<u>BETHUNE COLLEGE</u>

AFFILIATED TO

<u>UNIVERSITY OF CALCUTTA</u>

<u>B.Sc. SEMESTER VI (HONOURS) EXAMINATION 2022 (UNDER CBCS)</u>

<u>POPULATION PROJECTION AND DEMOGRAPHIC FEATURES STUDY OF CHINA</u>

REGISTRATION NO.: 221-1211-0227-19

ROLL NO.: 193221-11-0197

SUBJECT: STATISTICS (STSA)

PAPER: DSE-B2

TABLE OF CONTENTS

List Of Tables

List of Figures

1. EXECUTIVE SUMMARY

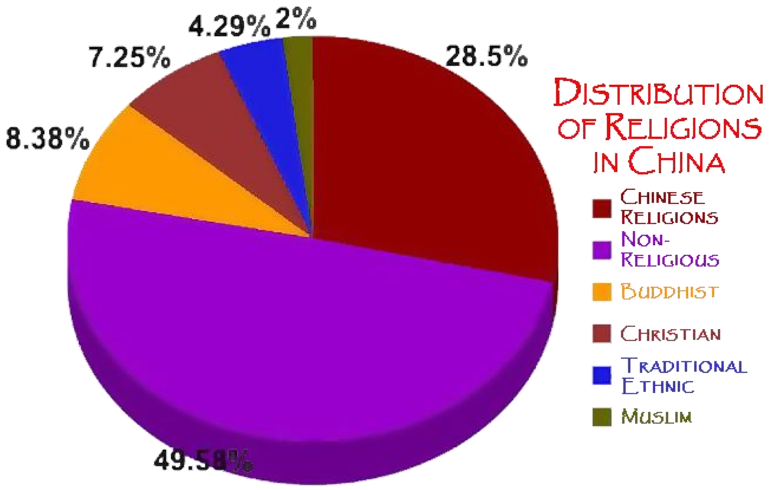
In this project work, the 2025 China population is projected based on the total population and sex. Considering two different approaches, firstly by considering the entire China population and secondly by considering male and female China population separately, the target is to establish the best fit model that is the model with the least percentage error. At the initial level the testing data points (2019 and 2020) were used to find out the best fit model for both the approaches by fitting the mathematical curve models which included linear, exponential, quadratic and S-Curve. The S-Curve model was found to have the least percentage error for both the approaches while comparing the testing data points with the actual population values. On obtaining the S-Curve model that fitted in the best way with the data set, the entire (training) data set was then fitted to the obtained best fit model (S-Curve model) to project 2025 China population. Since two approaches were followed, two different models were obtained – one is the total China population projection model for 2025, which projected the 2025 China population to be 1467124054 and the other is the combined male and female China population projection model for 2025, which projected the 2025 China population to be 1447250063. Finally, these two models are compared on the basis of percentage error. The model obtained on combining the male and the female China population projection had the least percentage error and is established as the best model for 2025 population of China. We have also calculated the sex ratios for the entire data set of China population in order to account for the gender balance in the country. Graphical representations and interpretations have also been presented.

2. INTRODUCTION

China, officially the People's Republic of China, located in East Asia, is the largest country in the world today. In January 2013, the Chinese Government released data confirming that the population of China was an impressive 1,354,040,000. India, the next largest country, has 120 million fewer people, for a population of 1.28 billion. The United States, the third largest country in the world, has a much smaller population of 323 million.

China is classified as an upper middle-income country by the World Bank, and its rapid growth over the decades has pulled hundreds of millions of its citizens out of poverty. Although 56 different ethnic groups are officially recognized in China, 91.51% of Chinese are Han Chinese. Other ethnic groups are growing at a higher rate than Han Chinese, but because of massive dominance of Han Chinese, this is not expected to dramatically alter China's ethnic composition.

China is officially an atheist state, thus doesn't survey its people on their religion. Because of this, no accurate figures based on religion is found. A survey taken in China showed that 85% of Chinese residents have some religious beliefs, while just 15% consider themselves to be atheists. China has seen an interesting syncing of the 3 religions, Taoism, Buddhism and Confucianism in the form of a folk religion that is common throughout the country. About 3% of the population is Islamic, with a Christian population estimated to be 5% and Buddhists of 10 to 18%.



China feeds about 22% of the world population on 7% of world's land. Per capita area of cultivated land is low for China's population size. Land area is 13.3 mu/capita or 25% of US land area. Land use in China is experiencing massive changes and impacts to the environment due to an unprecedented period of economic growth, which has catapulted it from one of the world's poorest countries 30 years ago to the world's second largest economy today. Based on trends in economic development, population growth, and land use, China's natural landscape will experience significant and increasing pressures well into the future. China has four times the population of the U.S., within roughly the same area. With 1.3 billion inhabitants, 20% of the world's population lives in China. While the rate of population growth has been declining for decades, the total number of inhabitants has been growing and is expected to do so through 2030. China's urban population is growing rapidly; between 1950 and 2009, the percentage of the population living in urban areas quadrupled from 14% to 48%. Meanwhile, the rural population is declining, opening the landscape in areas that are not urbanized. The forests and farming land of eastern China support far more people and major cities than do the grasslands, deserts, and high mountain regions of the west.

China is the world's most populous country with an astounding 1.44 billion citizens. Altogether, the size of the population of China is larger than nearly four regions combined: South America, Europe (excluding Russia), the U.S. & Canada, and Australia & New Zealand. The combined population of Australia, New Zealand, Europe (excluding Russia), South America, U.S. and Canada is 1.432 billion as compared to China's 1.439 billion.

Looking at history, the population of China has more than doubled since the 1950s. The country was the first in the world to hit one billion people in 1980.However, in 1979, in an attempt to control the burgeoning population, the infamous one-child policy was introduced, putting controls on how many children Chinese citizens could have. While the government eventually recognized the negative implications of this policy, it appeared to be too little, too late. The two-child policy was introduced in 2016, but it has not yet reversed the current slowdown in population growth. Despite these trends, however, China’s current population remains massive, constituting almost 20% of the world’s total population. Right now, 71% of the Chinese population is between the ages of 15 and 65 years old, meaning that the labor supply is still immense.

3. OBJECTIVE

China, the world's largest country, was the first to hit the population in billions in 1980 and is still growing. The objective of this project is to highlight some of the approaches by which the China population can be projected for 2025. Consequently, the target is to find out the best model with the least percentage error. To illustrate how these and future trends will reshape the nation in the decades to come and the pace of population change since the last data point (2020), the least square method for population projection is used for the year 2025. Moreover, to find out the gender balance of China, sex ratios are calculated for the entire data set. And then, the graphical representation and interpretations are presented.

4. Data

Population projections are influenced not only by the methods and assumptions used in their production, but also by the historical data series upon which they are based. Given below are the following tables for the dataset that has been used in the project work:

Table 4.1: Historical China population

YEAR	POPULATION	YEAR	POPULATION	YEAR	POPULATION	YEAR	POPULATION
1949	54,69,40,283	1967	76,25,81,176	1985	1,07,55,89,361	2003	1,31,53,03,521
1950	55,44,19,273	1968	78,40,74,709	1986	1,09,50,14,109	2004	1,32,30,84,641
1951	56,99,09,117	1969	80,59,85,939	1987	1,11,60,95,476	2005	1,33,07,76,380
1952	58,25,76,491	1970	82,76,01,394	1988	1,13,77,24,227	2006	1,33,84,08,647
1953	59,33,65,877	1971	84,87,59,710	1989	1,15,83,57,397	2007	1,34,59,93,888
1954	60,30,52,324	1972	86,94,85,964	1990	1,17,68,83,674	2008	1,35,35,69,484
1955	61,22,41,554	1973	88,94,85,372	1991	1,19,28,97,284	2009	1,36,11,69,419
1956	62,13,63,234	1974	90,84,64,198	1992	1,20,67,11,244	2010	1,36,88,10,615
1957	63,06,77,645	1975	92,62,40,885	1993	1,21,88,17,055	2011	1,37,64,97,639
1958	64,02,95,772	1976	94,26,85,411	1994	1,23,00,20,031	2012	1,38,42,06,401
1959	65,02,12,734	1977	95,78,91,272	1995	1,24,09,20,535	2013	1,39,18,83,330
1960	66,04,08,056	1978	97,22,05,442	1996	1,25,16,36,186	2014	1,39,94,53,965
1961	67,09,52,695	1979	98,61,32,202	1997	1,26,19,96,012	2015	1,40,68,47,870
1962	68,21,02,655	1980	1,00,00,89,235	1998	1,27,19,82,350	2016	1,41,40,49,351
1963	69,43,39,083	1981	1,01,40,22,212	1999	1,28,15,14,832	2017	1,42,10,21,791
1964	70,82,54,597	1982	1,02,79,48,987	2000	1,29,05,50,765	2018	1,42,76,47,786
1965	72,42,18,968	1983	1,04,24,31,412	2001	1,29,91,29,752	2019	1,43,37,83,686
1966	74,24,14,885	1984	1,05,81,71,976	2002	1,30,73,52,257	2020	1,43,93,23,776

Table 4.2: Historical China male population

YEAR	MALE POPULATION	YEAR	MALE POPULATION	YEAR	MALE POPULATION	YEAR	MALE POPULATION
1949	28,14,50,000	1967	39,11,50,000	1985	54,72,50,000	2003	66,55,60,000
1950	28,66,90,000	1968	40,22,60,000	1986	55,58,10,000	2004	66,97,60,000
1951	29,23,10,000	1969	41,28,90,000	1987	56,29,00,000	2005	67,37,50,000
1952	29,83,30,000	1970	42,68,60,000	1988	57,20,10,000	2006	67,72,80,000
1953	30,46,80,000	1971	43,81,90,000	1989	58,09,90,000	2007	68,04,80,000
1954	31,24,20,000	1972	44,81,30,000	1990	58,90,40,000	2008	68,35,70,000
1955	31,80,90,000	1973	45,87,60,000	1991	59,46,60,000	2009	68,64,66,800
1956	32,53,60,000	1974	46,72,70,000	1992	59,81,10,000	2010	68,74,84,557
1957	33,46,90,000	1975	47,56,40,000	1993	60,47,20,000	2011	69,06,80,000
1958	34,19,50,000	1976	48,25,70,000	1994	61,24,60,000	2012	69,39,50,000
1959	34,89,00,000	1977	48,90,80,000	1995	61,80,80,000	2013	69,72,80,000
1960	34,28,30,000	1978	49,56,70,000	1996	62,20,00,000	2014	70,46,75,333
1961	33,88,00,000	1979	50,19,20,000	1997	63,13,10,000	2015	70,86,47,978
1962	34,51,70,000	1980	50,78,50,000	1998	63,94,00,000	2016	71,25,69,385
1963	35,53,30,000	1981	51,51,90,000	1999	64,69,20,000	2017	71,67,28,147
1964	36,14,20,000	1982	52,35,20,000	2000	65,43,70,000	2018	71,99,05,214
1965	37,12,80,000	1983	53,15,20,000	2001	65,67,20,000	2019	72,22,64,717
1966	38,18,90,000	1984	53,84,80,000	2002	66,11,50,000	2020	72,36,83,488

Table 4.3: Historical China female population

YEAR	FEMALE POPULATION	YEAR	FEMALE POPULATION	YEAR	FEMALE POPULATION	YEAR	FEMALE POPULATION
1949	26,02,20,000	1967	37,25,30,000	1985	51,12,60,000	2003	66,55,60,000
1950	26,52,70,000	1968	38,30,80,000	1986	51,92,60,000	2004	66,97,60,000
1951	27,06,90,000	1969	39,38,20,000	1987	53,01,00,000	2005	67,37,50,000
1952	27,64,90,000	1970	40,30,60,000	1988	53,82,50,000	2006	67,72,80,000
1953	28,32,80,000	1971	41,41,00,000	1989	54,60,50,000	2007	68,04,80,000
1954	29,02,40,000	1972	42,36,40,000	1990	55,42,90,000	2008	68,35,70,000
1955	29,65,60,000	1973	43,33,50,000	1991	59,46,60,000	2009	68,64,66,800
1956	30,29,20,000	1974	44,13,20,000	1992	59,81,10,000	2010	68,74,84,557
1957	31,18,40,000	1975	44,85,60,000	1993	60,47,20,000	2011	69,06,80,000
1958	31,79,90,000	1976	45,46,00,000	1994	61,24,60,000	2012	69,39,50,000
1959	32,31,70,000	1977	46,06,60,000	1995	61,80,80,000	2013	69,72,80,000
1960	31,92,40,000	1978	46,69,20,000	1996	62,20,00,000	2014	70,46,75,333
1961	31,97,90,000	1979	47,35,00,000	1997	63,13,10,000	2015	70,86,47,978
1962	32,77,80,000	1980	47,92,00,000	1998	63,94,00,000	2016	71,25,69,385
1963	33,63,90,000	1981	48,55,30,000	1999	64,69,20,000	2017	71,67,28,147
1964	34,35,70,000	1982	49,30,20,000	2000	65,43,70,000	2018	71,99,05,214
1965	35,41,00,000	1983	49,85,60,000	2001	65,67,20,000	2019	72,22,64,717
1966	36,35,30,000	1984	50,50,90,000	2002	66,11,50,000	2020	72,36,83,488

Table 4.4: Calculated Sex Ratio of China

YEAR	SEX RATIO	YEAR	SEX RATIO	YEAR	SEX RATIO	YEAR	SEX RATIO
1949	1081.585	1967	1049.983	1985	1070.395	2003	1061.99
1950	1080.748	1968	1050.068	1986	1070.389	2004	1062.909
1951	1079.87	1969	1048.423	1987	1061.875	2005	1063.016
1952	1078.99	1970	1059.048	1988	1062.722	2006	1062.9
1953	1075.544	1971	1058.174	1989	1063.987	2007	1061.906
1954	1076.42	1972	1057.809	1990	1062.693	2008	1060.703
1955	1072.599	1973	1058.636	1991	1055.166	2009	1059.308
1956	1074.079	1974	1058.801	1992	1042.73	2010	1052.124
1957	1073.275	1975	1060.371	1993	1041.812	2011	1051.792
1958	1075.348	1976	1061.527	1994	1045.082	2012	1051.296
1959	1079.618	1977	1061.694	1995	1042.065	2013	1051.007
1960	1073.894	1978	1061.574	1996	1033.411	2014	1056.192
1961	1059.445	1979	1060.021	1997	1043.574	2015	1055.774
1962	1053.054	1980	1059.787	1998	1051.282	2016	1055.313
1963	1056.304	1981	1061.088	1999	1058.893	2017	1054.808
1964	1051.954	1982	1061.864	2000	1067.383	2018	1054.258
1965	1048.517	1983	1066.11	2001	1059.995	2019	1053.662
1966	1050.505	1984	1066.107	2002	1060.589	2020	1053.02

5. METHODOLOGY

There are several methods of population projection. Here the Mathematical Method is used which is one of the earliest methods used for population projection. In order to get the best fit model for population projection of USA, use of four well known mathematical curves – Linear, Exponential, Quadratic, S-Curve (Pearl –Reed Logistic) has been considered.

Fitting of straight line by least square method

Let the straight line trend between the given time-series values (y_t) and time t be given by the equation:

$$\hat{Y}_t = a + b*t \text{ ...... } (5.1)$$

Principle of least squares consists in minimizing the sum of squares of the deviations between the given values of y_t and their estimates given by (5.1). In other words, to find a and b such that for given values of y_t corresponding to n different values of t,

$$E = \sum (y_t - \hat{Y}_t)^2$$

$$= \sum (y_t - a - b*t)^2$$

is minimum. For a maxima or minima of E, for variations in a and b, one should have,

$$\begin{cases} \frac{\partial E}{\partial a} = 0 = -2\sum (y_t - a - bt) \\ \frac{\partial E}{\partial b} = 0 = -2\sum t(y_t - a - bt) \end{cases} \Rightarrow \begin{cases} \sum y_t = na + b\sum t \\ \sum y_t = a\sum t + b\sum t^2 \end{cases} \quad \text{... ... } (5.2)$$

which are the normal equations for estimating a and b.

The values of $\sum y_t$,$\sum t$, $\sum t^2$ obtained from the given data and the equations (5.2) can now be solved for a and b. With these values of a and b, the line (5.1) gives the desired trend line.

Fitting of Second Degree (Parabolic) Trend

Let the second degree, parabolic trend curve be:

$$y_t = a + bt + ct^2 \text{ .... } (5.3)$$

Proceeding similarly as in the case of a straight line, the normal equations for estimating a, b and c are given by:

$$\sum y_t = na + b \sum t + c \sum t^2$$

$$\sum ty_t = a \sum t + b \sum t^2 + c \sum t^3 \text{..........} \quad (5.4)$$

$$\sum t^2 y_t = a \sum t^2 + b \sum t^3 + c \sum t^4$$

the summation being taken over the values of the time series.

Fitting of Exponential Curve

Let the exponential trend curve be:

$$y_t = ab^t \text{ ......} \quad (5.5)$$

$$\Longrightarrow \log y_t = \log a + t \, logb \text{ ....} \quad (5.5a)$$

$$\Longrightarrow Y = A + Bt \; (say) \text{ ......} \quad (5.5b)$$

where $Y = log \; y_t$,$A = log \; a$,$B = log \; b$

(5.5a) is a straight line in t and Y and thus the normal equations for estimating A and B are

$$\sum Y = nA + B \sum t, \text{ ......} \quad (5.5c)$$

$$\sum tY = A \sum t + B \sum t^2$$

These equations can be solved for A and B and finally on using (5.5b), a and b are obtained as:

$$a = antilog(A) \; ; \; b = antilog(B)$$

Fitting of Logistic Curve by the method of Pearl and Reed

Let the Logistic Curve be given as:

$$P = \frac{L}{1+e^{r(\beta-1)}} \text{......} \quad (5.6)$$

To fit a logistic curve to a set of data, one has to estimate the constants L, r and β from the observed figures. It will be assumed that the population figures are given for L equidistant points of time, say, for t = 0,1,2……, N − 1. The population at time t will be denoted by P_t.

Since there are three unknown constants, these can be determined in such a way as to make the logistic curve pass through any three selected points (t ,$_t$).

These points should be so chosen that the whole range of observations is more or less evenly covered. It will be supposed that these three points are equidistant on the time scale, so that these may be denoted by (i, P_i), $(i + n, P_{i+n})$ and $(i + 2n, P_{i+2n})$ or, through a change of origin of t,

By $(0, P_0)$, (n, P_n) and $(2n, P_{2n})$. Since the curve is to pass through these points, we have,

$$(5.7) \ldots\ldots\ldots\ldots\ldots \begin{cases} \frac{1}{P_O} = \frac{1+e^{r\beta}}{L} \\ \frac{1}{P_n} = \frac{1+e^{r(\beta+n)}}{L} \\ \frac{1}{P_{2n}} = \frac{1+e^{r(\beta-2n)}}{L} \end{cases}$$

Writing

$$d_1 = \frac{1}{P_O} - \frac{1}{P_n}$$

and

$$d_2 = \frac{1}{P_n} - \frac{1}{P_{2n}}$$

from equation (5.7)

$$d_1 = \frac{1}{L}\, e^{r\beta(1-e^{-rn})}$$

and

$$d_2 = \frac{1}{L}\, e^{r(\beta-n)(1-e^{-rn})}$$

Whence, $e^{rn} = d_1/d_2$,

Or $$r = \frac{1}{n}(ln d_1 - ln d_2) \ldots\ldots. (5.8)$$

Further $$1 - \frac{d_2}{d_1} = Ld_1 / e^{r\beta}$$

$$\Longrightarrow \frac{{d_1}^2}{d_1 - d_2} = \frac{e^{r\beta}}{L} = \frac{1}{P_O} - \frac{1}{L}$$

$$\Longrightarrow \frac{1}{L} = \frac{1}{P_O} - \frac{{d_1}^2}{d_1 - d_2} \text{ ...... (5.9)}$$

Estimation of r and L is obtained from equation (5.8) and (5.9), respectively. Using these estimates and the relation,

$$(5.10) \text{ ....} \begin{cases} e^{r\beta} = \frac{L}{P_O} - 1 \\ \beta = \frac{1}{r} \ln \left(\frac{L}{P_O} - 1\right) \end{cases}$$

Finally, β is determined.

The values of L, β and r obtained in this way will, of course, be only rough estimates. Pearl and Reed suggest a method, based on the least-square principle, by which these can be improved upon.

Denoting the estimates found by the above 'method of three selected points' by L_O, r_O and β_O, one may write,

$$L = L_O + \delta_L,$$

$$r = r_O + \delta_r,$$

$$\beta = \beta_O + \delta_\beta,$$

where δ_L, δ_r and δ_β are the errors in the estimates. The population size P, regarded as a function of L, r and β, say,

$$f(L, r, \beta) = \frac{L}{1 + e^{r(\beta - t)}},$$

may then be written as

$$P = f(L_O, r_O, \beta_O) + \delta_L \left(\frac{\partial f}{\partial L}\right)_O + \delta_r \left(\frac{\partial f}{\partial r}\right)_O + \delta_\beta \left(\frac{\partial f}{\partial \beta}\right)_O$$

$$= f_O + \delta_L x + \delta_r y + \delta_\beta z, \text{ say}$$

The errors $\delta_{L,r}$ and δ_β may be estimated by the method of least squares, which yields the normal equations:

$\sum_t xi\ (P - f_{Oi}) = \delta_L \sum x_i^2 + \delta_r \sum x_i * y_i + \delta_\beta \sum x_i * z_i$

$\sum_t yi\ (P - f_{Oi}) = \delta_L \sum x_i * y_i + \delta_r \sum y_i^2 + \delta_\beta \sum y_i * z_i$

$\sum_t zi\ (P - f_{Oi}) = \delta_L \sum x_i * z_i + \delta_r \sum y_i * z_i + \delta_\beta \sum z_i^2$

where,

$$x_i = \left(\frac{\partial P_i}{\partial L}\right)_O = \frac{1}{1+e^{r_O(\beta_O - i)}},$$

$$y_i = \left(\frac{\partial P_i}{\partial r}\right)_O = \frac{L_O}{[1+e^{r_O(\beta_O - i)}]^2}.(\beta_o - i)\ e^{r_O(\beta_O - i)},$$

$$z_i = \left(\frac{\partial P_i}{\partial \beta}\right)_O = \frac{L_O r_O e^{r_O(\beta_O - i)}}{[1+e^{r_O(\beta_O - i)}]^2},$$

And the sums are taken over i = 0,1,2, ……, $N - 1$.

The process may be repeated to get still better estimates of L, r and β.

6. RESULTS AND ANALYSIS

The training data for the project work has been collected on population of China by spanning the time from 1949 to 2020. The population of China is taken as the dependent variable. To find the model that best fits the dataset, the last two years data points (2019, 2020), also known as testing data points has been kept aside. Several models like linear, exponential, quadratic and logistic are fitted with the training dataset, that is, with the remaining dataset and the last two year's population (2019 and 2020) have been projected. The projected population thus obtained by two different approaches (Approach 1 and Approach 2) is compared with the actual census population in 2019 and 2020 to find the best fit model.

Approach 1

In approach 1, the 2025 China population is projected by considering the entire population (1949-2020).

China Population Projection for 2019 and 2020

The following graphs for 2019 and 2020 China population projection by the four different models are shown below.

•LINEAR TREND MODEL

The linear trend model is:

Y_t = 532921323 + 13845773×t ………… (6.1)

Where, Y_t = China population at time t

On deleting the last two data points (2019,2020), the trend line is given as:

Y_t = 527660662 + 14063450×t ………..(6.2)

Fig. 6.1 depicts the trend analysis plot for China population using Linear Trend Model. The actual trend line (blue) shows a constant trend in the growth rate of population in the initial periods and then gradually increases. Whereas the fitted trend (red) increases monotonically for the entire time period taken into consideration. The points in green, in Fig. 6.2, gives the population projection for 2019 and 2020 as 1526165624 and 1540229074 respectively.

FIG. 6.1: LINEAR TREND FOR CHINA POPULATION

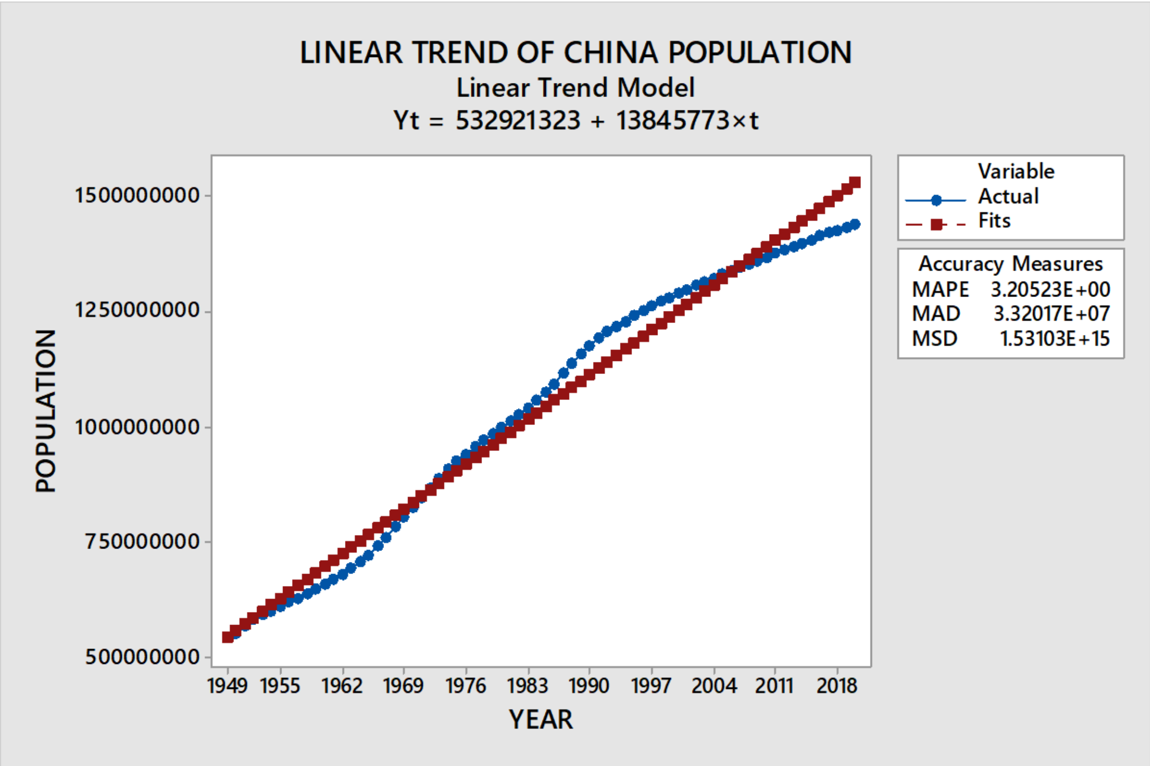
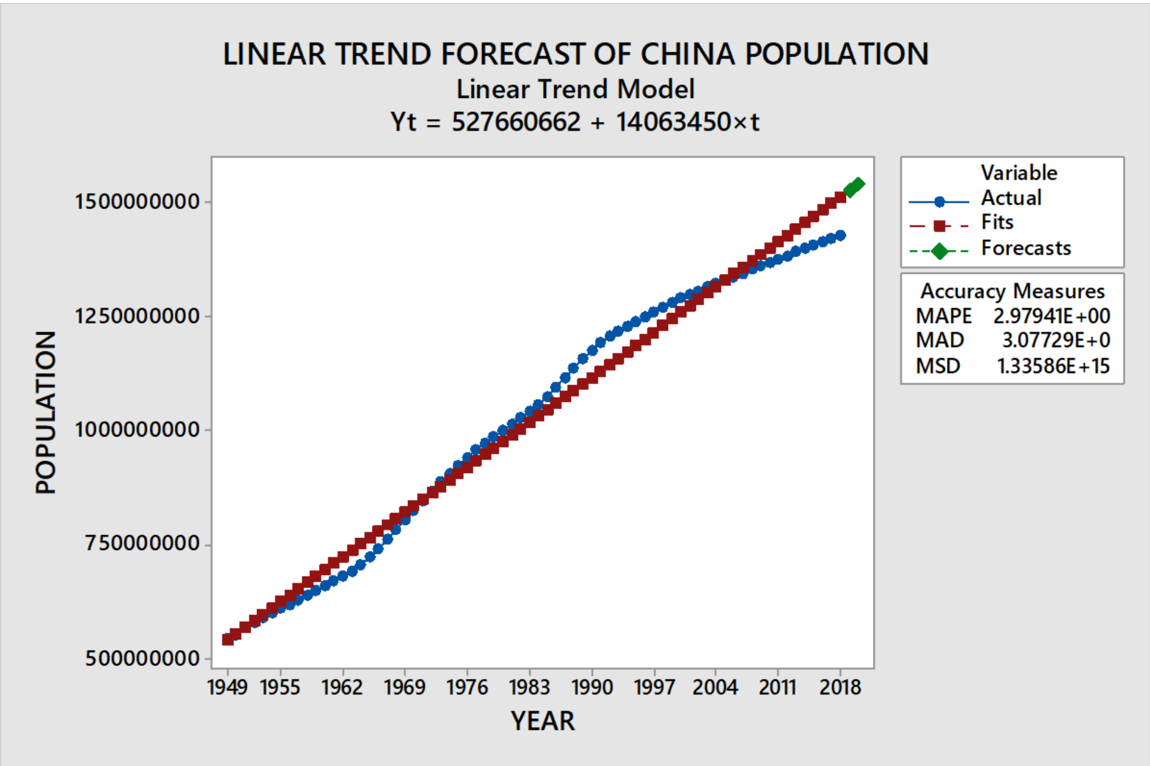


FIG. 6.2: FORECAST OF 2019 AND 2020 FOR LINEAR TREND OF CHINA POPULATION



•*GROWTH CURVE MODEL*

The Growth curve model is:

Y_t = 588569752 × (1.01445^t) .... (6.3)

Where, Y_t = China population at time t

On deleting the last two data points (2019,2020), the trend line is given as:

Y_t = 583800817 × (1.01479^t) .... (6.4)

Fig. 6.3 depicts the trend analysis plot for China population using Growth Curve Model. From the year 1966 the actual trend rises above the fitted trend line till 2005 and again moves below the fitted trend line thereafter. The points in green, in Fig 6.4, gives the population forecast for 2019 and 2020 as 1655970621 and 1680466744 respectively.

FIG. 6.3: GROWTH CURVE FOR CHINA POPULATION

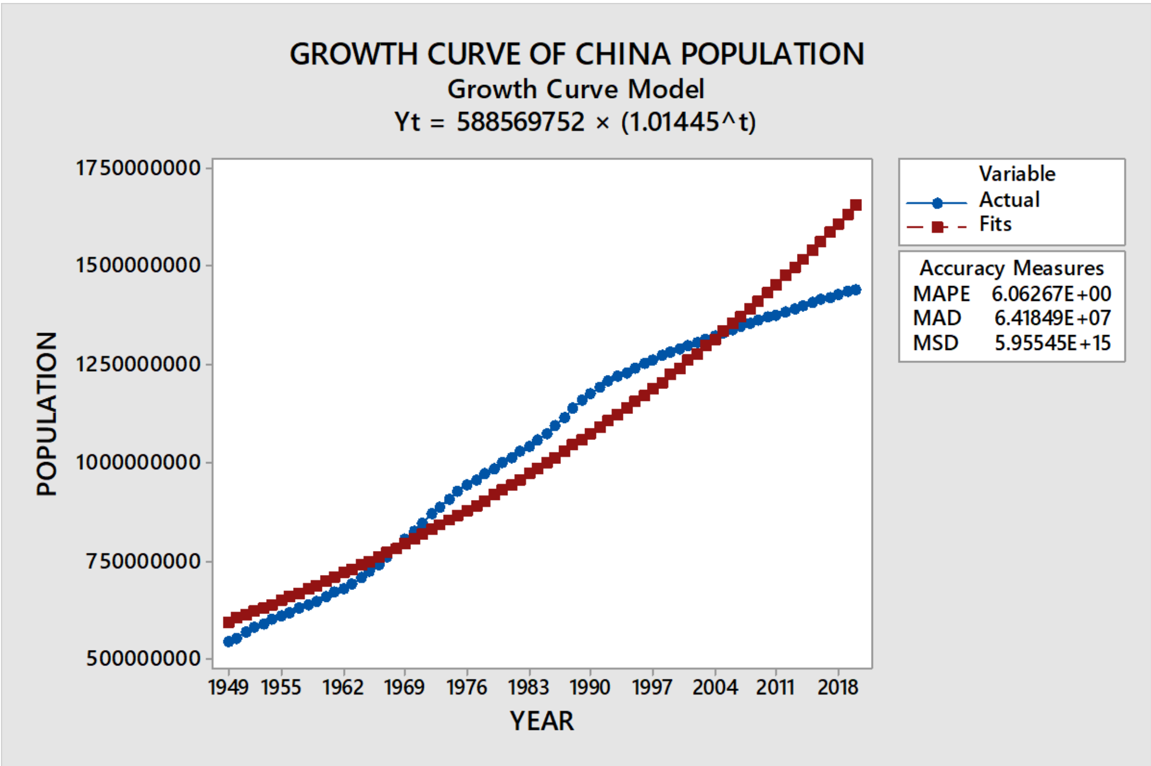
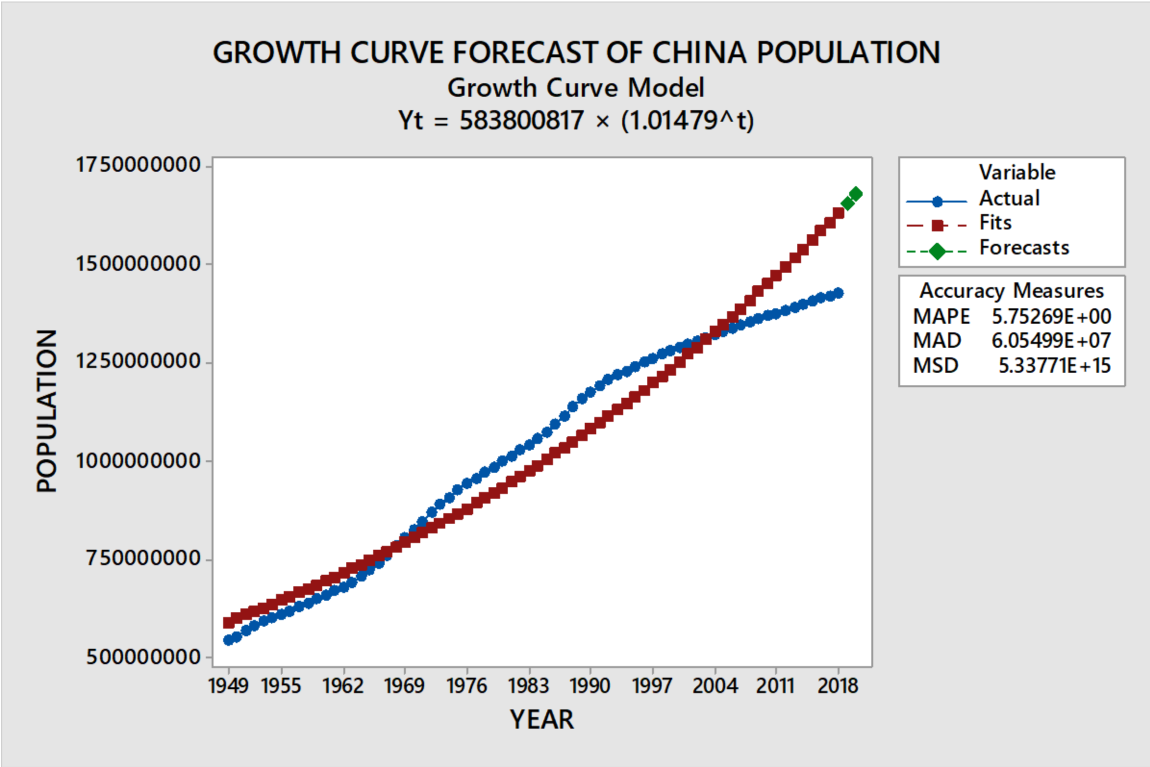


FIG. 6.4: FORECAST OF 2019 AND 2020 FOR GROWTH CURVE OF CHINA POPULATION



•S-CURVE MODEL

The S-Curve Model is:

Y_t = (10^10) / (6.23297 + 14.0777×(0.958819^t)) ……………. (6.5)

Where, Y_t = China population at time t

On deleting the last two data points (2019,2020), the trend line is given as:

Y_t= (10^10) / (6.22673 + 14.2249×(0.958624^t)) ……………. (6.6)

Fig. 6.5 depicts the trend analysis plot for China population using S-Curve Trend Model. Here the actual trend line (blue) completely coincides with the fitted trend line (red) from the time period 1949 till 2020. The points in green, in Fig. 6.6, gives the population forecast for 2019 and 2020 as 236032999 and 241064071 respectively.

FIG 6.5:S-CURVE TREND FOR CHINA POPULATION

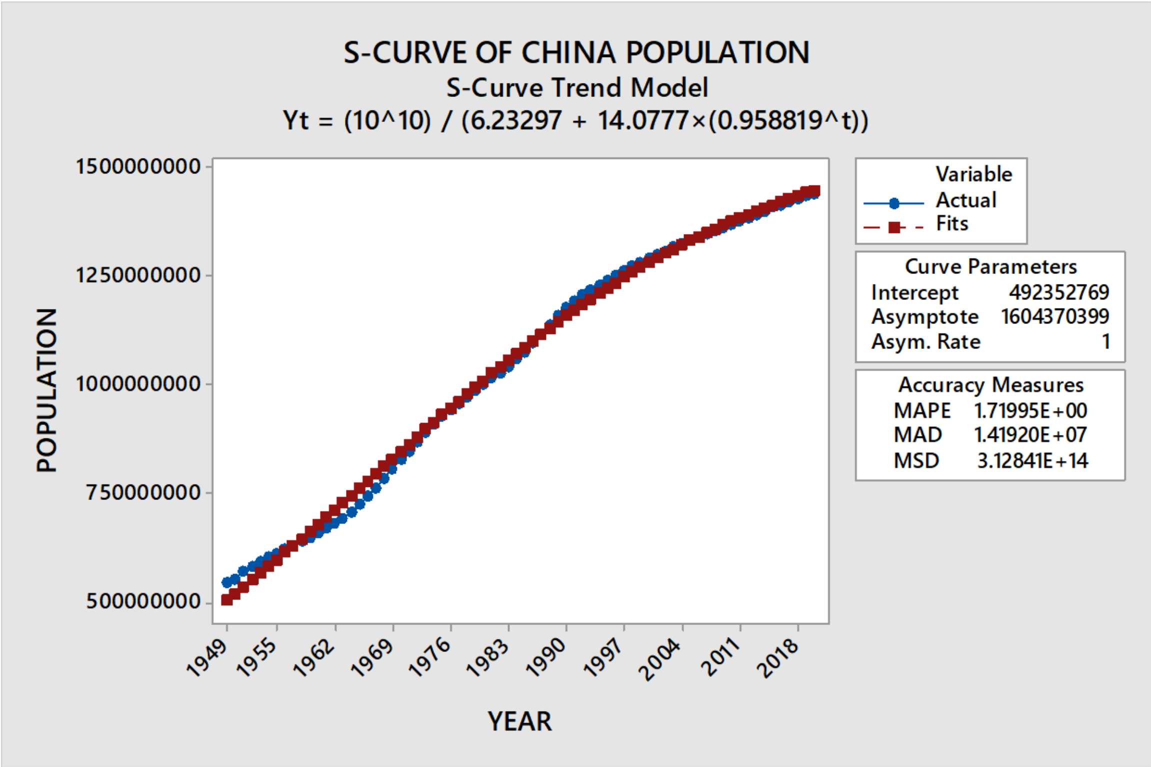
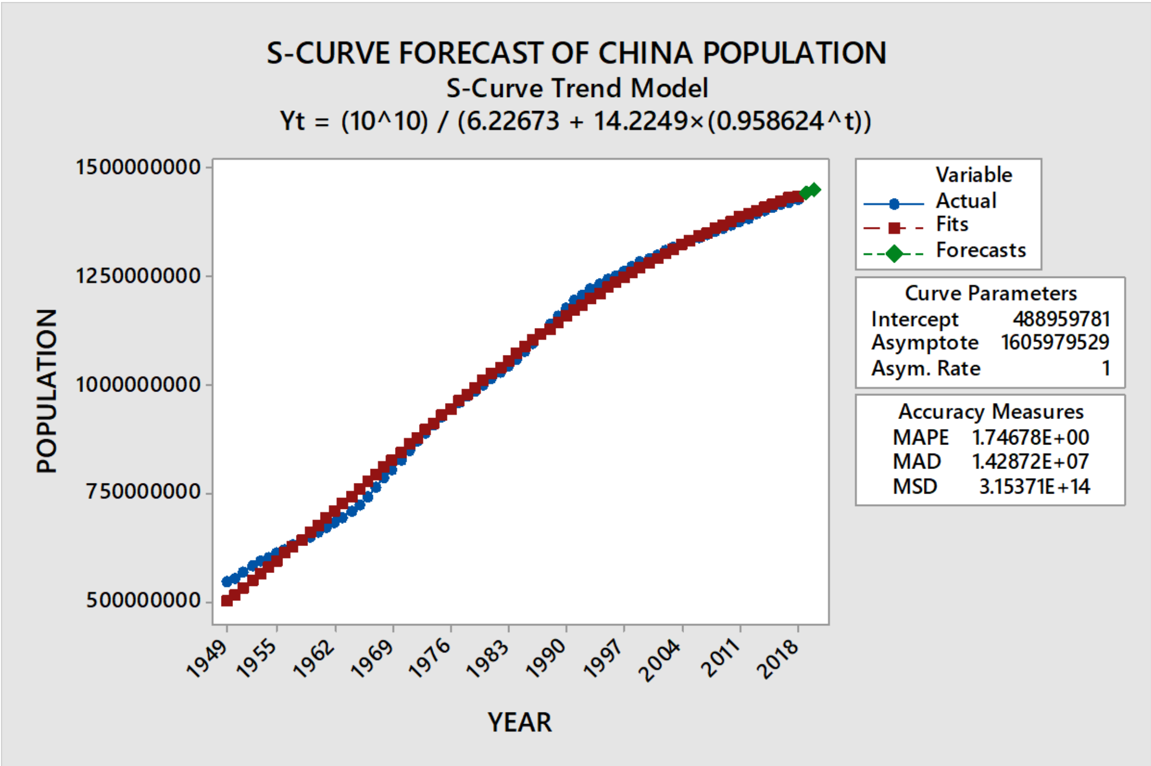


FIG.6.6: FORECAST OF 2019 AND 2020 FOR S-CURVE OF CHINA POPULATION



•*QUADRATIC TREND MODEL*

The Quadratic Trend Model is:

Y_t = 464423528 + 19399649×t - 76080×t^2 ………………… (6.7)

Where, Y_t = China population at time t

On deleting the last two data points (2019,2020), the trend line is given as:

Y_t= 467064922 + 19113095×t - 71122×t^2 ………………. (6.8)

Fig. 6.7 depicts the trend analysis plot for China population using Quadratic Trend Model. The actual trend line (blue) coincides with the fitted trend line (red) from 1971-1986 and again from 2002-2020, while in other periods it either rises above or goes down. The points in green, in Fig. 6.8, gives the population forecast for 2019 and 2020 as 15698718786 and 871687165 respectively.

FIG. 6.7: QUADRATIC MODEL FOR CHINA POPULATION

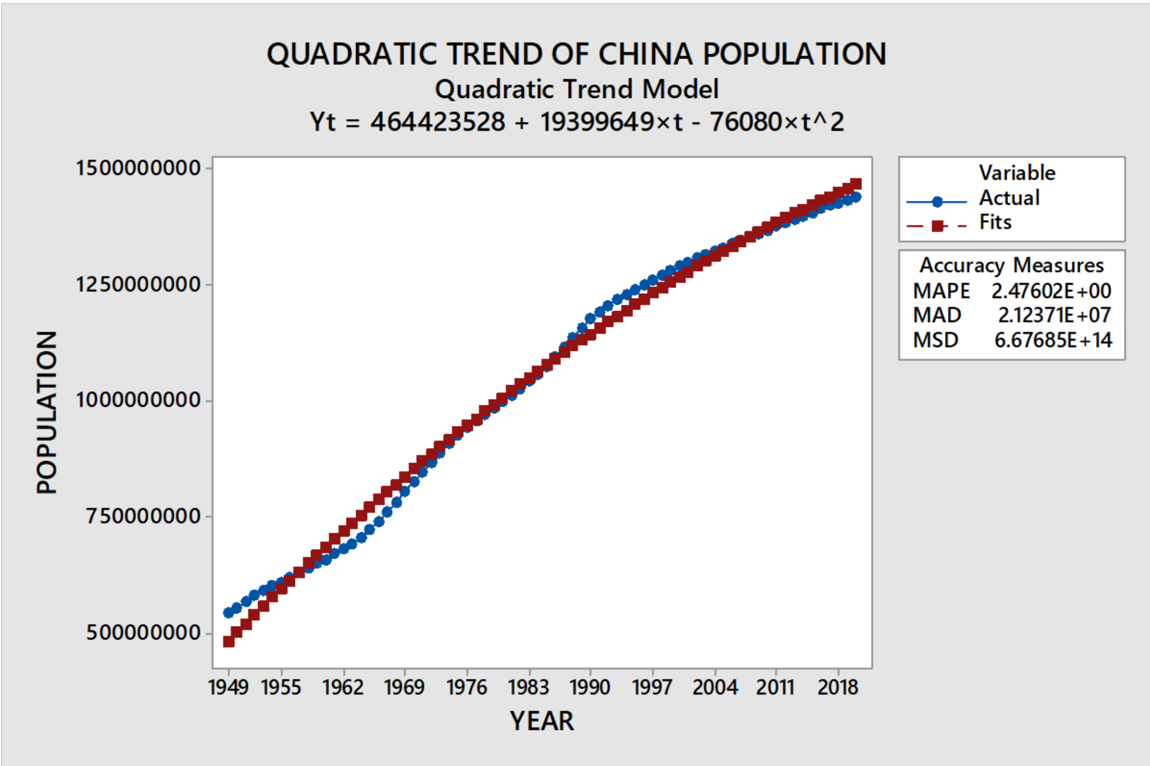
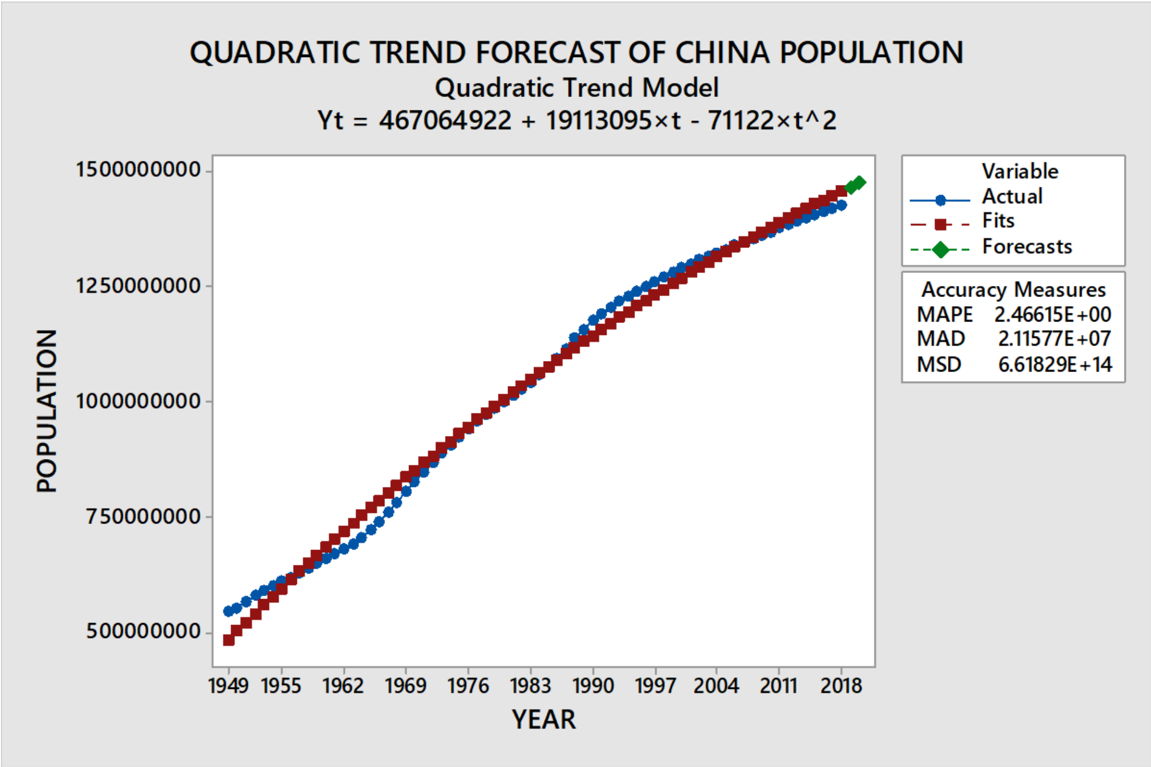


FIG. 6.8: FORECAST OF 2019 AND 2020 FOR QUADRATIC TREND OF CHINA POPULATION



After obtaining the China population projections for the years 2019 and 2020, a comparison of these projections with the observed populations is made with respect to percentage error. The formula for calculation of percentage error is given below:

$$\%\ error = \frac{(\text{observed Population} - \text{projected Population})}{\text{observed population}} * 100$$

The following tables show the percentage error on comparing the projected population with observed population for different models.

Table 6.1: Percentage error for Linear Model

YEAR	OBSERVED POPULATION	FORECAST POPULATION	% ERROR
2019	1433783686	1526165624	-6.44323
2020	1439323776	1540229074	-7.0106

Table 6.2: Percentage error for Growth Curve Model

YEAR	OBSERVED POPULATION	FORECAST POPULATION	% ERROR
2019	1433783686	1655970621	-15.4965
2020	1439323776	1680466744	-16.7539

Table 6.3: Percentage error for S-Curve Model

YEAR	OBSERVED POPULATION	FORECAST POPULATION	% ERROR
2019	1433783686	1442000205	-0.57307
2020	1439323776	1448118062	-0.611

Table 6.4: Percentage error for Quadratic Model

YEAR	OBSERVED POPULATION	FORECAST POPULATION	% ERROR
2019	1433783686	1474512567	-2.84066
2020	1439323776	1465569883	-1.8235

On comparing the above four tables (from Table 6.1 to Table 6.4) it can be seen that the S-curve model has the least % error among the other models.

A different approach to find the best fit model for the dataset is by comparing various measures such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE) and Mean Square Deviation (MSD) for different models. The table below provides the values of MAD, MAPE, and MSD for various models taken into consideration.

Table 6.5: Simultaneous comparison of MAD, MAPE, and MSD for the four model

Model	MAPE	MAD	MSD
Linear	2.97941	30772900	1335860000000000
Growth Curve	5.75269	60549900	5337710000000000
S-Curve	1.74678	14287200	315371000000000
Quadratic	5.75269	60549900	5337710000000000

From Table 6.5 it is evident that the S-curve model has the minimum MAD, MAPE, and MSD among the four models. So, the S-curve model fits in the best way with our dataset thereby claiming that the S-curve model is the best fit model.

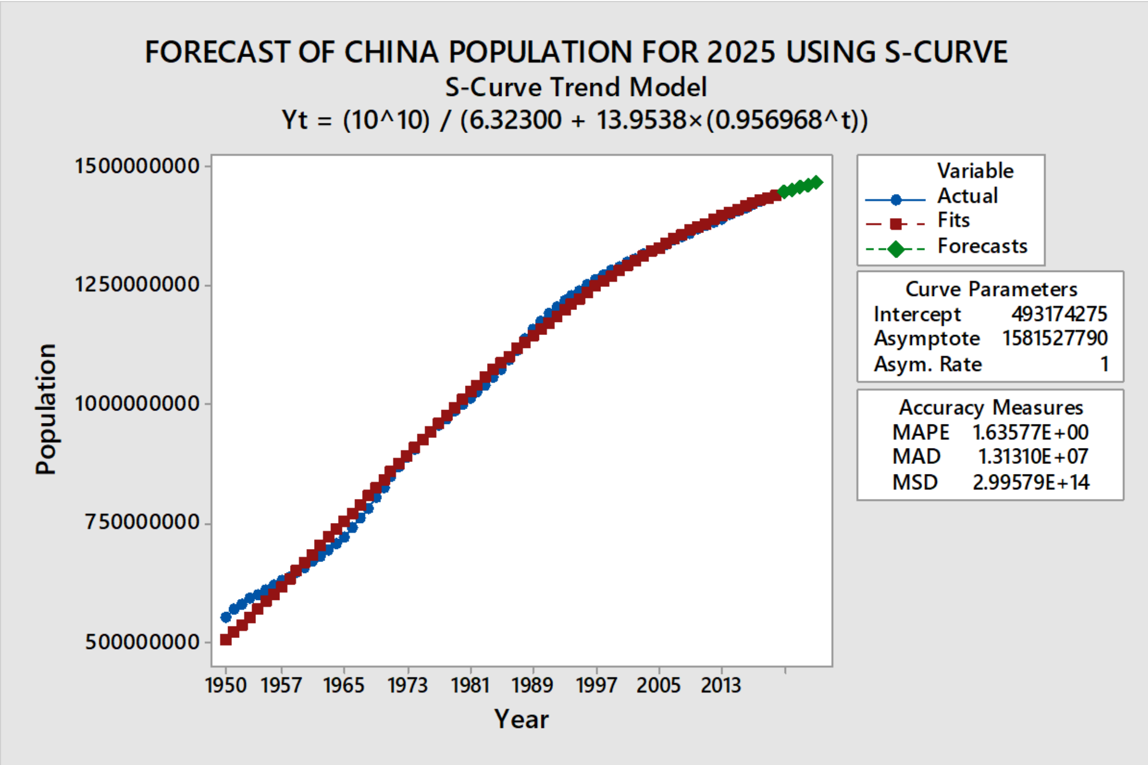
China Population Projection for 2025

Now as the S-curve model has been established as the best fit model, the entire dataset including the last two data points, that is, 2019 and 2020 is taken into consideration and the S-curve model is fitted to the dataset to project the China population for the census year 2025. The fitted graph (Fig: 6.9) for the 2025 China population projection by the S-curve model is given below. The China population for the year 2025 is projected to be 1467124054 by using the S-curve trend model. The S-curve Trend Model is:

Y_t = (10^10) / (6.23297 + 14.0777×(0.958819^t)) ………… (6.9)

Where, Y_t = China population at time t

Fig 6.9: Trend analysis plot for China population by S-curve model considering the data set (1949-2020)



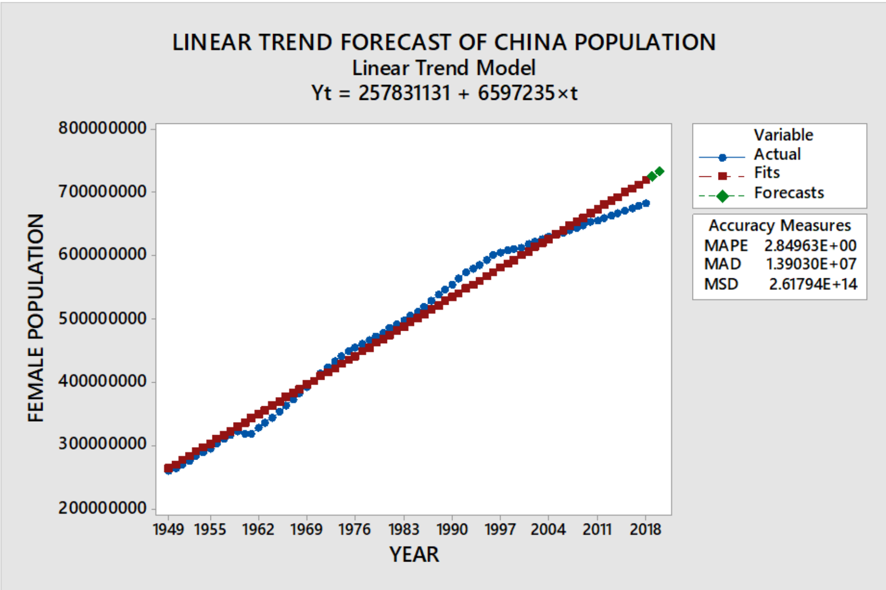
Approach 2

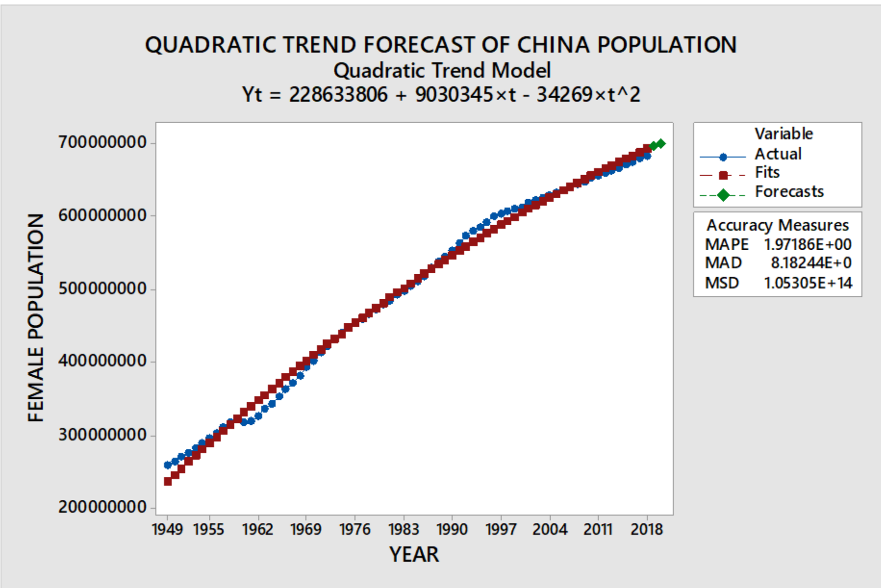
Approach 2 takes into consideration the male and female population data of China for different years. Then the exact same procedure is followed as mentioned above separately for male and female population. The data has been collected on male and female population of China separately by spanning from 1949 to 2020. The population of China is taken as the dependent variable. To find the model that best fits the dataset, the last two years data points (2019, 2020), also known as testing data points has been kept aside. Several models like linear, exponential, quadratic and logistic are fitted with the training dataset that is with the remaining dataset and the last two years male and female population (2019 and 2020) have been projected. The projected male and female population thus obtained is compared with the actual male and female population in 2019 and 2020 to find the best fit model.

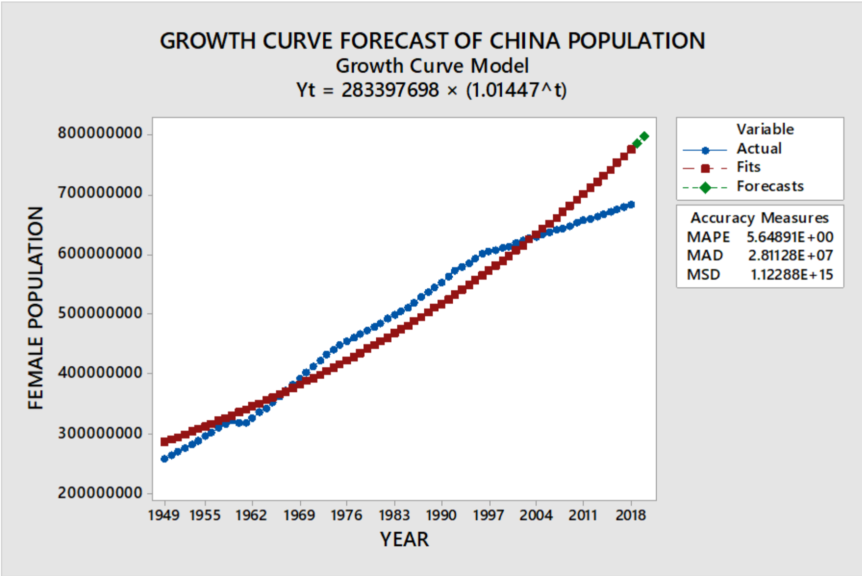
Finding out the best fit model for the female population of China

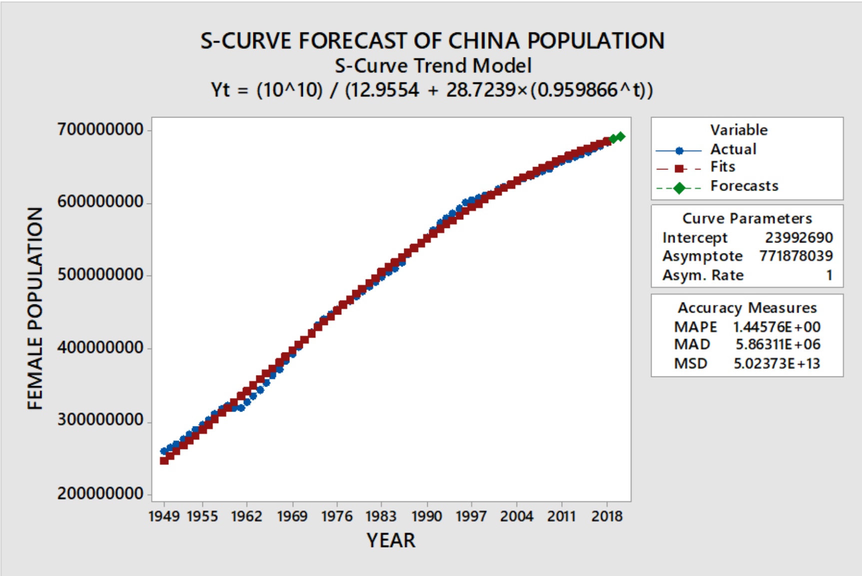
Trend analysis for the female population in China is done for the time period 1949 to 2018 to project the female population for the year 2019 and 2020. For finding the best fit model for the female China population, four models are considered namely Linear, Exponential, Quadratic and S-Curve. The corresponding graphs (Fig: 6.10) are shown below.

Fig 6.10: Trend Analysis Plots for females in China (using Linear, Quadratic, Exponential (Growth) and Logistic(S-Curve) model)









Thus Fig 6.10 depicts that among all the four models, S-curve model fits the female China population data (1949-2018) in the most appropriate way.

•Calculation and comparison of the percentage error of female China population

The percentage error of female population for the four different models is compared hereby. The table (Table 6.6) for percentage error of female population is given below.

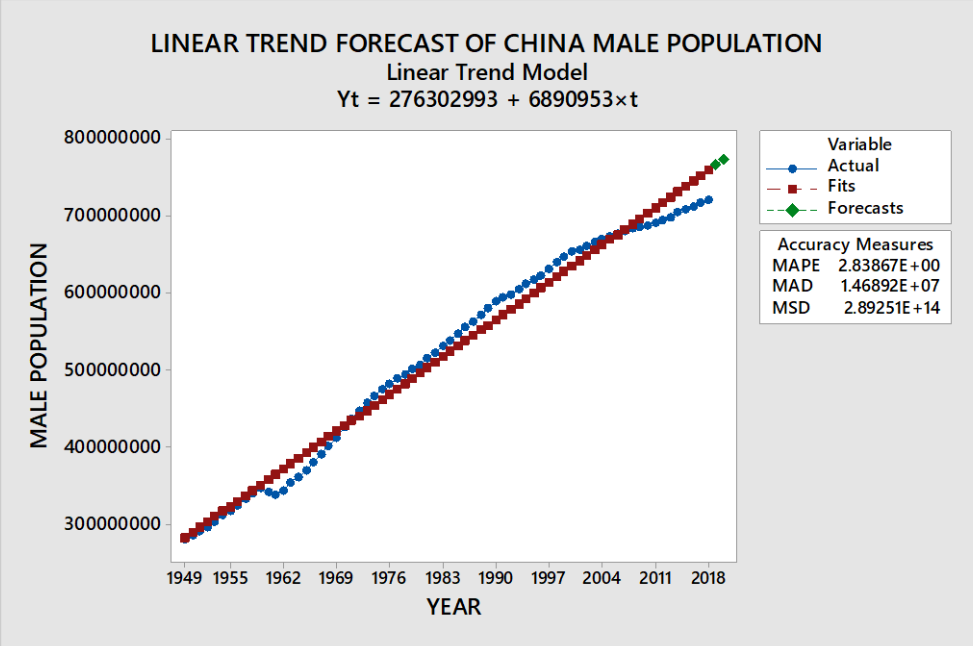
Table 6.6: Comparison of the percentage error of female China population

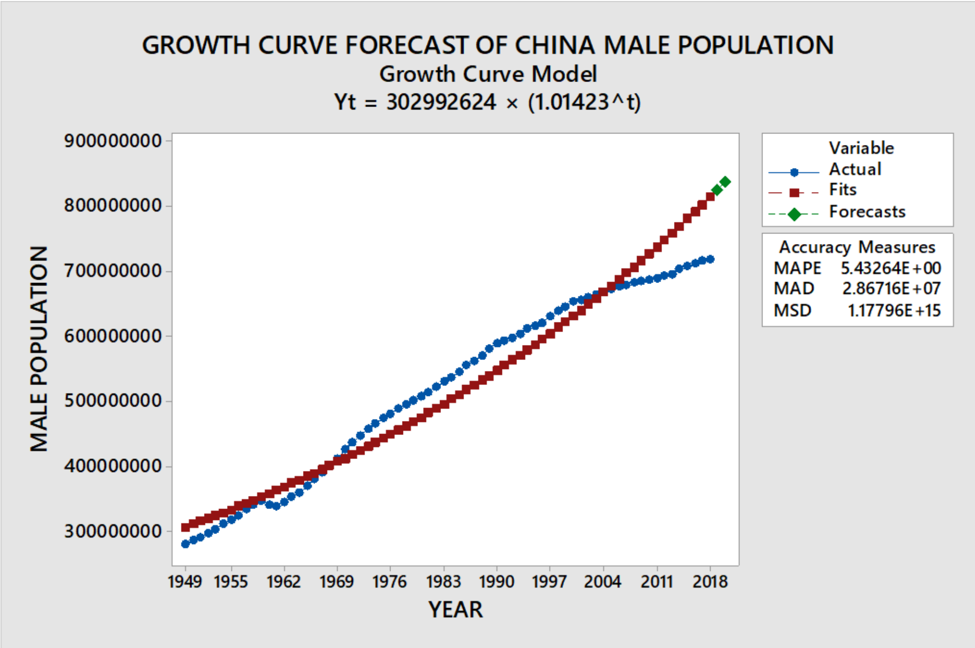
Model	Year	Observed Population	Forecast Population	% Error
Linear	2019	68,54,80,283	76234880	88.87862
	2020	68,72,45,874	732832035	-6.63317
Quadratic	2019	68,54,80,283	697037475	-1.686
	2020	68,72,45,874	701167330	-2.02569
Growth curve	2019	68,54,80,283	786067019	-14.6739
	2020	68,72,45,874	797443464	-16.0347
S-curve	2019	68,54,80,283	688568781	-0.45056
	2020	68,72,45,874	691564421	-0.62838

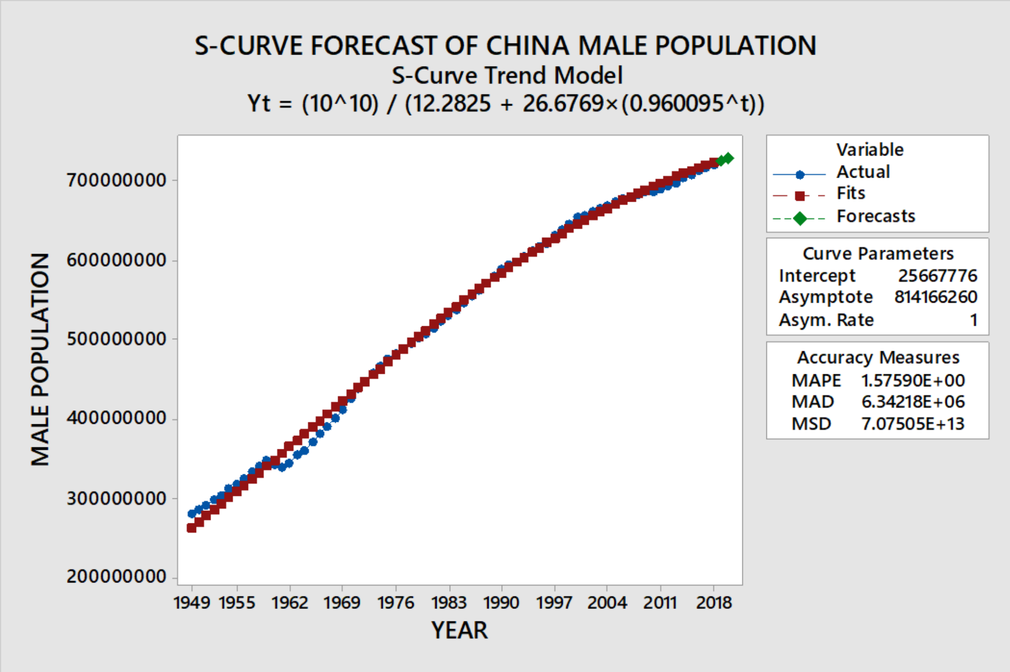
Finding out the best fit model for the male population of China

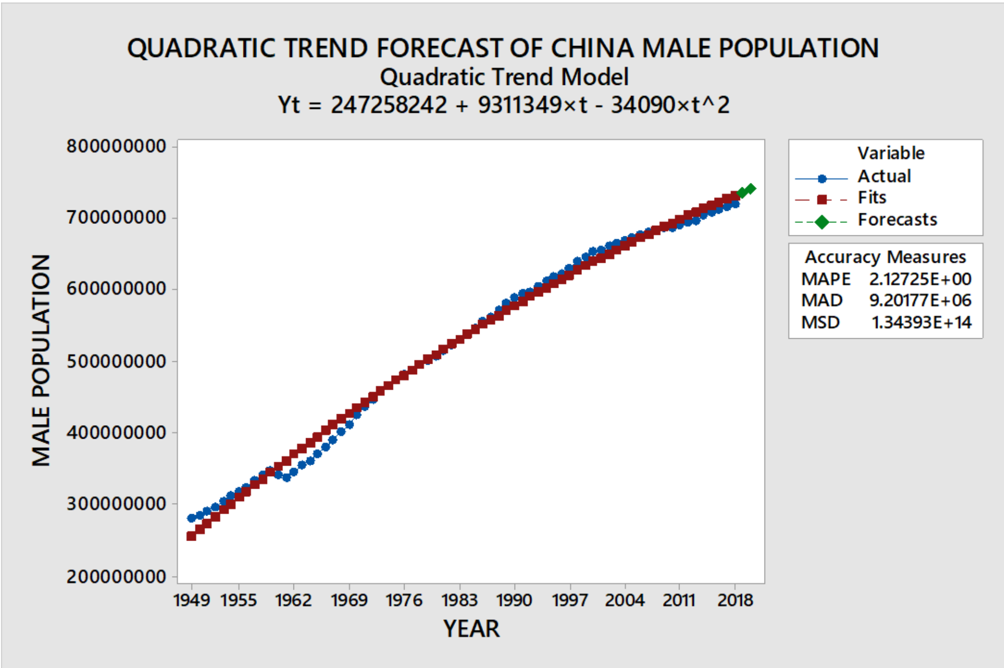
Similarly, trend analysis for the male population in China is done for the time period 1949 to 2018 to project the male population for the year 2019 and 2020. For finding the best fit model for the male China population, four models are considered namely Linear, Exponential, Quadratic and S-Curve. The corresponding graphs (Fig.: 6.11) are shown below.

Fig 6.11: Trend Analysis Plots for males in US (using Linear, Exponential (Growth), Logistic(S-Curve) and Quadratic trend model)









Thus Fig. 6.11 depicts that among all the four models, S-curve model fits the male China population data (1949-2018) in the most appropriate way.

•Calculation and comparison of the percentage error of male China population

The percentage error of male population for the four different models is compared hereby.

The table (Table 6.7) for percentage error of male population is given below.

Table 6.7: Comparison of the percentage error of male China population

MODEL	YEAR	OBSERVED POPULATION	FORECAST POPULATION	% ERROR
LINEAR	2019	72,22,64,717	765560647	-5.99447
	2020	72,36,83,488	772451600	-6.73887
QUADRATIC	2019	72,22,64,717	736515896	-1.97312
	2020	72,36,83,488	740952363	-2.38625
GROWTH CURVE	2019	72,22,64,717	826300245	-14.4041
	2020	72,36,83,488	838058979	-15.8046
S-CURVE	2019	72,22,64,717	726579457	-0.59739
	2020	72,36,83,488	729712055	-0.83304

From Table: 6.6 and Table: 6.7 it is obvious that the S-curve model for female as well as for male has the least % error among the other models.

A different method to find the best fit model for the dataset is by comparing various measures such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE) and Mean Square Deviation (MSD) for different models. The table below provides the values of MAD, MAPE, and MSD for various models taken into consideration for males and females separately.

Table 6.8: Simultaneous comparison of MAD, MAPE, and MSD for the four models

Model	MAPE		MAD		MSD	
	Male	Female	Male	Female	Male	Female
Linear	2.83867	2.84963	14689200	13903000	289251000000000	26179400000000
Quadratic	2.12725	1.97186	9201770	8182440	134393000000000	10530500000000
Exponential	5.43264	5.64891	28671600	28112800	117796000000000	11228800000000
S-curve	1.5759	1.44576	6342180	5863110	70750500000000	50237300000000

From Table 6.8 it is evident that the S-curve model has the minimum MAD, MAPE, and MSD among the four models for both males and females. So, the S-curve model fits in the best way with our dataset thereby claiming that the S-curve model is the best fit model.

China male and female Population Projection for 2025

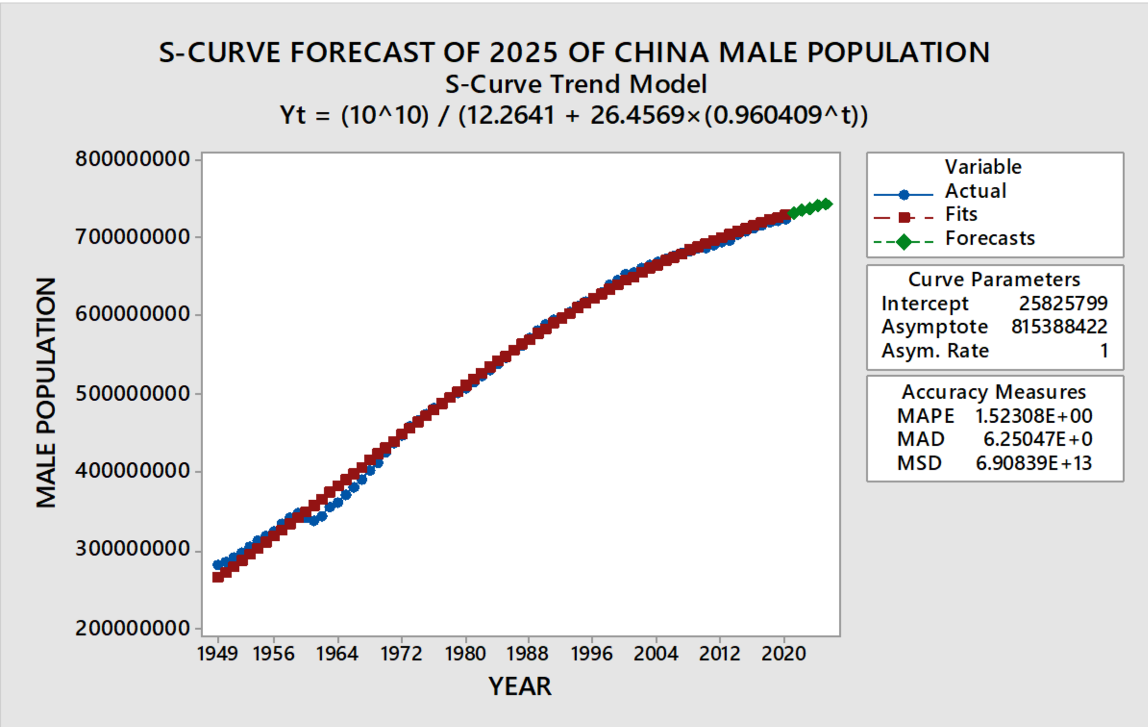
Now as the S-curve model has been established as the best fit model, the entire dataset including the last two years, that is, 2019 and 2020 is taken into consideration and the S-curve model is fitted to the dataset to project the China male and female population for the census year 2025. The fitted graphs (Fig: 6.12 and Fig: 6.13) for the 2025 China male and female population projection by the S-Curve model is given below. The China male population for the year 2025 is projected to be 743851659 and the China female population for the year 2025 is projected to be 703398404 by using the S-curve model.

The S-curve Model (Male):

$$Y_t = (10^{\wedge}10) / (12.2641 + 26.4569\times(0.960409^{\wedge}t)) \text{ .... (6.10)}$$

Where Y_t=China male population at time t

Fig 6.12: Trend Analysis Plots for males in China (using S-curve model) considering the data set from 1949-2020

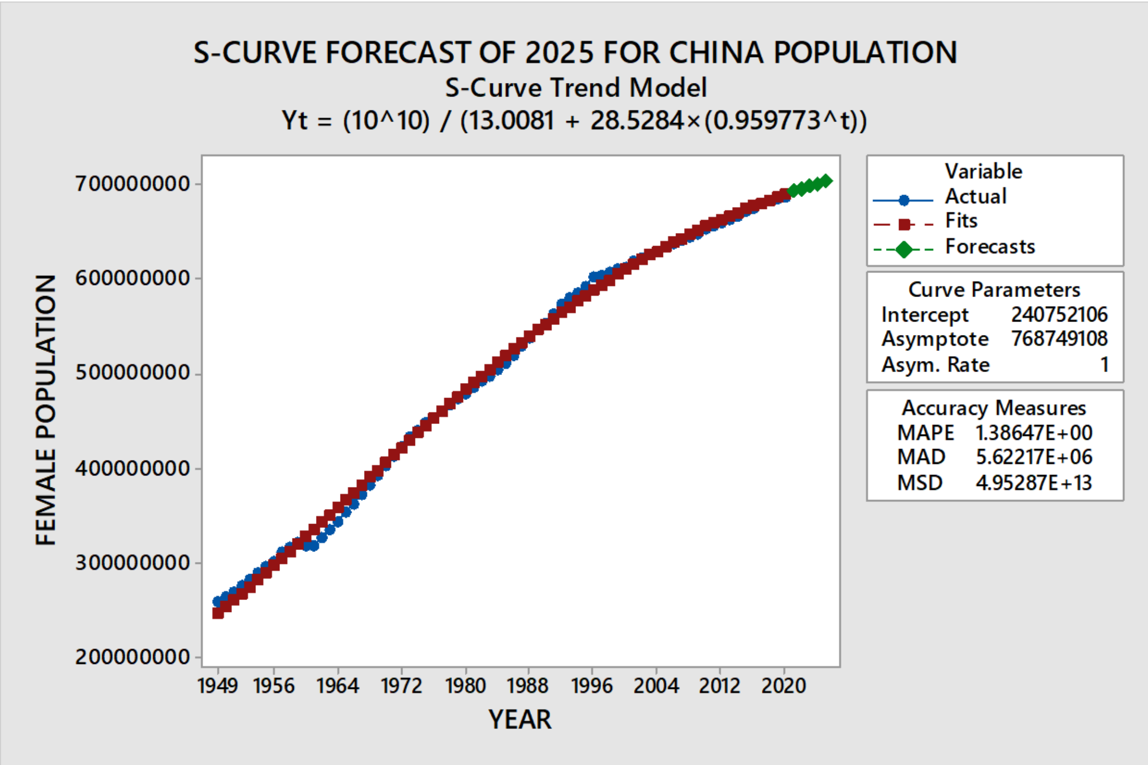


The S-Curve Model (Female):

Y_t = (10^10) / (13.0081 + 28.5284×(0.959773^t)) ………… (6.11)

Where, Y_t = China female population at time t

Fig 6.13: Trend Analysis Plots for females in China (using S-curve model) considering the data set from 1949-2020



•*Comparison of percentage error of the total population and the combined population with the observed population*

Here the percentage error of the total population is compared with the combined population (adding the male and female population) and is projected separately with the observed population which is represented in the following table (Table 6.9)

Table 6.9: Comparison of percentage error of the total population and the combined population with the observed population

	LINEAR		*EXPONENTIAL*	
Year	2019	2020	2019	2020
Observed Population	1,43,37,83,686	1,43,93,23,776	1,43,37,83,686	1,43,93,23,776
Projected Population by Approach 1	1526165624	1540229074	1655970621	1680466744
% Error	-6.44323	-7.0106	-15.4965	-16.7539
Projected Male Population	765560647	772451600	826300245	838058979
Projected Female Population	76234880	732832035	786067019	797443464
Projected combined (male + female) population by Approach 2	1491795447	1505283635	1612355184	1635488888
% Error of combined population	-5.97057	-6.68738	-14.5346	-15.9157

	QUADRATIC		*S-CURVE*	
Year	2019	2020	2019	2020
Observed Population	1,43,37,83,686	1,43,93,23,776	1,43,37,83,686	1,43,93,23,776
Projected Population by Approach 1	1474512567	1465569883	1442000205	1448118062
% Error	-2.84066	-1.8235	-0.57307	-0.611
Projected Male Population	736515896	740952363	726579457	729712055
Projected Female Population	697037475	701167330	688568781	691564421
Projected Combined (male + female) population by Approach 2	1433553371	1442119693	1415170620	1421299467
% Error of combined population	-1.83331	-2.21062	-0.52748	-0.73498

From Table 6.9 it is seen that the % error obtained on combining the male and female population is less than that of the total population.

Table 6.10: Comparison of 2025 China population projection by the two approaches

<table>
<tr><td colspan="2">Model</td><td colspan="2">2025 China Population Projection</td></tr>
<tr><td colspan="2">Total Population Model</td><td colspan="2">1467124054</td></tr>
<tr><td rowspan="2">Combined Model (Approach 2)</td><td>Model 6.10 (for male)</td><td>743851659</td><td rowspan="2">1447250063</td></tr>
<tr><td>Model 6.11 (for female)</td><td>703398404</td></tr>
</table>

Sex Ratio

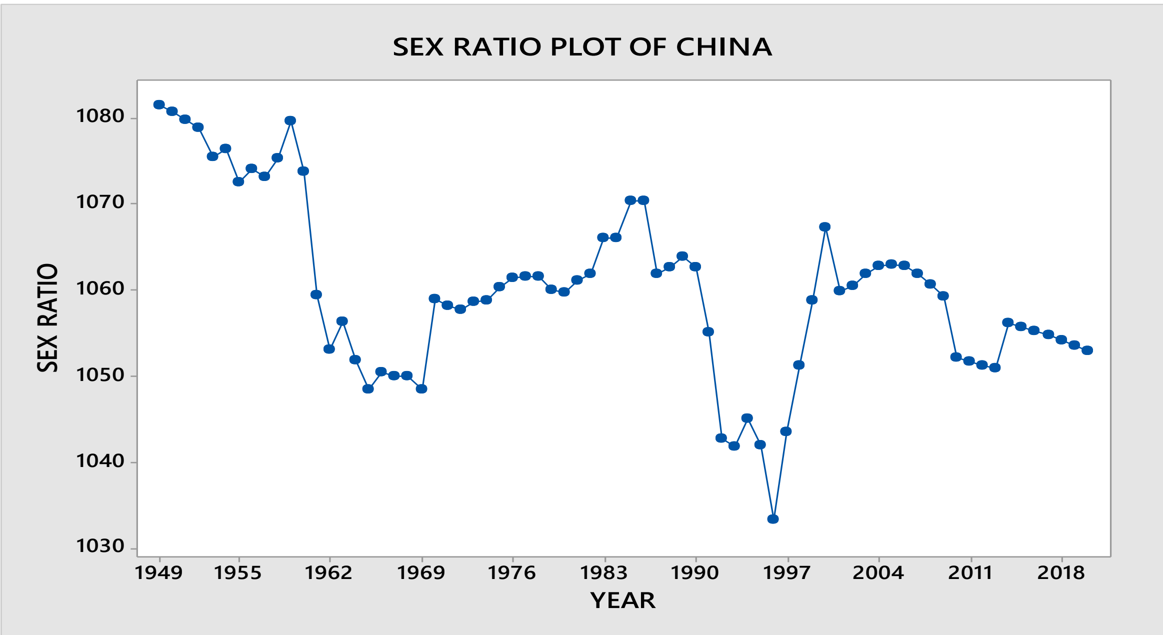
The sex ratio of a country is defined as the ratio of males to females in a population. For our given population of China (1949-2020), the sex ratio for each year has been obtained using the formula,

$$\text{Sex ratio}=\frac{\text{No.of males in the population}}{\text{No.of females in the population}}*100$$

and a time series plot is obtained of the sex ratio for the corresponding years.

The graph obtained is given below.

Fig. 6.14: Time Series Plot of sex ratio for China population (1949-2020)



From the above graph (Fig. 6.14), it is clearly seen that the sex ratio is mostly high, as seen in the period 1960-2000, with sometimes falling very low, as in 1962-1969 and 1992-1997, while sometimes reaching its peak, as in 1949-1959 or in 1985,1986,2000. But more or less, it remains constant within the range of value 1050 to 1070.

This signifies that the number of males per thousand females is very high, i.e., there is gender imbalance in the country. This indicates a serious and growing problem of missing female births in the country.

There are three possible explanations for these findings. The first is infanticide, the traditional method of disposing of unwanted births in feudal China and common in many premodern societies, which is very prevalent from the year 1959-1986 as seen from the graph. Second, the difference could be the result of abortions carried out after parents gained access to technologies to determine the gender of the foetus, which is seen from 2000-2020 in the graph. The third explanation is that the finding is the result of faulty statistical reporting: the missing female infants could have been safely born, and still living at home, but are now concealed by parents attempting to circumvent the national "one-child" family planning policy in their quest for a son. All three explanations imply important challenges to the health, safety, and welfare of girls and women in China.

But in the recent period from the year 2001, the graph is mostly showing a declining curve, which indicates a decrease in the sex ratio, that is, we can say that the number of females with respect to males has increased, thus, promoting gender balance. So, we can say that the government has certainly taken some useful measures to resolve this problem.

CONCLUSION

The results in this project work are based on the 2025 China population projection by demographic traits including sex.

• By approach 1, the S-Curve trend model has been established as the best fit model for the entire China population.

From equation (6.5), the S-Curve Model is:

$$Y_t = (10^{10}) / (6.23297 + 14.0777\times(0.958819^t))$$

Where, Y_t = China population at time t.

The total population projection for 2025 is obtained as 1467124054.

• By approach 2, the S-Curve trend model has been established as the best fit model for the male China population as well as for the female China population.

From equation (6.10), the S-Curve Model (Male) is:

$$Y_t = (10^{10}) / (12.2641 + 26.4569\times(0.960409^t))$$

Where, Y_t = China male population at time t.

From Fig. (6.11), the S-Curve Model (Female) is:

$$Y_t = (10^{10}) / (13.0081 + 28.5284\times(0.959773^t))$$

Where, Y_t = China female population at time t

The male population projection for 2025 is obtained as 743851659 and the female population projection for 2025 is obtained as 703398404. Therefore, on combining the male and female China population, 1447250063 is obtained as the total population projection for 2025. From the 2025 China population projection by the above two approaches, it is quite clear that both the models are robust in nature. However, precisely looking into the percentage errors of the testing dataset, the combined model (model for male China population + model for female China population) gives a better projection than that of the total population model.

Hence, the combined model (model (6.10) and model (6.11)) is the best model for the 2025 China population projection with the least percentage error.

•Lastly, we have also calculated the sex ratio as the number of males to number of females in the country and a time series plot is done. From the graph, we have seen the sex ratio is mostly high varying in between the range of 1050-1070. This signifies that there is gender imbalance in the country and the government must immediately take strict policies to solve the problem. But in the recent years, as from 2001 it seems the government did take certain measures as the curve is a declining one indicating balance towards gender equality in the upcoming years.

ACKNOWLEDGEMENT

The project work would not have been possible without the support of the professors of my Statistics Department. I would begin by thanking my supervisor, Dr. Sujan Chandra, without whose guidance, it would not have been possible to complete this project work. He has been my pillar from the very beginning of this project. From helping me channelize my thoughts to giving me innumerable feedbacks, he has put unimaginable efforts in making my project paper successful. Without his valuable remarks and comments, my project work would have never met with completion. He stood by me in every single doubt, in every way possible and his contribution to my project work is ineffable. I am really grateful to you, Sir.

Apart from my supervisor, I would definitely like to thank the Head of Department, Dr. Ajoy Kumar Biswas, for his cooperation. I am highly obliged to all these people for their great insights and contribution in making my project an insightful one.

REFERENCES

1. https://www.worldometers.info/world-population/china-population/
2. https://countryeconomy.com/demography/population/china
3. https://www.jstor.org/stable/1972529
4. Recent trends in sex ratios at birth in China, Terence H. Hull
5. https://www.visualcapitalist.com/the-population-of-china-compared-with-the-rest-of-the-world/
6. https://en.wikipedia.org/wiki/Land_use_in_China
7. Fundamentals of Statistics (volume two) by A.M.Gun, M.K.Gupta, B.Dasgupta.