DEPARTMENT OF CHEMICAL ENGINERING
CH5440: MULTIVARIATE DATA ANALYIS
ASSIGNMENT 2

1. The quality of 4898 white and 1599 red Portuguese wine samples were evaluated by three experts and their average quality ratings on a scale from 1 to 10 are reported along with several attributes of the wines (such as acidity, density, alcohol content etc.) in Excel file *winedata.xlsx (sheets Red Wine and White Wine)* (source: https://archive.ics.uci.edu/ml/datasets/Wine+Quality). Develop a regression model using (a) OLS and (b) TLS to predict the quality of the wine based on its attributes. Since the attributes are in different units, scale the data respectively, using the standard deviation of the measurements for each variable before applying OLS or TLS. For evaluating the performance of the regression models, use the first 3430 white wine samples and 1120 red wine samples to develop the regression model and the remaining as test samples. Compute the root mean square error (RMSE) between predicted and measured quality of the test samples as a measure of the regression model performance. Report the regression models, and RMSE values along with your conclusion. RMSE is defined as

$$RMSE = \sqrt{\sum_{i=1}^{N_{test}} \left(y_i - \hat{y}_i\right)^2 / N_{test}}$$

where y_i is the measured value of the output variable and $\hat{y}_i$ is the predicted value of the output variable.

Based on RMSE conclude whether the OLS or TLS model is better. Give reasons why one of the techniques is performing better.

2. The following gases carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O) and Ozone (O3) in the atmosphere are implicated in increasing global temperatures, and are known as greenhouse gases. The concentration of these gases in the atmosphere and corresponding global average temperatures obtained from the EPA website (https://www.epa.gov/climate-indicators/weather-climate) between the years 1984 to 2014 is given in the Excel file *temperature_global.xlsx* (units for different variables are also described in Excel sheet).

(a) Develop a linear regression model between global temperature (deviations) and concentrations of greenhouse gases using (a) OLS and (b) TLS. Before applying OLS or TLS scale the data using their respective standard deviation of measurements (also known as auto-scaling). Is the global temperature positively correlated with increase in the concentration of these gases?

(b) The effect of different gases on the global temperature is expressed in terms of CO2 equivalents or global warming potential (GWP). The GWP of different gases over a 20 year time horizon is as follows: CO2 (1), CH4 (86), N2O (289). Is it possible to make any inference regarding GWP of the gases from the regression coefficients? Which regression model (OLS or TLS) do you think is more reliable and why?

Notes: Water vapour, which is present in significant amount is the atmosphere is also a greenhouse gas, but it remains almost constant and is relatively unaffected by human activity. CFCs/HCFCs which are also greenhouse gases are however being monitored only in recent years.

3. A zoologist obtained measurements of the mass (in grams), the snout-vent length (SVL) and hind limb span (HLS) in mm of 25 lizards. **The mean and covariance matrix of the data about the mean** are given by

$$\bar{x} = \begin{bmatrix} 9 \\ 68 \\ 129 \end{bmatrix} \qquad S = \begin{bmatrix} 7 & 21 & 34 \\ 21 & 64 & 102 \\ 34 & 102 & 186 \end{bmatrix}$$

(a) The largest eigenvalue of the above covariance matrix is 250.4. Determine the normalized eigenvector corresponding to this eigenvalue. Also determine the remaining eigenvalues and corresponding mutually orthogonal eigenvectors.

(b) How many principal components should be retained, if at least 95% of the variance in the data has to be captured?

(c) Assuming that there are two linear relationships among the three variables, determine one possible set of these linear relations.

(d) Using the PCA model, determine the scores for a female lizard with the following measurements: mass = 10.1 gms, SVL = 73mm and HLS = 135.5mm.

(e) Using the PCA model, estimate the mass of a lizard whose measured SVL is 73mm

(f) Using the PCA model, estimate the mass of a lizard whose measured SVL is 73mm and measured HLS is 135.5 mm.

Note: The first two problems can be solved using MATLAB or R, while the last problem should be done manually (you can use MATLAB to verify your results). The MATLAB codes should be submitted along with the solution.