DEPARTMENT OF CHEMICAL ENGINERING
CH5440: MULTIVARIATE DATA ANALYIS
ASSIGNMENT 2

1. 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 *ghg-concentrations_1984-2014.xlsx* (units for different variables are also given in Excel sheet).

(a) Develop a multilinear regression model between global temperature (deviations) and concentrations of greenhouse gases using OLS. Is the global temperature positively correlated with increase in the concentration of these gases?

(b) Estimate the error variance in temperature measurements and confidence intervals (CIs) for all regression coefficients. Based on residual analysis, remove samples suspected of being outliers (one at a time) until there are no outliers.

(c) Improve the regression model obtained in step (b) by dropping unimportant (insignificant) variables (one at a time).

(d) The effect of different gases on the global temperature is expressed in terms of CO2 equivalents or global warming potential (GWP). Is it possible to make any inference regarding GWP of the gases from the regression coefficients? Compare the GWP obtained from regression coefficients to the values obtained over a 20 year time horizon: CO2 (1), CH4 (86), N2O (289).

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.

2. Consider the problem of developing a correlation between saturated pressure (P^{sat}) and saturated temperature T (boiling point). For pure components, the Antoine equation given below generally fits the data well

$$ln\, P^{sat} = A - B/(T + C)$$

For n-hexane, the values of the constants are A = 14.0568, B = 2825.42, and C = 230.44 where P^{sat} is given in kPa and T in deg C. Using this correlation a data set consisting of 100 samples have been generated in the temperature range 10 - 70 deg C. Gaussian measurements errors to both the true temperature and saturated pressures with standard deviations of 0.18 deg C and 2 kPa, respectively, have been added to generate the measurements (available in *vpdata.mat*)

(a) The Classius-Clapeyron equation is a theoretically derived model between P^{sat} and T and is given by

$$ln\, P^{sat} = A' - \frac{B'}{T}$$

Assuming that temperature measurements are noise-free and pressure measurements are noisy, use linear regression to obtain estimates of parameters A’ and B’.

(b) Assuming that temperature measurements are noise-free and pressure measurements are noisy, use nonlinear regression to obtain estimates of parameters A, B and C.

(c) Assuming both pressures and temperature measurements are noisy apply weighted total least squares obtain estimates of parameters A, B, and C. Use the inverse of standard deviation of errors as weights to set up the nonlinear optimization problem.

(d) For the models obtained in (a), (b), and (c) report the maximum error in predicting the saturated pressures using the identified model for the sample data.

Use MATLAB function lsqnonlin to estimate the nonlinear model parameters in (b) and (c)

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 and second problem can be solved using MATLAB, while the third problems should be solved manually and can be verified using MATLAB. Submit the MATLAB codes along with your solution