Statistical Computing and Simulation

Assignment 3, Due April 26/2024

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

Obs. #	x_1	x_2	x_3	y	Obs. #	x_1	x_2	x_3	y
1	80	27	89	42	11	58	18	89	14
2	80	27	88	37	12	58	17	88	13
3	75	25	90	37	13	58	18	82	11
4	62	24	87	28	14	58	19	93	12
5	62	22	87	18	15	50	18	89	8
6	62	23	87	18	16	50	18	86	7
7	62	24	93	19	17	50	19	72	8
8	62	24	93	20	18	50	19	79	8
9	58	23	87	15	19	50	20	80	9
10	58	18	80	14	20	56	20	82	15
					21	70	20	91	15

x_1 = air flow;
x_2 = cooling water inlet temperature;
x_3 = acid concentration;
y = stack loss.

2. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) both can be used to reduce the data dimensionality. Use the mortality data, 17 5-age groups for ages 0~4, 5~9, …, 80~84, in Taiwan area to demonstrate how these two methods work. The data of the years 1970-2000 are used as the "training" (in-sample) data and the years 2001-2005 are used as the "testing" (out-sample) data. You only need to perform one set of data, according to your gender.

3. (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)

4. Using simulation to construct critical values of the Mann-Whitney-Wilcoxon test in the case that $2 \leq n_1, n_2 \leq 10$, 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.)

5. Similar to what Efron did in the Law school data example, compute the bootstrap simulation for 50, 100, …, 10,000 replications. But, instead of using the original 15 observations, we want to know if the number of observations plays an important role. Randomly select 10, 15, 20, and 25 observations and then see if the bootstrap variance converges as the number of replications increases. (Note: You also need to compare your results with that of population.)

school	LSAT	GPA	school	LSAT	GPA	school	LSAT	GPA	school	LSAT	GPA
1	622	3.23	22	614	3.19	43	573	2.85	63	572	3.08
2	542	2.83	23	628	3.03	44	644	3.38	64	610	3.13
3	579	3.24	24	575	3.01	(45)	545	2.76	65	562	3.01
(4)	653	3.12	25	662	3.39	46	645	3.27	66	635	3.30
5	606	3.09	26	627	3.41	(47)	651	3.36	67	614	3.15
(6)	576	3.39	27	608	3.04	48	562	3.19	68	546	2.82
7	620	3.10	28	632	3.29	49	609	3.17	69	598	3.20
8	615	3.40	29	587	3.16	(50)	555	3.00	(70)	666	3.44
9	553	2.97	30	581	3.17	51	586	3.11	71	570	3.01
10	607	2.91	(31)	605	3.13	(52)	580	3.07	72	570	2.92
11	558	3.11	32	704	3.36	(53)	594	2.96	73	605	3.45
12	596	3.24	33	477	2.57	54	594	3.05	74	565	3.15
(13)	635	3.30	34	591	3.02	55	560	2.93	75	686	3.50
14	581	3.22	(35)	578	3.03	56	641	3.28	76	608	3.16
(15)	661	3.43	(36)	572	2.88	57	512	3.01	77	595	3.19
16	547	2.91	37	615	3.37	58	631	3.21	78	590	3.15
17	599	3.23	38	606	3.20	59	597	3.32	(79)	558	2.81
18	646	3.47	39	603	3.23	60	621	3.24	80	611	3.16
19	622	3.15	40	535	2.98	61	617	3.03	81	564	3.02
20	611	3.33	41	595	3.11	62	637	3.33	(82)	575	2.74
21	546	2.99	42	575	2.92						

Sampled schools have bold-faced school numbers.

6. To compare teaching, twenty schoolchildren were divided into two groups: ten taught by conventional methods and ten taught by an entirely new approach. The following are the test results:

Conventional	65	79	90	75	61	85	98	80	97	75
New	90	98	73	79	84	81	98	90	83	88

Are the two teaching methods equivalent in result? You need to use permutation test, (parametric and non-parametric) bootstrap, and parametric test, and then compare their differences in testing.