## Project: Population Regression

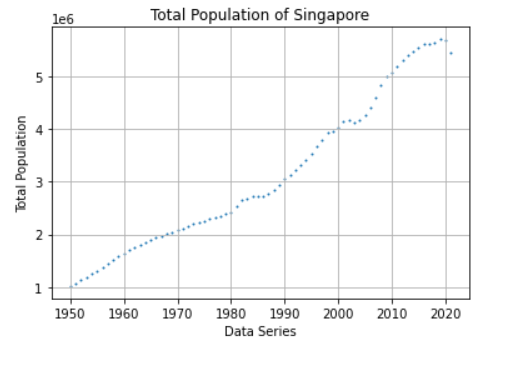
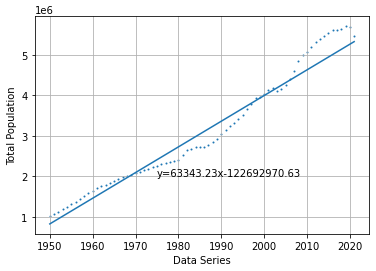
**Group Members: Yang Jingyuan, Zhao Yuan, Zhang Haonan, Chen Yuxuan, Jiang Fengcheng,Du Jiacheng, Chen Yunquan, Shi Jianchuang**

1. **Introduction**

In this project, we will look at the human population statistics collected by the various national governments and build a machine learning model to make population predictions.

1. **Data Source: Singapore Department of Statistics (SingStat)**

Let’s start with looking at the population statistics of Singapore. Download Singapore population data from 1950 to 2019 from: https://www.singstat.gov.sg/

* 1. Graph the total population vs year. 
  2. Use linear regression to build an estimator of the total population of Singapore in the future. Use the data for years 2013 and earlier as training data. (figure1)
  3. Performance metrics:
     1. What are the slope and y-intercept of the best fit line? Plot the best fit line over the empirical data.

**The slope is 63343.23**

**The y-intercept is -122692970.63**

* + 1. What is the *R2* coefficient for the best fit line? See Appendix for definition of the *R2* coefficient.

**The *R2* coefficient is 0.9614714944982808**

* + 1. What is the mean squared error (MSE) of the estimator on the training data?

**The MSE on the training data： 55496493545.96**

**Normalized MSE for train data： 0.03652981849117039**

* + 1. Use years greater than 2013 as test data and predict the population for those years.
    2. What is the MSE of the estimator on the test data? Hint: you may want to normalize the mean squared error for it to be meaningful.

**The MSE on the testing data： 257594689765.74**

**Normalized MSE for test data： 0.04254007327940072**

* 1. What is your estimate of Singapore’s population in 2030 and 2050? Do you think these estimates are reasonable? Explain your answer.

**Estimated total population of Singapore in 2030 is 5893779**

**Estimated total population of Singapore in 2050 is 7160643**

**No. Changes in population are influenced by many factors that are difficult to predict and quantify. Linear regression model cannot be perfect. Although ordinary linear regression can fit the regression curve, it is too coarse and will appear underfitting phenomenon.**

**This model does not take into account objective factors.**

* **The number of people of childbearing age**
* **People's willingness to have children**
* **The cost of raising children**
* **Ideology**
* **Some unexpected events(COVID-19)**
  1. What pattern do you expect for human population growth in Singapore?

**Modern Type**

**Basic features:**

**Birth rate of the population: began to decline significantly**

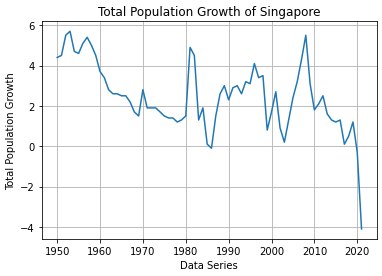
**Mortality rate: continued to decline and reached a low level**

**Natural growth rate: gradually decreases, and some developed countries even begin to experience negative or zero growth.**

**Pros: Effectively alleviate the trend of rapid population growth**

**Cons: Lack of social labor force: which is not conducive to industrial development and the economic development of the country; Population aging: increasing the burden of social welfare**

**Distribution: Most developed countries.**

(figure 2)

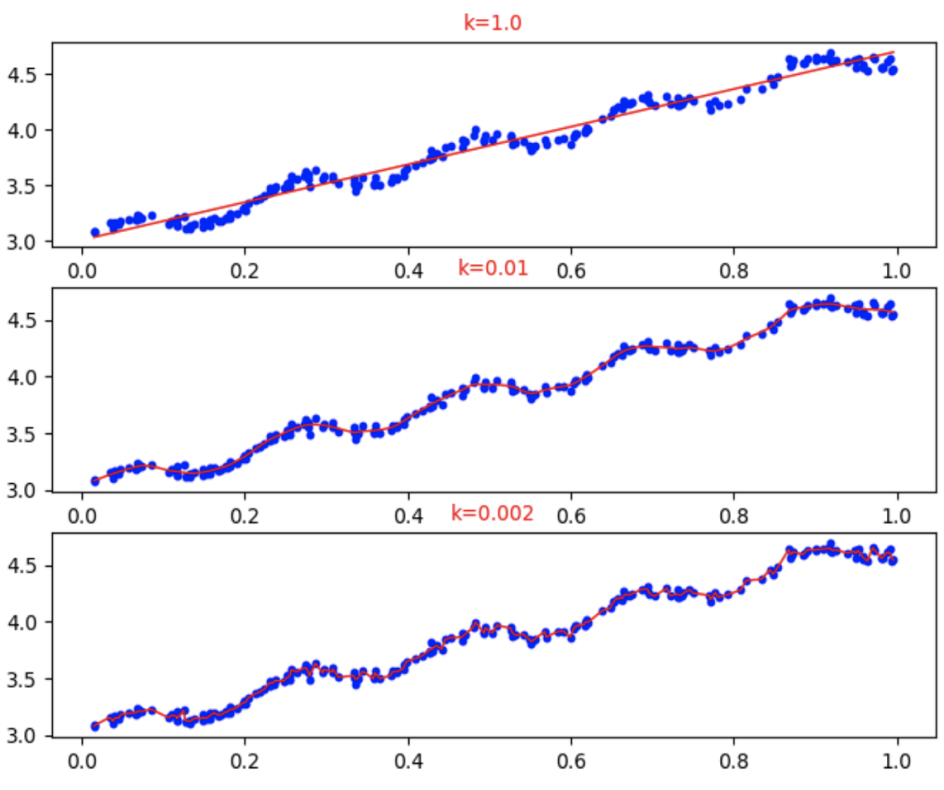
**Now that people's basic life has been guaranteed, future population growth will be more related to the fertility policy of government and unpredictable events like COVID-19, so this linear fitting result I think has little reference value.**

**We think Singapore's population growth model is fluctuating upward. Growth rate of it is constantly fluctuating every year, but judging from the curve we fit in Figure 1, the total population of Singapore has been rising steadily, from 1950 to now.**

* 1. How could you improve your estimates of the future population?

**Locally weighted linear regression (LWLR) reduces the mean square error of the prediction by introducing some bias in the estimates. The basic idea of this method is to assign a certain weight to each point near the point to be predicted, and store these weights in a new matrix W.**

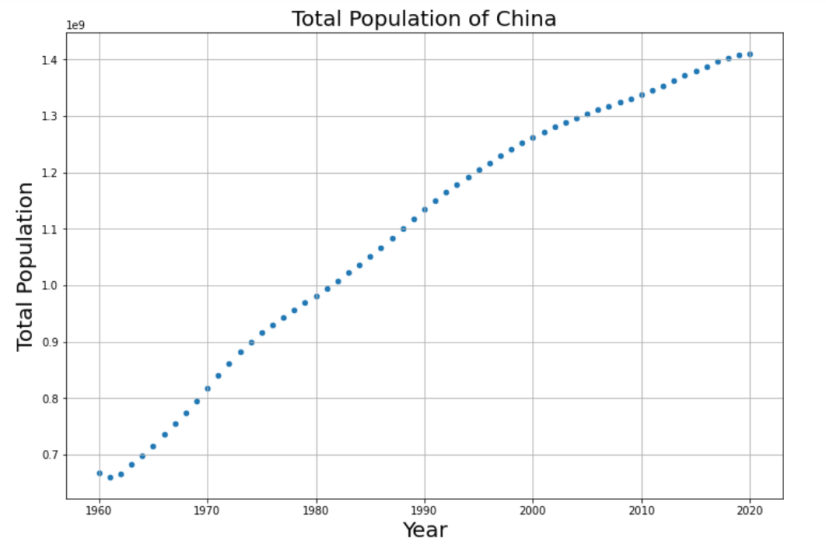
**Better prediction results can be obtained by local weighted regression, but adjusting the parameter k is the key, and k is too small may appear overfitting phenomenon.**



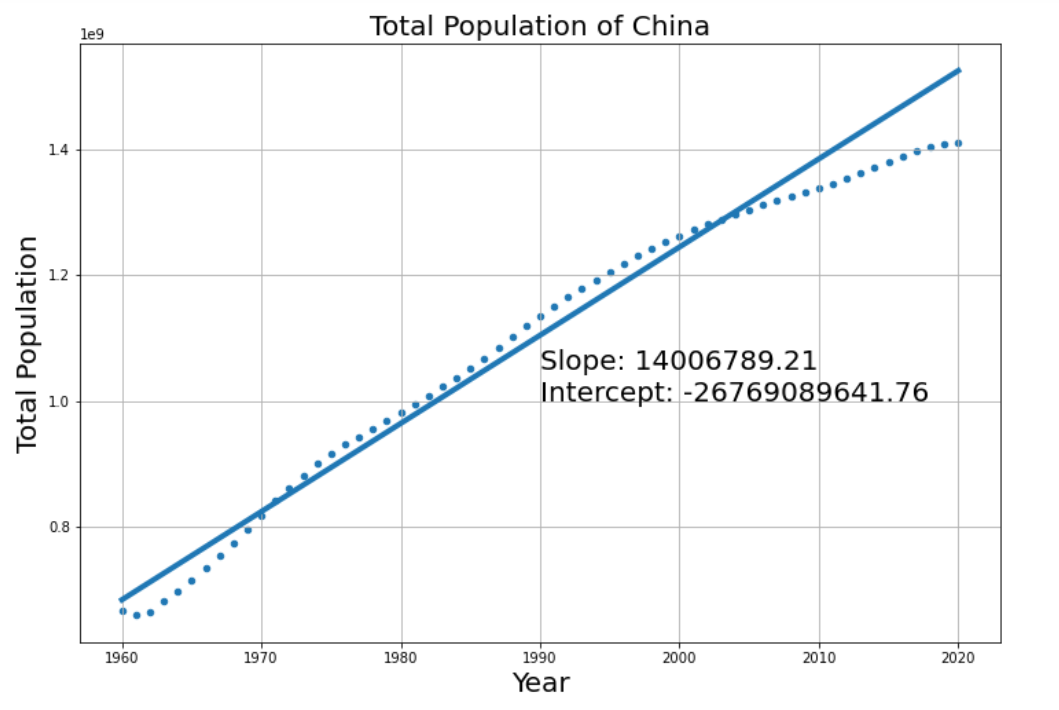
1. **Data Source: The World Bank (https://www.worldbank.org/)**

Download population data from: https://www.worldbank.org/. Note that you will have to search for the data from the World Bank website (sp.pop.totl).

* 1. You should be able to get an excel file with the population of every country from 1960 to 2019. First, verify that the data from the World Bank matches the Singapore population data you previously downloaded from SingStat.gov.sg.
  2. Use the total population data for China. Graph the total population vs year.



* 1. Use linear regression to build an estimator of the total population of China in the future. Use the data for years 2013 and earlier as training data.



* 1. Performance metrics:
     1. What are the slope and y-intercept of the best fit line? Plot the best fit line over the empirical data.

**The slope is 14006789.21**

**The y-intercept is -26769089641.76**

* + 1. What is the *R2* coefficient for the best fit line?

**The *R2* coefficient is 0.9685213495544637**

* + 1. What is the mean squared error (MSE) of the estimator on the training data?

**The MSE on the training data: 811985540736408.50**

**Normalized MSE for train data: 0.01624667590770196**

* + 1. Use years greater than 2013 as test data and predict the population for those years.
    2. What is the MSE of the estimator on the test data? Hint: you may want to normalize the mean squared error for it to be meaningful.

**The MSE on the testing data is 8088282817982782.00**

**Normalized MSE for test data: 0.08990166993014194**

* 1. What is your estimate of China’s population in 2030 and 2050? Do you think these estimates are reasonable? Explain your answer.

**Estimated total population of China in 2030 is 1590962712**

**Estimated total population of China in 2050 is 1833640011**

**No. Because the linear model does not take into account objective factors.**

1. **The number of people of childbearing age**
2. **People's willingness to have children**
3. **The cost of raising children**
4. **Ideology**
5. **Some unexpected events**
   1. What pattern do you expect for human population growth in China?

**By reviewing the relevant data, we can see that for China now, on the one hand, the uncertainty of future population growth has increased with the emergence of factors such as the aging of the population and the gradual decrease of the fertility rate.**

**On the other hand, according to the data, the maximum population carrying capacity of China should be controlled at about 1.6 billion, and the current population of China is already close to this carrying capacity, therefore, the future population growth rate is bound to become slower and slower, or even a trend of zero growth or negative growth.**

**(Reference: Study on the Productive Capacity of Land Resources and the Carrying Capacity of Population in China)**

* 1. How could you improve your estimates of the future population?

**We can apply** **Locally weighted linear regression (LWLR) and Time Series Model(ARIMA) discussed in 2.f.**

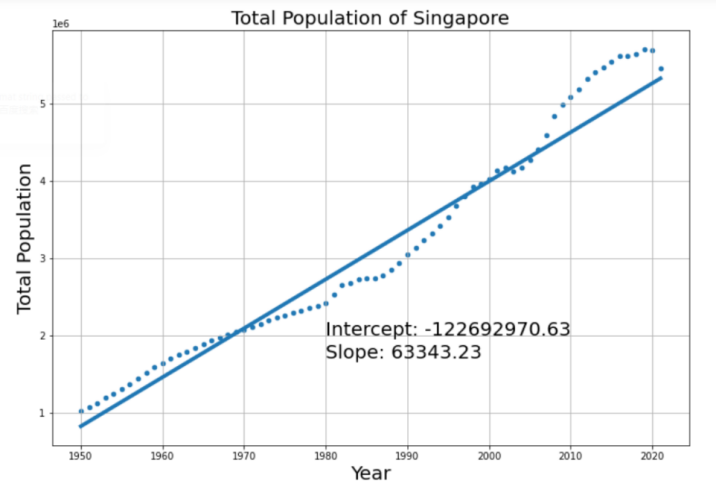
1. **Conclusion**

In this project, we use linear regression to build an estimator of the total population of Singapore in the future.

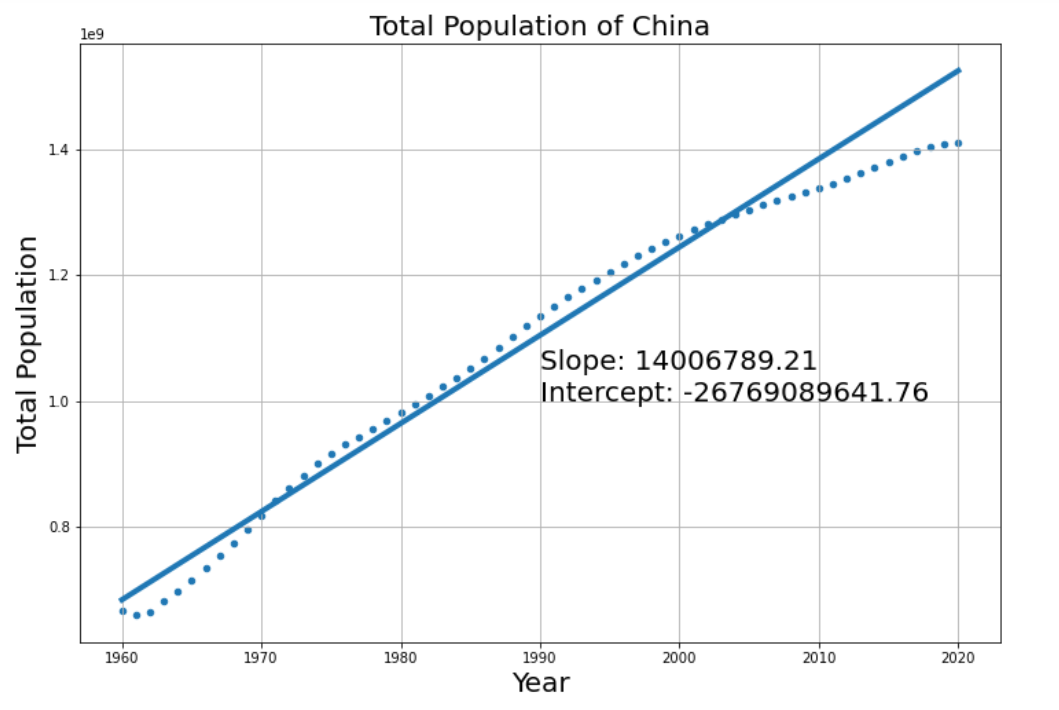
Data before 2013 : training data

Remaining data : testing data

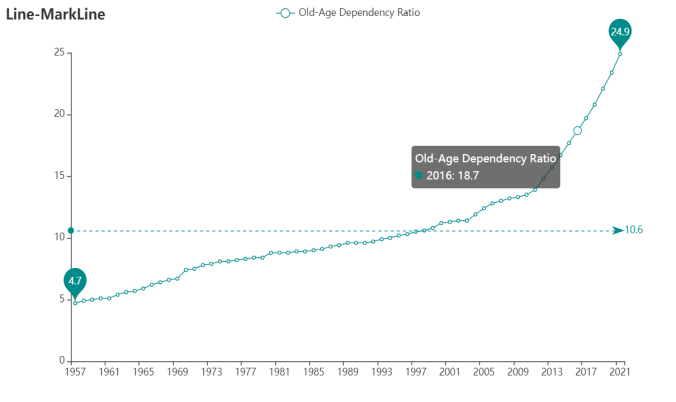
As shown in the figure, we can easily get the slope and y-intercept of the best fit line.



We also build an estimator for China’s Population



The pattern for population growth in Singapore & China: Modern Type



**Basic features:**

**Birth rate of the population:** began to decline significantly

**Mortality rate:** continued to decline and reached a low level

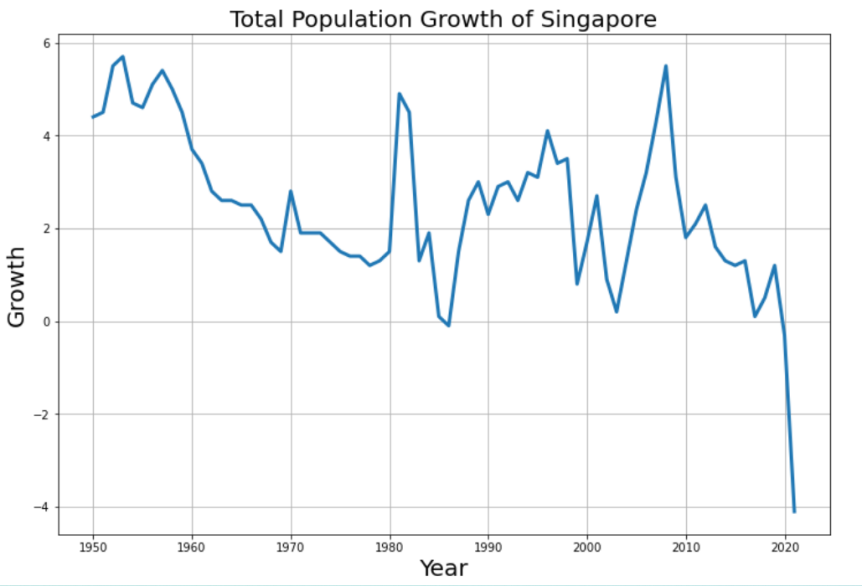
**Natural growth rate:** gradually decreases, and some developed countries even begin to experience negative or zero growth.

**Pros:** Effectively alleviate the trend of rapid population growth

**Cons:** Lack of social labor force: which is not conducive to industrial development and the economic development of the country; Population aging: increasing the burden of social welfare

**Distribution:** Most developed countries.

**How to improve the accuracy?**



Changes in population are influenced by many factors that are difficult to predict and quantify. Linear regression model cannot be perfect。

Although ordinary linear regression can fit the regression curve, it is too coarse and will appear underfitting phenomenon.

Locally weighted linear regression (LWLR) reduces the mean square error of the prediction by introducing some bias in the estimates.

The basic idea of this method is to assign a certain weight to each point near the point to be predicted, and store these weights in a new matrix W.

Better prediction results can be obtained by local weighted regression, but adjusting the parameter k is the key, and k is too small may appear overfitting phenomenon.