Data Analysis in Geophysics

Activity 4 – MATLAB – Data Fitting

During this activity you will perform linear and non-linear regression analyses on sedimentary rock data from the eastern and western U.S. The data tables include just two simple columns: the age of the rock in kyr and the burial depth in meters. By correlating these two quantities, geologist can determine the history of the rock, and potential exhumation processes.

Part I – Linear Correlations

1. Load the data set for the eastern U.S. (**sediment_eastUS.txt**) The data file contains sedimentary depths and age within a two-column data table. You can use: readtable() to load the data
2. Create a scatter plot and label x and y axis (plot() or scatter())
3. Compute the correlation coefficient and significance of correlation using: corrcoeff()
4. Fit the data using a linear relationship and determine the confidence bounds! (use: polyfit(), polyval(), polyconf())
5. Plot the best-fitting solution together with the confidence interval. (plot())
6. Plot a histogram of the residuals between model and observation. Are the residuals normally distributed? (histogram(), chi2gof() or makedist() and kstest())

7. The **coefficient of determination** (R^2) is commonly used to measure the goodness-of-fit of a linear model. It is computed using:

$$R^2 = 1 - \frac{RSS}{TSS},$$

Where RSS = sum of squares of residuals between model and data, and TSS = total variation in the data. Thus, R^2 provides a fractional estimate of the total explained data variation by the model. Determine R2 for both data sets!

(use: fitnlm(data, model_fct, par0))

8. Repeat 1 to 6 for the second data set (**sediment_westUS.txt**)! Describe the differences between the data sets.

Bootstrap resampling to get uncertainties and confidence intervals

9. Perform a bootstrap analysis and plot the distribution of Pearson *r*-values for N=1000 bootstraps. What does the histogram tell you about the applicability of the linear model? (use: bootstrp() and histogram())
10. Plot the PDFs of the regression coefficients and report the expected 95% confidence intervals (histogram(), prctile()).
11. Remove any outliers and re-do the linear fitting (rmoutliers()).

Note that your code in Part I can be used for any higher-order polynomial!

Part II – Fitting Non-Linear Functions to Observations

Fitting Non-linear data

1. Load the data file: 'exp_growth.mat'! Which variables are contained within the data file?
2. Use the following equation to perform a non-linear least-square fit on the data:

$$f(t) = N_o \exp(rt))$$

3. Describe the characteristics of the residuals between model and observation. Do you think the model describes the data well? Why or why not?
4. Report the best-fit parameters N_o and *r* as well as their confidence bounds!

Part III – Nonlinear and Weighted Regression

1. Consider the following data vectors:
   a. x = [1, 2, 3, 5, 7, 10] and
   b. y = [109, 149, 149, 191, 213, 224]
   c. with measurement error, ε_y=[10, 10, 2, 2, 2, 2]
2. Fit the data with the following exponential function:
   a. $f(x) = b_1(1 - \exp(-b_2 x))$
3. Compare the results between a standard and weighted least-squares fit using:

```
>>fitnlm( __, __, __, 'Weights', w)
```