MA2252 Introduction to Computing

Lecture 13
Least Squares Regression

Sharad Kumar Keshari

School of Computing and Mathematical Sciences

University of Leicester

Learning outcomes

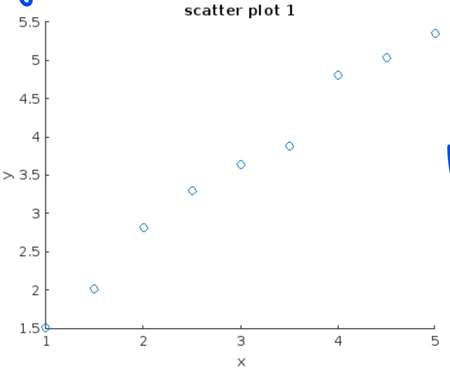
At the end of lecture, students will be able to

- understand scatter plot and regression
- understand theory of Least Squares Regression
- solve basic regression problems in MATLAB

Scatter plot

A **scatter plot** plots two different sets of data using dots. Unlike line plot, the dots are not connected by a curve.

2D scatterplot plots data of two different variables as dots.



plot(x,y)

scatter(x,y)

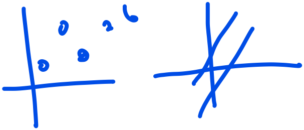
plot(x,y,'o')

Scatter plot (contd.)

MATLAB's scatter() function can be used to create a scatter plot.

Example: The code below creates the scatter plot shown earlier.


```
x=1:0.5:5; %create data for vector x
s=size(x); %find size of vector x
y=x+rand(s); %create a vector y=x+'some random values'
scatter(x,y) %create scatter plot
title('scatter plot 1')
xlabel('x')
ylabel('y')
```



Scatter plot (contd.)

Demo

Regression

Regression is a statistical technique used to find the 'best-fit curve' that describes a scatter plot.

curve which is closest approximation to the data

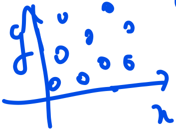
Depending on the data trend in a scatter plot, one may use

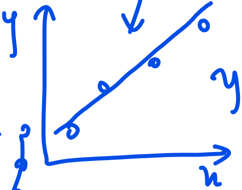
- ~~Linear Regression~~

Linear Regression curve

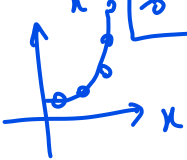
- ~~Non-linear Regression~~

Non-linear Regression curve





$y = mx + c$



curve

Linear Regression example

regression line

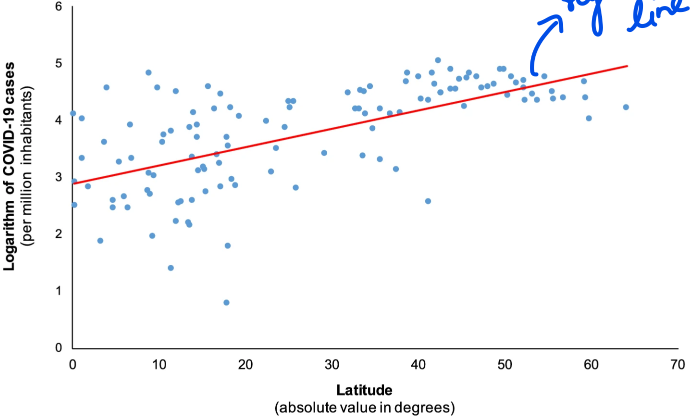


Figure: Scatter plot showing linear trend [1]

Non-linear Regression curve example

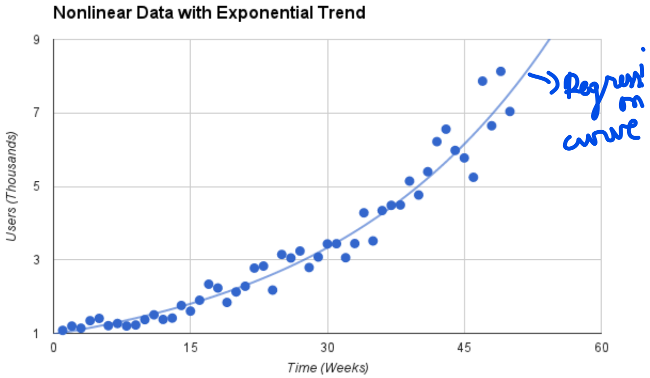


Figure: Data showing number of active users on a website with time [2]

[1] Chen, S., Prettner, K., Kuhn, M. et al. Climate and the spread of COVID-19. Sci Rep 11, 9042 (2021). https://doi.org/10.1038/s41598-021-87692-z

[2] http://sam-koblenski.blogspot.com

Regression model

linear function
or nonlinear
function

A regression model provides a function to describe the relationship between one (or more) independent variables and a dependent variable.

A basic regression model is the 'Least Squares Regression model'.

$m >> n$

Here, the relationship between dependent data points $y_i (i = 1, 2, ...m)$ and independent data points x_i is modelled as

For m data points,

$$\hat{y}(x) = \sum_{i=1}^{n} \alpha_i f_i(x) \tag{1}$$

$\alpha_1 f_1(x) + \alpha_2 f_2(x) + \cdots f_n(x)$

will give m eqns.

plot your regression curve

n parameters

$\hat{y}(x_1) = \sum_{i=1}^{n} \alpha_i f_i(x_1)$

$\hat{y}(x_2) = \cdots f_i(x_2)$

$\vdots$

$\hat{y}(x_m) \cdots$

where

- $\hat{y}(x)$ is an estimation function
- α_i are parameters of estimation function
- $f_i(x)$ are linearly independent basis functions

The parameters are then found by minimising the total squared error E.

$$E = \sum_{i=1}^{m} (\hat{y} - y_i)^2 \tag{2}$$

Substituting (1) in (2) gives

$$E = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_j f_j(x_i) - y_i \right)^2 \tag{3}$$

E is a function of n variables namely $\alpha_j (j = 1, 2, \cdots, n)$.

$$\frac{\partial E}{\partial \alpha_j} = 0$$

Least Squares Regression (contd.)

The solution for n parameters α_j which minimise the total squared error E is given as

$$\beta = pinv(A) * Y \tag{4}$$

Here,

- β is a column vector with n entries α_j
- A is a mxn matrix with entries $A(i,j) = f_j(x_i)$
- $pinv(A)$ is the pseudo-inverse of A
- Y is a column vector with m entries y_i

Least Squares Regression (contd.)

Derivation of (4)**:** Please refer book and lecture recording

Example: Perform a least squares regression for the scatter plot created before using estimation function $\hat{y}(x) = \alpha_1 x + \alpha_2$.

$f_1(x) = x \quad , \quad f_2(x) = 1$

handle

```
openfig('scatter_plot_1.fig') %opens figure scatter_plot_1.fig
a = get(gca,'Children');
x = get(a, 'XData'); %extract x-data points
y = get(a, 'YData'); %extract y-data points
A=[x',ones(size(x'))]; %create the matrix A of basis functions
beta=pinv(A)*y'; %evaluate vector beta containing values of parameters

%plot the regression line
hold on
plot(x,beta(1)*x+beta(2))
```

$$\begin{array}{cc} F_1(x) & \\ \end{array}$$

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & \vdots \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}_{m \times 2} \qquad \beta = \mathrm{pinv}(A) * Y$$

$$\beta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

$\alpha_1 = \mathrm{Beta}(1) \; , \; \alpha_2 = \mathrm{Beta}(2)$

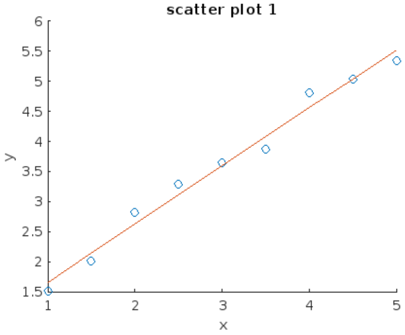


Figure: Regression line

polyfit and polyval functions

When the estimation function $\hat{y}(x)$ is a polynomial, MATLAB's polyfit and polyval functions can be used.

For vectors x and y containing x-data and y-data points respectively,

- p=polyfit(x,y,n) creates a vector p of the coefficients of regression polynomial p(x) of degree n
- polyval(p,x) calculates the values of p(x)

$y = \alpha_1 x + \alpha_2$

$n = 1$

$p = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$ $p = \text{polyfit}(x, y, 1)$

polyfit and polyval functions (contd.)

Example: This code creates the same regression line as was shown before.

```
openfig('scatter_plot_1.fig') %opens figure scatter_plot_1.fig
a = get(gca,'Children');
x = get(a, 'XData'); %extract x-data points
y = get(a, 'YData'); %extract y-data points

p= polyfit(x,y,1); %creates coefficients of regression polynomial of degree 1
Y=polyval(p,x); %evaluates the value of polynomial at x-data points
hold on
plot(x,polyval(p,x)) %plot the regression polynomial
```

polyfit and polyval functions (contd.)

Demo

$\frac{1}{\alpha x+\beta}$

Sometimes, a nonlinear estimation function provides the best fit for a scatter plot. This means we require

non-linear in α_j

$y = \frac{1}{x}$, $\sin x$, $\log x$, 2^x

$$\hat{y}(x) = g(\alpha_1, \alpha_2, \cdots, \alpha_n, x) \tag{5}$$

$\cancel{\hat{y} = \sum_{i=1}^{m} \alpha_j f_j(x_i)}$

where g is some nonlinear function.

In some special cases, a transformation such as

$$\tilde{y}(x) = h(\hat{y}(x)) \tag{6}$$

$\hat{y} = \alpha_1 x^2 + \alpha_2 x + \alpha_3$

$f_1(x) = x^2$, $f_2(x) = x$, $f_3(x) = 1$

can linearise the equation (5) into (1).

h is a clever transformation

$\alpha_1(\quad) + \alpha_2(\quad)$

Example: Consider the estimation function

$\beta = \text{pinv}A \cdot Y$

$$\hat{y}(x) = \alpha_1 e^{\alpha_2 x} \tag{7}$$

$\log \hat{y} = \log(\alpha_1 e^{\alpha_2 x})$

Applying the transformation

$\beta = \begin{bmatrix} \tilde{\alpha}_1 \\ \alpha_2 \end{bmatrix}$

$$\tilde{y}(x) = log(\hat{y}(x)) \tag{8}$$

$= \log \alpha_1 + \log(e^{\alpha_2 x})$

converts (7) into

$$\tilde{y}(x) = \tilde{\alpha_1} + \alpha_2 x \tag{9}$$

$= \log \alpha_1 + \alpha_2 x$

where we define $\tilde{\alpha_1} = log(\alpha_1)$. Now, least squares regression can be applied to equation (9). The parameter α_1 can be found using $\alpha_1 = e^{\tilde{\alpha_1}}$.

$\alpha_1 = e^{\tilde{\alpha}_1}$

$f_1(x) = 1, \quad f_2(x) = x$

End of Lecture 13

Please provide your feedback here