basic Fitting UI” on page 2-14. You also can use the MATLAB `polyfit` and `polyval` functions to fit your data to a model that is linear in the coefficients. For an example, see “Programmatic Fitting” on page 2-28.

If you need to fit data with a nonlinear model, transform the variables to make the relationship linear. Alternatively, try to fit a nonlinear function directly using either the Statistics and Machine Learning Toolbox `nlinfit` function, the Optimization Toolbox™ `lsqcurvefit` function, or by applying functions in the Curve Fitting Toolbox™.

This topic explains how to:

- Perform simple linear regression using the `\` operator.
- Use correlation analysis to determine whether two quantities are related to justify fitting the data.
- Fit a linear model to the data.
- Evaluate the goodness of fit by plotting residuals and looking for patterns.
- Calculate measures of goodness of fit  $R^2$  and adjusted  $R^2$

## Simple Linear Regression

This example shows how to perform simple linear regression using the `accidents` dataset. The example also shows you how to calculate the coefficient of determination  $R^2$  to evaluate the regressions. The `accidents` dataset contains data for fatal traffic accidents in U.S. states.

Linear regression models the relation between a dependent, or response, variable  $y$  and one or more independent, or predictor, variables  $x_1, \dots, x_n$ . Simple linear regression considers only one independent variable using the relation

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where  $\beta_0$  is the y-intercept,  $\beta_1$  is the slope (or regression coefficient), and  $e$  is the error term.

Start with a set of  $n$  observed values of  $x$  and  $y$  given by  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Using the simple linear regression relation, these values form a system of linear equations. Represent these equations in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, B = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

The relation is now  $Y = XB$ .

In MATLAB, you can find  $B$  using the `mldivide` operator as  $B = X \backslash Y$ .

From the dataset `accidents`, load accident data in `y` and state population data in `x`. Find the linear regression relation  $y = \beta_1 x$  between the accidents in a state and the population of a state using the `\` operator. The `\` operator performs a least-squares regression.

```
load accidents
x = hwydata(:,14); %Population of states
y = hwydata(:,4); %Accidents per state
format long
b1 = x \ y

b1 =
    1.372716735564871e-04
```

`b1` is the slope or regression coefficient. The linear relation is  $y = \beta_1 x = 0.0001372x$ .

Calculate the accidents per state `yCalc` from `x` using the relation. Visualize the regression by plotting the actual values `y` and the calculated values `yCalc`.

```
yCalc1 = b1*x;
scatter(x,y)
hold on
plot(x,yCalc1)
xlabel('Population of state')
ylabel('Fatal traffic accidents per state')
title('Linear Regression Relation Between Accidents & Population')
grid on
```





Improve the fit by including a y-intercept  $\beta_0$  in your model as  $y = \beta_0 + \beta_1x$ . Calculate  $\beta_0$  by padding  $x$  with a column of ones and using the  $\backslash$  operator.

```
X = [ones(length(x),1) x];
b = X\y
```

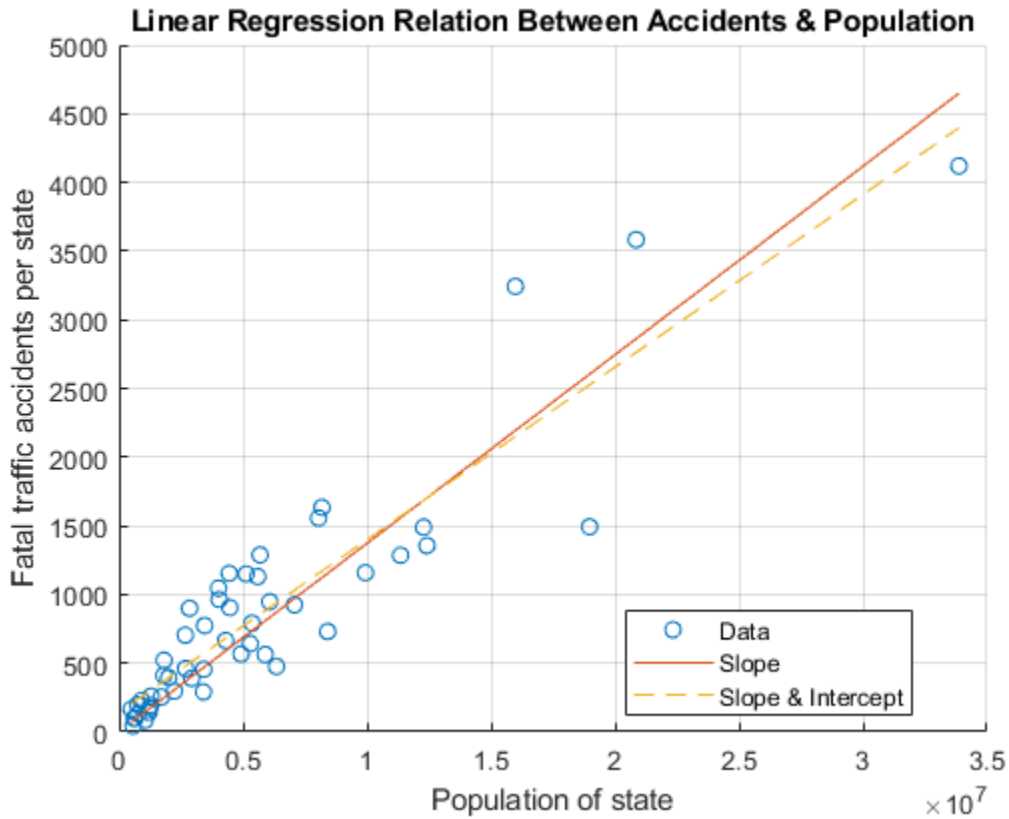
```
b = 2×1
    102 ×
```

```
    1.427120171726537
    0.000001256394274
```

This result represents the relation  $y = \beta_0 + \beta_1x = 142.7120 + 0.0001256x$ .

Visualize the relation by plotting it on the same figure.

```
yCalc2 = X*b;
plot(x,yCalc2,'--')
legend('Data','Slope','Slope & Intercept','Location','best');
```



From the figure, the two fits look similar. One method to find the better fit is to calculate the coefficient of determination,  $R^2$ .  $R^2$  is one measure of how well a model can predict the data, and falls between 0 and 1. The higher the value of  $R^2$ , the better the model is at predicting the data.

Where  $\hat{y}$  represents the calculated values of  $y$  and  $\bar{y}$  is the mean of  $y$ ,  $R^2$  is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Find the better fit of the two fits by comparing values of  $R^2$ . As the  $R^2$  values show, the second fit that includes a  $y$ -intercept is better.

$$Rs_{q1} = 1 - \frac{\text{sum}((y - y_{\text{Calc1}}).^2)}{\text{sum}((y - \text{mean}(y)).^2)}$$

$$Rs_{q1} = 0.822235650485566$$

$$Rs_{q2} = 1 - \frac{\text{sum}((y - y_{\text{Calc2}}).^2)}{\text{sum}((y - \text{mean}(y)).^2)}$$

$$Rs_{q2} = 0.838210531103428$$

## Residuals and Goodness of Fit

Residuals are the difference between the *observed* values of the response (dependent) variable and the values that a model *predicts*. When you fit a model that is appropriate for your data, the residuals approximate independent random errors. That is, the distribution of residuals ought not to exhibit a discernible pattern.

Producing a fit using a linear model requires minimizing the sum of the squares of the residuals. This minimization yields what is called a least-squares fit. You can gain insight into the “goodness” of a fit by visually examining a plot of the residuals. If the residual plot has a pattern (that is, residual data points do not appear to have a random scatter), the randomness indicates that the model does not properly fit the data.

Evaluate each fit you make in the context of your data. For example, if your goal of fitting the data is to extract coefficients that have physical meaning, then it is important that your model reflect the physics of the data. Understanding what your data represents, how it was measured, and how it is modeled is important when evaluating the goodness of fit.

One measure of goodness of fit is the coefficient of determination, or  $R^2$  (pronounced r-square). This statistic indicates how closely values you obtain from fitting a model match the dependent variable the model is intended to predict. Statisticians often define  $R^2$  using the residual variance from a fitted model:

$$R^2 = 1 - SS_{\text{resid}} / SS_{\text{total}}$$

$SS_{\text{resid}}$  is the sum of the squared residuals from the regression.  $SS_{\text{total}}$  is the sum of the squared differences from the mean of the dependent variable (*total sum of squares*). Both are positive scalars.

To learn how to compute  $R^2$  when you use the Basic Fitting tool, see “ $R^2$ , the Coefficient of Determination” on page 2-19. To learn more about calculating the  $R^2$  statistic and its multivariate generalization, continue reading here.

### Example: Computing $R^2$ from Polynomial Fits

You can derive  $R^2$  from the coefficients of a polynomial regression to determine how much variance in  $y$  a linear model explains, as the following example describes:

- 1 Create two variables,  $x$  and  $y$ , from the first two columns of the count variable in the data file `count.dat`:

```
load count.dat
x = count(:,1);
y = count(:,2);
```

- 2 Use `polyfit` to compute a linear regression that predicts  $y$  from  $x$ :

```
p = polyfit(x,y,1)

p =
    1.5229    -2.1911
```

`p(1)` is the slope and `p(2)` is the intercept of the linear predictor. You can also obtain regression coefficients using the Basic Fitting UI on page 2-13.

- 3 Call `polyval` to use `p` to predict  $y$ , calling the result `yfit`:

```
yfit = polyval(p,x);
```

Using `polyval` saves you from typing the fit equation yourself, which in this case looks like:

```
yfit = p(1) * x + p(2);
```

- 4 Compute the residual values as a vector of signed numbers:

```
yresid = y - yfit;
```

- 5 Square the residuals and total them to obtain the residual sum of squares:

```
SSresid = sum(yresid.^2);
```

- 6 Compute the total sum of squares of `y` by multiplying the variance of `y` by the number of observations minus 1:

```
SStotal = (length(y)-1) * var(y);
```

- 7 Compute  $R^2$  using the formula given in the introduction of this topic:

```
rsq = 1 - SSresid/SStotal
```

```
rsq =  
0.8707
```

This demonstrates that the linear equation  $1.5229 * x - 2.1911$  predicts 87% of the variance in the variable `y`.

### Computing Adjusted R2 for Polynomial Regressions

You can usually reduce the residuals in a model by fitting a higher degree polynomial. When you add more terms, you increase the coefficient of determination,  $R^2$ . You get a closer fit to the data, but at the expense of a more complex model, for which  $R^2$  cannot account. However, a refinement of this statistic, adjusted  $R^2$ , does include a penalty for the number of terms in a model. Adjusted  $R^2$ , therefore, is more appropriate for comparing how different models fit to the same data. The adjusted  $R^2$  is defined as:

$$R^2_{\text{adjusted}} = 1 - (SS_{\text{resid}} / SS_{\text{total}}) * ((n-1)/(n-d-1))$$

where  $n$  is the number of observations in your data, and  $d$  is the degree of the polynomial. (A linear fit has a degree of 1, a quadratic fit 2, a cubic fit 3, and so on.)

The following example repeats the steps of the previous example, “Example: Computing R2 from Polynomial Fits” on page 2-9, but performs a cubic (degree 3) fit instead of a linear (degree 1) fit. From the cubic fit, you compute both simple and adjusted  $R^2$  values to evaluate whether the extra terms improve predictive power:

- 1 Create two variables, `x` and `y`, from the first two columns of the `count` variable in the data file `count.dat`:

```
load count.dat  
x = count(:,1);  
y = count(:,2);
```

- 2 Call `polyfit` to generate a cubic fit to predict `y` from `x`:

```
p = polyfit(x,y,3)  
  
p =  
-0.0003    0.0390    0.2233    6.2779
```

$p(4)$  is the intercept of the cubic predictor. You can also obtain regression coefficients using the Basic Fitting UI on page 2-13.

- 3 Call `polyval` to use the coefficients in `p` to predict `y`, naming the result `yfit`:

```
yfit = polyval(p,x);
```

`polyval` evaluates the explicit equation you could manually enter as:

```
yfit = p(1) * x.^3 + p(2) * x.^2 + p(3) * x + p(4);
```

- 4 Compute the residual values as a vector of signed numbers:

```
yresid = y - yfit;
```

- 5 Square the residuals and total them to obtain the residual sum of squares:

```
SSresid = sum(yresid.^2);
```

- 6 Compute the total sum of squares of `y` by multiplying the variance of `y` by the number of observations minus 1:

```
SStotal = (length(y)-1) * var(y);
```

- 7 Compute simple  $R^2$  for the cubic fit using the formula given in the introduction of this topic:

```
rsq = 1 - SSresid/SStotal
```

```
rsq =  
0.9083
```

- 8 Finally, compute adjusted  $R^2$  to account for degrees of freedom:

```
rsq_adj = 1 - SSresid/SStotal * (length(y)-1)/(length(y)-length(p))
```

```
rsq_adj =  
0.8945
```

The adjusted  $R^2$ , 0.8945, is smaller than simple  $R^2$ , .9083. It provides a more reliable estimate of the power of your polynomial model to predict.

In many polynomial regression models, adding terms to the equation increases both  $R^2$  and adjusted  $R^2$ . In the preceding example, using a cubic fit increased both statistics compared to a linear fit. (You can compute adjusted  $R^2$  for the linear fit for yourself to demonstrate that it has a lower value.) However, it is not always true that a linear fit is worse than a higher-order fit: a more complicated fit can have a lower adjusted  $R^2$  than a simpler fit, indicating that the increased complexity is not justified. Also, while  $R^2$  always varies between 0 and 1 for the polynomial regression models that the Basic Fitting tool generates, adjusted  $R^2$  for some models can be negative, indicating that a model that has too many terms.

Correlation does not imply causality. Always interpret coefficients of correlation and determination cautiously. The coefficients only quantify how much variance in a dependent variable a fitted model removes. Such measures do not describe how appropriate your model—or the independent variables you select—are for explaining the behavior of the variable the model predicts.

## Fitting Data with Curve Fitting Toolbox Functions

The Curve Fitting Toolbox software extends core MATLAB functionality by enabling the following data-fitting capabilities:

- Linear and nonlinear parametric fitting, including standard linear least squares, nonlinear least squares, weighted least squares, constrained least squares, and robust fitting procedures
- Nonparametric fitting
- Statistics for determining the goodness of fit
- Extrapolation, differentiation, and integration
- Dialog box that facilitates data sectioning and smoothing
- Saving fit results in various formats, including MATLAB code files, MAT-files, and workspace variables

For more information, see the Curve Fitting Toolbox documentation.

## Interactive Fitting

### In this section...

“Basic Fitting UI” on page 2-13

“Preparing for Basic Fitting” on page 2-13

“Opening the Basic Fitting UI” on page 2-13

“Example: Using Basic Fitting UI” on page 2-14

## Basic Fitting UI

The MATLAB Basic Fitting UI allows you to interactively:

- Model data using a spline interpolant, a shape-preserving interpolant, or a polynomial up to the tenth degree
- Plot one or more fits together with data
- Plot the residuals of the fits
- Compute model coefficients
- Compute the norm of the residuals (a statistic you can use to analyze how well a model fits your data)
- Use the model to interpolate or extrapolate outside of the data
- Save coefficients and computed values to the MATLAB workspace for use outside of the dialog box
- Generate MATLAB code to recompute fits and reproduce plots with new data

---

**Note** The Basic Fitting UI is only available for 2-D plots. For more advanced fitting and regression analysis, see the Curve Fitting Toolbox documentation and the Statistics and Machine Learning Toolbox documentation.

---

## Preparing for Basic Fitting

The Basic Fitting UI sorts your data in ascending order before fitting. If your data set is large and the values are not sorted in ascending order, it will take longer for the Basic Fitting UI to preprocess your data before fitting.

You can speed up the Basic Fitting UI by first sorting your data. To create sorted vectors `x_sorted` and `y_sorted` from data vectors `x` and `y`, use the MATLAB `sort` function:

```
[x_sorted, i] = sort(x);  
y_sorted = y(i);
```

## Opening the Basic Fitting UI

To use the Basic Fitting UI, you must first plot your data in a figure window, using any MATLAB plotting command that produces (only) `x` and `y` data.

To open the Basic Fitting UI, select **Tools > Basic Fitting** from the menus at the top of the figure window.

## Example: Using Basic Fitting UI

This example shows how to use the Basic Fitting UI to fit, visualize, analyze, save, and generate code for polynomial regressions.

### Load and Plot Census Data

The file `census.mat` contains U.S. population data for the years 1790 through 1990 at 10 year intervals.

To load and plot the data, type the following commands at the MATLAB prompt:

```
load census
plot(cdate,pop, 'ro')
```

The `load` command adds the following variables to the MATLAB workspace:

- `cdate` — A column vector containing the years from 1790 to 1990 in increments of 10. It is the predictor variable.
- `pop` — A column vector with U.S. population for each year in `cdate`. It is the response variable.

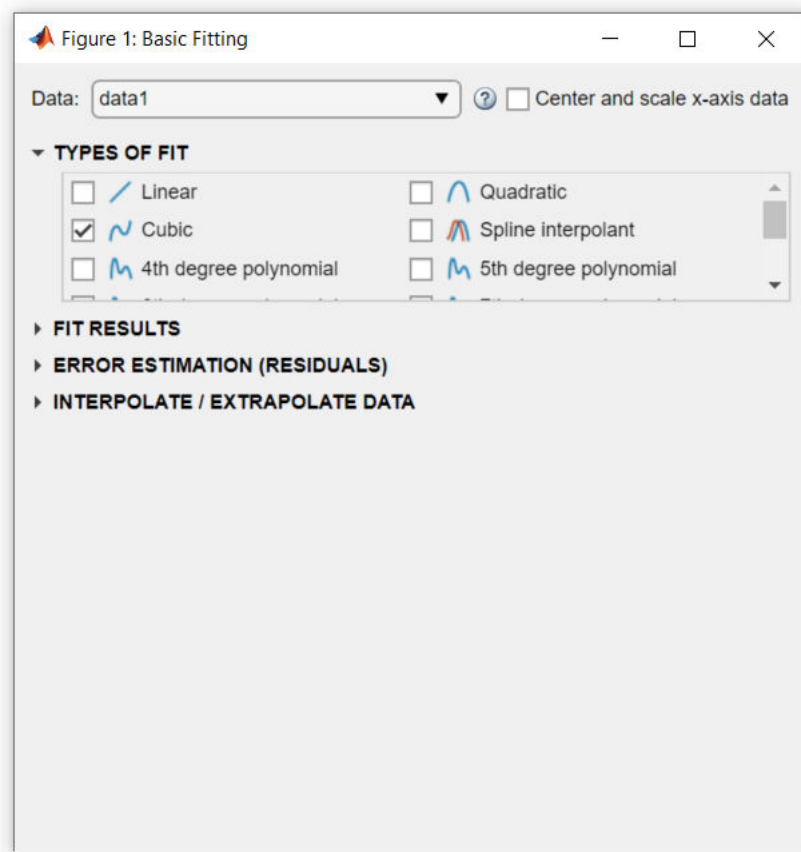
The data vectors are sorted in ascending order, by year. The plot shows the population as a function of year.

Now you are ready to fit an equation the data to model population growth over time.

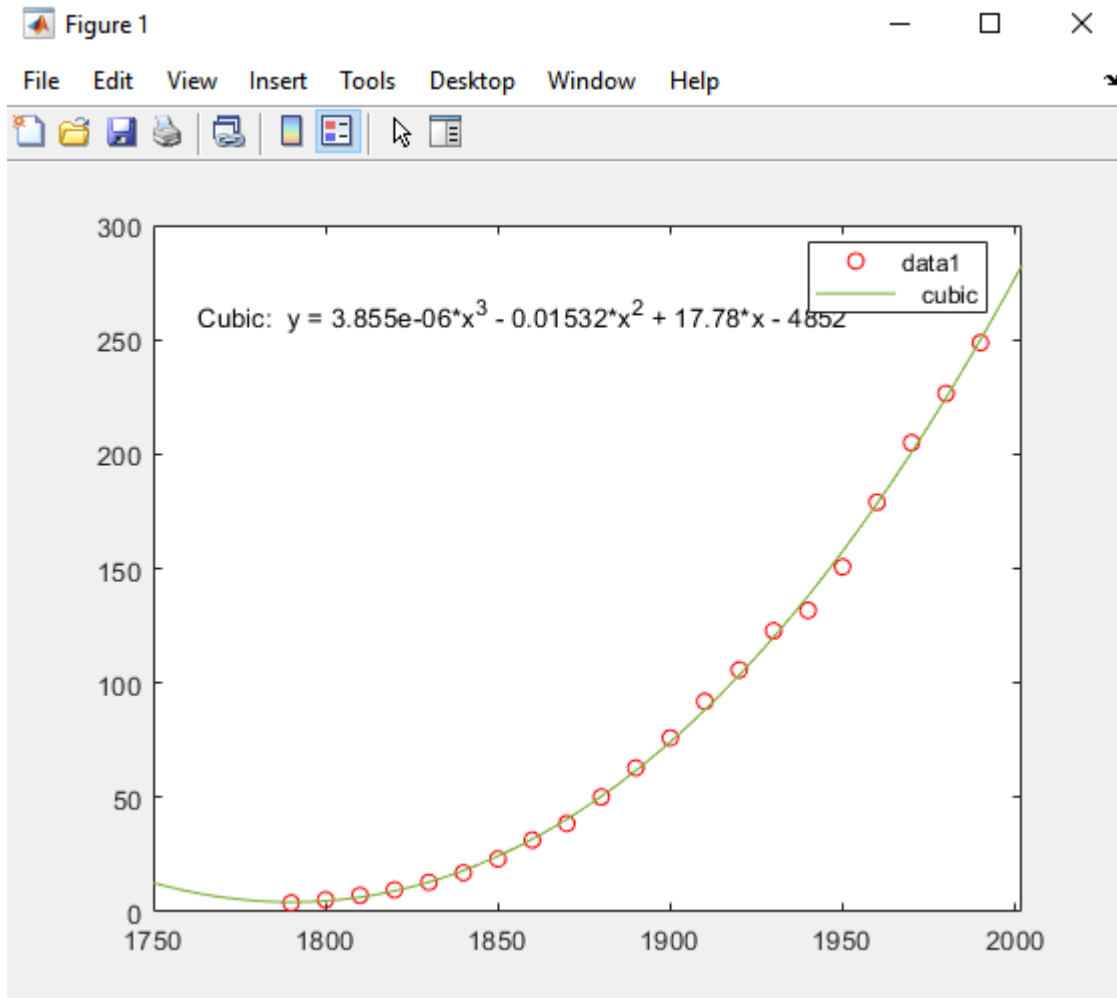
### Predict the Census Data with a Cubic Polynomial Fit

- 1 Open the Basic Fitting dialog box by selecting **Tools > Basic Fitting** in the Figure window.
- 2 In the **TYPES OF FIT** area of the Basic Fitting dialog box, select the **Cubic** check box to fit a cubic polynomial to the data.

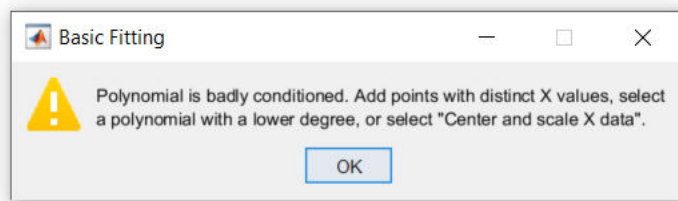




MATLAB uses your selection to fit the data, and adds the cubic regression line to the graph as follows.



In computing the fit, MATLAB encounters problems and issues the following warning:

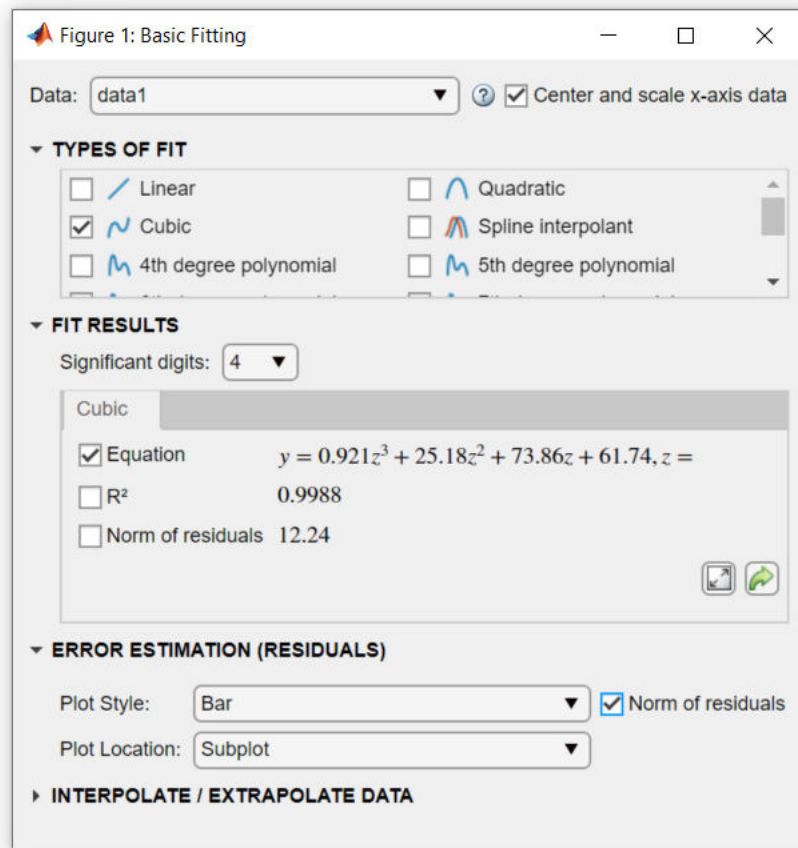


This warning indicates that the computed coefficients for the model are sensitive to random errors in the response (the measured population). It also suggests some things you can do to get a better fit.

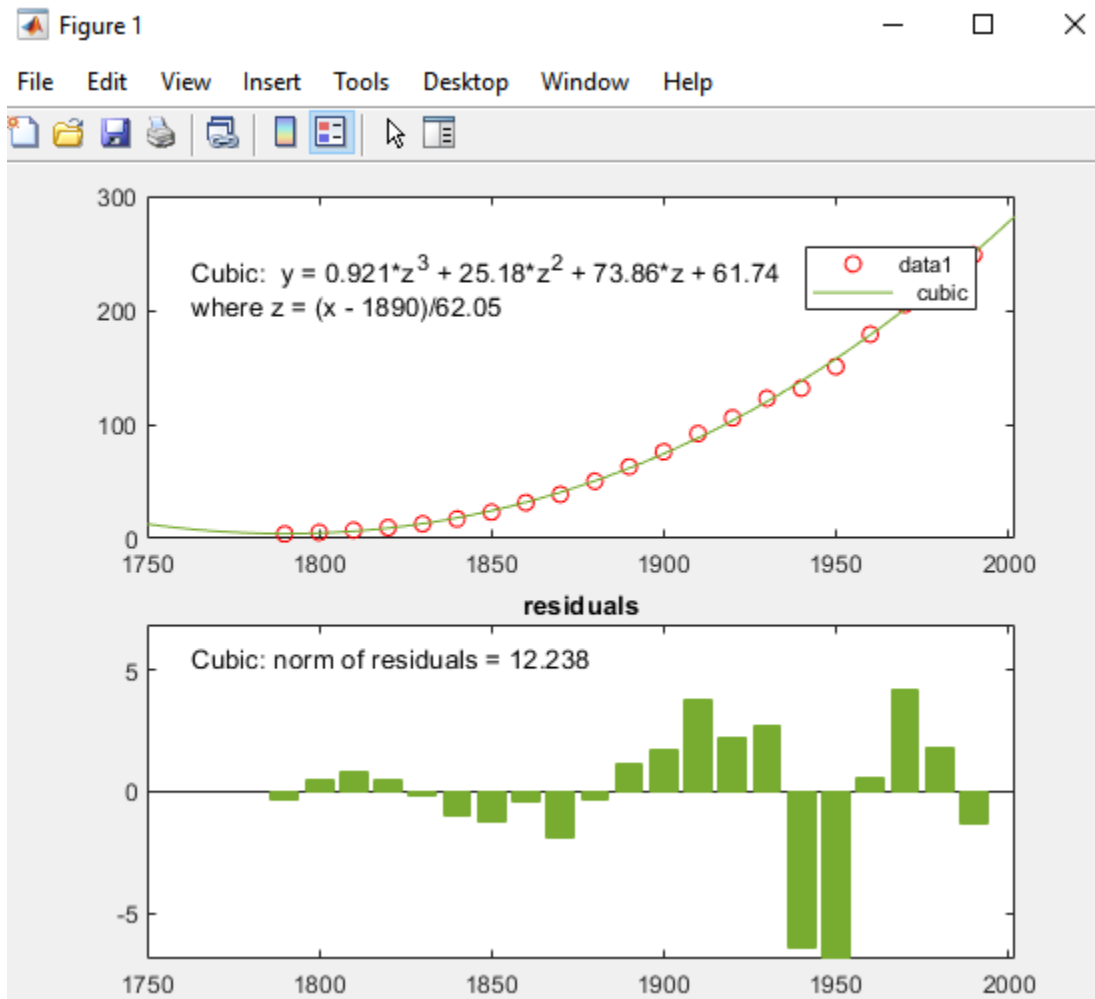
- 3 Continue to use a cubic fit. As you cannot add new observations to the census data, improve the fit by transforming the values you have to *z-scores* before recomputing a fit. Select the **Center and scale x-axis data** check box in the top right of the dialog box to make the Basic Fitting tool perform the transformation.

To learn how centering and scaling data works, see "Learn How the Basic Fitting Tool Computes Fits" on page 2-23.

- 4 Under **ERROR ESTIMATION (RESIDUALS)**, select the **Norm of residuals** check box. Select **Bar** as the **Plot Style**.



Selecting these options creates a subplot of residuals as a bar graph.

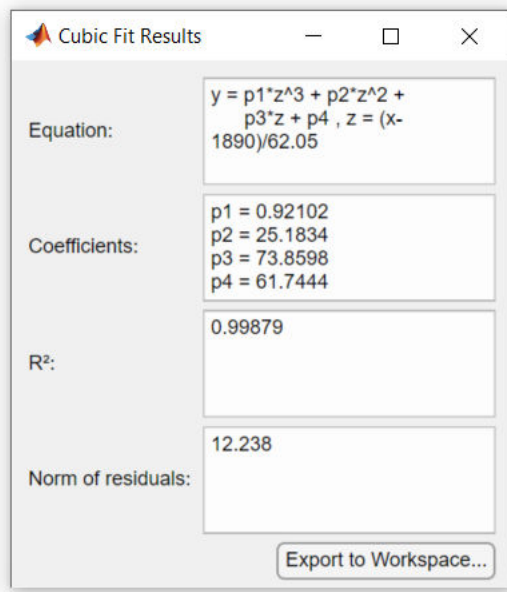


The cubic fit is a poor predictor before the year 1790, where it indicates a decreasing population. The model seems to approximate the data reasonably well after 1790. However, a pattern in the residuals shows that the model does not meet the assumption of normal error, which is a basis for the least-squares fitting. The **data 1** line identified in the legend are the observed  $x$  (cdate) and  $y$  (pop) data values. The **Cubic** regression line presents the fit after centering and scaling data values. Notice that the figure shows the original data units, even though the tool computes the fit using transformed  $z$ -scores.

For comparison, try fitting another polynomial equation to the census data by selecting it in the **TYPES OF FIT** area.

### View and Save the Cubic Fit Parameters

In the Basic Fitting dialog box, click the **Expand Results** button  to display the estimated coefficients and the norm of residuals.



Save the fit data to the MATLAB workspace by clicking the **Export to Workspace** button on the Numerical results panel. The Save Fit to Workspace dialog box opens.

With all check boxes selected, click **OK** to save the fit parameters as a MATLAB structure `fit`:

```
fit
fit =
    struct with fields:
        type: 'polynomial degree 3'
        coeff: [0.9210 25.1834 73.8598 61.7444]
```

Now, you can use the fit results in MATLAB programming, outside of the Basic Fitting UI.

## R<sup>2</sup>, the Coefficient of Determination

You can get an indication of how well a polynomial regression predicts your observed data by computing the coefficient of determination, or R-square (written as R<sup>2</sup>). The R<sup>2</sup> statistic, which ranges from 0 to 1, measures how useful the independent variable is in predicting values of the dependent variable:

- An R<sup>2</sup> value near 0 indicates that the fit is not much better than the model  $y = \text{constant}$ .
- An R<sup>2</sup> value near 1 indicates that the independent variable explains most of the variability in the dependent variable.

R<sup>2</sup> is computed from the residuals, the signed differences between an observed dependent value and the value your fit predicts for it.

$$\text{residuals} = y_{\text{observed}} - y_{\text{fitted}} \quad (2-1)$$

The R<sup>2</sup> number for the cubic fit in this example, 0.9988, is located under **FIT RESULTS** in the Basic Fitting dialog.

To compare the  $R^2$  number for the cubic fit to a linear least-squares fit, select **Linear** under **TYPES OF FIT** and obtain the  $R^2$  number, 0.921. This result indicates that a linear least-squares fit of the population data explains 92.1% of its variance. As the cubic fit of this data explains 99.9% of that variance, the latter seems to be a better predictor. However, because a cubic fit predicts using three variables ( $x$ ,  $x^2$ , and  $x^3$ ), a basic  $R^2$  value does not fully reflect how robust the fit is. A more appropriate measure for evaluating the goodness of multivariate fits is adjusted  $R^2$ . For information about computing and using adjusted  $R^2$ , see “Residuals and Goodness of Fit” on page 2-9.

### **Interpolate and Extrapolate Population Values**

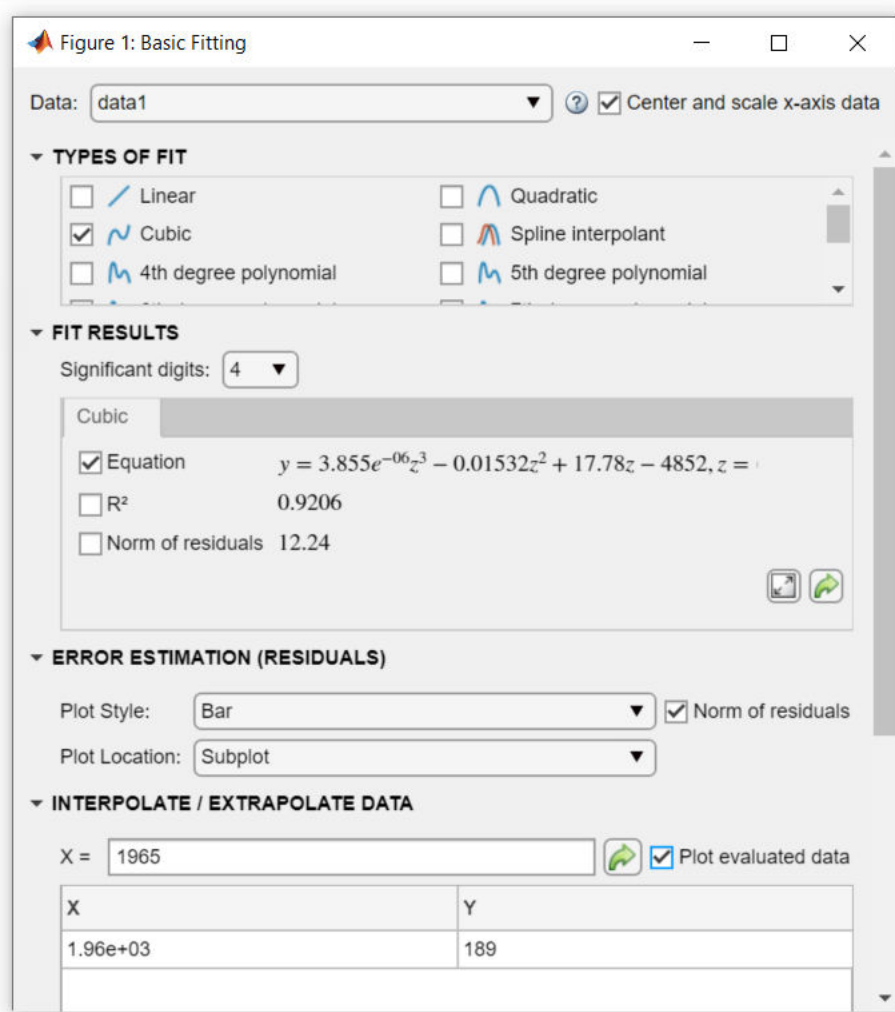
Suppose you want to use the cubic model to interpolate the U.S. population in 1965 (a date not provided in the original data).

In the Basic Fitting dialog box, under **INTERPOLATE / EXTRAPOLATE DATA**, enter the **X** value 1965 and check the **Plot evaluated data** box.

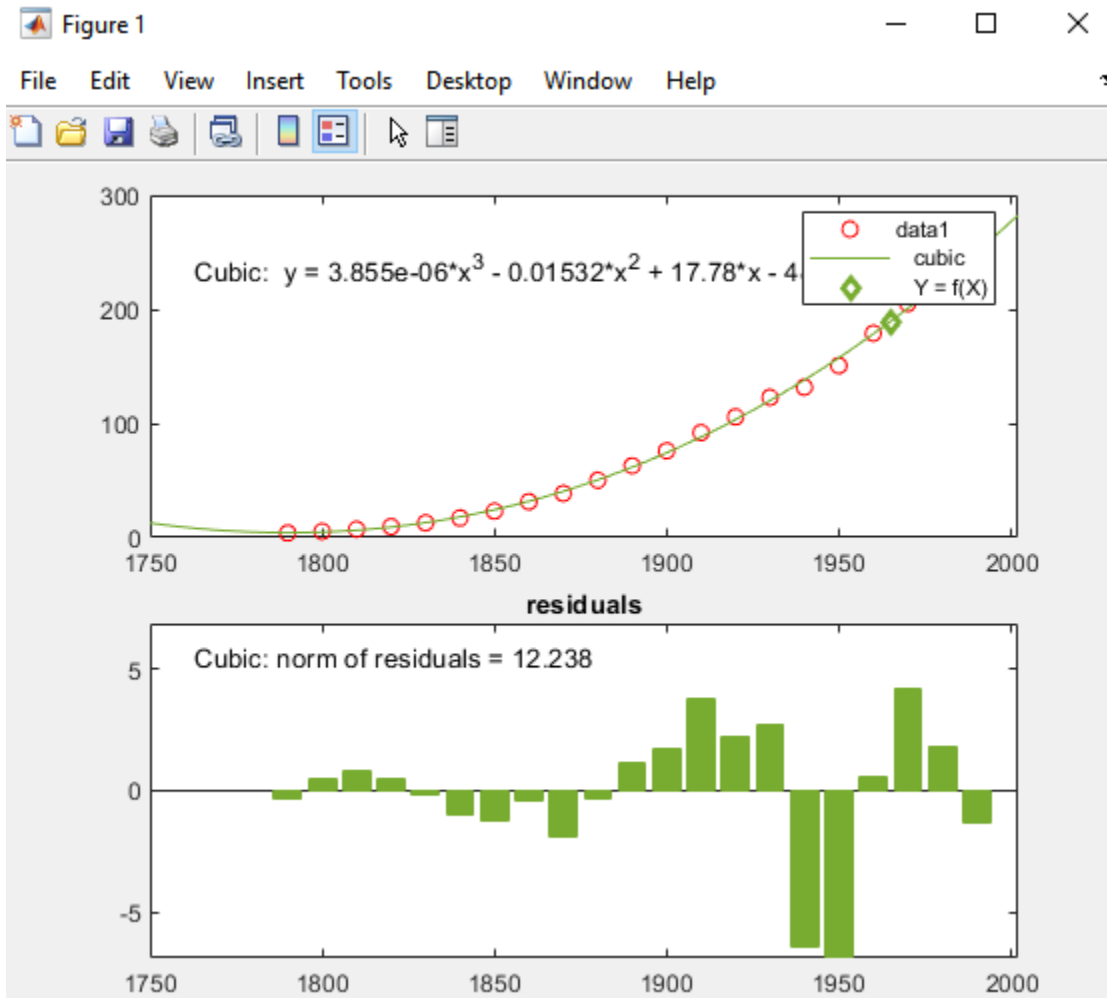
---

**Note** Use unscaled and uncentered  $X$  values. You do not need to center and scale first, even though you selected to scale  $X$  values to obtain the coefficients in “Predict the Census Data with a Cubic Polynomial Fit” on page 2-14. The Basic Fitting tool makes the necessary adjustments behind the scenes.

---



The X values and the corresponding values for  $f(X)$  are computed from the fit and plotted as follows:



### Generate a Code File to Reproduce the Result

After completing a Basic Fitting session, you can generate MATLAB code that recomputes fits and reproduces plots with new data.

- 1 In the Figure window, select **File > Generate Code**.

This creates a function and displays it in the MATLAB Editor. The code shows you how to programmatically reproduce what you did interactively with the Basic Fitting dialog box.

- 2 Change the name of the function on the first line from `createfigure` to something more specific, like `censusplot`. Save the code file to your current folder with the file name `censusplot.m`. The function begins with:

```
function censusplot(X1, Y1, valuesToEvaluate1)
```

- 3 Generate some new, randomly perturbed census data:

```
rng('default')
randpop = pop + 10*randn(size(pop));
```

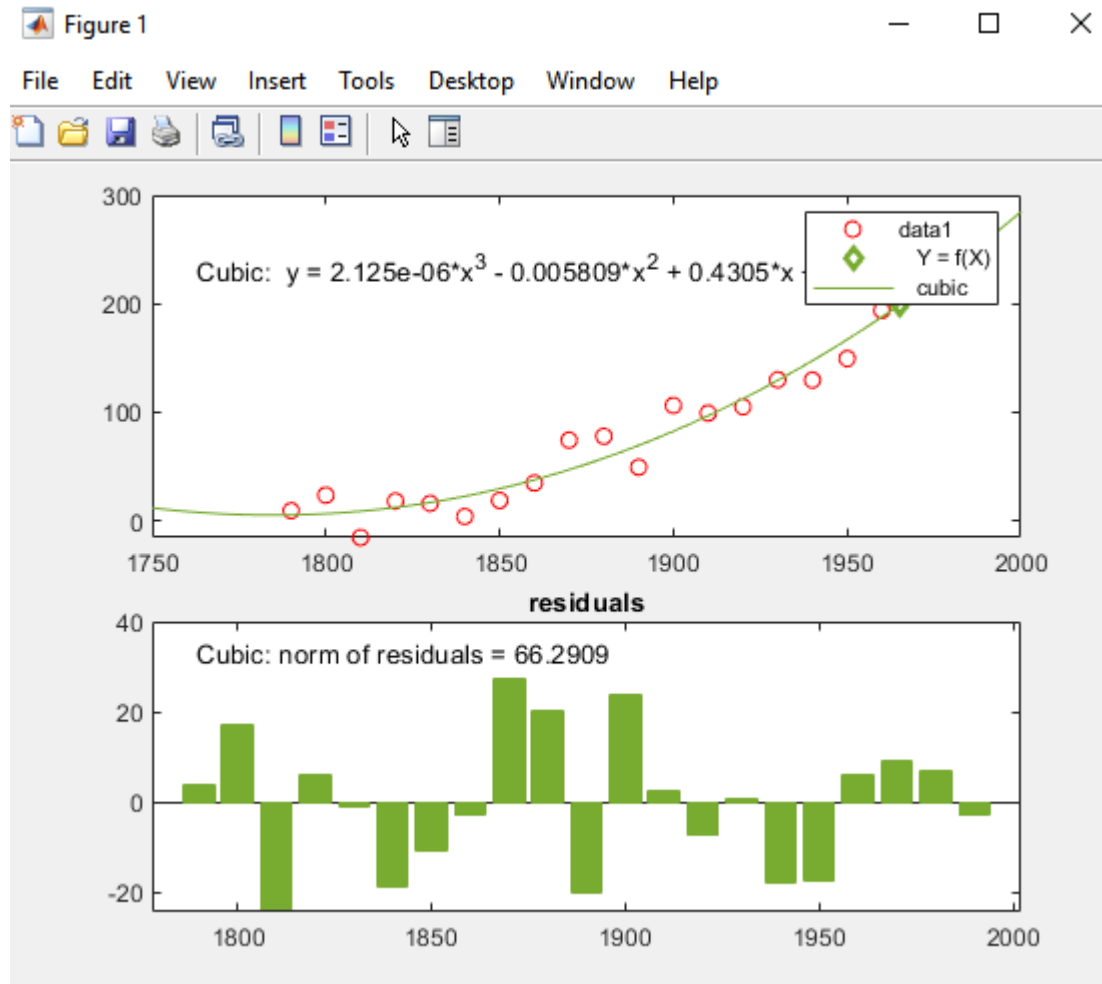
- 4 Reproduce the plot with the new data and recompute the fit:

```
censusplot(cdate, randpop, 1965)
```



You need three input arguments:  $x,y$  values (data 1) plotted in the original graph, plus an  $x$ -value for a marker.

The following figure displays the plot that the generated code produces. The new plot matches the appearance of the figure from which you generated code except for the  $y$  data values, the equation for the cubic fit, and the residual values in the bar graph, as expected.



### Learn How the Basic Fitting Tool Computes Fits

The Basic Fitting tool calls the `polyfit` function to compute polynomial fits. It calls the `polyval` function to evaluate the fits. `polyfit` analyzes its inputs to determine if the data is well conditioned for the requested degree of fit.

When it finds badly conditioned data, `polyfit` computes a regression as well as it can, but it also returns a warning that the fit could be improved. The Basic Fitting example section “Predict the Census Data with a Cubic Polynomial Fit” on page 2-14 displays this warning.

One way to improve model reliability is to add data points. However, adding observations to a data set is not always feasible. An alternative strategy is to transform the predictor variable to normalize its center and scale. (In the example, the predictor is the vector of census dates.)

The `polyfit` function normalizes by computing  $z$ -scores:

$$z = \frac{x - \mu}{\sigma}$$

where  $x$  is the predictor data,  $\mu$  is the mean of  $x$ , and  $\sigma$  is the standard deviation of  $x$ . The  $z$ -scores give the data a mean of 0 and a standard deviation of 1. In the Basic Fitting UI, you transform the predictor data to  $z$ -scores by selecting the **Center and scale x-axis data** check box.

After centering and scaling, model coefficients are computed for the  $y$  data as a function of  $z$ . These are different (and more robust) than the coefficients computed for  $y$  as a function of  $x$ . The form of the model and the norm of the residuals do not change. The Basic Fitting UI automatically rescales the  $z$ -scores so that the fit plots on the same scale as the original  $x$  data.

To understand the way in which the centered and scaled data is used as an intermediary to create the final plot, run the following code in the Command Window:

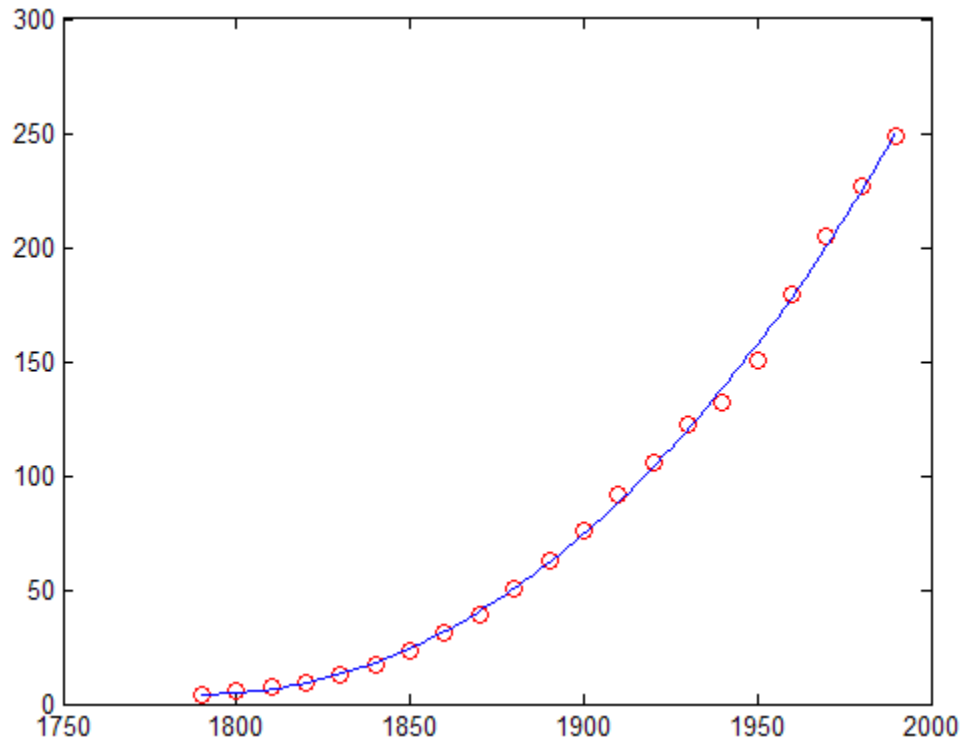
```
close
load census
x = cdate;
y = pop;
z = (x-mean(x))/std(x); % Compute z-scores of x data

plot(x,y,'ro') % Plot data as red markers
hold on % Prepare axes to accept new graph on top

zfit = linspace(z(1),z(end),100);
pz = polyfit(z,y,3); % Compute conditioned fit
yfit = polyval(pz,zfit);

xfit = linspace(x(1),x(end),100);
plot(xfit,yfit,'b-') % Plot conditioned fit vs. x data
```

The centered and scaled cubic polynomial plots as a blue line, as shown here:



In the code, computation of `z` illustrates how to normalize data. The `polyfit` function performs the transformation itself if you provide three return arguments when calling it:

```
[p,S,mu] = polyfit(x,y,n)
```

The returned regression parameters, `p`, now are based on normalized `x`. The returned vector, `mu`, contains the mean and standard deviation of `x`. For more information, see the `polyfit` reference page.

## Programmatic Fitting

### In this section...

“MATLAB Functions for Polynomial Models” on page 2-26

“Linear Model with Nonpolynomial Terms” on page 2-26

“Multiple Regression” on page 2-27

“Programmatic Fitting” on page 2-28

## MATLAB Functions for Polynomial Models

Two MATLAB functions can model your data with a polynomial.

### Polynomial Fit Functions

Function	Description
<code>polyfit</code>	<code>polyfit(x, y, n)</code> finds the coefficients of a polynomial $p(x)$ of degree $n$ that fits the $y$ data by minimizing the sum of the squares of the deviations of the data from the model (least-squares fit).
<code>polyval</code>	<code>polyval(p, x)</code> returns the value of a polynomial of degree $n$ that was determined by <code>polyfit</code> , evaluated at $x$ .

If you are trying to model a physical situation, it is always important to consider whether a model of a specific order is meaningful in your situation.

## Linear Model with Nonpolynomial Terms

This example shows how to fit data with a linear model containing nonpolynomial terms.

When a polynomial function does not produce a satisfactory model of your data, you can try using a linear model with nonpolynomial terms. For example, consider the following function that is linear in the parameters  $a_0$ ,  $a_1$ , and  $a_2$ , but nonlinear in the  $t$  data:

$$y = a_0 + a_1e^{-t} + a_2te^{-t}.$$

You can compute the unknown coefficients  $a_0$ ,  $a_1$ , and  $a_2$  by constructing and solving a set of simultaneous equations and solving for the parameters. The following syntax accomplishes this by forming a *design matrix*, where each column represents a variable used to predict the response (a term in the model) and each row corresponds to one observation of those variables.

Enter  $t$  and  $y$  as column vectors.

```
t = [0 0.3 0.8 1.1 1.6 2.3]';
y = [0.6 0.67 1.01 1.35 1.47 1.25]';
```

Form the design matrix.

```
X = [ones(size(t)) exp(-t) t.*exp(-t)];
```

Calculate model coefficients.

```
a = X\y
```

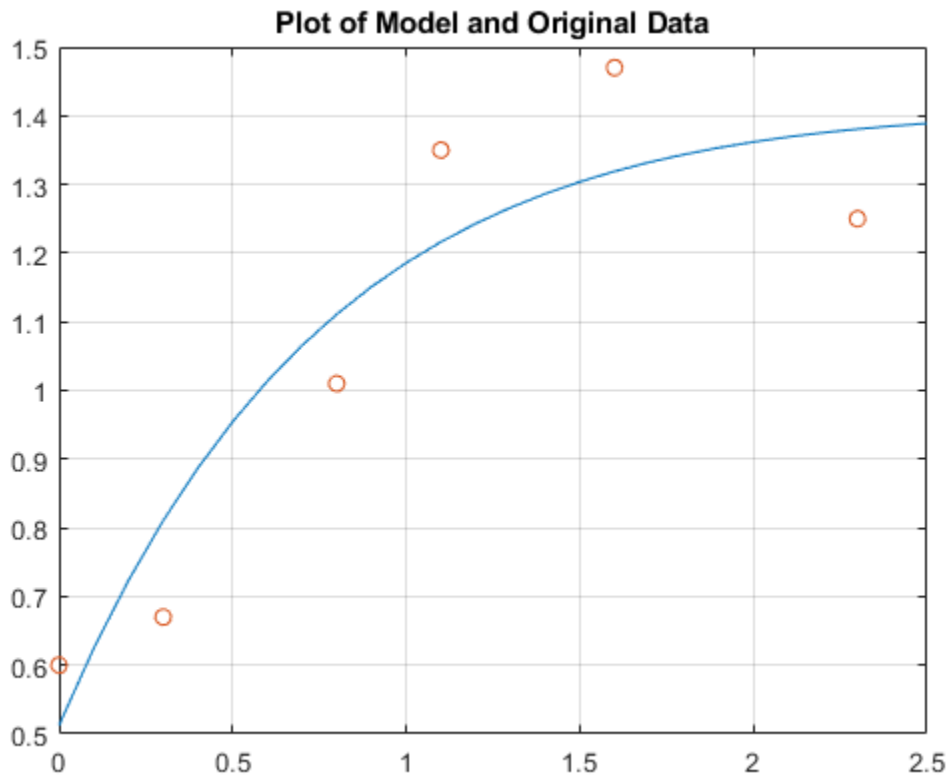
```
a = 3×1
    1.3983
   -0.8860
    0.3085
```

Therefore, the model of the data is given by

$$y = 1.3983 - 0.8860e^{-t} + 0.3085te^{-t}.$$

Now evaluate the model at regularly spaced points and plot the model with the original data.

```
T = (0:0.1:2.5)';
Y = [ones(size(T)) exp(-T) T.*exp(-T)]*a;
plot(T,Y,'-',t,y,'o'), grid on
title('Plot of Model and Original Data')
```



## Multiple Regression

This example shows how to use multiple regression to model data that is a function of more than one predictor variable.

When  $y$  is a function of more than one predictor variable, the matrix equations that express the relationships among the variables must be expanded to accommodate the additional data. This is called *multiple regression*.

Measure a quantity  $y$  for several values of  $x_1$  and  $x_2$ . Store these values in vectors  $x_1$ ,  $x_2$ , and  $y$ , respectively.

```
x1 = [.2 .5 .6 .8 1.0 1.1]';  
x2 = [.1 .3 .4 .9 1.1 1.4]';  
y = [.17 .26 .28 .23 .27 .24]';
```

A model of this data is of the form

$$y = a_0 + a_1x_1 + a_2x_2.$$

Multiple regression solves for unknown coefficients  $a_0$ ,  $a_1$ , and  $a_2$  by minimizing the sum of the squares of the deviations of the data from the model (least-squares fit).

Construct and solve the set of simultaneous equations by forming a design matrix,  $X$ .

```
X = [ones(size(x1)) x1 x2];
```

Solve for the parameters by using the backslash operator.

```
a = X\y
```

```
a = 3×1
```

```
    0.1018  
    0.4844  
   -0.2847
```

The least-squares fit model of the data is

$$y = 0.1018 + 0.4844x_1 - 0.2847x_2.$$

To validate the model, find the maximum of the absolute value of the deviation of the data from the model.

```
Y = X*a;  
MaxErr = max(abs(Y - y))
```

```
MaxErr = 0.0038
```

This value is much smaller than any of the data values, indicating that this model accurately follows the data.

## Programmatic Fitting

This example shows how to use MATLAB functions to:

- “Calculate Correlation Coefficients” on page 2-29
- “Fit a Polynomial to the Data” on page 2-30
- “Plot and Calculate Confidence Bounds” on page 2-31

Load sample census data from `census.mat`, which contains U.S. population data from the years 1790 to 1990.

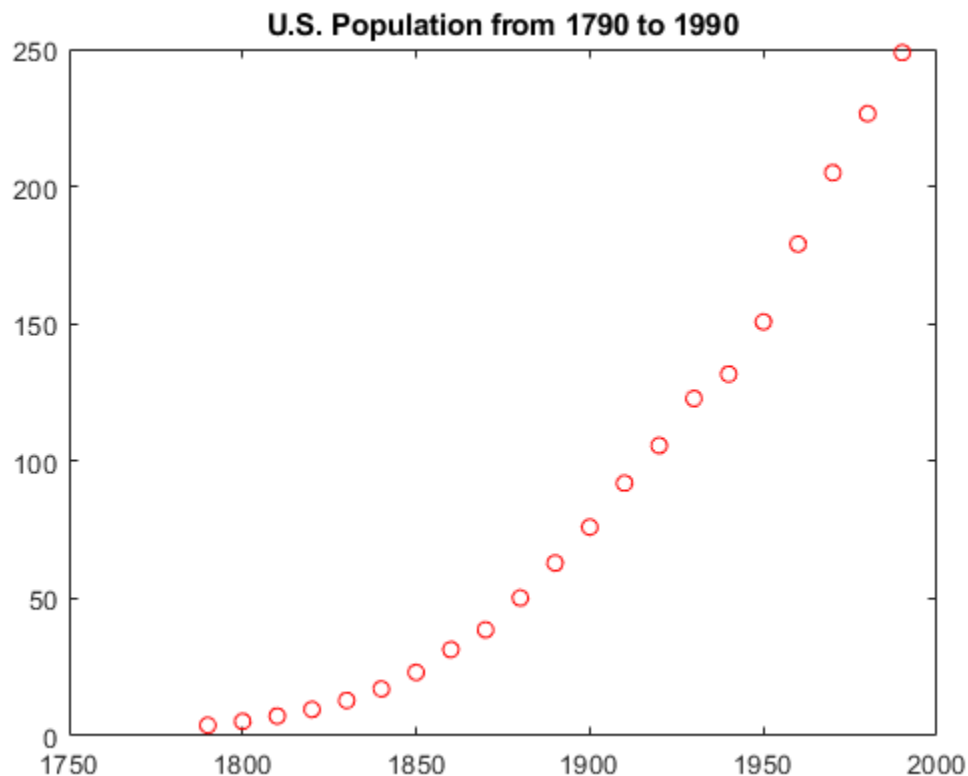
```
load census
```

This adds the following two variables to the MATLAB workspace.

- `cdate` is a column vector containing the years 1790 to 1990 in increments of 10.
- `pop` is a column vector with the U.S. population numbers corresponding to each year in `cdate`.

Plot the data.

```
plot(cdate,pop,'ro')
title('U.S. Population from 1790 to 1990')
```



The plot shows a strong pattern, which indicates a high correlation between the variables.

### Calculate Correlation Coefficients

In this portion of the example, you determine the statistical correlation between the variables `cdate` and `pop` to justify modeling the data. For more information about correlation coefficients, see “Linear Correlation” on page 2-2.

Calculate the correlation-coefficient matrix.

```
corrcoef(cdate,pop)
```

```
ans = 2x2
```

```
    1.0000    0.9597
```

```
0.9597    1.0000
```

The diagonal matrix elements represent the perfect correlation of each variable with itself and are equal to 1. The off-diagonal elements are very close to 1, indicating that there is a strong statistical correlation between the variables `cdate` and `pop`.

### Fit a Polynomial to the Data

This portion of the example applies the `polyfit` and `polyval` MATLAB functions to model the data.

Calculate fit parameters.

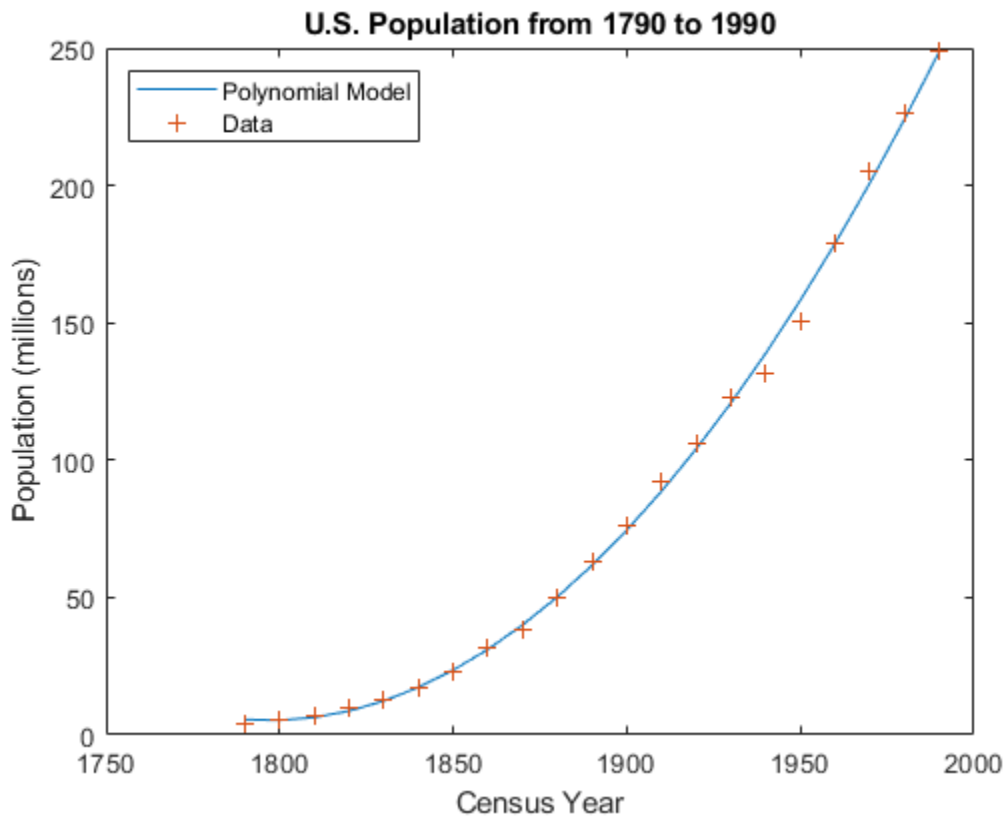
```
[p,ErrorEst] = polyfit(cdate,pop,2);
```

Evaluate the fit.

```
pop_fit = polyval(p,cdate,ErrorEst);
```

Plot the data and the fit.

```
plot(cdate,pop_fit,'-',cdate,pop,'+');
title('U.S. Population from 1790 to 1990')
legend('Polynomial Model','Data','Location','NorthWest');
xlabel('Census Year');
ylabel('Population (millions)');
```

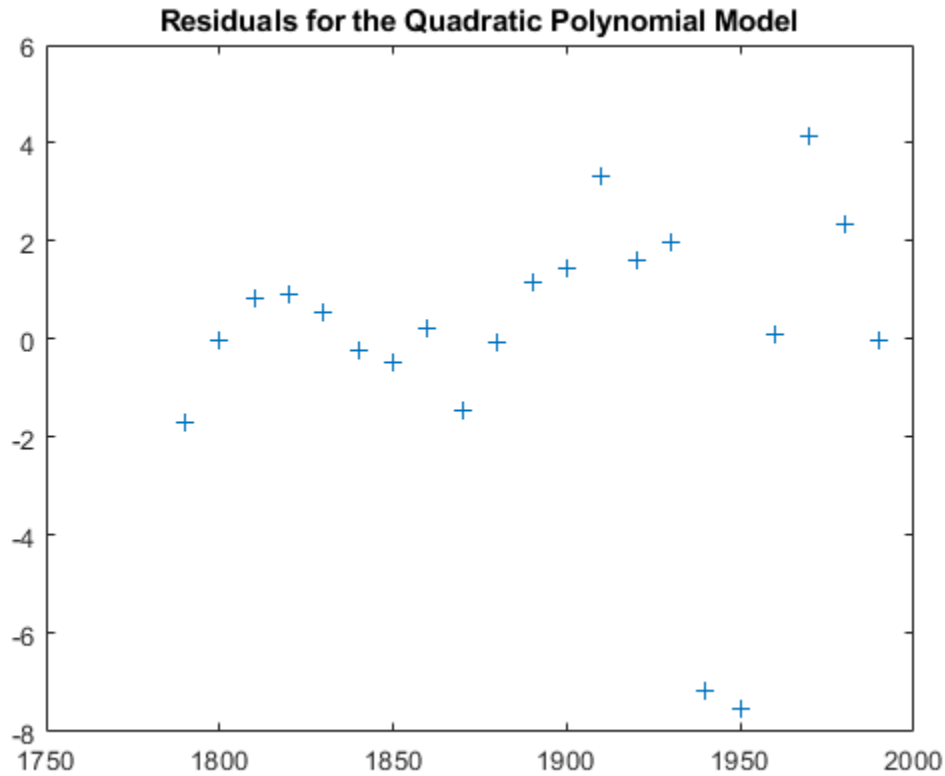


The plot shows that the quadratic-polynomial fit provides a good approximation to the data.



Calculate the residuals for this fit.

```
res = pop - pop_fit;
figure, plot(cdate,res,'+')
title('Residuals for the Quadratic Polynomial Model')
```



Notice that the plot of the residuals exhibits a pattern, which indicates that a second-degree polynomial might not be appropriate for modeling this data.

### Plot and Calculate Confidence Bounds

Confidence bounds are confidence intervals for a predicted response. The width of the interval indicates the degree of certainty of the fit.

This portion of the example applies `polyfit` and `polyval` to the census sample data to produce confidence bounds for a second-order polynomial model.

The following code uses an interval of  $\pm 2\Delta$ , which corresponds to a 95% confidence interval for large samples.

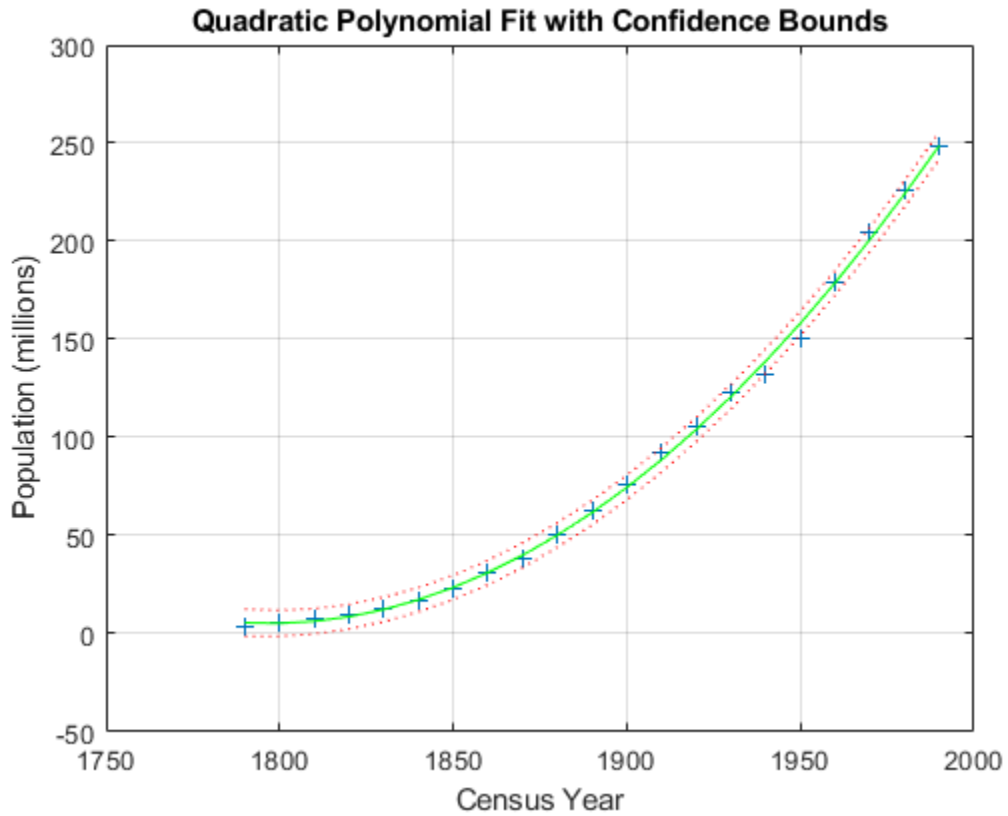
Evaluate the fit and the prediction error estimate (delta).

```
[pop_fit,delta] = polyval(p,cdate>ErrorEst);
```

Plot the data, the fit, and the confidence bounds.

```
plot(cdate,pop,'+',...
      cdate,pop_fit,'g-',...
      cdate,delta,'r-',...
      cdate,-delta,'r-',...)
```

```
cdate,pop_fit+2*delta,'r:',...  
cdate,pop_fit-2*delta,'r:');  
xlabel('Census Year');  
ylabel('Population (millions)');  
title('Quadratic Polynomial Fit with Confidence Bounds')  
grid on
```



The 95% interval indicates that you have a 95% chance that a new observation will fall within the bounds.

# Time Series Analysis

---

- “What Are Time Series?” on page 3-2
- “Time Series Objects and Collections” on page 3-3

## What Are Time Series?

Time series are data vectors sampled over time, in order, often at regular intervals. They are distinguished from randomly sampled data, which form the basis of many other data analyses. Time series represent the time-evolution of a dynamic population or *process*. The linear ordering of time series gives them a distinctive place in data analysis, with a specialized set of techniques.

Time series analysis is concerned with:

- Identifying patterns
- Modeling patterns
- Forecasting values

Several dedicated MATLAB functions perform time series analysis. This section introduces objects and interactive tools for time series analysis.

## Time Series Objects and Collections

### In this section...

“Types of Time Series and Their Uses” on page 3-3  
 “Time Series Data Sample” on page 3-3  
 “Example: Time Series Objects and Methods” on page 3-5  
 “Time Series Constructor” on page 3-12  
 “Time Series Collection Constructor” on page 3-12

### Types of Time Series and Their Uses

MATLAB time series objects are of two types:

- `timeseries` — Stores data and time values, as well as the metadata information that includes units, events, data quality, and interpolation method
- `tscollection` — Stores a collection of `timeseries` objects that share a common time vector, convenient for performing operations on synchronized time series with different units

This section discusses the following topics:

- Using time series constructors to instantiate time series classes
- Modifying object properties using `set` methods or dot notation
- Calling time series functions and methods

To get a quick overview of programming with `timeseries` and `tscollection` objects, follow the steps in “Example: Time Series Objects and Methods” on page 3-5.

### Time Series Data Sample

To properly understand the description of `timeseries` object properties and methods in this documentation, it is important to clarify some terms related to storing data in a `timeseries` object—the difference between a *data value* and a *data sample*.

A *data value* is a single, scalar value recorded at a specific time. A *data sample* consists of one or more values associated with a specific time in the `timeseries` object. The number of data samples in a time series is the same as the length of the time vector.

For example, consider data that consists of three sensor signals: two signals represent the position of an object in meters, and the third represents its velocity in meters/second.

To enter the data matrix, type the following at the MATLAB prompt:

```
x = [-0.2 -0.3 13;
     -0.1 -0.4 15;
      NaN  2.8 17;
      0.5  0.3 NaN;
     -0.3 -0.1 15]
```

The NaN value represents a missing data value. MATLAB displays the following 5-by-3 matrix:

```
x=  
-0.2000   -0.3000   13.0000  
-0.1000   -0.4000   15.0000  
   NaN     2.8000   17.0000  
  0.5000    0.3000    NaN  
-0.3000   -0.1000   15.0000
```

The first two columns of `x` contain quantities with the same units and you can create a multivariate `timeseries` object to store these two time series. For more information about creating `timeseries` objects, see “Time Series Constructor” on page 3-12. The following command creates a `timeseries` object `ts_pos` to store the position values:

```
ts_pos = timeseries(x(:,1:2), 1:5, 'name', 'Position')
```

MATLAB responds by displaying the following properties of `ts_pos`:

```
timeseries  
  
Common Properties:  
    Name: 'Position'  
    Time: [5x1 double]  
    TimeInfo: [1x1 tsdata.timemetadata]  
    Data: [5x2 double]  
    DataInfo: [1x1 tsdata.datametadata]
```

More properties, Methods

The Length of the time vector, which is 5 in this example, equals the number of data samples in the `timeseries` object. Find the size of the data sample in `ts_pos` by typing the following at the MATLAB prompt:

```
getdatasamplesize(ts_pos)
```

```
ans =  
  
    1    2
```

Similarly, you can create a second `timeseries` object to store the velocity data:

```
ts_vel = timeseries(x(:,3), 1:5, 'name', 'Velocity');
```

Find the size of each data sample in `ts_vel` by typing the following:

```
getdatasamplesize(ts_vel)
```

```
ans =  
  
    1    1
```

Notice that `ts_vel` has one data value in each data sample and `ts_pos` has two data values in each data sample.

---

**Note** In general, when the time series data is an  $M$ -by- $N$ -by- $P$ -by-... multidimensional array with  $M$  samples, the size of each data sample is  $N$ -by- $P$ -by-... .

---

If you want to perform operations on the `ts_pos` and `ts_vel` timeseries objects while keeping them synchronized, group them in a time series collection. For more information, see “Time Series Collection Constructor Syntax” on page 3-12.

## Example: Time Series Objects and Methods

- “Creating Time Series Objects” on page 3-5
- “Modifying Time Series Units and Interpolation Method” on page 3-6
- “Defining Events” on page 3-6
- “Creating Time Series Collection Objects” on page 3-7
- “Resampling a Time Series Collection Object” on page 3-8
- “Adding a Data Sample to a Time Series Collection Object” on page 3-8
- “Removing and Interpolating Missing Data” on page 3-9
- “Removing a Time Series from a Time Series Collection” on page 3-10
- “Displaying Time Vector Values as Date Strings” on page 3-10
- “Plotting Time Series Collection Members” on page 3-11

### Creating Time Series Objects

This portion of the example illustrates how to create several `timeseries` objects from an array. For more information about the `timeseries` object, see “Time Series Constructor” on page 3-12.

Import the sample data from `count.dat` to the MATLAB workspace.

```
load count.dat
```

This adds the 24-by-3 matrix, `count`, to the workspace. Each column of `count` represents hourly vehicle counts at each of three town intersections.

View the `count` matrix.

```
count
```

Create three `timeseries` objects to store the data collected at each intersection.

```
count1 = timeseries(count(:,1), 1:24, 'name', 'intersection1');
count2 = timeseries(count(:,2), 1:24, 'name', 'intersection2');
count3 = timeseries(count(:,3), 1:24, 'name', 'intersection3');
```

---

**Note** In the above construction, `timeseries` objects have both a variable name (e.g., `count1`) and an internal object name (e.g., `intersection1`). The variable name is used with MATLAB functions. The object name is a property of the object, accessed with object methods. For more information on `timeseries` object properties and methods, see “Time Series Properties” on page 3-12 and “Time Series Methods” on page 3-12.

---

By default, a time series has a time vector having units of seconds and a start time of 0 sec. The example constructs the `count1`, `count2`, and `count3` time series objects with start times of 1 sec, end times of 24 sec, and 1-sec increments. You will change the time units to hours in “Modifying Time Series Units and Interpolation Method” on page 3-6.

**Note** If you want to create a `timeseries` object that groups the three data columns in `count`, use the following syntax:

```
count_ts = timeseries(count, 1:24,'name','traffic_counts')
```

This is useful when all time series have the same units and you want to keep them synchronized during calculations.

---

### Modifying Time Series Units and Interpolation Method

After creating a `timeseries` object, as described in “Creating Time Series Objects” on page 3-5, you can modify its units and interpolation method using dot notation.

View the current properties of `count1`.

```
get(count1)
```

MATLAB displays the current property values of the `count1` `timeseries` object.

View the current `DataInfo` properties using dot notation.

```
count1.DataInfo
```

Change the data units for `count1` to 'cars'.

```
count1.DataInfo.Units = 'cars';
```

Set the interpolation method for `count1` to zero-order hold.

```
count1.DataInfo.Interpolation = tsdata.interpolation('zoh');
```

Verify that the `DataInfo` properties have been modified.

```
count1.DataInfo
```

Modify the time units to be 'hours' for the three time series.

```
count1.TimeInfo.Units = 'hours';  
count2.TimeInfo.Units = 'hours';  
count3.TimeInfo.Units = 'hours';
```

### Defining Events

This portion of the example illustrates how to define events for a `timeseries` object by using the `tsdata.event` auxiliary object. Events mark the data at specific times. When you plot the data, event markers are displayed on the plot. Events also provide a convenient way to synchronize multiple time series.

Add two events to the data that mark the times of the AM commute and PM commute.

Construct and add the first event to all time series. The first event occurs at 8 AM.

```
e1 = tsdata.event('AMCommute',8);  
e1.Units = 'hours'; % Specify the units for time  
count1 = addevent(count1,e1); % Add the event to count1  
count2 = addevent(count2,e1); % Add the event to count2  
count3 = addevent(count3,e1); % Add the event to count3
```



Construct and add the second event to all time series. The second event occurs at 6 PM.

```
e2 = tsdata.event('PMCommute',18);
e2.Units = 'hours';           % Specify the units for time
count1 = addevent(count1,e2); % Add the event to count1
count2 = addevent(count2,e2); % Add the event to count2
count3 = addevent(count3,e2); % Add the event to count3
```

Plot the time series, count1.

```
figure
plot(count1)
```

When you plot any of the time series, the plot method defined for time series objects displays events as markers. By default markers are red filled circles.

The plot reflects that count1 uses zero-order-hold interpolation.

Plot count2.

```
plot(count2)
```

If you plot time series count2, it replaces the count1 display. You see its events and that it uses linear interpolation.

Overlay time series plots by setting hold on.

```
hold on
plot(count3)
```

### Creating Time Series Collection Objects

This portion of the example illustrates how to create a `tscollection` object. Each individual time series in a collection is called a *member*. For more information about the `tscollection` object, see “Time Series Collection Constructor” on page 3-12.

---

**Note** Typically, you use the `tscollection` object to group synchronized time series that have different units. In this simple example, all time series have the same units and the `tscollection` object does not provide an advantage over grouping the three time series in a single `timeseries` object. For an example of how to group several time series in one `timeseries` object, see “Creating Time Series Objects” on page 3-5.

---

Create a `tscollection` object named `count_coll` and use the constructor syntax to immediately add two of the three time series currently in the MATLAB workspace (you will add the third time series later).

```
tsc = tscollection({count1 count2}, 'name', 'count_coll')
```

---

**Note** The time vectors of the `timeseries` objects you are adding to the `tscollection` must match.

---

Notice that the `Name` property of the `timeseries` objects is used to name the collection members as `intersection1` and `intersection2`.

Add the third `timeseries` object in the workspace to the `tscollection`.

```
tsc = addts(tsc, count3)
```

All three members in the collection are listed.

### Resampling a Time Series Collection Object

This portion of the example illustrates how to resample each member in a `tscollection` using a new time vector. The resampling operation is used to either select existing data at specific time values, or to interpolate data at finer intervals. If the new time vector contains time values that did not exist in the previous time vector, the new data values are calculated using the default interpolation method you associated with the time series.

Resample the time series to include data values every 2 hours instead of every hour and save it as a new `tscollection` object.

```
tsc1 = resample(tsc,1:2:24)
```

In some cases you might need a finer sampling of information than you currently have and it is reasonable to obtain it by interpolating data values.

Interpolate values at each half-hour mark.

```
tsc1 = resample(tsc,1:0.5:24)
```

To add values at each half-hour mark, the default interpolation method of a time series is used. For example, the new data points in `intersection1` are calculated by using the zero-order hold interpolation method, which holds the value of the previous sample constant. You set the interpolation method for `intersection1` as described in “Modifying Time Series Units and Interpolation Method” on page 3-6.

The new data points in `intersection2` and `intersection3` are calculated using linear interpolation, which is the default method.

Plot the members of `tsc1` with markers to see the results of interpolating.

```
hold off % Allow axes to clear before plotting
plot(tsc1.intersection1,'-xb','Displayname','Intersection 1')
```

You can see that data points have been interpolated at half-hour intervals, and that `Intersection 1` uses zero-order-hold interpolation, while the other two members use linear interpolation.

Maintain the graph in the figure while you add the other two members to the plot. Because the `plot` method suppresses the axis labels while `hold` is on, also add a legend to describe the three series.

```
hold on
plot(tsc1.intersection2,'-.xm','Displayname','Intersection 2')
plot(tsc1.intersection3,':xr','Displayname','Intersection 3')
legend('show','Location','NorthWest')
```

### Adding a Data Sample to a Time Series Collection Object

This portion of the example illustrates how to add a data sample to a `tscollection`.

Add a data sample to the `intersection1` collection member at 3.25 hours (i.e., 15 minutes after the hour).

```
tsc1 = addsampletocollection(tsc1, 'time', 3.25, ...
    'intersection1', 5);
```

There are three members in the `tsc1` collection, and adding a data sample to one member adds a data sample to the other two members at 3.25 hours. However, because you did not specify the data values for `intersection2` and `intersection3` in the new sample, the missing values are represented by NaNs for these members. To learn how to remove or interpolate missing data values, see “Removing Missing Data” on page 3-9 and “Interpolating Missing Data” on page 3-10.

### tsc1 Data from 2.0 to 3.5 Hours

Hours	Intersection 1	Intersection 2	Intersection 3
2.0	7	13	11
2.5	7	15	15.5
3.0	14	17	20
3.25	5	NaN	NaN
3.5	14	15	14.5

To view all `intersection1` data (including the new sample at 3.25 hours), type

```
tsc1.intersection1
```

Similarly, to view all `intersection2` data (including the new sample at 3.25 hours containing a NaN value), type

```
tsc1.intersection2
```

### Removing and Interpolating Missing Data

Time series objects use NaNs to represent missing data. This portion of the example illustrates how to either remove missing data or interpolate values for it by using the interpolation method you specified for that time series. In “Adding a Data Sample to a Time Series Collection Object” on page 3-8, you added a new data sample to the `tsc1` collection at 3.25 hours.

As the `tsc1` collection has three members, adding a data sample to one member added a data sample to the other two members at 3.25 hours. However, because you did not specify the data values for the `intersection2` and `intersection3` members at 3.25 hours, they currently contain missing values, represented by NaNs.

#### Removing Missing Data

Find and remove the data samples containing NaN values in the `tsc1` collection.

```
tsc1 = delsamplefromcollection(tsc1, 'index', ...
    find(isnan(tsc1.intersection2.Data)));
```

This command searches one `tscollection` member at a time—in this case, `intersection2`. When a missing value is located in `intersection2`, the data at that time is removed from *all* members of the `tscollection`.

---

**Note** Use dot-notation syntax to access the `Data` property of the `intersection2` member in the `tsc1` collection:

```
tsc1.intersection2.Data
```

For a complete list of `timeseries` properties, see “Time Series Properties” on page 3-12.

### Interpolating Missing Data

For the sake of this example, reintroduce NaN values in `intersection2` and `intersection3`.

```
tsc1 = addsampletocollection(tsc1, 'time', 3.25, ...
    'intersection1', 5);
```

Interpolate the missing values in `tsc1` using the current time vector (`tsc1.Time`).

```
tsc1 = resample(tsc1, tsc1.Time);
```

This replaces the NaN values in `intersection2` and `intersection3` by using linear interpolation—the default interpolation method for these time series.

**Note** Dot notation `tsc1.Time` is used to access the `Time` property of the `tsc1` collection. For a complete list of `tscollection` properties, see “Time Series Collection Properties” on page 3-13.

To view `intersection2` data after interpolation, for example, type

```
tsc1.intersection2
```

### New `tsc1` Data from 2.0 to 3.5 Hours

Hours	Intersection 1	Intersection 2	Intersection 3
2.0	7	13	11
2.5	7	15	15.5
3.0	14	17	20
3.25	5	16	17.3
3.5	14	15	14.5

### Removing a Time Series from a Time Series Collection

Remove the `intersection3` time series from the `tscollection` object `tsc1`.

```
tsc1 = removets(tsc1, 'intersection3')
```

Two time series as members in the collection are now listed.

### Displaying Time Vector Values as Date Strings

This portion of the example illustrates how to control the format in which numerical time vector display, using MATLAB date strings. For a complete list of the MATLAB date-string formats supported for `timeseries` and `tscollection` objects, see the definition of time vector definition in the `timeseries` reference page.

To use date strings, you must set the `StartDate` field of the `TimeInfo` property. All values in the time vector are converted to date strings using `StartDate` as a reference date.

Suppose the reference date occurs on December 25, 2009.

```
tsc1.TimeInfo.Units = 'hours';
tsc1.TimeInfo.StartDate = '25-DEC-2009 00:00:00';
```

Similarly to what you did with the `count1`, `count2`, and `count3` time series objects, set the data units to of the `tsc1` members to the string `'car count'`.

```
tsc1.intersection1.DataInfo.Units = 'car count';
tsc1.intersection2.DataInfo.Units = 'car count';
```

### Plotting Time Series Collection Members

To plot data in a time series collection, you plot its members one at a time.

First graph `tsc1` member `intersection1`.

```
hold off
plot(tsc1.intersection1);
```

When you plot a member of a time series collection, its time units display on the x-axis and its data units display on the y-axis. The plot title is displayed as `'Time Series Plot:<member name>'`.

If you use the same figure to plot a different member of the collection, no annotations display. The time series `plot` method does not attempt to update labels and titles when `hold` is on because the descriptors for the series can be different.

Plot `intersection1` and `intersection2` in the same figure. Prevent overwriting the plot, but remove axis labels and title. Add a legend and set the `DisplayName` property of the line series to label each member.

```
plot(tsc1.intersection1, '-xb', 'Displayname', 'Intersection 1')
hold on
plot(tsc1.intersection2, '-.xm', 'Displayname', 'Intersection 2')
legend('show', 'Location', 'NorthWest')
```

The plot now includes the two time series in the collection: `intersection1` and `intesection2`. Plotting the second graph erased the labels on the first graph.

Finally, change the date strings on the x-axis to hours and plot the two time series collection members again with a legend.

Specify time units to be `'hours'` for the collection.

```
tsc1.TimeInfo.Units = 'hours';
```

Specify the format for displaying time.

```
tsc1.TimeInfo.Format = 'HH:MM';
```

Recreate the last plot with new time units.

```
hold off
plot(tsc1.intersection1, '-xb', 'Displayname', 'Intersection 1')

% Prevent overwriting plot, but remove axis labels and title.
hold on
plot(tsc1.intersection2, '-.xm', 'Displayname', 'Intersection 2')
legend('show', 'Location', 'NorthWest')
```

```
% Restore the labels with the |xlabel| and |ylabel| commands and overlay a  
% data grid.  
xlabel('Time (hours)')  
ylabel('car count')  
grid on
```

For more information on plotting options for time series, see `timeseries`.

## Time Series Constructor

Before implementing the various MATLAB functions and methods specifically designed to handle time series data, you must create a `timeseries` object to store the data. See `timeseries` for the `timeseries` object constructor syntax.

For an example of using the constructor, see “Creating Time Series Objects” on page 3-5.

### Time Series Properties

See `timeseries` for a description of all the `timeseries` object properties. You can specify the `Data`, `IsTimeFirst`, `Name`, `Quality`, and `Time` properties as input arguments in the constructor. To assign other properties, use the `set` function or dot notation.

---

**Note** To get property information from the command line, type `help timeseries/tsprops` at the MATLAB prompt.

---

For an example of editing `timeseries` object properties, see “Modifying Time Series Units and Interpolation Method” on page 3-6.

### Time Series Methods

For a description of all the time series methods, see `timeseries`.

## Time Series Collection Constructor

- “Introduction” on page 3-12
- “Time Series Collection Constructor Syntax” on page 3-12
- “Time Series Collection Properties” on page 3-13
- “Time Series Collection Methods” on page 3-14

### Introduction

The MATLAB object, called `tscollection`, is a MATLAB variable that groups several time series with a common time vector. The `timeseries` objects that you include in the `tscollection` object are called *members* of this collection, and possess several methods for convenient analysis and manipulation of `timeseries`.

### Time Series Collection Constructor Syntax

Before you implement the MATLAB methods specifically designed to operate on a collection of `timeseries` objects, you must create a `tscollection` object to store the data.

The following table summarizes the syntax for using the `tscollection` constructor. For an example of using this constructor, see “Creating Time Series Collection Objects” on page 3-7.

### Time Series Collection Syntax Descriptions

Syntax	Description
<code>tsc = tscollection(ts)</code>	<p>Creates a <code>tscollection</code> object <code>tsc</code> that includes one or more <code>timeseries</code> objects.</p> <p>The <code>ts</code> argument can be one of the following:</p> <ul style="list-style-type: none"> <li>• Single <code>timeseries</code> object in the MATLAB workspace</li> <li>• Cell array of <code>timeseries</code> objects in the MATLAB workspace</li> </ul> <p>The <code>timeseries</code> objects share the same time vector in the <code>tscollection</code>.</p>
<code>tsc = tscollection(Time)</code>	<p>Creates an empty <code>tscollection</code> object with the time vector <code>Time</code>.</p> <p>When time values are date strings, you must specify <code>Time</code> as a cell array of date strings.</p>
<code>tsc = tscollection(Time, TimeSeries, 'Parameter', Value, ...)</code>	<p>Optionally enter the following parameter-value pairs after the <code>Time</code> and <code>TimeSeries</code> arguments:</p> <ul style="list-style-type: none"> <li>• <code>Name</code> (see “Time Series Collection Properties” on page 3-13)</li> </ul>

### Time Series Collection Properties

This table lists the properties of the `tscollection` object. You can specify the `Name`, `Time`, and `TimeInfo` properties as input arguments in the `tscollection` constructor.

### Time Series Collection Property Descriptions

Property	Description
Name	<code>tscollection</code> object name entered as a string. This name can differ from the name of the <code>tscollection</code> variable in the MATLAB workspace.
Time	<p>A vector of time values.</p> <p>When <code>TimeInfo.StartDate</code> is empty, the numerical <code>Time</code> values are measured relative to 0 in specified units. When <code>TimeInfo.StartDate</code> is defined, the time values represent date strings measured relative to <code>StartDate</code> in specified units.</p> <p>The length of <code>Time</code> must match either the first or the last dimension of the <code>Data</code> property of each <code>tscollection</code> member.</p>
TimeInfo	<p>Uses the following fields to store contextual information about <code>Time</code>:</p> <ul style="list-style-type: none"> <li>• <code>Units</code> — Time units with the following values: 'weeks', 'days', 'hours', 'minutes', 'seconds', 'milliseconds', 'microseconds', and 'nanoseconds'</li> <li>• <code>Start</code> — Start time</li> <li>• <code>End</code> — End time (read-only)</li> <li>• <code>Increment</code> — Interval between two subsequent time values. The increment is NaN when times are not uniformly sampled.</li> <li>• <code>Length</code> — Length of the time vector (read-only)</li> <li>• <code>Format</code> — String defining the date string display format. See the MATLAB <code>datestr</code> function reference page for more information.</li> <li>• <code>StartDate</code> — Date string defining the reference date. See the MATLAB <code>setabstime</code> function reference page for more information.</li> <li>• <code>UserData</code> — Stores any additional user-defined information</li> </ul>

### Time Series Collection Methods

- “General Time Series Collection Methods” on page 3-14
- “Data and Time Manipulation Methods” on page 3-15

### General Time Series Collection Methods

Use the following methods to query and set object properties, and plot the data.

### Methods for Querying Properties

Method	Description
<code>get</code>	Query <code>tscollection</code> object property values.
<code>isempty</code>	Evaluate to <code>true</code> for an empty <code>tscollection</code> object.
<code>length</code>	Return the length of the time vector.
<code>plot</code>	Plot the time series in a collection.
<code>set</code>	Set <code>tscollection</code> property values.
<code>size</code>	Return the size of a <code>tscollection</code> object.



**Data and Time Manipulation Methods**

Use the following methods to add or delete data samples, and manipulate the `tscollection` object.

**Methods for Manipulating Data and Time**

Method	Description
<code>addts</code>	Add a <code>timeseries</code> object to a <code>tscollection</code> object.
<code>addsampletocollection</code>	Add data samples to a <code>tscollection</code> object.
<code>delsamplefromcollection</code>	Delete one or more data samples from a <code>tscollection</code> object.
<code>getabstime</code>	Extract a date-string time vector from a <code>tscollection</code> object into a cell array.
<code>getsamplusingtime</code>	Extract data samples from an existing <code>tscollectionobject</code> into a new <code>tscollection</code> object.
<code>gettimeseriesnames</code>	Return a cell array of time series names in a <code>tscollection</code> object.
<code>horzcat</code>	Horizontal concatenation of <code>tscollection</code> objects. Combines several <code>timeseries</code> objects with the same time vector into one time series collection.
<code>removets</code>	Remove one or more <code>timeseries</code> objects from a <code>tscollection</code> object.
<code>resample</code>	Select or interpolate data in a <code>tscollection</code> object using a new time vector.
<code>setabstime</code>	Set the time values in the time vector of a <code>tscollection</code> object as date strings.
<code>settimeseriesnames</code>	Change the name of the selected <code>timeseries</code> object in a <code>tscollection</code> object.
<code>vertcat</code>	Vertical concatenation of <code>tscollection</code> objects. Joins several <code>tscollection</code> objects along the time dimension.

