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

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

Contents

EXAMPLE 1: Load the Darwin Finch beak size parentage data

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

Questions	Answers
What does the `beaks` variable hold?	The `beaks` variable holds an array of the values of the data file `DaphneIsland.txt`. Each line of the file corresponds to one row of `beaks`. The columns of `beaks` are the individual tab-separated numbers on each line.
What if the file contains a mixture of text and numeric values in text format?	The `load` and `csvread` functions won't work for those types of files. MATLAB provides a variety of other functions such as `textscan` and `importdata` to handle these more complicated situations.

EXAMPLE 2: Define variables for child and parent beak sizes

```
child = beaks(:, 1);

parent = mean(beaks(:, 2:3), 2);
```

EXAMPLE 3: Calculate and output the beak size correlations

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

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

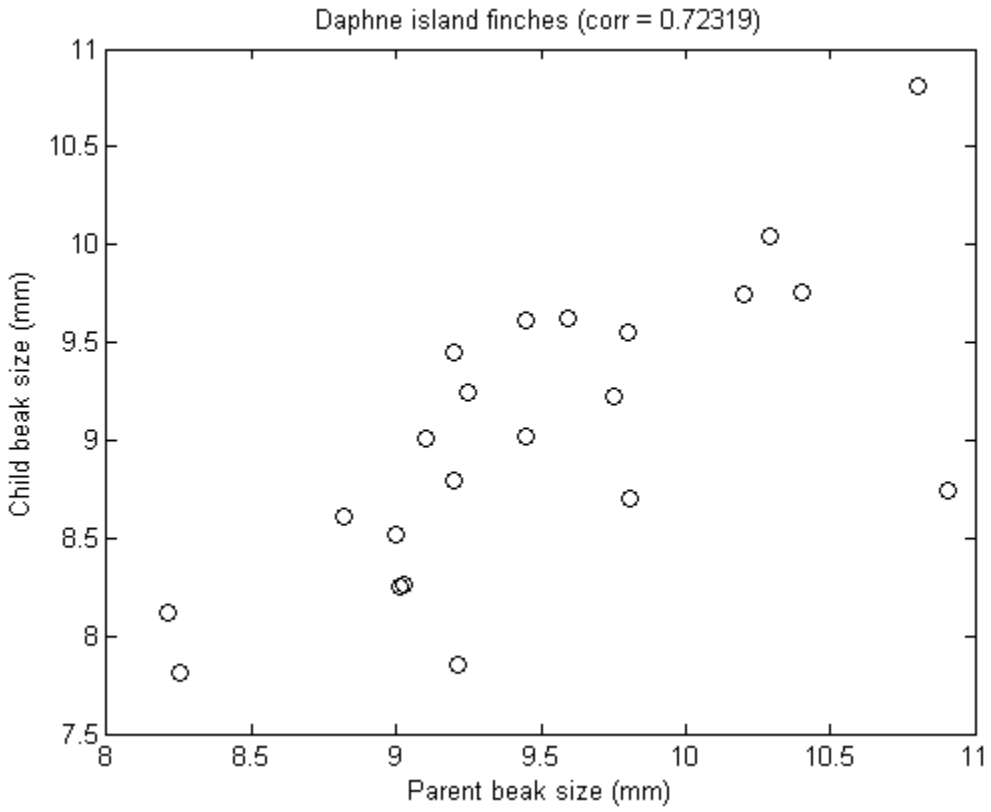
```
Parent-child beak correlation: 0.72319
```

Questions	Answers
What does the `corr` function do?	The `corr` function computes the correlation between two variables. Correlation is a standard statistical measure of how linearly related two variables are. If the corresponding values of the two variables go up and down in a similar fashion, the correlation will be close to 1. If the values go in opposite directions, the correlation will be close to -1. If there is not much relationship, the correlation will be close to 0.
Can the value of the correlation ever go above 1?	No, the value of the correlation is always between -1 and 1 inclusive.
Can `corr(x, y)` and `corr(y, x)` be different values?	No, the values are the same. The `corr` function is said to be *symmetric* in its arguments.
Do the column vectors `x` and `y` have to have the same number of elements in order to compute `corr(x, y)`?	Yes, `x` and `y` have to have the same number of elements. (When the arguments of `corr` are arrays, the situation is more complicated.)
Can I compute the correlation between two variables that are of vastly different magnitudes?	Yes, `corr` scales the values as part of the computation.
Why can't I find the correlation of two numbers?	The correlation between two numbers (e.g., `corr(1, 3)`) is undefined and `corr` returns `NaN` (which stands for "not a number"). The underlying reason is that correlation divides each column by its standard deviation as part of the scaling. The standard deviation of a single value is 0, so this scaling results in division by zero. A more intuitive answer is that you can't possible estimate trend without comparing the values at at least two different points.
Does `corr` work for matrices?	Yes, if `X` and `Y` are matrices with the same number of rows, `corr(X, Y)` returns a matrix of correlations. The (i, j)-th entry in the result is the correlation between column i of `X` and column j of `Y`.

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

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

figure('Name', tString)

plot(parent, child, 'ko')
```

```
xlabel('Parent beak size (mm)');

ylabel('Child beak size (mm)');

title(tString);
```



Questions	Answers
What is a scatter plot?	A scatter plot graphs two variables against each other without displaying the connecting lines. The purpose of a scatter plot is to reveal relationships between variables.
How does a scatter plot differ from a line graph?	In a line graph, successive pairs of points are connected by straight lines. The data corresponding to the horizontal axes should be ordered or the plot will probably not make sense. All of the line graphs that we have looked at so far use time for the first coordinate and rely on the data points being sorted either in increasing or decreasing order by time. In contrast, the data for scatter plots can be ordered in any way because no relationship is assumed between successive pairs of points.
Does MATLAB have a special function for drawing scatter plots?	The MATLAB `scatter` function specifically draws a scatter plot. The `plot` function includes all of the capability of `scatter`, so we just use `plot` in this lesson.

EXAMPLE 5: Edit the figure of EXAMPLE 4

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);

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

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

Questions	Answers
What does the `1` stand for in expression `pPoly = polyfit(parent, child, 1)`?	The `1` specifies the degree of the polynomial to fit to the data.
What is in the variable `pMother` after `pPoly = polyfit(parent, child, 1)` is evaluated?	The `pPoly(1)` is the slope and pPoly(2) is the intercept of the best linear fit of `mother` versus `child`.
How would I fit a quadratic to a set of data points (x, y)?	Use `q = polyfit(x, y, 2)`.

EXAMPLE 7: Predict the child beak sizes from its parent beak sizes

```
pPred = polyval(pPoly, parent);   % Evaluate linear fit of parent equation
```

Questions	Answers
What does `pPred` represent?	The `pPred` values are the predictions of the child beak size from a linear model derived from the parent's beak sizes.
What does the `polyval` function do?	The `polyval` function evaluates a polynomial.
How is the polynomial to be evaluated specified?	The first argument of `polyval` is a vector containing the coefficients of the polynomial. The coefficients are ordered with the coefficient of the highest power term being first.
How can I evaluate the polynomial y = 5x^2 - 3x + 6 for x = -3?	The polynomial y = 5x^2 - 3x + 6 is represented by the coefficient vector `p = [5, -3, 6]`. The evaluation is `polyval(p, -3)`.
How many coefficients does a cubic	Polynomials have one more coefficient than their degree (due to the presence of the constant

Questions	Answers
polynomial have?	term). A cubic or degree 3 polynomial has 4 coefficients..
Suppose A is a 3 x 4 array and p contains the coefficients of a first degree polynomial. What size is B after the statement `B = polyval(p, A)` is executed?	The `polyval` function returns an array that is the same size as A regardless of how big p is. That is, a polynomial evaluation produces a single y value for each x. Hence, B is a 3 x 4 array.

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

```
pError = child - pPred;
```

Questions	Answers
Does `pError` represent the error in measurement of the parent's beak sizes?	No. The `pError` variable represents the errors made when predicting the child beak sizes from the parent beak sizes.
Are `pError` always greater than zero?	No. The values of the elements of `pError` may be positive or negative depending on whether the individual predictions overestimate or underestimate the actual child beak size.

EXAMPLE 9: Find mean square error (MSE)

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

EXAMPLE 10: Find root mean squared error (RMS)

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

Questions	Answers
What effect does squaring the error have?	The obvious effect is that squaring makes all the values non-negative so it is easier to compare magnitudes. Squaring also puts more weight on larger values.
Why take the square root?	Without the square root, the error is called the mean-squared error (MSE). With the square it is called the root mean squared error (RMS). RMS error has the same units as the original data, while MSE does not. For example the error in beak sizes has units of mm, since the beak size measurements are in mm. The MSE has units of mm^2. Taking the square root brings us back to mm. Thus, we can make reasonable comparisons between the RMS error and the original beak sizes.

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.*

Published with MATLAB® 7.14