INDIAN INSTITUTE OF SCIENCE

STOCHASTIC HYDROLOGY

Lecture -22

Course Instructor : Prof. P. P. MUJUMDAR

Department of Civil Engg., IISc.

Summary of the previous lecture

- Case study -5: Sakleshpur rainfall data
  - Plots of Time series, Correlogram, Partial Autocorrelation function and Power spectrum
  - Candidate ARMA models
  - Validation tests
- Summary of all the case studies

MARKOV CHAINS

Markov Chains

- A Markov chain is a stochastic process with the property that value of process X_t at time t depends on its value at time t-1 and not on the sequence of other values (X_{t-2} , X_{t-3},……. X_0) that the process passed through in arriving at X_{t-1}.

$$P\left[X_t / X_{t-1}, X_{t-2}, \ldots.. X_0\right] = P\left[X_t / X_{t-1}\right]$$

Single step Markov chain

Markov Chains

$$P\left[X_t = a_j / X_{t-1} = a_i\right]$$

- This conditional probability gives the probability that at time t, the process will be in state 'j', given that the process was in state 'i' at time t-1.
- The conditional probability is independent of the states occupied prior to t-1.
- For example, if X_{t-1} is a dry day, we would be interested in the probability that X_t is a dry day or a wet day.
- This probability is commonly called as transition probability

Markov Chains

$$P\left[X_t = a_j / X_{t-1} = a_i\right] = P_{ij}^t$$

- Usually written as P_{ij}^t indicating the transition of the process from state a_i at time t-1 to a_j at time t.
- If P^t_{ij} is independent of time, then the Markov chain is said to be homogeneous.

i.e., $P_{ij}^t = P_{ij}^{t+\tau}$ $\forall$ t and τ

The transition probabilities remain same across time.

Markov Chains

Transition Probability Matrix(TPM):

$$
\begin{array}{l}
\qquad\qquad t \rightarrow 1 \qquad 2 \qquad 3 \qquad . \qquad . \qquad m\\
\qquad t-1\\
\qquad \downarrow
\end{array}
$$

$$
P = \begin{array}{c} 1\\ 2\\ 3\\ .\\ .\\ m \end{array}
\begin{bmatrix}
P_{11} & P_{12} & P_{13} & . & . & P_{1m}\\
P_{21} & P_{22} & P_{23} & . & . & P_{2m}\\
P_{31} & & & & & \\
. & & & & & \\
. & & & & & \\
P_{m1} & P_{m2} & & & & P_{mm}
\end{bmatrix}_{m \times m}
$$

Markov Chains

$$\sum_{j=1}^{m} P_{ij} = 1 \quad \forall \text{ i}$$

- Elements in any row of TPM sum to unity
- TPM can be estimated from observed data by enumerating the number of times the observed data went from state 'i' to 'j'

$$\hat{P}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{m} n_{ij}}$$

- $P_j^{(n)}$ is the probability of being in state 'j' in time step 'n'.

Markov Chains

- $p_j^{(0)}$ is the probability of being in state 'j' in period t = 0.

$$p^{(0)} = \begin{bmatrix} p_1^{(0)} & p_2^{(0)} & . & . & p_m^{(0)} \end{bmatrix}_{1\times m}$$ …. Probability vector at time 0

$$p^{(n)} = \begin{bmatrix} p_1^{(n)} & p_2^{(n)} & . & . & p_m^{(n)} \end{bmatrix}_{1\times m}$$ …. Probability vector at time 'n'

- If $p^{(0)}$ is given and TPM is given

$$p^{(1)} = p^{(0)} \times P$$

Markov Chains

$$p^{(1)} = \begin{bmatrix} p_1^{(0)} & p_2^{(0)} & . & . & p_m^{(0)} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} & . & . & P_{1m} \\ P_{21} & P_{22} & P_{23} & . & . & P_{2m} \\ P_{31} & & & & & \\ . & & & & & \\ P_{m1} & P_{m2} & & & & P_{mm} \end{bmatrix}$$

$$\Big[p_1^{(0)} P_{11} + p_2^{(0)} P_{21} + ... + p_m^{(0)} P_{m1}$$ …. Probability of going to state 1

$$p_1^{(0)} P_{12} + p_2^{(0)} P_{22} + ... + p_m^{(0)} P_{m2}$$ …. Probability of going to state 2

$$\Big]$$

Markov Chains

Therefore

$$p^{(1)} = \begin{bmatrix} p_1^{(1)} & p_2^{(1)} & \cdot & \cdot & p_m^{(1)} \end{bmatrix}_{1\times m}$$

$$\begin{aligned} p^{(2)} &= p^{(1)} \times P \\ &= p^{(0)} \times P \times P \\ &= p^{(0)} \times P^2 \end{aligned}$$

In general,

$$p^{(n)} = p^{(0)} \times P^n$$

Markov Chains

- As the process advances in time, $p_j^{(n)}$ becomes less dependent on $p^{(0)}$
- The probability of being in state 'j' after a large number of time steps becomes independent of the initial state of the process.
- The process reaches a steady state at large n

$$p = p \times P$$

- As the process reaches steady state, the probability vector remains constant

Example – 1

Consider the TPM for a 2-state first order homogeneous Markov chain as

$$TPM = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

State 1 is a non-rainy day and state 2 is a rainy day

Obtain the

1. probability that day 1 is a non-rainy day given that day 0 is a rainy day
2. probability that day 2 is a rainy day given that day 0 is a non-rainy day
3. probability that day 100 is a rainy day given that day 0 is a non-rainy day

Example – 1 (contd.)

1. probability that day 1 is a non-rainy day given that day 0 is a rainy day

$$TPM = \begin{array}{c c} & \begin{array}{cc} \text{No rain} & \text{rain} \end{array} \\ \begin{array}{r} \text{No rain} \\ \text{rain} \end{array} & \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

The probability is 0.4

2. probability that day 2 is a rainy day given that day 0 is a non-rainy day

$$p^{(2)} = p^{(1)} \times P$$

$p^{(1)}$, in this case is [0.7 0.3] because it is given that day 0 is a non-rainy day.

Example – 1 (contd.)

$$p^{(2)} = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.61 & 0.39 \end{bmatrix}$$

The probability is 0.39

3. probability that day 100 is a rainy day given that day 0 is a non-rainy day

$$p^{(n)} = p^{(0)} \times P^n$$

Example – 1 (contd.)

$$P^2 = P \times P$$

$$= \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix}$$

$$P^4 = P^2 \times P^2 = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}$$

$$P^8 = P^4 \times P^4 = \begin{bmatrix} 0.5715 & 0.4285 \\ 0.5714 & 0.4286 \end{bmatrix}$$

$$P^{16} = P^8 \times P^8 = \begin{bmatrix} 0.5714 & 0.4286 \\ 0.5714 & 0.4286 \end{bmatrix}$$

Example – 1 (contd.)

Steady state probability

$$p = \begin{bmatrix} 0.5714 & 0.4286 \end{bmatrix}$$

For steady state,

$$p = p \times P^n$$

$$= \begin{bmatrix} 0.5714 & 0.4286 \end{bmatrix} \begin{bmatrix} 0.5714 & 0.4286 \\ 0.5714 & 0.4286 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5714 & 0.4286 \end{bmatrix}$$

Markov Chains

Difficulties in using Markov chains in hydrology

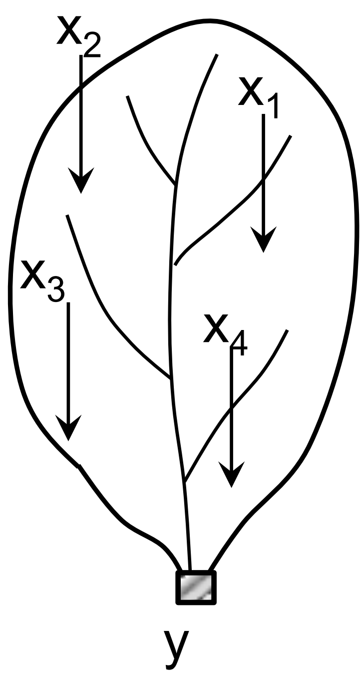
- Determining the number of states to use.
- Determining the intervals of the variable under study to associate with each state.
- Assigning a number to the magnitude of an event once the state is determined.
- Estimating the large number of parameters involved in even a moderate size Markov chain model.
- Handling situations where some transitions are dependent on several previous time periods while others are dependent on only one prior time period.

Ref: Statistical methods in Hydrology by C.T.Haan, Iowa state university press

MULTIPLE LINEAR REGRESSION

Multiple Linear Regression

- A variable (y) is dependent on many other independent variables, x_1, x_2, x_3, x_4 and so on.
- For example, the runoff from the water shed depends on many factors like rainfall, slope of catchment, area of catchment, moisture characteristics etc.



- Any model for predicting runoff should contain all these variables

Multiple Linear Regression

A general linear model of the form is

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots.. + \beta_p x_p$$

y is dependent variable,

$x_1, x_2, x_3, \ldots\ldots, x_p$ are independent variables and

$\beta_1, \beta_2, \beta_3, \ldots\ldots, \beta_p$ are unknown parameters

- 'n' observations are required on y with the corresponding 'n' observations on each of the 'p' independent variables.

Multiple Linear Regression

- 'n' equations are written for each observation as

$$y_1 = \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots\ldots.. + \beta_p x_{1,p}$$
$$y_2 = \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots\ldots.. + \beta_p x_{2,p}$$
$$.$$
$$.$$
$$y_n = \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots\ldots.. + \beta_p x_{n,p}$$

- Solving 'n' equations for obtaining the 'p' parameters.
- 'n' must ne equal to or greater than 'p', in practice 'n' must be at least 3 to 4 times large as 'p'.

Multiple Linear Regression

- If y_i is the i^{th} observation on y and $y_{i,j}$ is the i^{th} observation on the j^{th} independent variable, the generalized form of the equations can be written as

$$y_i = \sum_{j=1}^{p} \beta_j x_{i,j}$$

- The equation can be written in matrix notation as

$$Y_{(n\times 1)} = X_{(n\times p)} \times \mathrm{B}_{(p\times 1)}$$

Multiple Linear Regression

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & . & . & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & . & . & x_{2,p} \\ x_{3,1} & & & & & \\ . & & & & & \\ . & & & & & \\ x_{n,1} & x_{n,1} & & & & x_{n,p} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ \beta_p \end{bmatrix}$$

Y is an nx1 vector of observations on the dependent variable, X is an nxp matrix with n observations on each p independent variables, Β is a px1 vector of unknown parameters.

Multiple Linear Regression

- To have an intercept term, it is assumed that $x_{i,1}=1$ for i=1 to n, therefore β_1 is the intercept
- The linear regression model discussed previously (y=ax+b) is a special form of the multiple regression model with $x_{i,1}=1$, $x_{i,2}=x$, $\beta_1=a$ and $\beta_2=b$
- The same methodology used for solving the parameters in simple linear regression is adopted here. i.e., the unknown parameters are estimated by minimizing the sum of square errors (e_i) where

$$e_i = y_i - \hat{y}_i$$

Multiple Linear Regression

In matrix notation,

$$\sum e_i^2 = E^{'}E$$
$$= \left(Y - X\hat{\mathrm{B}}\right)^{'}\left(Y - X\hat{\mathrm{B}}\right)$$

The equation is differentiated with respect to $\hat{\mathrm{B}}$ and equated to zero

$$0 = -2X^{'}\left(Y - X\hat{\mathrm{B}}\right)$$
$$X^{'}Y = X^{'}X\hat{\mathrm{B}}$$

Multiplying with $\left(X^{'}X\right)^{-1}$ on both the sides,

Multiple Linear Regression

$$X'Y\left(X'X\right)^{-1} = X'X\left(X'X\right)^{-1}\hat{\mathrm{B}}$$

$$X'Y\left(X'X\right)^{-1} = \hat{\mathrm{B}}$$

or

$$\hat{\mathrm{B}} = \left(X'X\right)^{-1}X'Y$$

- $\left(X'X\right)$ is a pxp matrix and rank is p
- $\left(X'X\right)^{-1}$ is made up of sum of squares and cross products of the independent variables and
- This matrix plays an important role in estimating $\hat{\mathrm{B}}$

Multiple Linear Regression

- Suppose if no. of regression coefficients are 3, then matrix $(X'X)$ is as follows

$$(X'X) = \begin{bmatrix} \sum_{i=1}^{n} x_{i,1}^2 & \sum_{i=1}^{n} x_{i,2}x_{i,1} & \sum_{i=1}^{n} x_{i,3}x_{i,1} \\ \sum_{i=1}^{n} x_{i,1}x_{i,2} & \sum_{i=1}^{n} x_{i,2}^2 & \sum_{i=1}^{n} x_{i,3}x_{i,2} \\ \sum_{i=1}^{n} x_{i,1}x_{i,3} & \sum_{i=1}^{n} x_{i,2}x_{i,3} & \sum_{i=1}^{n} x_{i,3}^2 \end{bmatrix}$$

Multiple Linear Regression

- Assumption: An independent variable cannot be a perfect linear function of any other independent variable
- For the rank of $\left(X'X\right)$ to be p, an independent variable cannot be linearly dependent on any linear function of the remaining independent variables.
- If there is a linear dependence, the calculation of $\left(X'X\right)^{-1}$ may involve round off errors or loss of significance leading to non-logical estimates for $\hat{\mathrm{B}}$

Multiple Linear Regression

- A multiple coefficient of determination, R^2 (as in case of simple linear regression) is defined as

$$R^2 = \frac{Sum\ of\ squares\ due\ to\ regression}{Sum\ of\ squares\ about\ the\ mean}$$

$$= \frac{\mathrm{B}'X'Y - n\overline{y}^2}{Y'Y - n\overline{y}^2}$$

Example – 1

In a watershed, the mean annual flood (Q) is considered to be dependent on area of watershed (A) and rainfall(R). The table gives the observations for 12 years. Obtain regression coefficients and R^2 value.

Q in thousand cms	0.44	0.24	2.41	2.97	0.7	0.11	0.05	0.51	0.25	0.23	0.1	0.054
A in thousand hectares	324	226	1474	2142	420	45	38	363	77	84	46	38
Rainfall in cm	43	53	48	50	43	61	81	68	74	71	71	69

Example – 1 (Contd.)

The regression model is as follows

$$Q = \beta_1 + \beta_2 A + \beta_3 R$$

Where Q is the mean flood in thousand m^2/sec,

A is the watershed area in thousand hectares and

R is the average annual daily rainfall in mm

This is represented in matrix form as

$$Y_{(12\times1)} = X_{(12\times3)} \times \mathrm{B}_{(3\times1)}$$

Example – 1 (Contd.)

To obtain coefficients this equation is to be solved

$$\begin{bmatrix} 0.44 \\ 0.24 \\ 2.41 \\ 2.97 \\ 0.7 \\ 0.11 \\ 0.05 \\ 0.51 \\ 0.25 \\ 0.23 \\ 0.1 \\ 0.054 \end{bmatrix} = \begin{bmatrix} 1 & 324 & 43 \\ 1 & 226 & 53 \\ 1 & 1474 & 48 \\ 1 & 2142 & 50 \\ 1 & 420 & 43 \\ 1 & 45 & 61 \\ 1 & 38 & 81 \\ 1 & 363 & 68 \\ 1 & 77 & 74 \\ 1 & 84 & 71 \\ 1 & 46 & 71 \\ 1 & 38 & 69 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Example – 1 (Contd.)

The coefficients are obtained from

$$\hat{\mathrm{B}} = \left(X^{'} X \right)^{-1} X^{'} Y$$

$$\left(X^{'} X \right) = \begin{bmatrix} \sum_{i=1}^{n} x_{i,1}^{2} & \sum_{i=1}^{n} x_{i,2} x_{i,1} & \sum_{i=1}^{n} x_{i,3} x_{i,1} \\ \sum_{i=1}^{n} x_{i,1} x_{i,2} & \sum_{i=1}^{n} x_{i,2}^{2} & \sum_{i=1}^{n} x_{i,3} x_{i,2} \\ \sum_{i=1}^{n} x_{i,1} x_{i,3} & \sum_{i=1}^{n} x_{i,2} x_{i,3} & \sum_{i=1}^{n} x_{i,3}^{2} \end{bmatrix}$$

Example – 1 (Contd.)

$$\left(X^{'}X\right) = \begin{bmatrix} 12 & 5277 & 732 \\ 5277 & 7245075 & 269879 \\ 732 & 269879 & 46536 \end{bmatrix}$$

The inverse of this matrix is

$$\left(X^{'}X\right)^{-1} = \begin{bmatrix} 3.35 & -6.1\times10^{-4} & -0.05 \\ -6.1\times10^{-4} & 2.9\times10^{-7} & 7.9\times10^{-6} \\ -0.05 & 7.9\times10^{-6} & 7.5\times10^{-4} \end{bmatrix}$$

Example – 1 (Contd.)

$$\left(X'Y\right) = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i,2} y_i \\ \sum_{i=1}^{n} x_{i,3} y_i \end{bmatrix} = \begin{bmatrix} 8.06 \\ 10642 \\ 417 \end{bmatrix}$$

Example – 1 (Contd.)

$$\hat{\mathrm{B}} = \left(X^{'}X\right)^{-1} X^{'}Y$$

$$= \begin{bmatrix} 3.35 & -6.1\times10^{-4} & -0.05 \\ -6.1\times10^{-4} & 2.9\times10^{-7} & 7.9\times10^{-6} \\ -0.05 & 7.9\times10^{-6} & 7.5\times10^{-4} \end{bmatrix} \times \begin{bmatrix} 8.06 \\ 10642 \\ 417 \end{bmatrix}$$

$$= \begin{bmatrix} 0.0351 \\ 0.0014 \\ 5.0135\times10^{-5} \end{bmatrix}$$

Example – 1 (Contd.)

Therefore the regression equation is as follows

$$Q = 0.0351 + 0.0014A + 5.0135*10^{-5}R$$

From this equation, the estimated Q and the corresponding errors are tabulated

Example – 1 (Contd.)

Q	A	I	$\hat{Q}$	e
0.44	324	43	0.49	-0.05
0.24	226	53	0.35	-0.11
2.41	1474	48	2.10	0.31
2.97	2142	50	3.04	-0.07
0.7	420	43	0.63	0.07
0.11	45	61	0.10	0.01
0.05	38	81	0.09	-0.04
0.51	363	68	0.55	-0.04
0.25	77	74	0.15	0.10
0.23	84	71	0.16	0.07
0.1	46	71	0.10	0.00
0.054	38	69	0.09	-0.04

Example – 1 (Contd.)

Multiple coefficient of determination, R^2 :

$$R^2 = \frac{\mathrm{B}'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

$$= \frac{15.64 - 5.42}{15.77 - 5.42}$$

$$= 0.99$$