GEOS 397/597

Lecture #20: Linear fitting and statistics

use **practical_20.m**

1) Update from github

2) polyfit, polyval, corrcoef

MATLAB provides several built-in functions to fit curves

- Many require the “Curve Fitting Toolbox”, or other toolboxes.
- We will only use the basic curve fitting functions that are part of standard MATLAB
- We will focus on:

1) polyfit

2) polyval

5) corrcoef

3) Review of polynomials with polyval, polyfit

See practical exercises on noise-free and noisy data.

4) Goodness of fit when linearly fitting data

When we use polyfit to fit data with a **linear** (1st order) polynomial, we would like to know how well we actually fit the data. When we fit data in this way, we call it *linear regression* and we are interested in a 'goodness of fit' measure. We would like to know how well does the regression equation truly represent the data?

There are a variety of ways to compare data and model fits; takes Dr. HP Marshall's geostatistics course if you would like more knowledge in this area. In this class you will move beyond **univariate** statistics and into **multivariate** statistics (e.g. covariance).

The measure we will consider in this class if the *linear correlation coefficient* r. You will often see people cite the *coefficient of determination* r^2 value, and it is important to build an intuitive sense for what this value means.

r^2 identifies the percentage (%) of the data's variance that is explained by a linear fit. For example, r^2 =0.95 means that 95% of the data's variance is explained by a linear relationship.

Let's go a little deeper here and make sure we understand what we mean when we say variance.

First we need to consider the *mean* value $\bar{y}$ of the data. In MATLAB this is the function **mean()**.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The variance σ^2 is then the average of the distances of each point from the mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

In MATLAB the variance is computed with the function **var()**.

The standard deviation σ is simply the square root of the variance. In MATLAB the standard deviation is computed with the function **std()**.

Linear correlation coefficient

The **linear** correlation coefficient r is sometimes referred to as the *Pearson correlation coefficient* because of the developer Karl Pearson.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

where the summation is over the n data points (x being the independent variable and y being the dependent variable).

- The value of r is normalized such that $-1 < r < +1$. The $+$ and $-$ signs indicate positive linear correlations and negative linear correlations, respectively.
- **Positive correlation:** If x and y have a strong positive linear correlation, r is close to $+1$. An r value of exactly $+1$ indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- **Negative correlation:** If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- **No correlation:** If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.
- A **perfect correlation** of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.
- Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

In MATLAB we can use the function **corrcoef()**.

Coefficient of determination

- The coefficient of determination, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in **making predictions** from a certain model/graph.
- The *coefficient of determination* is the ratio of the explained variation to the total variation.
- The *coefficient of determination* is normalized such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y.
- The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.
- The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

The total variation SS_{tot} in y is closely related to the variance σ^2.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The most general definition of the *coefficient of determination* is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where the *residual sum of squares* SS_{res} is defined as

$$SS_{res} = \sum_{i=1}^{n}(y_i - f_i)^2$$

with f the predicted data.

---

6) Student exercises in class

Some information for this lecture was taken from mathbits.com.