COMPREHENSIVE WRITTEN EXAMINATION, PAPER III
PART 2: FRIDAY AUGUST 12, 2022 1:00 P.M.–5:00 P.M.
STOR 664 Data Analysis Question (50 points)

Table 1 (see end of exam) is part of a data base documenting deaths in the salmon population along the Columbia River in Washington State, as a function of time and various pollutants. The observations are as follows:

Y1	Number of dead fish found
SP	Indicator for spring quarter
SU	Indicator for summer quarter
FA	Indicator for fall quarter
YR	Year since start of study
PH	Level of phosphorus in the river
NT	Level of nitrogen in the river
SO2	Level of atmospheric sulfur dioxide
PM10	Level of atmospheric particulate matter

There is no indicator for the winter quarter because this is collinear with the indicators for spring, summer and fall.

The data may be downloaded at: http://rls.sites.oasis.unc.edu/salmon.csv

Answer the following questions. Your answers should include enough computer code to show how you got your answers, but please take care also to write complete verbal answers to the questions.

(a) We propose an analysis in which some transformation of Y_1 is regressed against some combination of the other variables, using standard normal-theory linear regression. Using plots and numerical diagnostics, as appropriate, find an appropriate transformation of Y_1. [**5 points**]

From now on, assume that the transformation adopted in (a) is $y = \sqrt{Y_1}$. This is not necessarily the correct answer to (a).

(b) Using standard variable selection methods, find a suitable model for regressing y against some subset of the remaining 8 variables. You should consider at least the following model selection criteria: AIC, BIC, forward selection, stepwise selection and backward selection. Explain your reasoning and justify your final selection. [**10 points**]

(c) Using your preferred model from (b), give as complete a discussion as you can about all the aspects of model diagnostics, including (i) points of high leverage, (ii) residuals and outliers, (iii) influential observations by all the standard criteria for assessing influence, (iv) multicollinearity. [**10 points**]

(d) Fit the same data using principal components regression. Use cross-validation or some other technique (which you should describe) to determine the number of components. [**5 points**]

(e) Fit the same data using Lasso regression. Use cross-validation or some other technique (which you should describe) to determine the optimal tuning parameter. [**5 points**]

(f) Briefly summarize your conclusions from this whole exercise. Do all the methods lead to comparable answers? Would any one of them stand out as best in your opinion? [**5 points**]

(g) The last four rows in Table 1 are dummy rows (missing variable in Y_1). Suggest a reason why these rows might have been included, and state your best predictions of these variables based on your analyses in the earlier sections (you don't have to consider all the preceding methods for this part: it will suffice to pick one and discuss the predictions in detail. **[10 points]**

TABLE 1

Y1	SP	SU	FA	YR	PH	NT	SO2	PM10
51	1	0	0	1	4.7	25.1	9.3	40.4
133	0	1	0	1	10.1	17.5	8.3	18.9
12	0	0	1	1	4.6	17.6	7.1	44.6
41	0	0	0	1	15.9	22.0	29.1	28.3
38	1	0	0	2	4.3	15.6	12.9	21.4
101	0	1	0	2	5.5	15.4	7.1	51.9
21	0	0	1	2	5.0	22.9	5.0	11.8
19	0	0	0	2	7.9	19.3	13.1	13.9
42	1	0	0	3	1.6	16.9	24.0	27.3
139	0	1	0	3	8.8	26.2	14.7	22.2
11	0	0	1	3	14.7	19.0	10.1	26.5
16	0	0	0	3	14.1	20.5	14.1	13.0
58	1	0	0	4	10.9	17.8	24.3	36.8
103	0	1	0	4	9.7	19.2	9.5	48.1
17	0	0	1	4	8.0	18.5	9.4	7.1
294	1	0	0	5	96.9	19.4	26.4	16.9
19	0	0	0	4	12.0	19.1	12.6	35.6
138	0	1	0	5	12.6	19.4	23.1	78.3
15	0	0	1	5	2.1	20.1	35.4	90.6
15	0	0	0	5	18.6	28.7	18.9	25.5
33	1	0	0	6	13.8	20.3	10.0	73.7
101	0	1	0	6	10.1	22.2	14.7	32.8
13	0	0	1	6	9.4	13.9	17.8	12.4
10	0	0	0	6	3.6	24.5	6.8	38.2
19	1	0	0	7	8.8	24.3	9.7	9.7
60	0	1	0	7	1.8	23.6	12.4	16.8
9	0	0	1	7	6.6	22.0	12.1	59.8
7	0	0	0	7	16.7	14.6	5.9	69.4
29	1	0	0	8	4.2	22.1	14.0	73.7
65	0	1	0	8	9.5	17.3	7.4	75.3
6	0	0	1	8	4.9	25.2	11.0	71.6
13	0	0	0	8	4.0	19.6	7.8	90.0
35	1	0	0	9	25.7	21.0	12.6	37.7
56	0	1	0	9	3.4	22.3	17.5	27.7
10	0	0	1	9	34.7	15.9	15.5	32.1
4	0	0	0	9	4.6	23.3	7.2	23.6
30	1	0	0	10	2.2	22.7	21.1	13.6
97	0	1	0	10	35.8	19.7	14.1	43.3
9	0	0	1	10	3.1	21.4	10.1	113.9
6	0	0	0	10	14.2	18.3	17.1	46.0
NA	1	0	0	11	12.128	20.36	13.98	40.51
NA	0	1	0	11	12.128	20.36	13.98	40.51
NA	0	0	1	11	12.128	20.36	13.98	40.51
NA	0	0	0	11	12.128	20.36	13.98	40.51