CH 5440 Multivariate Data Analysis

Assignment 1

Due Date: 30/1/17

1. (a) Let $x_1, x_2, \cdots, x_N$ and $y_1, y_2, \cdots, y_N$ be a set of N measurements of two variables x and y which are linearly related. We are interested in determining the linear regression parameter m where $y = mx + c$. Assume that the measurements of x and y contain errors, with standard deviations σ_δ and σ_ε, respectively. If the ratio of the error variances $\gamma = \sigma_\varepsilon^2 / \sigma_\delta^2$ is known, derive the weighted TLS (WTLS) estimates of m and c in terms of $s_{xx}, s_{xy}, s_{yy}, \bar{x}, \bar{y}, \gamma$. Prove that the standard OLS estimates and inverse OLS estimates for m and c are obtained in the limit as γ tends to ∞ and 0, respectively. (b) How will the solution for m change if it is already known that the constant c is known to be 0?

   Note: The WTLS regression problem when the error variances are known is the solution of the following minimization problem. Multiply the objective function by σ_ε^2 and replace the ratio of the error variances by γ. Differentiate the objective function with respect to the decision variables and solve resulting set of nonlinear algebraic equations for obtaining the parameters m and c.

$$\underset{m,c,\hat{x}_i}{Min} \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)^2 / \sigma_\varepsilon^2 + (x_i - \hat{x}_i)^2 / \sigma_\delta^2$$

2. Obtain the parameters m and c for the above WTLS problem using the eigenvector of the covariance matrix of measurements and verify whether they are the same as estimates obtained in problem 1.

3. Anscombe (1973) has provided four synthetic data sets consisting of two variables x and y (data in file anscombe.xls). Find the best fit linear model for the four data sets using standard OLS. What do you observe? For which of the four data sets do you think that a linear model is appropriate and why?