Lab Assignment 4
Linear Regression and Correlation

ACS II

Spring 2020

Assigned: January 30

Due: February 6

Introduction

Given a set of independent and dependent variables one often wishes to find some relationship between the values of the independent variables and the values of the dependent variables. In the real world, this is often hampered by the presence of noise in the data, which can obfuscate the underlying relation and prevent there from being a function of the desired form which exactly reproduces the data. In such circumstances, the goal shifts from being the exact recovery of the response function that describes the data to the similar goal of recovery of the optimal response function for the data.

Linear Least Squares Fitting

Suppose you are given n measurements, $\{x_i, y_i\}_{i=1}^n$ with one independent variable x_i and one dependent variable y_i. The goal of regression is to find a set of parameters that best fit the given data, i.e.

$$y_i \approx \beta_1 x_i + \beta_0 \quad i = 1, \ldots, n$$

where β_0 and β_1 are the parameters to be determined. An important aspect of fitting is to describe what one means by a "best" fit. The method of least-squares defines this criteria as minimizing the sum of squares of the residuals, i.e.

$$\text{minimize} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2$$

There are several methods of accomplishing this task.

The most common method of least-squares regression is known as ordinary least squares (OLS), whereby the parameters of a linear function are chosen. Note that the method is not restricted by the number of parameters β_i, and in fact many models may be more appropriately fit by several such parameters. In such a case, the previous example could instead be written as

$$y_i \approx \beta_{m-1} x_i^{m-1} + \cdots + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 \quad i = 1, \ldots, n$$

Such a system is conveniently described using a system of linear equations

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{m-1} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{m-1} \end{bmatrix}}_{\boldsymbol{\beta}},$$

where $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^m$, and $X \in \mathbb{R}^{n \times m}$. This type of system is nearly always over-determined, and likely will have no solution. However, the normal equations (the sum of residuals from earlier) which are now written terms of matrices

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} X_{i,j} \beta_j|^2 = ||\mathbf{y} - X\boldsymbol{\beta}||^2,$$

may be minimized. It turns out that this minimization problem does have a unique solution, given by

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}.$$

To obtain a measurement of how good of a fit one has to the observed data, a coefficient of determination may be computed. Given your data, the mean of the observed data $\overline{y}$, and your model (fitted) values f_i, the coefficient of determination is computed as

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

where SS_{res} is the sum of squares of residuals,

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

and SS_{tot} is the total sum of squares,

$$SS_{res} = \sum_{i}^{n} (y_i - \overline{y})^2.$$

Deliverable

Your task for this section is as follows:

- Take the data in `sample1.dat` and `sample2.dat`, and perform linear regression on each with $m = 2$. Plot the original data and your fitted curve for each set of samples (two plots).
- Repeat this process with $m = 3$.
- Compute the coefficient of determination for each data set and for each model and discuss the results.

You must use `C++` or `Fortran` for this part, but visualization may be done using `Matlab` or `Python`, as usual. For the matrix computations, you may use any library of your choosing. You are welcome to use any matrix solver routines (Jacobi, Gauss-Seidel, etc.) that you may have written in the past.

Correlation between Physical Data Sets

In this section you will look into the covariance of data sets from certain physical measurements. We're going to use some of the data that's available about weather conditions here in Leon County, at `https://leon.weatherstem.com/data`. Select `Leon County, Florida` and `FSU WeatherSTEM` for the station. We will want to look at the measured temperature, solar radiation, and rain rate. So select `Thermometer`, `Solar Radiation Sensor`, and `Rain Rate` from the list of measurements. For the time period, select from `July 1, 2018` to `July 31, 2018`. For output format, select `.CSV`, for the interval select `Hour`, and for the operation select `Average`. In order to compute the correlation between two sets of data $\{t_i, x_i\}_{i=1}^{n}$ and

$\{t_i, y_i\}_{i=1}^{n}$ we first compute the covariance as

$$Cov(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n}.$$

Then finally the correlation is computed using the Pearson correlation coefficient as

$$\rho_{\mathbf{x},\mathbf{y}} = \frac{Cov(\mathbf{x}, \mathbf{y})}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}}$$

Deliverable

Your task is to read the `.CSV` files into your `C++` or `Fortran` code and do the following:

- Compute the correlation between the solar radiation and the temperature data.
- Compute the correlation between the solar radiation and the rain rate data.
- Interpret the results of each and comment on your findings.

Submission and Grading

Please submit a folder containing the following in Canvas:

- your source code files
- Your report as a single file in '.pdf' format, including results from your work and relevant discussion of your observations, results, and conclusions.

This information must be received by 11:59pm, February 6 through Canvas. As stated in the course syllabus, late assignment submissions will be subject to a 10% point penalty per 24 hours past the due date at time of submission, to a maximum reduction of 50%, according to the formula:

*[final score] = [raw score] - min(0.5, 0.1 * [# of days past due]) * [maximum score]*