Assignment #4— Due Saturday, 27 Nov.

*This homework covers Chapter 7 (*Problem* 1-3) and Chapter 8 (*Problem* 4). Submit your homework on Canvas or send it to our TA, Mr. LYU Zhongyuan (zlyuab@connect.ust.hk).

*No late homework will be accepted for credit.

*Append the R codes you used to your submission. *If the problem does not need R or is not explicitly stated to complete in R, then you should just do it by hand with a calculator.*

*In case of rounding error, keep 3 figures after the decimal point.

Problem 1 (Use R) The following table shows the annual world crude oil production in millions of barrels for the period 1880-1988, stored in the file *Crude-Oil-Production.txt*.

Annual World Crude Oil Production in Millions of Barrels (1880–1988)

Year	OIL	Year	OIL	Year	OIL
1880	30	1940	2,150	1972	18,584
1890	77	1945	2,595	1974	20,389
1900	149	1950	3,803	1976	20,188
1905	215	1955	5,626	1978	21,922
1910	328	1960	7,674	1980	21,722
1915	432	1962	8,882	1982	19,411
1920	689	1964	10.310	1984	19,837
1925	1,069	1966	12,016	1986	20,246
1930	1,412	1968	14,104	1988	21,338
1935	1,655	1970	16,690		

Take the log transformation on OIL(barrels) and fit a linear regression on Year.

(a) Draw the residual versus index plot, do you believe the errors are autocorrelated? Explain

(b) Compute the Durbin-Watson statistic d. What conclusion regarding the presence of autocorrelation would you draw from d?

(c) Compare the number of runs to their expected value and standard deviation. What conclusions regarding the presence of autocorrelation would you draw from this comparison?

(d) Try Cochrane and Orcutt procedure for one iteration, and report the result and draw your conclusion.

Problem 2 (Use R) In the analysis of the Advertising data (stored in the file *Advertising.txt*) in Section 7.5 on lecture slides, it is suggested that the regression of sales S_t against E_t and three of the remaining four variables $(A_t, P_t, A_{t-1}, S_{t-1})$ may resolve the collinearity problem. Run the four, i.e. $\binom{4}{3} = 4$, suggested regressions and, for each of them, examine the resulting VIF_j's to see if collinearity has been eliminated.

Problem 3 (Use R) Use again the Advertising data, and fit the model $S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t$ by Ridge regression.

(a) Plot the Ridge trace and visually propose an appropriate value of the bias parameter. (*Hint: to obtain the Ridge trace, you might need to write a loop function in R, gather the results similarly as in Table 7.15 on lecture slides, and then draw the plot.*)

(b) Use the iterative method to get the value of bias parameter, report the final estimates by Ridge regression and compare them to the simple OLS (ordinary least squares).

Problem 4 (Use R) Gasoline Consumption: To study the factors that determine the gasoline consumption of cars, data were collected on 30 models of cars. Besides the gasoline consumption (Y), measured in miles per gallon for each car, 11 other measurements representing physical and mechanical characteristics are made. Definitions of variables and the data are given in the following two tables, stored in the file *Gasolin-Consumption.txt*.

Variables for the Gasoline Consumption Data

Variable	Definition
Y	Miles/gallon
X_1	Displacement (cubic inches)
X_2	Horsepower (feet/pound)
X_3	Torque (feet/pound)
X_4	Compression ratio
X_5	Rear axle ratio
X_6	Carburetor (barrels)
X_7	Number of transmission speeds
X_8	Overall length (inches)
X_9	Width (inches)
X_{10}	Weight (pounds)
X_{11}	Type of transmission (1 = automatic; 0 = manual)

(a) Given the nature of all these predictors, would you include all the variables to predict the gasoline consumption of the cars? Explain.

(b) Six alternative models have been suggested:

(a) Regression of Y on X_1.

(b) Regression of Y on X_{10}.

(c) Regression of Y on X_1 and X_{10}

(d) Regression of Y on X_2 and X_{10}

(e) Regression of Y on X_8 and X_{10}

(f) Regression of Y on X_8 and X_5, and X_{10}

Compare the above models using the criterions R_a^2, Mallow's C_p, AIC and BIC, and reports the model you prefer the most. Note that Mallow's C_p needs a common full model, and it is up to your choice, just report the details.

(c) Plot Y against X_1, X_2, X_8, X_{10} (one at a time). Do the plots suggest that the relationship between Y and the 11 predictor variables may not be linear?

(d) Using only the predictor variables $X_1, X_2, X_5, X_8, X_{10}$, run the forward selection procedure, and report your results.

Gasoline Consumption and Automotive Variables

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
18.9	350.0	165	260	8.00	2.56	4	3	200.3	69.9	3910	1
17.0	350.0	170	275	8.50	2.56	4	3	199.6	72.9	3860	1
20.0	250.0	105	185	8.25	2.73	1	3	196.7	72.2	3510	1
18.3	351.0	143	255	8.00	3.00	2	3	199.9	74.0	3890	1
20.1	225.0	95	170	8.40	2.76	1	3	194.1	71.8	3365	0
11.2	440.0	215	330	8.20	2.88	4	3	184.5	69.0	4215	1
22.1	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3020	1
21.5	262.0	110	200	8.50	2.56	2	3	179.3	65.4	3180	1
34.7	89.7	70	81	8.20	3.90	2	4	155.7	64.0	1905	0
30.4	96.9	75	83	9.00	4.30	2	5	165.2	65.0	2320	0
16.5	350.0	155	250	8.50	3.08	4	3	195.4	74.4	3885	1
36.5	85.3	80	83	8.50	3.89	2	4	160.6	62.2	2009	0
21.5	171.0	109	146	8.20	3.22	2	4	170.4	66.9	2655	0
19.7	258.0	110	195	8.00	3.08	1	3	171.5	77.0	3375	1
20.3	140.0	83	109	8.40	3.40	2	4	168.8	69.4	2700	0
17.8	302.0	129	220	8.00	3.00	2	3	199.9	74.0	3890	1
14.4	500.0	190	360	8.50	2.73	4	3	224.1	79.8	5290	1
14.9	440.0	215	330	8.20	2.71	4	3	231.0	79.7	5185	1
17.8	350.0	155	250	8.50	3.08	4	3	196.7	72.2	3910	1
16.4	318.0	145	255	8.50	2.45	2	3	197.6	71.0	3660	1
23.5	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3050	1
21.5	360.0	180	290	8.40	2.45	2	3	214.2	76.3	4250	1
31.9	96.9	75	83	9.00	4.30	2	5	165.2	61.8	2275	0
13.3	460.0	223	366	8.00	3.00	4	3	228.0	79.8	5430	1
23.9	133.6	96	120	8.40	3.91	2	5	171.5	63.4	2535	0
19.7	318.0	140	255	8.50	2.71	2	3	215.3	76.3	4370	1
13.9	351.0	148	243	8.00	3.25	2	3	215.5	78.5	4540	1
13.3	351.0	148	243	8.00	3.26	2	3	216.1	78.5	4715	1
13.8	360.0	195	295	8.25	3.15	4	3	209.3	77.4	4215	1