INDIAN INSTITUTE OF TECHNOLOGY MADRAS
Department of Chemical Engineering

CH5019 Mathematical Foundations of Data Science (Jan. - May 2024)
Course Project

Due: May 5, 2024 11:00PM

Q1.

Linear regression has been one of the most important statistical data analysis tools. Given the independent and identically distributed (i.i.d) observations $(x_i, y_i), i = 1, ..., n$, in order to understand how the response y_i's are related to the covariates x_i's, we traditionally assume the following linear regression model:

$$y_i = x_i^T \beta + \epsilon_i,$$

where β is an unknown $p \times 1$ vector, and the ϵ_i's are i.i.d and independent of x_i with $E(\epsilon_i | x_i) = 0$. The most commonly used estimate for β is the ordinary least-square (OLS) estimate that minimizes the sum of squared residuals

$$\sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

However, it is well known that the OLS estimate is highly sensitive to the outliers. A single outlier can have large effect on the OLS estimate. Among the several methods, Least Median of Squares (LMS) estimates and Least Trimmed Squares (LTS) estimates are of interest in this project.

Least Median of Squares:
The LMS estimates are found by minimizing the median of the squared residuals

$$\hat{\beta} = \arg\min_{\beta} \text{Med} \left\{ (y_i - x_i^T \beta)^2 \right\}$$

Least Trimmed Squares:
The LTS estimate is defined as

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{q} r_{(i)}(\beta)^2$$

where $r_{(1)}(\beta)^2 \leq \cdots \leq r_{(q)}(\beta)^2$ are ordered squared residuals, $q = (\frac{n}{2} + 1)$ and n is the number of samples.

(i) Write a MATLAB/Python program to implement OLS, LMS, and LTS methods using gradient descent.

(ii) Generate synthetic data as per below equation:

$$Y = 3X_1 + 5X_2 + \epsilon$$

Let X_i follow a standard normal distribution $\mathcal{N}(0,1)$, where $i = 1, 2$, and noise ϵ drawn from a normal distribution $\mathcal{N}(0,1)$. Estimate the above model's parameters using the three methods OLS, LMS and LTS with codes written in question (i) for N = 20, and N = 100 observations with realizations $R = 200$. Compare the parameter estimates using the metrics given below and report the metrics in a tabular form separately for N = 20 and N = 100.
Metrics for Comparison:
Mean Square Error (MSE), Robust Bias (RB) and Median Absolute Deviation (MAD) are defined as follows:

$$MSE_i = bias(\hat{\beta}_i)^2 + var(\hat{\beta}_i)$$
$$RB_i = \text{median}(\hat{\beta}_i) - \beta_i$$
$$MAD_i = \text{median}(|\hat{\beta}_i - \beta_i|)$$

where, $\hat{\beta}_i$ and β_i are parameter estimate and true parameter respectively and $i = 0, 1, ..., p$ are number of regressors.

(iii) **Real Dataset**: The medical insurance dataset **medical_insurance.csv** with regressors $X_i = \{\text{age}, \text{sex}, \text{bmi}, \text{children}, \text{smoker}, \text{region}\}$ and response $Y_i = \text{charges}$, encompasses various factors influencing medical expenses such as age, sex, BMI, smoking status, number of children, and region.

(a) Split the data into training and test sets(test data = 20% of dataset) and fit a linear model using OLS, LMS, and LTS methods on the training data.
(b) Give your conclusion for choice of method based on performance(Mean Squared Error) on test dataset. (**Note:** Mean Squared Error in this question is given by $\frac{1}{N}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$

Q2.

In the course we have assumed that time-series is obtained at regular intervals in time. However, it is quite common to encounter the case of uneven sampling, i.e., data being obtained at irregular intervals of time, especially in the fields of astronomy, seismology, climatology, genetics and to a lesser extent in the field of process engineering. In such cases, it is not possible to directly use the standard methods of TSA including both non-parametric (e.g., ACF, spectral analysis) and parametric types of analyses (e.g., ARIMA modelling). There exist several approaches to the analysis and reconstruction of irregular time-series.

Lomb-Scale periodogram is a method for spectral analysis of unevenly sampled time-series. It is also known as **Least Square periodigram**. To implement LS periodogram use the following objective function to estimate the parameters a and b:

$$L = \min_{a_i, b_i} \sum \left(y_k - a_i \cos(\Omega_m t_k) + b_i \sin(\Omega_m t_k)\right)^2$$

where $i \; \epsilon$ the set of frequencies in the signal

Write a function to implement the LSP (please note that you are NOT allowed to use existing packages or scripts in the open literature) using gradient descent algorithm. Also, make sure that the signal is mean-centered before estimating the coefficients (and hence the periodogram).

Data sets for testing / analysis:
(a) Synthetic data: The signal for testing is testing is a sum of sines containing two frequencies, 10 Hz and 17 Hz (continuous-time frequencies). Choosing a sampling frequency of $Fs = 50Hz$ to

generate the full-length discrete-time signal of $N = 500$ observations. Now randomly sample this regular discrete-time sine wave for a given percentage of missing data (e.g., 10%). Finally, add noise to the resulting irregular series such that the SNR is approximately 10.

(b) Stock Price: The given data **tesla_stock_price.csv** contains (irregularly sampled) daily closing price of Tesla Stock for the period of 2010 to 2020. The objective is to develop a model to forecast the closing stock price. Split your data set into training and testing. Validate the choice of the model using Normalized Mean Square Error and Mean Absolute Percentage Error metrics on the test data. Compare the test and training results by using the following modeling techniques for forecasting of tesla stock price:

(i) Lomb-Scale periodogram [**hint**: Choose a suitable range of frequencies for modelling].

(ii) ARIMA model.