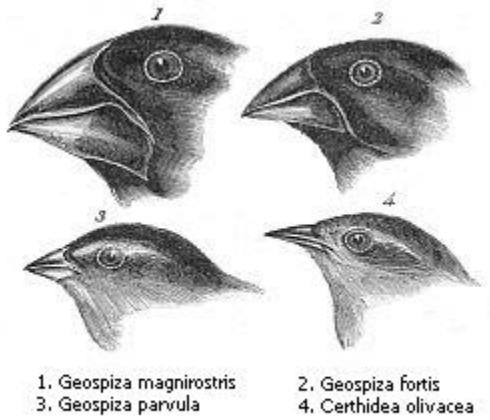
LESSON: Linear models, Scatter plots, curve fitting and correlation

FOCUS QUESTION: How can I determine whether two variables are related?

This lesson shows you how to determine whether two variables are related by fitting a linear model (straight line) and by calculating the correlation between the two sets of values. This lesson also introduces scatter plots as a way of visualizing this relationship.

In this lesson you will:

- Compute the correlation between two data sets.
- Compare two data sets by plotting them against each other in a scatter plot.
- Add a linear fit line to a scatter plot using the plot tools.
- Compute the linear fit line directly (linear model).
- Compute the error between linear predictions and actual data.
- Create computed title strings based on specifics of the data.



Finches from Galapagos Archipelago

Contents

- DATA FOR THIS LESSON
- SETUP FOR LESSON
- EXAMPLE 1: Load the Darwin Finch beak size parentage data
- EXAMPLE 2: Define variables for child and parent beak sizes
- EXAMPLE 3: Calculate and output the beak size correlations
- EXAMPLE 4: Plot the average parent beak size against the child beak size
- EXAMPLE 5: Edit the figure of EXAMPLE 4
- EXAMPLE 6: Calculate and output the best fit lines for beak parentage
- EXAMPLE 7: Predict the child beak sizes from its parent beak sizes (linear model)
- EXAMPLE 8: Calculate the error between child beak sizes and model predictions
- EXAMPLE 9: Find mean square error (MSE)
- EXAMPLE 10: Find root mean squared error (RMS)
- SUMMARY OF SYNTAX

DATA FOR THIS LESSON

File	Description

DaphneIsland.txt	▪ The data consists of family beak size data for Darwin ground finches on Daphne Island:
	  ▪ The first column contains the chick beak size in mm
	  ▪ The second column contains the mother's beak size in mm
	  ▪ The third column contains the father's beak size in mm.
	▪ The data was extracted from a data set distributed with the case study Natural Selection and Darwin's Finches by Martin Wikelski available on the web at http://wps.prenhall.com/esm_freeman_evol_3/0,7017,749374-,00.html.
	▪ The original data is summarized in the article: "The classical case of character release: Darwin's finches (Geospiza) on Isla Daphne Major, Galápagos" by P. T. Boag and P. R. Grant that appeared in *Biological Journal of the Linnean Society* 22:243-277 (1974). See http://en.wikipedia.org/wiki/Peter_and_Rosemary_Grant for additional information.
	A brief discussion of their work on evolution and natural selection can be found at PBS Evolution Home Library Finch Beak Data Sheet.

SETUP FOR LESSON

- Create an `LinearModels` directory on your `V:` drive and make it your comment directory.
- Download the `NYCDiseases.mat` and the `DaphneIsland.txt` to your `LinearModels` directory.
- Create a `LinearModelLesson` script file in your `LinearModels` directory.

EXAMPLE 1: Load the Darwin Finch beak size parentage data

Create a new cell in which you type and execute:

```
beaks = load('DaphneIsland.txt');
```

You should see a `beaks` variable in your Workspace.

EXERCISE 1: Draw a picture of `beaks` and label it.

EXAMPLE 2: Define variables for child and parent beak sizes

Create a new cell in which you type and execute:

```
child = beaks(:, 1);                % Size of offspring beaks

parent = mean(beaks(:, 2:3), 2);    % Average the parent beak sizes
```

You should 2 additional variables in your Workspace Browser:

- `child` - a column vector containing the offspring beak sizes
- `parent` - a column vector containing the average of the parents' beak sizes

EXAMPLE 3: Calculate and output the beak size correlations

Create a new cell in which you type and execute:

```
pCorr = corr(parent, child);

fprintf('Parent-child beak correlation: %g\n', pCorr)
```

You should see the `pCorr` variable in your Workspace Browser and the following output in your Command Window:

```
Parent-child beak correlation: 0.72319
```

EXERCISE 2: Do you think that parent beak sizes have a causal relationship to child beak sizes? Why or why not?

EXERCISE 3: Calculate and output the correlation between mumps and chicken pox in the NYC diseases data set.

EXERCISE 4: Do you think mumps and chicken pox have a causal relationship? Why or why not? What other factors might be relevant?

EXERCISE 5: Calculate and output the correlation between mumps and measles in the NYC diseases data set. Why do you think the mumps-chicken pox correlation is higher than the mumps-measles correlation?

EXERCISE 6: Why do you think the mumps-chicken pox correlation is higher than the mumps-measles correlation?

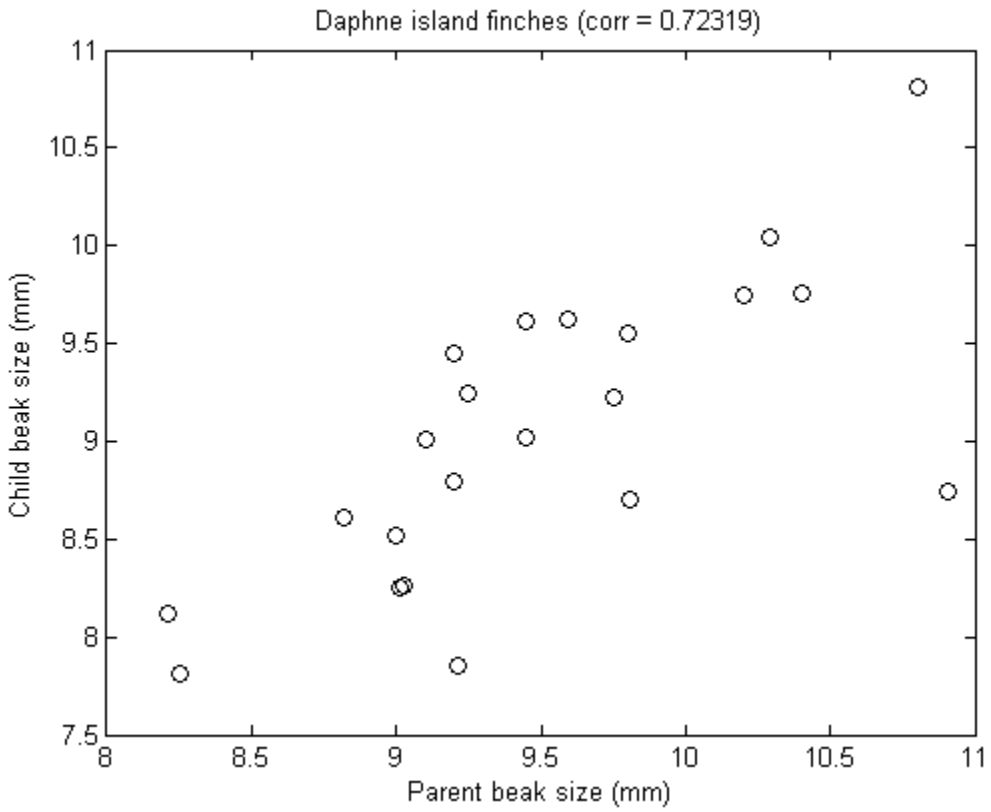
EXAMPLE 4: Plot the average parent beak size against the child beak size

Create a new cell in which you type and execute:

```
tString = ['Daphne island finches (corr = ', num2str(pCorr) ')'];

figure('Name', tString)                    % Put title on the window

plot(parent, child, 'ko')                  % Plot a scatter plot

xlabel('Parent beak size (mm)');           % Label the x-axis

ylabel('Child beak size (mm)');            % Label the y-axis
```

```
    title(tString);                              % Put title on the graph
```

You should see a Figure Window with the following plot:



EXAMPLE 5: Edit the figure of EXAMPLE 4

- Make the plot editable. (Use `Tools->Edit Plot` from the Figure Window menubar.)
- Add a linear fit line. (Use `Tools->Basic Fitting` from the Figure Window menubar.)
- Save the figure as `BeakSize.fig`.

EXAMPLE 6: Calculate and output the best fit lines for beak parentage

Create a new cell in which you type and execute:

```
    pPoly = polyfit(parent, child, 1);   % Linear fit of parent vs child

    fprintf('Model: child = %g*parent + %g\n',  pPoly(1), pPoly(2))
```

You should see the `pPoly` variable in your Workspace Browser and the following output in your Command Window:

```
Model: child = 0.766283*parent + 1.763
```

EXERCISE 7: Calculate and output the best fit line between measles and mumps in the NYC diseases data set.

EXAMPLE 7: Predict the child beak sizes from its parent beak sizes (linear model)

Create a new cell in which you type and execute:

```
pPred = polyval(pPoly, parent);   % Find parent-child relationship
```

You should see the following variable in the Workspace Browser:

- `pPred` - prediction from linear model for the offspring's beak given the mother's

EXAMPLE 8: Calculate the error between child beak sizes and model predictions

Create a new cell in which you type and execute:

```
pError = child - pPred;     % Actual - predicted by parent's size
```

You should see the following variable in the Workspace Browser:

- `pError` - difference between offspring and mother's prediction

EXAMPLE 9: Find mean square error (MSE)

Create a new cell in which you type and execute:

```
pMSE = mean(pError.*pError);
```

You should see the following variable in the Workspace Browser:

- `pMSE` - mean squared prediction error

EXAMPLE 10: Find root mean squared error (RMS)

Create a new cell in which you type and execute:

```
pRMS = sqrt(pMSE);
```

You should see the following variable in the Workspace Browser:

- pRMS - root mean squared prediction error

SUMMARY OF SYNTAX

MATLAB syntax	Description
`rho = corr(x, y)`	Calculate the correlation between the column vectors x and y. The variable rho contains a single value between -1 and 1 indicating how linearly related the values of x and y are. If the correlation is close to 1, the values of x and y go up and down together. If the correlation is close to -1, the values of x and y go in opposite directions (if one goes up, the other tends to go down). A correlation value close to 0 indicates that the values of x and y are not related. The column vectors x and y must have the same same number of elements.
`p = polyfit(x, y, n)`	Calculate the coefficients of the *best* polynomial of degree n that fits the curve x versus y. This polynomial minimizes the RMS error, and is sometimes called the *least-squared error* approximation. The coefficients of the polynomial appear in the vector p, such that p(1) has the coefficient of the highest term in the polynomial.
`Y = polyval(p, X)`	Evaluate the polynomial whose coefficients are in the vector p at the points contained in the array X. The array Y holds the results of this evaluation.*This lesson was written by Kay A. Robbins of the University of Texas at San Antonio and last modified on 29-Sep-2013. Please contact krobbins@cs.utsa.edu with comments or suggestions. The drawing was done by John Gould before 1772 and was cataloged in Darwin's finches or Galapagos finches. Darwin, 1745. Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N. 2d edition. The copyright has expired on this image.* *Published with MATLAB® 7.14*