A S S I G N M E N T 4

Due date: Dec 10,2021

1. Selecting the best one-variable predictor. There are six independent variables, x_1, x_2, x_3, x_4, x_5, and x_6 , that might be useful in predicting a response y . A total of $n = 50$ observations are available, and it is decided to employ stepwise regression to help in selecting the independent variables that appear to be useful. The computer fits all possible one-variable models of the form

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_i$$

where x_i is the ith independent variable, $i = 1, 2, \ldots, 6$. The information in the table is provided from the computer printout.

Independent Variable	$\widehat{\beta}_i$	$s_{\widehat{\beta}_i}$
x_1	1.6	.42
x_2	-.9	.01
x_3	3.4	1.14
x_4	2.5	2.06
x_5	-4.4	.73
x_6	.3	.35

   (a) Which independent variable is declared the best one-variable predictor of y? Explain.

   (b) Would this variable be included in the model at this stage? Explain.

   (c) Describe the next phase that a stepwise procedure would execute.

2. Cooling method for gas turbines. Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high-pressure inlet fogging method for a gas turbine engine. A number of independent variables were used to predict the heat rate (kilojoules per kilowatt per hour) for each in a sample of 67 gas turbines augmented with high-pressure inlet fogging. The independent variables available are engine type (traditional, advanced, or aeroderivative), number of shafts, cycle speed (revolutions per minute), cycle pressure ratio, inlet temperature (°C), exhaust gas temperature (°C), air mass flow rate (kilograms per second), and horsepower (Hp units). The data are saved in the GASTURBINE file. (The first and last five observations are listed in the table.)

GASTURBINE (Data for first and last five gas turbines shown)

ENGINE	SHAFTS	RPM	CPRATIO	INLET-TEMP	EXH-TEMP	AIRFLOW	POWER	HEATRATE
Traditional	1	27245	9.2	1134	602	7	1630	14622
Traditional	1	14000	12.2	950	446	15	2726	13196
Traditional	1	17384	14.8	1149	537	20	5247	11948
Traditional	1	11085	11.8	1024	478	27	6726	11289
Traditional	1	14045	13.2	1149	553	29	7726	11964
...								
Aeroderiv	2	18910	14.0	1066	532	8	1845	12766
Aeroderiv	3	3600	35.0	1288	448	152	57930	8714
Aeroderiv	3	3600	20.0	1160	456	84	25600	9469
Aeroderiv	2	16000	10.6	1232	560	14	3815	11948
Aeroderiv	1	14600	13.4	1077	536	20	4942	12414

Source: Bhargava, R., and Meher-Homji, C. B. "Parametric analysis of existing gas turbines with inlet evaporative and overspray fogging," *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005.

(a) Use stepwise regression (with stepwise selection) to find the "best" predictors of heat rate.

(b) Use stepwise regression (with backward elimination) to find the "best" predictors of heat rate.

(c) Use all-possible-regressions-selection to find the "best" predictors of heat rate.

(d) Compare the results, parts a–c. Which inde- pendent variables consistently are selected as the "best" predictors?

(e) Explain how you would use the results, parts a–c, to develop a model for heat rate.

3. Consider the data set *t*ass4_q1.txt.

(a) Generate the correlation matrix and see if there are any strong correlation between the independent variables.

(b) Fit the data to the first order model. Do the F-test in the ANOVA table and the individual t-tests suggest multicollinearity?

(c) Do the VIF values suggest multicollinearity? State the rule of thumb and check the VIF values.

(d) Do the conditional indices and proportions of variation suggest multicollinearity? State the rule of thumb and check the indices and proportions.

(e) Plot the ridge trace in which the ridge parameter goes from 0.0 to 0.2 by 0.005.

(f) From the ridge trace and the VIF values, suggest a value for the ridge parameter. Explain your choice.

(g) Report the parameter estimates of the ridge regression with the parameter you chose above.

4. Consider the data set *t*ass4_q2.txt.

(a) Fit the data to the first order model. Plot the residuals against each independent variable x_i, the fitted value $\widehat{y}$ and the observed value of y. Identify any problems in these plots and suggest what can be done if there is a problem?

(b) Fit the data to a more appropriate model suggested by the residual plots, and then produce the same residuals as above. Identify any problems in these plots and suggest what can be done if there is a problem?

(c) Find the optimal Box-Cox transformation.

(d) Transform the data using the optimal Box-Cox transformation (without the geometric mean term) and then plot the residuals against the predicted values.

5. Consider the data set *t*ass4_q3.txt.

(a) Generate scatterplots and identify any leverage points and outliers.

(b) Calculate the studentised residuals. State the rule of thumb and check whether there are outliers.

(c) Calculate the leverages. State the rule of thumb and check whether there are leverage points.

(d) Calculate Cook's distances. State the rule of thumb and check whether there are influential points.

(e) Calculate *DFBETAS*. State the rule of thumb and check whether there are influential points.

(f) Calculate *DFFITS*. State the rule of thumb and check whether there are influential points.

(g) Calculate *COVRATIO*. State the rule of thumb and check whether there are influential points.

(h) Produce the normal probability plot for the studentised residuals. Do the studentised residuals follow the normal distribution? If not, how does the distribution deviate from the normal distribution? (Heavy-tailed? Light-tailed? Positively skewed? Negative skewed? etc.)

6. Consider the data set tass4_q4.txt.

(a) Draw a scatterplot.

(b) Suppose the data came from a study of population growth, and the theory suggests that the relationship is a logistic one:

$$\mathbb{E}(Y) = \frac{\beta_1}{1 + \beta_2 \mathrm{e}^{\beta_3 x}}. \tag{1}$$

Does the scatterplot agree with the theory?

(c) To use the SAS procedure or R procedure, we need some initial guesses for the parameter values.

i. Explain from the scatterplot why β_3 is probably negative.

ii. Suppose β_3 is negative, explain from the scatterplot why β_1 is approximately 100 (corrected to two significant figures).

iii. Suppose $\beta_1 = 100$, explain from the scatterplot why β_2 is approximately 3 or 4 (corrected to the nearest integer).

iv. Suppose $\beta_1 = 100$, explain how to estimate an initial value for β_3 (I would guess that it is approximately -0.3, corrected to the first decimal place).