DEPARTMENT OF CHEMICAL ENGINERING

CH5440: MULTIVARIATE DATA ANALYIS

ASSIGNMENT 2

1. The quality of Portuguese 4898 white and 1599 red wines samples were evaluated by three experts and their average quality on a scale from 1 to 10 is 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)*. 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.

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 autoscaling). 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

(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? Ratio eigen value to sum of all eigen value

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

(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.