Descriptive Statistics With R Software

Fitting of Linear Models

::

Least Squares Method – R Commands and More than One Variables

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Fitting Linear Model through Least Squares Estimates

R Command

Fitting Linear Models `lm`

Description

`lm` **is used to fit linear models.**

Usage

```
lm(formula, data, subset, weights, na.action,
method = "qr", model = TRUE, x = FALSE, y =
FALSE, qr = TRUE, singular.ok = TRUE, contrasts
= NULL, offset, ...)
```

Fitting Linear Model through Least Squares Estimates

Arguments

`formula` **an object of class "`formula`" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.**

`data` **an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model.**

`subset` **an optional vector specifying a subset of observations to be used in the fitting process.**

Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41

Example

Solving it for the given data on `marks` and `hours`, we get the values of α and β as follows:

$$\overline{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 389.9, \qquad \overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 35.1$$

$$\hat{\beta} = \frac{\sum_{i=1}^{20}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{20}(x_i - \overline{x})^2} = 6.3,$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 168.65$$

Model: marks = 168.65 + 6.3*hours

Example

```
marks =
c(337,316,327,340,374,330,352,353,370,380,384,
398,413,428,430,438,439,479,460,450)

hours =
c(23,25,26,27,30,26,29,32,33,34,35,38,39,42,43
,44,45,46,44,41)
```

Fitting Linear Model through Least Squares Estimates

Example

R Command

```
> lm(marks ~ hours)

Call:
lm(formula = marks ~ hours)

Coefficients:
(Intercept)        hours
    168.647        6.304
```

Fitting Linear Model through Least Squares Estimates

Example

R Console

```
> hours
 [1] 23 25 26 27 30 26 29 32 33 34 35 38 39 42 43 44
[17] 45 46 44 41
> marks
 [1] 337 316 327 340 374 330 352 353 370 380 384 398
[13] 413 428 430 438 439 479 460 450
> lm(marks~hours)

Call:
lm(formula = marks ~ hours)

Coefficients:
(Intercept)        hours
    168.647        6.304
```

Fitting Linear Model through Least Squares: More than One Variables

$$y = \alpha + \beta x + e$$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

Relationship between y and $x_1, x_2, \ldots, x_p$ is linear.

Matrix plots are useful in graphically verifying the linearity.

Conduct the experiment and obtain *n* tuples of observations on dependent variable (*y*) and independent variables $x_1, x_2, \ldots, x_p$.

Fitting Linear Model through Least Squares: More than One Variables

$$y_1 = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1$$

$$y_2 = \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$y_n = \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$y = X\beta + e$$

Fitting Linear Model through Least Squares: More than One Variables

How to find parameters?

Use principle of least squares

$\hat{\beta} = (X'X)^{-1} X'y$ Least squares estimator

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Example with Two Variables

Following data is obtained on the delivery time taken in delivering the parcels and corresponding distance travelled by a courier person.

Delivery Time Data			
Obs. number	Delivery time(in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)
1	16.68	7	560
2	11.5	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.5	5	605
11	40.33	16	688
12	21	10	215

Delivery Time Data			
Obs. number	Delivery time(in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)
13	13.5	4	255
14	19.75	6	462
15	24	9	448
16	29	10	776
17	16.35	6	200
18	19	7	132
19	9.5	3	36
20	35.1	17	770
21	17.9	10	140
22	52.32	26	817
23	18.75	9	450
24	19.83	8	635
25	10.75	4	450

Example with Two Variables

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \dots, 25$$

```
deltime =
c(16.68,11.5,12.03,14.88,13.75,18.11,8,17.83,
79.24,21.5,40.33,21,13.5,19.75,24,29,16.35,19
,9.5,35.1,17.9,52.32,18.75,19.83,10.75)

parcelno =
c(7,3,3,4,6,7,2,7,30,5,16,10,4,6,9,10,6,7,3,1
7,10,26,9,8,4)

distance =
c(560,220,340,80,150,330,110,210,1460,605,688
,215,255,462,448,776,200,132,36,770,140,817,4
50,635,450)
```

Matrix Plot

`pairs(x, ...)` produces a matrix of scatterplots.

```
pairs(formula, data = NULL, ..., subset,
      na.action = stats::na.pass)
```

Arguments:

`x` coordinates of points given as numeric columns of a matrix or data frame.

`formula` a formula, such as `~ x + y + z`. Each term will give a separate variable in the pairs plot, so terms should be numeric vectors.

`data` a data.frame (or list) from which the variables in `formula` should be taken.

Matrix Plot

Arguments

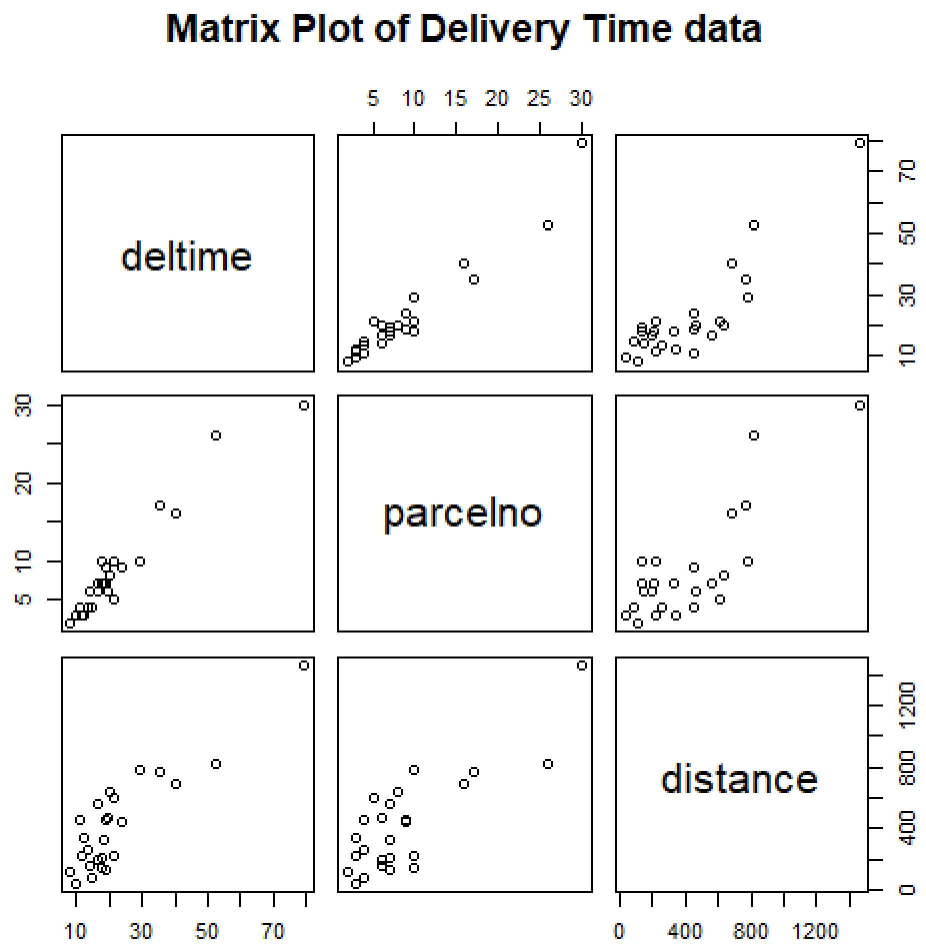
`subset` an optional vector specifying a subset of observations to be used for plotting.

`data` a data.frame (or list) from which the variables in `formula` should be taken.

Example with Two Variables

Matrix Plot

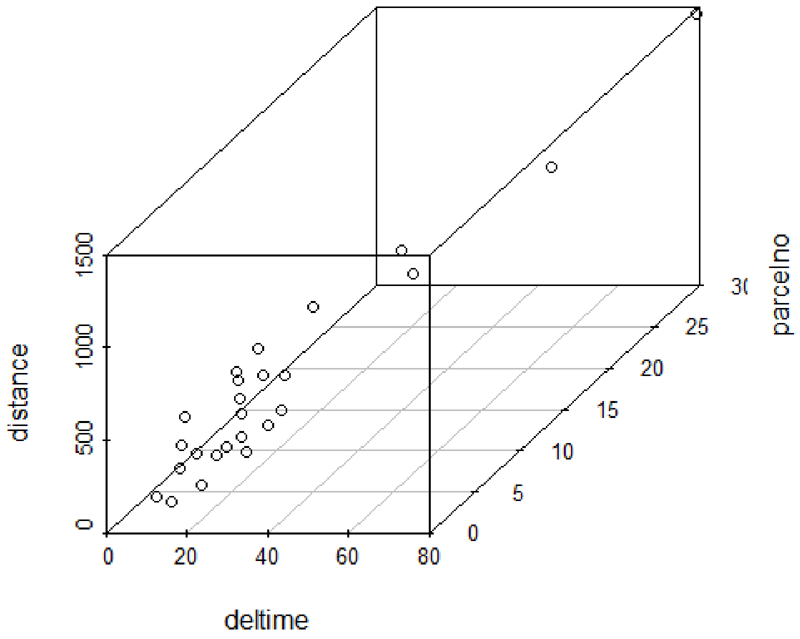
```
> pairs(~deltime + parcelno + distance,
main="Matrix Plot of Delivery Time data")
```

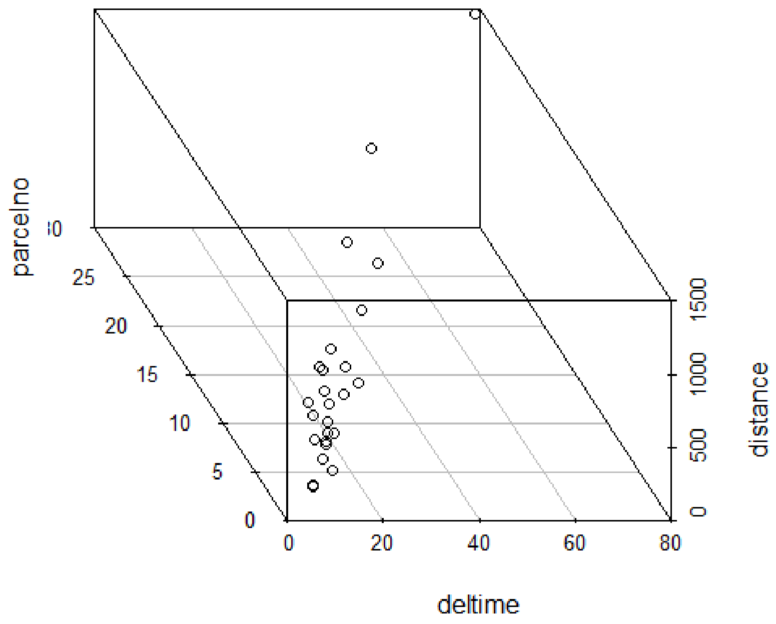


Example with Two Variables

```
>  library(scatterplot3d)
```

`scatterplot3d(deltime, parcelno, distance)`	`scatterplot3d(deltime, parcelno, distance,angle=120)`





Example with Two Variables

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \ldots, 25$$

```
>lm(deltime ~ parcelno + distance)

Call:
lm(formula = deltime ~ parcelno + distance)

Coefficients:
(Intercept)     parcelno     distance
    2.19579      1.67803      0.01311
```

Model:

deltime = 2.196 + 1.68 * parcelno + 0.013 * distance

Example with Two Variables

R Console

```
> deltime
 [1] 16.68 11.50 12.03 14.88 13.75 18.11  8.00 17.83 79.24 21.50 40.33
[12] 21.00 13.50 19.75 24.00 29.00 16.35 19.00  9.50 35.10 17.90 52.32
[23] 18.75 19.83 10.75
> parcelno
 [1]  7  3  3  4  6  7  2  7 30  5 16 10  4  6  9 10  6  7  3 17 10 26  9
[24]  8  4
> distance
 [1]  560  220  340   80  150  330  110  210 1460  605  688  215  255
[14]  462  448  776  200  132   36  770  140  817  450  635  450
> lm(deltime~parcelno+distance)

Call:
lm(formula = deltime ~ parcelno + distance)

Coefficients:
(Intercept)     parcelno     distance
    2.19579      1.67803      0.01311
```