Statistical Computing and Simulation

Spring 2023

Assignment 3, Due April 21/2023

1. Given the following data, use the orthogonalization methods such as Cholesky or QR to perform regression analysis, including the parameter estimates and their standard errors via the sweep operator. (You may use the functions of matrix computation built in R, but not the function "*lm*" or "*glm*".) Compare your results with those from statistical software, such as SAS, SPSS, and Mintab.

x_1 Reactor temperature (^{0}C)	x_2 Ratio of H_2 to n-heptane (mole ratio)	x_3 Contact time (sec.)	y Conversion of n-heptane to acetylene(%)
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	29.5

It is anticipated that an equation of the following form would fit the data

$$E(Y) = \beta_0 + \sum \beta_i X_i + \sum \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j \text{, and } Var(Y) = \sigma^2$$

2. Figure a way to find the parameters of AR(1) and AR(2) models for the data "lynx" in R. Also, apply statistical software (e.g., R, SAS, SPSS, & Minitab) to get estimates for the AR(1) and AR(2) model and compare them to those from your program.

3. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) both can be used to reduce the data dimensionality. Please go to the webpage of Ministry of Interior and download Taiwan mortality data,17 five-age groups for ages 0~4, 5~9, …, 80~84 in 2001-2020, and use these data to demonstrate how these two methods work. The data of the years 2001-2015 are used as the "training" (in-sample) data and the years 2016-2020 are used as the "testing" (out-sample) data. Comments on your findings.

4. (a) Write a small program to perform the "Permutation test" and test your result on the correlation of DDT vs. eggshell thickness in class, and the following data:

X	585	1002	472	493	408	690	291
Y	0.1	0.2	0.5	1.0	1.5	2.0	3.0

Check your answer with other correlation tests, such as regular Pearson and Spearman correlation coefficients.

(b) Simulate a set of two correlated normal distribution variables, with zero mean and variance 1. Let the correlation coefficient be 0.2 and 0.8. (Use Cholesky!) Then convert the data back to Uniform(0,1) and record only the first decimal number. (亦即只取小數第一位，0至9的整數) Suppose the sample size is 10. Apply the permutation test, Pearson and Spearman correlation coefficients, and records the p-values of these three methods. (10,000 simulation runs)

5. Using simulation to construct critical values of the Mann-Whitney-Wilcoxon test in the case that $2 \le n_1, n_2 \le 10$, where n_1 and n_2 are the number of observations in two populations. (Note: The number of replications shall be at least 10,000.)

6. This assignment is to test parametric vs. nonparametric bootstrap, i.e., sensitivity of distribution assumption. Suppose 25 observations are drawn from N(0,1) and t(5). The goal is to give a 95% confidence interval for mean via both parametric and nonparametric bootstrap simulations. Assuming that observations are all from normal distribution for the parametric bootstrap. Conduct the at least 500 bootstrap simulations each case (parametric vs. nonparametric, normal vs. t) for 1,000 times and comment on the results.

7. (Bonus!) The task is to use SVD to compress the lion image, similar to the handout on my webpage. In addition to the SVD, you need to apply other non-linear data reduction methods and compare their differences.