EE300 Project Final

Simulation & Prediction of Population Growth in China

Leyu,Lin
EE300
San Diego State University
San Diego, United Sate

***Abstract*—This project will analyze trend of growth rate of China population and predict the next year of China population via three models: Polynomial model, Logistics model, exponential model and its seems logistics provide the most close result to the data. Please do not copy my code .All code will be will well commented.**

All files also provide in my git-hub please scan code on right.

I. Introduction

In this project, I will analyze the population growth of China. Since China has tremendous amount of people. I am curious about the trend of next year of the population. And the growth rate of population of each year. I research on Google scholar and Find out 10-50 yrs data of China population from the official china statistics government website (The links in reference) . I try three times for this project to find the best fit model for this population data. And the result turns out is the logistic model. The first approach is polynomial model and it very hard to fit and it looks distort and fit part of the data or few years. Then my second approach is use exponential model to match the data. This approach much better than the first approach. However it seems population keeps growth without stop. Then the last approach I struggled a bit and I got recommends form Interest to use logistic model. The result turns out its very good. Finally, I used this model for predicting next few years population in China. More details will be discussed in Simulation& data analyze section.

II. Data

Total Population(10^9)(05-14)	13.08	13.14	13.21	13.38	13.34	13.39	13.47	13.54	13.61	13.68
Sum Population(10k)	136782	136072	135404	134735	134091	133450	132802	132129	131448	130756
male	70079	69728	69395	69068	68748	68647	68357	68048	67728	67375
female	66703	66344	66009	65667	65343	64803	64445	64081	63720	63381
Urban	74916	73111	71182	69079	66978	64512	62403	60633	58288	56212
Rural	61866	62961	64222	65656	67113	68938	70399	71496	73160	74544

Table. I. The total population and majority population areas data [4]

1959-1968 population sum (unit:10^9)	6.72	6.62	6.59	6.73	6.91	7.04	7.25	7.45	7.63	7.85
1969-1978	8.07	8.30	8.52	8.71	8.92	9.09	9.24	9.37	9.5	9.63
1979-1988	9.75	9.87	10.01	10.17	10.30	10.44	10.59	10.75	10.93	11.10
1989-1998	11.27	11.43	11.58	11.71	11.85	11.98	12.11	12.23	12.36	12.47
1999-2008	12.57	12.67	12.76	12.85	12.92	13.00	13.08	13.14	13.21	13.38

Table. II. The total population 1959-2008 data in units of hundred million [4]

1959-1968 population growth rate	0.1350	0.1150	0.0575	0.1783	0.1517	0.1667	0.2133	0.1858	0.1983	0.2225
1969-1978	0.2258	0.2283	0.2025	0.2008	0.1933	0.1583	0.1383	0.1283	0.1308	0.1250
1979-1988	0.1175	0.1283	0.1542	0.1458	0.1317	0.1450	0.1542	0.1717	0.176	0.1700
1989-1998	0.1658	0.1558	0.1383	0.1342	0.1358	0.1300	0.1242	0.1258	0.1217	0.1033
1999-2008	0.1000	0.0950	0.0908	0.0792	0.0733	0.0825	0.0692	0.0633	0.0750	0.0700

Table. III. The population growth rate from 1959-2008 [4]

III. MATH

A. Equations

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \quad \text{E01}$$

$$f'(x_j) \approx \frac{1}{2h}[-3f(x_0) + 4f(x_1) - f(x_2)] \quad \text{E02}$$

$$f'(x_2) \approx \frac{1}{2h}[f(x_0) - 4f(x_1) + 3f(x_2)] \quad \text{E03}$$

$$f'(x_j) \approx \frac{1}{12h}[f_{j-2}) - 8f_{j-1} + 8fx_{j+1} - f_{j+2}] \quad \text{E04}$$

$$\varepsilon = (actual\_val - result\_val) / actual\_val \quad \text{E05}$$

$$x(t) = x_0 e^{rt} \quad \text{E06}$$

$$\begin{aligned} &y = \ln(EQ6) => y = a_1 t + a_2 \\ &\ln(EQ6) => r = a_1, x_0 = e^{a_2} \end{aligned} \quad \text{E07}$$

$$\begin{cases} x(0) = x_0 \\ \dfrac{dx}{dt} = rx(1 - \dfrac{x}{x_m}) \end{cases} \quad \text{E08}$$

$$x(t) = \frac{1}{1 + (\frac{1}{x_0} - 1)e^{-rt}} <=> \frac{a_1}{1 + (\frac{a_1}{x_0} - 1)e^{-a_2 t}} \quad \text{E09}$$

IV. SIMULATION & ANALYZISIS

Since it was tremendous amount of data. I was try to change the unit and scale the data. Then I use Hundred million as unit instead of 10k got Table.II. I curious about the whats population of next few years. My first approach was using the polynomial up to 9th terms to model the population. My result is figure.I (this figure is model of mount of the population) and II. Figure 1 seems hard to analyze and look not right. So I calculate the growth rate of each year try other way to analyze the data. Figure 2 is based on the growth rate.

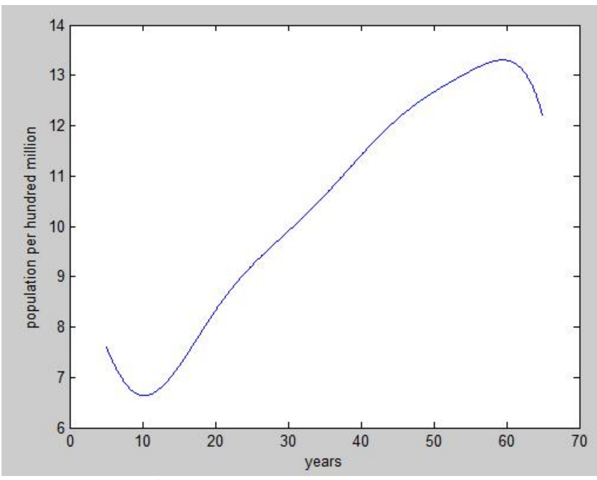

FIGURE.1. The 9th term polynomial model of the Population of Table.I result

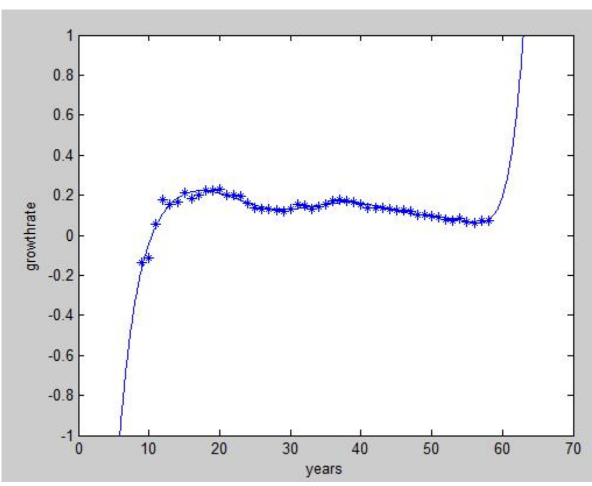

FIGURE.2. The curvefit of discrete value plot of Table .III polynomial model of Growth rate of Population.

In figure. II. The graph is the model result of growth rate of population and compare with the real data the Table.III. as we can see population growth is stable form 20 which is 1970 to 60 which is 2010.In the Table.I the population nearly 13.345. Which is the actual_val. Use Eq5 to estimate the difference is 0.03243. It is very close to actual data. However, after it the slope is very high. This obvious wrong. The population cant growth like that. This first approach modeling failed. (for this approach EQ1-5 was used and code details in CODE section).

Second approach I try to use exponential model with EQ6 and EQ7 and then get figure 3 result.

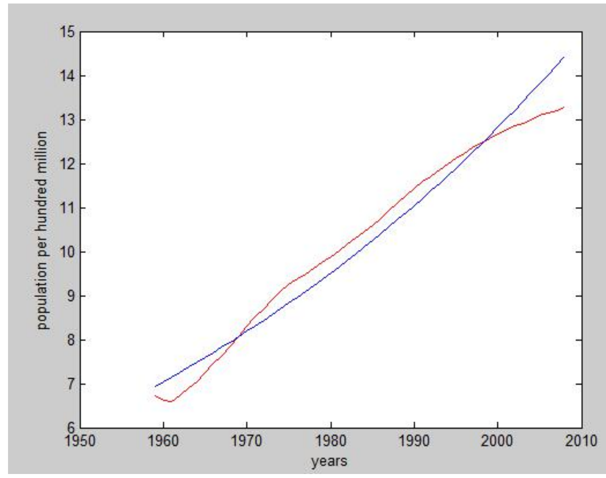

FIGURE.3. The exponential model of Population of Table.II.

This approach seems much better than previous model. The population growth nearly linear. Then I try estimate whats population after 2008. For example 2020 so I enter t =2020 for my function. The result is

```
Columns 2006 through 2010

  13.1400    13.2100    13.2800          0    16.4486
```

FIGURE.4. The mat-lab estimation result of exponential model at 2010.

According to figure 4 and figure 3 the 2010 estimated population is 16.4486. This seems growth too fast. We can clearly see the right ends Our model (blue line) is much greater than the actual data in Table.I. This much greater than the actual data which is 13.39 in Table.I first row last four column. I think the actual population should be limiter by some factors such as food resource or spaces. According professional journal, “based on GIS system. necessary to consider these factors separately when modeling population density using DMSP/OLS data” [2] This based on GIS so is much accurate than my data. And they consider a lot factors. I can believe we do need consider other factors when we analyze the population data. This model still seem not good enough for this population data. However, it is better than the first approach .

Third approach I see the best model recommendation on the Internet is logistic model. Therefore in this last approach I try use logistic model to do the simulation. According professional research “we set up a hierarchical model on the K vectors of parameters,K. The hierarchical model allows us to estimate the population variability and also improves the estimates for the individual subjects.” [3] They set up a new vector K; however, my model may not as much as complicated as theirs but one thing can be confirmed is that logistics model is the great model to analyze population because their research result very close to actual data. I use EQ8 for this simulation the result is figure.5

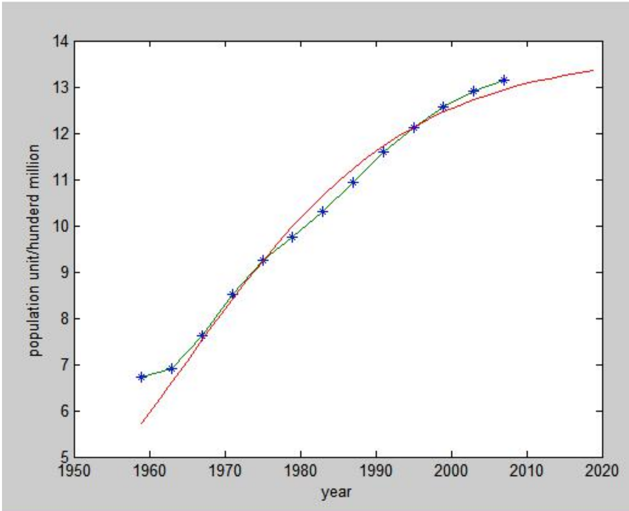

FIGURE.5. The logistic model of population with discrete plot of Table.II And prediction of 2008 to 2020.

Wow! This model so far is the best compare with other two models. Compare with table.I data from 2005 to 2014 the first row. 2005 is the first year china has 13 hundred million people This also a well-know fact. According to figure 5 this looks very close close form 2005 to 2014 between 13 and 14. And we can see after 2014 the curve is saturated. This is exactly we expect to see and this is also the drawbacks of second model.

Now we have our great model and data set lets compared with those in details.

V. Data & Model comparison

Use EQ5 to check the difference. First lets compare 2014 between model and data. Then predict 2020.

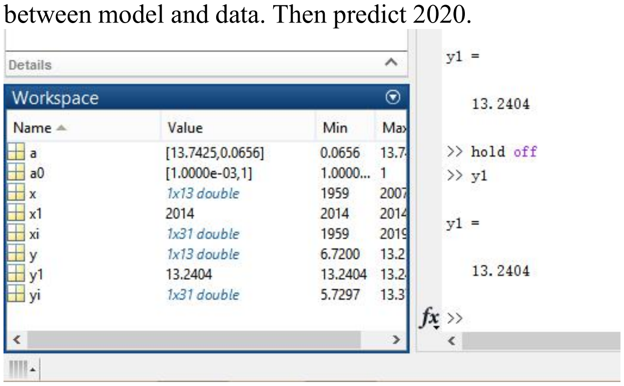

FIGURE.6.The mat-lab stimulation result of values of logistic model

As we can see 2014 population prediction of matlab result is 13.2404 (result_val). According to Table.I first row last column is 13.68 (actual_val). Lets use equations get difference

$$\varepsilon = (13.68 - 13.2404) / 13.68 \approx 0.032 \qquad \text{EQ5}$$

According to figure.V the highest different is after 2010 so the rest years should have difference less tan 0.032. I think is close but not prefect. But it is good enough for my prediction So I enter x2= 2020 to get 2020's population

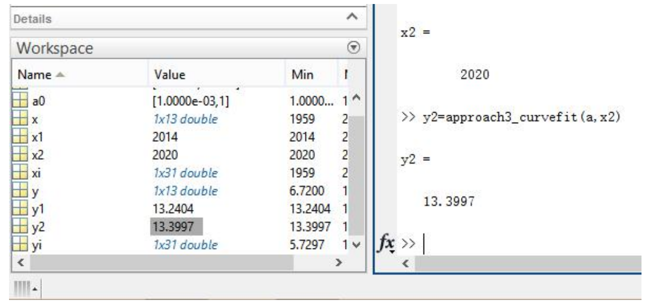

FIGURE.7.The mat-lab stimulation result of values of logistic model of prediction of 2020

My result is nearly 13.4 so 2020's population is around this number. This seems kind of lower than I expected.

VI.CODE

```
%For Table II. Is from 1959 to 2008 so I use set up t to present those year This is first approach.
t = 1959 :2008
% the population p is those values on Table.II
p= [6.72 6.62 6.59 6.73 6.91 7.04 7.25 7.45 7.63 7.85 8.07 8.30 8.52 8.71 8.92 9.09 9.24 9.37 9.5 9.63 9.75 9.87 10.01 10.17 10.30 10.44 10.59 10.75 10.93 11.10 11.27 11.43 11.58 11.71 11.85 11.98 12.11 12.23 12.36 12.47 12.57 12.67 12.76 12.85 12.92 13.00 13.08 13.14 13.21 13.28]
% make the year scale start from 1950
t = t -1950
%define initial values lets say 1for 50 yrs 1959 -2008
v = ones (1,50)
%Then model with polynomial 9 times
m = polyfit (t,p,9)
% more accurate  points
n =linspace (5,65,1000)
% seek for the result
r = polyval (m,n)
% plot and label also hold the plot
plot (n,r)
xlabel ('years'),ylabel('population per hundred million')
% now type equation in for two ends x0 and x2 first end is 0 1 2 last end is 48 49 50 three terms in EQ2 & 3 create function also EQ4  for rest part file equations.m
% 50years
for n = 1:50
%first end EQ2
if n<=2
v(n) = (-3*p(n)+4*p(n+1)-p(n+2))/(2*(t(n+1)-t(n)))
%last end EQ3
else if n<=48
 v(n) = (p(n-2)-8*p(n-1)+8*p(n+1)-p(n+2))/(12*(t(n)-t(n-1)))
%other case
    else
        v(n) = (3*p(n)-4*p(n-1)+p(n-2))/(2*(t(n)-t(n-1)))
    end
```

```
  end
end
%now ready to plot
disp ([t',v'])
% discrete plot as stars
plot (t,v,'*');
xlabel ('years'),ylabel('growthrate')
% now we can model with our equation see how well it looks
hold on
a = polyfit(t,v,9)
T=linspace(5,65,1000)
V= polyval(a,T)
plot(T,V)
%limited axis X form 0to 70 y from -1 to 1
axis([0,70,-1,1]);
%clear for next model
clc
clear all
% the exponential model so the year is the same 1950 to 2008
t = 1959:2008
% enter the value of the function the mount of population
x(t)= [6.72 6.62 6.59 6.73 6.91 7.04 7.25 7.45 7.63 7.85 8.07
8.30 8.52 8.71 8.92 9.09 9.24 9.37 9.5 9.63 9.75 9.87 10.01
10.17 10.30 10.44 10.59 10.75 10.93 11.10 11.27 11.43 11.58
11.71 11.85 11.98 12.11 12.23 12.36 12.47 12.57 12.67 12.76
12.85 12.92 13.00 13.08 13.14 13.21 13.28]
% enter EQ7 EQ6
y = log(x(t)), a=polyfit(t,y,1), r=a(1),x0=exp(a(2))
x1=x0.*exp(r.*t)
%ready to plot the red line is data the blue line is the function
Lets use these two colors here. More easier to compare
plot(t,x(t),'r',t,x1,'b')
%label
xlabel ('years'),ylabel('population per hundred million')
%for 2010 estimation r=0.015 x0=1.3248e-12 and we have
completer equation
t =2010
x0 = 1.3248e-12
x(t)=x0*exp(0.0150*t)
% The last model I will sample every 4 yrs from 1959until
2007 so the data will be 6.72 6.91 7.63 8.52 9.24 9.75 10.30
10.93 11.58 12.11 12.57 12.92 13.14
x=1959:4:2007
y=[6.72 6.91 7.63 8.52 9.24 9.75 10.30 10.93 11.58 12.11
12.57 12.92 13.21]

%function file approach3_curvefit
function f=approach3_curvefit ( a,t )
%Based EQ9
   f=a(1)./(1+(a(1)/3.9-1)*exp(-a(2)*(t-1950)));
end
%plot those data as discrete value and connect as a line
plot(x,y,'*',x,y)
%hold it later for ploy our model
hold on
% set up a0 initial value
a0=[0.001,1]
%non-linear curve-fit the function file is at bottom
a= lsqcurvefit('approach3_curvefit',a0,x,y)
%display the result
disp(['a=' num2str(a)])
%predict from 1959 to 2020 every 2years
xi=1959:2:2020
%pass value into the function get result
yi=approach3_curvefit(a,xi)
%ready for plot
plot (xi,yi,'r')
%label
xlabel ('year'),ylabel ('population unit/hunderd million')
```

REFERENCES

[1]. Refer to TABLE.I TABLE.II DATA.I [2].based on GIS system. necessary to consider these factors separately when modelling population density using DMSP/OLS data." [3]."we set up a hierarchical model on the K vectors of parameters,K. The hierarchical model allows us to estimate the population variability and also improves the estimates for the individual subjects." ". after log-transformation and appropriate scaling, are modeled as normal with a population mean "

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] L. Zhuo, T. Ichinose, J. Zheng, J. Chen, P. J. Shi, and X. Li, "Modelling the population density of china at the pixel level based on DMSP/OLS non - radiance - calibrated night - time light images," International Journal of Remote Sensing, vol. 30, no. 4, pp. 1003 - 1018, Feb. 2009

[3] A. Gelman, F. Bois, and J. Jiang, "Physiological Pharmacokinetic analysis using population modeling and informative prior distributions," Journal of the American Statistical Association, vol. 91, no. 436, p. 1400, Dec. 1996.

[4]"National bureau of statistics of china >> annual data," in NBC, 1996. [Online]. Available: http://www.stats.gov.cn/english/Statisticaldata/AnnualData/. Accessed: Nov. 10, 2016.