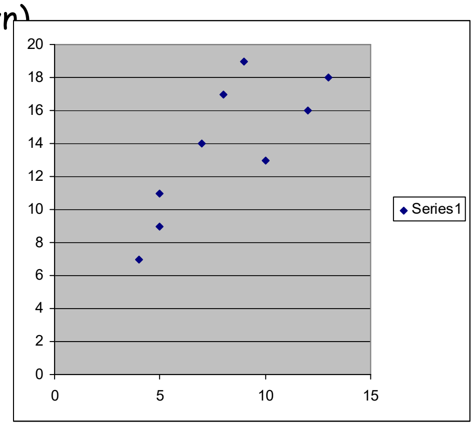
DATA ANALYSIS

Let N set of measurements are made for X and Y

(x1, y1) (x2, y2) ....(xn, yn)

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19



Assume a linear function is to be fitted to this data

Y = A * X + B

Linear Regression

YEST (1) = A * X(1) + B
YEST (2) = A * X(2) + B
YEST (3) = A * X(3) + B
.....
YEST (n) = A * X(n) + B

YEST (1) = 5 A + B
YEST (2) = 12 A + B
YEST (3) = 13 A + B
.....
YEST (9) = 9 A + B

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

Linear Regression

YEST (1) = A * X(1) + B
YEST (2) = A * X(2) + B
YEST (3) = A * X(3) + B
.....
YEST (n) = A * X(n) + B

DIF (1) = Y(1) – YEST (1)
DIF (2) = Y(2) – YEST (2)
DIF (3) = Y(3) – YEST (3)
....
DIF (n) = Y(n) – YEST (n)

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

DIF (1) = 11 – YEST (1)
DIF (2) = 16 – YEST (2)
DIF (3) = 18 – YEST (3)
....
DIF (9) = 19 – YEST (9)

Linear Regression

YEST (1) = A * X(1) + B
YEST (2) = A * X(2) + B
YEST (3) = A * X(3) + B
.....
YEST (n) = A * X(n) + B

DIF (1) = Y(1) – YEST (1)
DIF (2) = Y(2) – YEST (2)
DIF (3) = Y(3) – YEST (3)
....
DIF (n) = Y(n) – YEST (n)

Min Z = $[DIF(1)]^2 + [DIF(2)]^2 + \ldots\ldots + [DIF(n)]^2$

One can substitute the above equalities and solve for A & B

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad \text{veya} \quad b = \frac{\sum y_i - a \sum x_i}{n}$$

$$a = \frac{n \sum y_i x_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Correlation coeficient r, (changes between -1 and 1)

$$r = \frac{n \sum y_i x_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2}\ \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Linear Regression

Minimize Z = $[DIF(1)]^2 + [DIF(2)]^2 + \ldots + [DIF(n)]^2$

YEST (1) = 5 A + B
YEST (2) = 12 A + B
YEST (3) = 13 A + B
.....
YEST (9) = 9 A + B

DIF (1) = 11 – YEST (1)
DIF (2) = 16 – YEST (2)
DIF (3) = 18 – YEST (3)
....
DIF (9) = 19 – YEST (9)

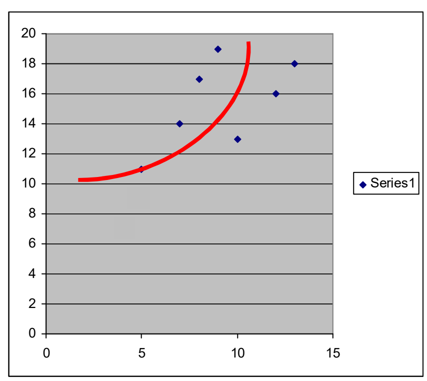
N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

NONLINEAR REGRESSION

Let N set of measurements are made for X and Y

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

If these two measurements were wrong



A non linear relationship between these two variables may be

$$Y = A * X^2 + B * X + C$$

"The best" solution is the one closest to data points.

MULTIPLE Linear REGRESSION

Let N sets of measurements were made for X, Y & Z

(x1, y1, z1) (x2, y2, z2)...

A linear relationship may be seeked as

Z = A1 * X + A2 * Y + B

Similiarly the "best" solution may be find by minimizing the error terms.

N	X	Y	z
1	5	11	14
2	12	16	17
3	13	18	21
4	4	7	9
5	5	9	11
6	10	13	18
7	7	14	13
8	8	17	15
9	9	19	12

$$\text{Min } Z = [DIF(1)]^2 + [DIF(2)]^2 + \ldots + [DIF(n)]^2$$

ZEST (1) = A1 * X(1) + A2 * Y(1) + B
ZEST (2) = A1 * X(2) + A2 * Y(2) + B
ZEST (3) = A1 * X(3) + A2 * Y(3) + B
.....
ZEST (n) = A1 * X(n) + A2 * Y(n) + B

DIF (1) = Z(1) – ZEST (1)
DIF (2) = Z(2) – ZEST (2)
DIF (3) = Z(3) – ZEST (3)
....
DIF (n) = Z(n) – ZEST (n)

TIME SERIES

A series of measuremets are made for an event

Example : Temperature, flow, precipitation mesurements

T	X_t	X_{t-1}
1	5	-
2	12	5
3	13	12
4	4	13
5	5	4
6	10	5
7	7	10
8	8	7
9	9	8

A relationship can be seeked with time

X = A * T + B

However this may not be too appropriate

But instead, a better approach may be

$$X_t = A * X_{t-1} + B$$

These series are called "autoregressive" series..

TIME SERIES

T	X_t	X_{t-1}	X_{t-2}
1	5	-	-
2	12	5	-
3	13	12	5
4	4	13	12
5	5	4	13
6	10	5	4
7	7	10	5
8	8	7	10
9	9	8	7

$$X_t = A1 * X_{t-1} + A2 * X_{t-2} + B$$

Second degree "autoregressive" series.

TIME SERIES

T	X_t	X_{t-1}
1	5	-
2	12	5
3	13	12
4	4	13
5	5	4
6	10	5
7	7	10
8	8	7
9	9	8

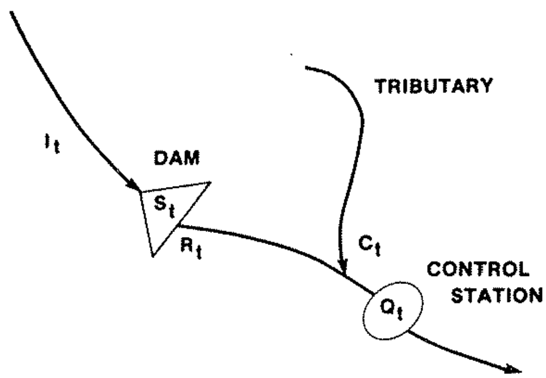
$$X_t = A * X_{t-1} + B$$

IF X_{ave} is the average of X data, one can look for

$$(X_t - X_{ave}) = A * (X_{t-1} - X_{ave}) + B$$

These series are called "moving average" series.

TIME SERIES (Example)



$$Q_t - \alpha_1 R_t - \alpha_2 R_{t-1} = \alpha_3 C_t$$

SAMPLE A relationship in the form of $Y = a X^3 + b$ is assumed to exist between variables X & Y.

Formulate a regression model that will estimate the coefficients of the above equation using the observed data given below, if the "best" fit is defined as the minimization of sum of square of error terms. Explain the solution technique.

Note : Do NOT solve.

N	X	Y
1	2	11
2	3	32
3	1	5
4	2	9

YEST (1) = a * X(1)^3 + b
YEST (2) = a * X(2)^3 + b
YEST (3) = a * X(3)^3 + b
YEST (4) = a * X(n)^3 + b

DIF (1) = Y(1) – YEST (1)
DIF (2) = Y(2) – YEST (2)
DIF (3) = Y(3) – YEST (3)
DIF (4) = Y(n) – YEST (n)

MIN Z = DIF(1)^2+ DIF(2)^2 + DIF(3)^2 + DIF(4)^2

Linear Regression

YEST (1) = A * X(1) + B
YEST (2) = A * X(2) + B
YEST (3) = A * X(3) + B
.....
YEST (n) = A * X(n) + B

DIF (1) = Y(1) – YEST (1)
DIF (2) = Y(2) – YEST (2)
DIF (3) = Y(3) – YEST (3)
....
DIF (n) = Y(n) – YEST (n)

Min Z = $[DIF(1)]^2 + [DIF(2)]^2 + \ldots + [DIF(n)]^2$

One can substitute the above equalities and solve for A & B

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad \text{veya} \quad b = \frac{\sum y_i - a \sum x_i}{n}$$

$$a = \frac{n \sum y_i x_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Correlation coeficient r, (changes between -1 and 1)

$$r = \frac{n \sum y_i x_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2}\ \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Solution via LP

YEST (1) = A * X(1) + B
YEST (2) = A * X(2) + B
YEST (3) = A * X(3) + B
.....
YEST (n) = A * X(n) + B

DIF (1) = Y(1) – YEST (1)
DIF (2) = Y(2) – YEST (2)
DIF (3) = Y(3) – YEST (3)
....
DIF (n) = Y(n) – YEST (n)

Min Z = ABS[DIF (1)] + ABS[DIF (2)] +..+ ABS[DIF(n)]

DIF (n) is not restricted.
For every DIF (n) introduce two new nonnegative variables

DIF (1) = DIFP(1) – DIFN(1)

Solution via LP

YEST (1) = A * X(1) + B	DIFP (1) – DIFN (1) = Y(1) – YEST (1)
YEST (2) = A * X(2) + B	DIFP (2) – DIFN (2) = Y(2) – YEST (2)
YEST (3) = A * X(3) + B	DIFP (3) – DIFN (3) = Y(3) – YEST (3)
.....	....
YEST (n) = A * X(n) + B	DIFP (n) – DIFN (n) = Y(1) – YEST (n)

MIN Z= DIFP(1)+DIFN(1)+DIFP(2)+DIFN(2)..+DIFP(n)+DIFN(n)

Solution via LP

MIN Z= DIFP(1) + DIFN(1) + DIFP(2) + DIFN(2)
.....+ DIFP(9) + DIFN(9)

YEST (1) = 5 A + B
YEST (2) = 12 A + B
YEST (3) = 13 A + B
.....
YEST (9) = 9 A + B

DIFP(1) – DIFN(1) = 11 – YEST (1)
DIFP(2) – DIFN(2) = 16 – YEST (2)
DIFP(3) – DIFN (3) = 18 – YEST (3)
....
DIFP(9) – DIFN (9) = 19 – YEST (9)

All variables non-negative

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

If A and/or B can take negative values

Replace A & B with difference of
two non-negative variables

A = AP – AN

B = BP - BN

Solution via LP

MIN Z= DIFP(1) + DIFN(1) + DIFP(2) + DIFN(2)
.....+ DIFP(9) + DIFN(9)

YEST (1) = 5 AP – 5 AN + BP - BN
YEST (2) = 12 AP – 12 AN + BP - BN
YEST (3) = 13 AP -13 AN + BP - BN
.....
YEST (9) = 9 AP – 9 AN + BP - BN

DIFP(1) – DIFN(1) = 11 – YEST (1)
DIFP(2) – DIFN(2) = 16 – YEST (2)
DIFP(3) – DIFN (3) = 18 – YEST (3)
....
DIFP(9) – DIFN (9) = 19 – YEST (9)

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19

All variables non-negative

Minimization of Maximum Deviation

MIN MAX Problems

MIN Z= MAXDIF

DIFP (i) $\leq$ MAXDIF for all i
DIFN(i) $\leq$ MAXDIF for all i

YEST (1) = 5 AP – 5 AN + BP - BN
YEST (2) = 12 AP – 12 AN + BP - BN
YEST (3) = 13 AP -13 AN + BP - BN
.....
YEST (9) = 9 AP – 9 AN + BP - BN

DIFP(1) – DIFN(1) = 11 – YEST (1)
DIFP(2) – DIFN(2) = 16 – YEST (2)
DIFP(3) – DIFN (3) = 18 – YEST (3)
....
DIFP(9) – DIFN (9) = 19 – YEST (9)

All variables non-negative

N	X	Y
1	5	11
2	12	16
3	13	18
4	4	7
5	5	9
6	10	13
7	7	14
8	8	17
9	9	19