MSc Mathematical and Computational Finance

Introduction to Statistics - Problem Sheet 'Week 1'

1. Data Analysis Report - restrict answers for this session to no more than 2 pages, clearly stating the problem you are examining, describing the data, labeling your graphs and backing your conclusions with evidence. Your research will be extensive, however, the report should focus on describing the models considered, summarising your key findings with relevant data. The report should not be reduced to a 'cut-and-paste' exercise of the outputs of all your code.

   - Download the files 'AEMOPriceDataSimple.csv' and 'SydTemp.csv', reading the data into a panda DataFrame in Python.
   - The AEMO file contains daily average Regional Reference Price (RRP) for New South Wales (Australia), and is adapted from https://www.aemo.com.au/energy-systems/electricity. The data gives the daily average spot price over 2013. The SydTemp file gives daily maximum and minimum temperatures at Observatory Hill, Sydney, since 1859, and is taken from the main weather observatory http://www.bom.gov.au/climate/data-services/.
   - Your task is to investigate the relationship between temperature and electricity prices, taking account of possible nonlinearities and day-of-the-week effects. Be careful to investigate the data for possible outliers (there was a large electrical storm in Sydney on 20 December 2013), and deal with any appropriately. Write a summary of your findings, including what models were investigated, how you validated your assumptions, the statistical significance of your findings, what temperature and day of the week effects you detected. Present this information in a useful way, using tables, graphs and textual discussion. Also give predictions (with estimated errors) for the electricity prices in the first week of 2014, when temperatures were:

Date	MaxTemp	MinTemp
1/1/14	27.1	20.5
2/1/14	36.5	22.3
3/1/14	26.2	21.0
4/1/14	27.5	20.1
5/1/14	26.0	20.8
6/1/14	32.0	19.6
7/1/14	21.6	18.3

2. In the coming lectures, piecewise linear regressions will be examined. Write a function `pielinear` in Python which finds the model matrix for a standard one-dimensional a piecewise linear regression. Show your Python code as a picture, with some examples. That is, when given an array `x` (which will later be the set of observations $\{x_i\}_{i=1}^n$ and a set of nodes `knots`, $\{k_j\}_{j=1}^m$, it returns values of the piecewise linear function of this data, ready to be used for the regression model:

$$\begin{bmatrix} 1 & \mathrm{x}_1 & (\mathrm{x}_1 - k_1)^+ & ... & (\mathrm{x}_1 - k_m)^+ \\ 1 & \mathrm{x}_1 & (\mathrm{x}_2 - k_1)^+ & ... & (\mathrm{x}_2 - k_m)^+ \\ \vdots & & \ddots & & \vdots \end{bmatrix}$$

3. Consider $\{X_i\}_{i=1}^n$ a sequence of i.i.d. random variables, with mean μ, variance σ^2 and excess kurtosis $\gamma = E[(X-\mu)^4]/\sigma^4 - 3 < \infty$. Consider an estimator θ for σ^2, of the form

$$\theta = \alpha \sum_i (X - \bar{X})^2$$

   (a) For a general α, find the mean square error of θ as an estimator of σ^2. You may use the fact

$$Var\left[\frac{1}{n-1}\sum_i (X-\bar{X})^2\right] = \left(\frac{\gamma}{n} + \frac{2}{n-1}\right)$$

   (b) Find the value of α, as a function of γ,σ^2 and n, such that the mean squared error of θ as an estimator of σ^2 is minimized.

   (c) For the case when the X_i come from a normal distribution, calculate the bias of the minimal MSE estimator, and the difference in MSE between the optimal and unbiased estimators.