

DSBS7290 Regression Analysis
Assignment 2

1. Consider to fit the data set $\{(x_i, y_i), i = 1, \dots, n\}$ by the regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^3 + \varepsilon_i, \quad (1)$$

where β_0, β_1 and β_2 are unknown constants, $\{\varepsilon_i\}$ are i.i.d. with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. The model (1) can be expressed as the following standard form

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \varepsilon \quad (2)$$

(a) Write down $\mathbf{y}, \mathbf{X}, \mathbf{B}$ and ε .

(b) Find $E(\varepsilon)$ and $\text{Cov}(\varepsilon)$.

(c) Find $E(\mathbf{y})$ and $\text{Cov}(\mathbf{y})$.

2. Consider to fit the data set $\{(x_i, y_i), i = 1, \dots, n\}$ by the regression model

$$y_i = \beta x_i^2 + \varepsilon_i,$$

Find the least squares estimator of β .

3. Show that the residuals from a linear regression model can be expressed as

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Furthermore, calculate the expectation and covariance matrix of the residual vector \mathbf{e} , i.e., $E(\mathbf{e})$ and $\text{Cov}(\mathbf{e})$.

4. The data shown below present the average number of surviving bacteria in a canned food product and the minutes of exposure to 300° F heat.

Number of Bacteria	Minutes of Exposure
175	1
108	2
95	3
82	4
71	5
50	6
49	7
31	8
28	9
17	10
16	11
11	12

- Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?
- Fit the straight-line model. Compute the summary statistics and the residual plots. What are your conclusions regarding model adequacy?
- Identify an appropriate transformed model for these data. Fit this model to the data and conduct the usual tests of model adequacy.

5. Fitness data give various measures of heart and pulse rate taken on men in a physical fitness course. The goal is to predict the rate of oxygen consumption (which is difficult and expensive to measure) from the other variables. The following factors were considered:

X_1 : Age in year

X_2 : Weight in kilograms

X_3 : Time to run $1\frac{1}{2}$ miles

X_4 : Resting pulse rate

X_5 : Pulse rate at begin of run

X_6 : Pulse rate at end of run

Y : Oxygen consumption in milliliters (ml) per kilogram (kg) body weight per minute.

The data is given in the table below.

Individual	y	x_1	x_2	x_3	x_4	x_5	x_6
1	44.609	44	89.47	11.37	62	178	182
2	45.313	40	75.07	10.07	62	185	185
3	54.297	44	85.84	8.65	45	156	168
4	59.571	42	68.15	8.17	40	166	172
5	49.874	38	89.02	9.22	55	178	180
6	44.811	47	77.45	11.63	58	176	176
7	45.681	40	75.98	11.95	70	176	180
8	49.091	43	81.19	10.85	64	162	170
9	39.442	44	81.42	13.08	63	174	176
10	60.055	38	81.87	8.63	48	170	186
11	50.541	44	73.03	10.13	45	168	168
12	37.388	45	87.66	14.03	56	186	192
13	44.754	45	66.45	11.12	51	176	176
14	47.273	47	79.15	10.60	47	162	164
15	51.855	54	83.12	10.33	50	166	170
16	49.156	49	81.42	8.95	44	180	185
17	40.836	51	69.63	10.95	57	168	172
18	46.672	51	77.91	10.00	48	162	168
19	46.774	48	91.63	10.25	48	162	164
20	50.388	49	73.37	10.08	76	168	168
21	39.407	57	73.37	12.63	58	174	176
22	46.080	54	79.38	11.17	62	156	165
23	45.441	52	76.32	9.63	48	164	166
24	54.625	50	70.87	8.92	48	146	155
25	45.118	51	67.25	11.08	48	172	172
26	39.203	54	91.63	12.88	44	168	172
27	45.790	51	73.71	10.47	59	186	188
28	50.545	57	59.08	9.93	49	148	155
29	48.673	49	76.32	9.40	56	186	188
30	47.920	48	61.24	11.50	52	170	176
31	47.467	52	82.78	10.50	53	170	172

(a) Consider the following multiple linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon, \quad (3)$$

where the random error $\varepsilon \sim N(0, \sigma^2)$. Estimate the coefficients of the model and σ^2 by the least squares method.

(b) Applying the forward selection, backward elimination, stepwise regression to model (3) and give your recommended models by using other criteria.

(c) Consider the following quadratic regression model:

$$y = \beta_0 + \sum_{i=1}^6 \beta_i x_i + \sum_{i=1}^6 \sum_{j=1}^i \beta_{ij} x_i x_j + \varepsilon, \quad (4)$$

where the random error $\varepsilon \sim N(0, \sigma^2)$. Estimate the coefficients of the model and σ^2 by the least squares method.

- (d) Applying the forward selection, backward elimination, stepwise regression to model (4) and give your recommended models by using other criteria.