Univariate Models
4.1 Use of Nonlinear Models in Sugarcane Yield Forecasting:

Nonlinear model plays a very important role in understanding of complex interrelationship among variables. A ‘nonlinear model’ is one in which at least one of the parameters appears nonlinear.

Those models, which describe growth behaviour over time, are applied in many fields. In the area of population biology, growth occurs in plants, animals, organisms etc. The type of model needed in a specific situation depends on type of growth that occurs. Growth models are mechanistic in nature, rather than empirical. In the mechanistic model, parameters have meaningful biological interpretation; whereas empirical model is just like a ‘black-box’ where output is obtained by just giving some input values. A mechanistic model usually arises as a result of making assumptions about the type of growth, writing down differential or difference equations that represent these assumptions, and then solving these equations to obtain a growth model. The utility of such models is that, they help us to gain insight into underlying mechanism of the system and they are of immense help in efficient management. Some well known nonlinear growth models are:

(i) **Monomolecular Model:** This model describes the progress of a growth situation in which it is believed that the rate of growth at any time is proportional to the resources yet to be achieved, i.e.

\[
\frac{dy}{dt} = r(K-y), \quad \text{(4.1)}
\]
Integrating Equation (4.1),

\[ y(t) = K - (K-b) \exp(-r t) \] ....(4.2)

Where, \( y(t) \) denotes sugarcane production during the time ‘t’, 
‘r’ denotes the intrinsic growth rate, 
‘b’ denotes the different functions of the initial value \( y(0) \), 
‘K’ denotes the carrying capacity of the model

(ii) Logistic Model: This model is represented by the differential equation (4.3)

\[ \frac{dy}{dt} = r \, y \left(1 - \frac{y}{K}\right) \] ...(4.3)

Integrating Equation 4.3,

\[ y(t) = \frac{K}{1 + \left(K/b - 1\right) \exp(-r t)} \] ...(4.4)

The graph of \( y(t) \) versus ‘t’ is elongated S-shaped and curve is symmetrical about its point of inflexion.

(iii) Gompertz Model: This is another model having a sigmoid type of behaviour and found to be quite useful in biological work. Unlike the logistic model, this is not symmetric about its point of inflexion. The differential equation for this model is

\[ \frac{dy}{dt} = r \, y \ln\left(\frac{K}{y}\right) \] ...(4.5)

Integrating of this equation yields

\[ y(t) = K \exp\left[\ln\left(b/K\right) \exp(-r t)\right] \] ...(4.6)
4.1.1 Fitting of Nonlinear Models:

The above models are deterministic. As these models are unrealistic, they are replaced by statistical models with the addition of error term on right hand side and making appropriate assumptions about them which results in a ‘Nonlinear statistical model’. As in linear regression, parameter estimates are obtained by the ‘Method of least squares’. Minimization of residual sum of squares yield normal equations which are nonlinear in parameters. Since it is not possible to solve nonlinear equations exactly, the next alternative is to obtain approximate analytic solutions by employing iterative procedures. Three main methods considered are:

(i) Linearization (or Taylor Series) method
(ii) Steepest Descent method
(iii) Levenberg-Marquardt’s method

The details of these methods along with their merits and demerits are given in Draper and Smith (1998). The linearization method uses the results of linear least square theory in succession of stages. Neither this method nor the steepest descent method is ideal. The latter method is able to converge on true parameter values even though initial trial values are far from true parameter values, but this convergence tends to be very slow at later stages of iterative process. On the other hand, linearization method will converge very rapidly provided the vicinity of true parameter values has been reached. but if initial trial values are too far removed, convergence does not occur at all.

The most widely used method for computing nonlinear least square estimators is Levenberg-Marquardt’s method. Levenberg-Marquardt method is a compromise between other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not ‘slow down’ at latter part of iterative process. The method has been discussed along with some details. Consider the model
where \( x_i \) and \( y_i \) are respectively the \( i^{th} \) observations of explanatory and response variables, \( \theta_0 = (\theta_1, \theta_2, \ldots, \theta_p) \) are parameters, and error terms \( \varepsilon_i \) are independent and follow \( N(0, \sigma^2) \). The residual sum of squares is

\[
S(\theta) = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2
\]

\[(4.8)\]

Let \( \theta_0 = (\theta_1, \theta_2, \ldots, \theta_p) \) be the vector of initial parameter values. Then the algorithm for obtaining successive estimates is essentially given as

\[
(H + \tau I)(\theta_0 - \theta_1) = g,
\]

\[(4.9)\]

where, \( g = \left. \frac{\partial S(\theta)}{\partial \theta} \right|_{\theta = \theta_0} \), \( H = \left. \frac{\partial^2 S(\theta)}{\partial \theta \partial \theta'} \right|_{\theta = \theta_0} \), \( I \) is the identity matrix and \( \tau \) is a suitable multiplier.

**Choice of Initial Values:** All the procedures for nonlinear estimation require initial values of parameters and choice of good initial values are very critical. There is no standard procedure for getting initial estimates. The most obvious method for making initial guess is to make use of prior information. Estimates calculated from previous experiments, known values of similar systems, values computed from theoretical considerations all these form ideal initial guess. Some other methods are:

(i) **Linearization.** After ignoring the error term, the form of the model is checked to find if it could be transformed into a linear form by means of some transformation. In such cases linear regression is used to obtain initial values.
(ii) **Solving a system of equations.** If there are \( p \) parameters, \( p \) sets are substituted for observations into the model ignoring the error. These equations are solved for estimating the parameters. Widely separated \( x \) values are used.

(iii) **Using properties of the model.** The response function having the limits of \( x_i \) as zero or infinity is considered. It is substituted for observations that match the conditions in scale and context of the problem. And the resulting equations are solved.

(iv) **Graphical method.** Visual estimates are obtained through plotting the data graphically.

### 4.1.2 Goodness of Fit:

This is generally assessed by the coefficient of determination, \( R^2 \). Frequently mistakes occur when the fit of a linear and non-linear model are compared by using same \( R^2 \) expression for different variables. A power model or exponential model is linearised by using logarithmic transformation and then fitted to data using ordinary least square method. The \( R^2 \)-value is then calculated using data points \((\log_e y_i, \log_e \hat{y}_i)\). The \( R^2 \) is generally interpreted as a measure of goodness for fit of the original non-linear model, which is incorrect.

The coefficient of determination \( R^2 \) which is quite appropriate even for non-linear models is given as

\[
R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}
\]  \hspace{1cm} (4.10)

The other measures of goodness of fit which has been used to judge the adequacy of model developed are:
Mean Absolute Error (MAE)  \[= \frac{\sum |y_i - \hat{y}_i|}{n} \tag{4.11} \]

Mean Squared Error (MSE)  \[= \frac{\sum (y_i - \hat{y}_i)^2}{n-p} \tag{4.12} \]

Root Mean Square Error (RMSE)  \[= \left( \frac{\sum (y_i - \hat{y}_i)^2}{n} \right)^{1/2} \tag{4.13} \]

where \( n \) denotes total number of observed values and \( p \) denotes the number of model parameters.

4.1.3 Examination of Residuals:

Further the analysis of residuals for each model has been carried out as a measure for goodness of fit. The main assumptions made in the models are

(i) Errors are random.

(ii) Errors are normally distributed.

The randomness of residuals is tested using Run test. Further to test whether residuals are normally distributed Shapiro-Wilks test is performed. The procedures for carrying out Run test and Shapiro-Wilks test is explained further.

Test for randomness (Run test): The test is based on number of runs (\( r \)), where a run is defined as sequence of symbols of one kind separated by symbols of another kind. The residuals are replaced by '+' or '-' accordingly they are positive or negative. The Null hypothesis \( H_0 \) is given as

\( H_0 \): The residuals are random against

\( H_1 \): Residuals are not random.

Let 'm' be the number of positive signs

'n' be the number of negative signs and

'\( r \)' be the number of runs

The mean is given as  
\[ \mu_r = \frac{2mn}{m + n} + 1 \]

Variance is given as  
\[ \sigma^2 = \frac{2mn(2mn-m-n)(m+n)^2(m+n-1)}{(m+n)^2} \]
For large sample, the test statistic is
\[ Z = \frac{r + h - \mu}{\sigma} \sim N(0, 1) \]

Where
\[ H = \begin{cases} 
0.5 & \text{if } r < \mu \\
-0.5 & \text{if } r > \mu 
\end{cases} \]

\( H_0 \) is rejected at level \( \alpha \) if \( |z| > Z_{\alpha/2} \) where
\[ Z_{\alpha} = \Phi^{-1}(\frac{1}{2}) \]

Test for Normality (Shapiro-Wilk test): The Null hypothesis:
\( H_0: \) The residuals are normally distributed against
\( H_1: \) The residuals are not normally distributed

The test statistic is given as
\[ W = S^2/b \]

Where
\[ S^2 = \sum a(k)x_{(n+1-k)x(k)} \]
\[ b = \sum(x_i - \bar{x})^2 \]

In the above parameter ‘k’ takes values
\[ k = \{1, 2, \ldots, n/2 \} \text{ when } n \text{ is even} \]
\[ \{1, 2, \ldots, (n-1)/2 \} \text{ when } n \text{ is odd} \]

and \( x_{(k)} \) is the \( k^{th} \) order statistic of the set of residuals. \( H_0 \) is rejected at level \( \alpha \) if \( W \) is less than the tabulated value.

### 4.1.4 Computation of Compound Growth Rates:

Compound growth rates are widely employed in the field of agriculture as these have important policy implications. The response variables particularly in area of agriculture, do not behave in pattern required for valid computation of compound growth
rates, resulting in faulty conclusions. A more appropriate procedure is to use non linear growth models for computation of compound growth rates.

4.1.5 Existing Method and its Drawbacks:

The Malthus model is given as
\[ y_t = y_0 (1 + r)^t \] ...(4.14)
where \( y_t \) denotes observation (agricultural production, productivity or area) at time \( t \) and \( r \) is compound growth rate.

The usual method is to assume a multiplicative error term \( \varepsilon \) in equation (4.14) so that the model is linearised by means of logarithmic transformation, giving equation (4.15)
\[ \ln(y_t) = A + Bt + \varepsilon \] ...(4.15)
where, \( A = \ln(y_0) \), and \( B = \ln(1+r) \). The parameters are estimated by 'method of least squares' and goodness of fit is assessed through coefficient of determination \( R^2 \). Finally compound growth rates are estimated as
\[ \hat{r} = \exp(\hat{B}) - 1 \] ...(4.16)

According to Prajneshu and Chandran (2005) the above method has following difficulties
(i) Under Malthus model the response variable \( y_t \) tends to infinity as \( t \) tends to infinity, which cannot happen in reality.
(ii) the assumption of multiplicative error term does not hold good in reality.
(iii) the third drawback is that \( R^2 \), which is a measure of goodness of fit and standard error, is found only for new linearised version of model and not to original model.
4.1.6 Alternative Method:

To overcome the above said drawbacks, following methodology suggested by Prajneshu and Chandran (2005), is adopted. The exponential model given through Equation (4.14) is replaced by more realistic growth models like Logistic, Gompertz and Monomolecular. A special feature of these models is that they are ‘Mechanistic’ in nature where the parameters have specific biological interpretation. A mechanistic model usually arises as a result of making assumptions about the type of growth, writing down differential or difference equations which represent these assumptions, and then solving Equations (4.2), (4.4) and (4.6) to obtain a growth model. Non linear estimation procedures, like ‘Levenberg-Marquardt’ algorithm (Refer Section 4.1.2) is employed for fitting the models. Finally, goodness of fit of models is examined as discussed in Section 4.1.3.

After successfully identifying the best model for a data set, last step is to compute compound growth rate which is given as

\[
d \left[ \log(y) \right]/dt, \text{ which is rewritten as } \\
y^{-1} \left( dy/dt \right)
\]

For Monomolecular, Logistic and Gompertz models, compound growth rates pertaining to the period \((t_i, t_{i+1})\), \(i = 0, 1, \ldots, n-1\), where \(n\) denotes number of data points is given as,

\[
R_i^{M} = r \left[ K/y(t) - 1 \right] \quad \text{ and } \quad \ldots (4.17)
\]

\[
R_i^{L} = r \left[ 1 - y(t)/K \right] \quad \text{ and } \quad \ldots (4.18)
\]

\[
R_i^{G} = r \log \left[ K/y(t) \right] \quad \ldots (4.19)
\]

The required compound growth rate over a given period of time is obtained through computing arithmetic mean.
4.2 Use of Double Exponential Smoothing for Yield Prediction:

Performance of double exponential smoothing model for Coimbatore district has been compared with ARIMA models. Sugarcane yield forecast for the district has been predicted and the results are presented in Chapter – 6. The model used for the purpose is described in the following equations.

\[ L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \]  \hspace{1cm} (4.20)

\[ b_t = \beta (L_t - L_{t-1}) + (1 - \beta)b_{t-1} \]  \hspace{1cm} (4.21)

\[ F_{t+m} = L_t + b_t m \]  \hspace{1cm} (4.22)

where, \( L_t \) is level of the series at time \( t \)

\( b_t \) is slope of the series at time \( t \)

\( \alpha \) and \( \beta \) (=0.1, 0.2, ..., 0.9) are the smoothing and trend parameters.

4.3 ARIMA Model for Sugarcane Yield Forecast:

ARIMA models were popularized by Box and Jenkins (1970) in the early 1970’s and their names have frequently been used synonymously with general ARIMA models applied to time series analysis and forecasting. Box and Jenkins effectively put together in a comprehensive manner the relevant information required to understand and use univariate time series ARIMA models. The basis of the Box-Jenkins approach to modeling time series is summarized in Figure 4.3.1 and consists of three phases: identification, estimation and testing and application.
Phase I  
Data preparation  
- Transform data to stabilize variance  
- Difference data to obtain stationary series  

Model selection  
- Examine data, ACF and PACF to identify potential models  

Phase II  
Estimation and testing  

Estimation  
- Estimate parameters in potential models  
- Select best model using suitable criterion  

Diagnostics  
- Check ACF / PACF of residuals  
- Do port manteau test of residuals  
- Are the residuals white noise?  

Yes  

Forecasting  
- Use model to forecast  

No  

Figure 4.3.1 Schematic representation of the Box-Jenkins methodology for time series modeling
4.3.1. Description of the Model:

In general, an ARIMA model is characterized by the notation ARIMA \((p, d, q)\) where \(p, d, q\) denote orders of auto-regression, integration (differencing) and moving average respectively. In ARIMA, time series is a linear function of past actual values and random shocks. A stationary ARIMA \((p, q)\) process is defined by the equation

\[
Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t + \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q} \quad \ldots (4.23)
\]

where,

\[
Y_t = \text{response (dependant) variable at time } t.
\]

\[
Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p} = \text{response (dependant) variable at time lags } t-1, t-2, \ldots, t-p \text{ respectively; these } Y \text{’s are independent variables.}
\]

\[
\Phi_1, \Phi_2, \ldots, \Phi_p = \text{coefficients to be estimated.}
\]

\[
\varepsilon_t = \text{error term at time } t \text{ that represents the effects of variables not explained by the model; assumptions about the error term are same as those for standard regression model.}
\]

\[
\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q} = \text{error term that represents the effect of variables not explained by the model. The assumptions about the error term are same as those for standard regression model.}
\]

\[
\omega_1, \omega_2, \ldots, \omega_q = \text{coefficients to be estimated.}
\]

4.3.2 ARIMA Model Building:

**Identification:** The foremost step in process of modeling is to check for stationarity of the series, as estimation procedures are available only for stationary series. There are two kinds of, viz., stationarity in ‘mean’ and stationarity in ‘variance’. Visual examination of data and structure of autocorrelation, and partial correlation coefficients helps to check the presence of stationarity. Another way of checking for stationarity is to
fit first order autoregressive model for raw data and test whether the coefficient ' $\Phi_1$ ' is less than one. If the model is found to be non-stationary, stationarity is achieved by differencing the series.

If ' $Y_t$ ' denotes the original series, the non-seasonal difference of first order is

$$ Y_t = X_t - X_{t-1} \quad \text{(4.24)} $$

The next step in identification process is to find initial values for orders of non-seasonal parameters, $p$ and $q$. They are obtained through significant autocorrelation and partial autocorrelation coefficients. There are no strict rules in choosing initial values. Though sample autocorrelation coefficients are poor estimates for population autocorrelation coefficients, still they are used as initial values while final models are achieved after going through the stages repeatedly.

**Estimation:** At the identification stage, one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then precise estimates of parameters for the model are obtained by least squares. Standard computer packages are available for finding the estimates of relevant parameters using iterative procedures.

**Diagnostics:** Different models are obtained for various combinations of Auto Regressive and Moving Average individually and collectively. The best model is selected based on following diagnostics:

a) Low Akaike Information Criteria (AIC)
b) Insignificance of auto correlations for residuals (Q-tests)
c) Significance of the parameters

a) **Low AIC:** AIC is given by $\text{AIC} = (-2 \log L + 2m)$ where $m = p+q$ and $L$ is the likelihood function. Since $-2 \log L$ is approximately equal to $\frac{n(1 + \log 2\pi) + n \log \sigma^2}{2}$
where $\sigma^2$ is the model MSE, AIC is written as $AIC = n(1 + \log 2\pi) + n\log \sigma^2 + 2m$ and the first term in this equation is a constant, so it is omitted while comparing between models. As an alternative to AIC, sometimes SBC is also used which is given as $SBC = \log \sigma^2 + (m \log n)/n$.

b) Insignificance of auto correlations for residuals (Q-tests): After tentative model is fitted to the data, it is important to perform diagnostic checks to test adequacy of the model and, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of chosen model by examining a quantity $Q$ known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and computed as follows:

$$Q = n \sum r^2(j) \quad \ldots (4.25)$$

where summation extends from 1 to $k$ with $k$ as the maximum lag considered, $n$ denotes number of observations in the series, $r(j)$ is the estimated autocorrelation at lag $j$: $k$ is a positive integer and is usually around 20. $Q$ follows Chi-square with $(k - m_I)$ degrees of freedom where $m_I$ is number of parameters estimated in the model. A modified $Q$ statistic is the Ljung-Box statistic which is given as

$$Q = n(n + 2) \sum r^2(j)/(n - j) \quad \ldots (4.26)$$

The $Q$ statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and $Q$ should be small (the probability value should be large). A significant value indicates that the chosen model does not fit well.
4.4 Data Description:

Tamilnadu state data on sugarcane production (million tonnes), productivity (tonnes/hectare) and area (million hectares) cultivated for a period 57 years (1950-51 to 2007-08) has been used for developing the univariate models (Refer Appendix – I). The sugarcane data has been collected from various volumes of ‘Cooperative Sugar’ (CSJ, 1980; CSJ, 2007) and ‘Indian Sugar’ (ISJ, 1985; ISJ, 2009) journals.

The data for a period of 55 years (1950-2005) has been used for model building and remaining three years (2006-08) data has been used for validation. Based on the models developed, sugarcane production, productivity and area have been forecasted for future.

Univariate models have also been developed for Coimbatore district data on sugarcane yield. Data for a period of 44 years (1961-2004) has been used in the study (Refer Appendix – IV). Data for a period of 40 years has been used for model building and the remaining for validation. The results of various univariate models are presented and compared in Chapter-6.