**Project: Population Regression**

**Instructions**: Please work in your preassigned groups to complete and submit your work to the appropriate folder in LumiNUS. Project submissions are due on

Please submit all the following documents as a single zip file named Group-X-Project.zip:

1. Powerpoint slides named as Group-X-Project.pptx (20 slides max)
2. Completed Word file named as Group-X-Project.docx (with all results)
3. Print preview of ipynb file named as Group-X-Project.pdf (with all results)
4. Your working ipynb file named as Group-X-Project.ipynb
5. Your data files (either csv or excel).
6. **Introduction and Reading Assignment**

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.

Please read the following article from Nature Education.

An Introduction to Population Growth

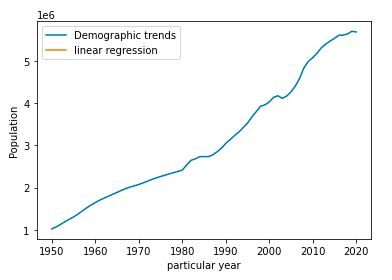
By: Sunny B. Snider (College of Agriculture, California State University, Chico) & Jacob N. Brimlow (College of Agriculture, California State University, Chico)

<https://www.nature.com/scitable/knowledge/library/an-introduction-to-population-growth-84225544/>

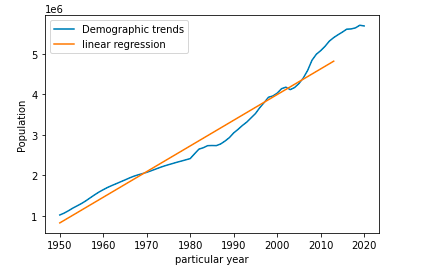
1. **Data Source: Singapore Department of Statistics (SingStat)**

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.



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



slope: 0.011

y-intercept: -21.691

best fit line: y=0.011 x-21.691 (Normalized)

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

R2 coefficient: 0.961

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

mean squared error (Normalization): 0.0017

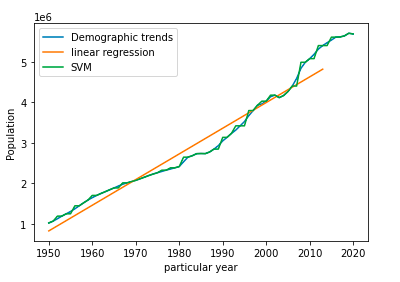
* + 1. Use years greater than 2013 as test data and predict the population for those years.

|  |  |  |
| --- | --- | --- |
| Year | Actual population | Prediction |
| 2014 | 5469724 | 4880287 |
| 2015 | 5535002 | 4943630 |
| 2016 | 5607283 | 5006974 |
| 2017 | 5612253 | 5070317 |
| 2018 | 5638676 | 5133660 |
| 2019 | 5703569 | 5197003 |
| 2020 | 5685807 | 5260346 |

* + 1. 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.

mean squared error (Normalization): 0.0090

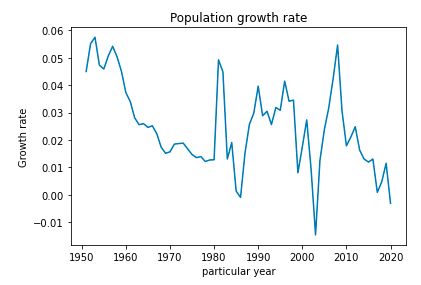
* 1. What is your estimate of Singapore’s population in 2030 and 2050? Do you think these estimates are reasonable? Explain your answer.



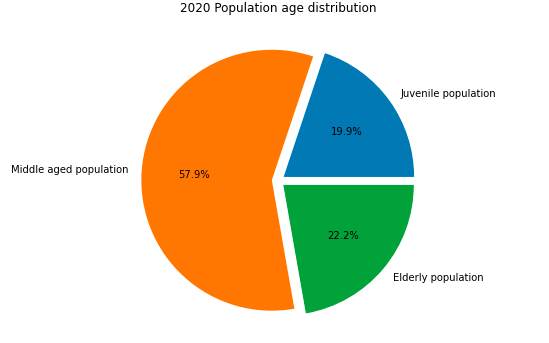
2030 Singapore population: 6527211

2050 Singapore population: 7037872

I think the population forecast is reasonable after I analyzing the data of population for the past years.



* Population proportion: in the pie chart below, I analyzed the population composition of Singapore in 2020 and found that the proportion of its elderly population is too large, which indicates that the natural growth rate of Singapore's population will continue to decline in the future.



* Policy analysis: according to the 2020 Singapore population white paper, the Singapore government announced that there is still a gap of 1 million people, which indicates that a large number of immigrants will be approved to increase the national population.
* International environmental analysis: according to the briefing, the main factor leading to the population decline is that the non resident population fell by 2.1%, about 40000. in 2020, Singapore's population declined for the first time because of the rampant COVID-19, which led to a decrease in the number of migrant workers.

So I am optimistic about Singapore's future population growth.

* 1. What pattern do you expect for human population growth in Singapore?

I think Singapore's future population growth model will be dominated by immigrants, supplemented by natural growth.With the aging of the population, the Singapore government will reduce the conditions of immigration to adjust the proportion of the population.In addition, because of COVID-19's gradual control, Singapore's population will grow steadily in the next 30 years, and the rate of increase will not slow down.

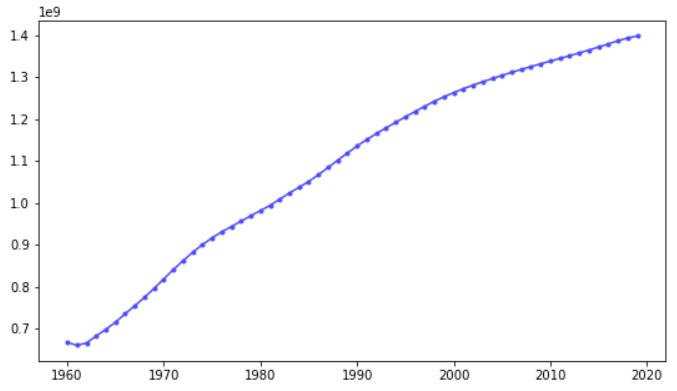
* 1. How could you improve your estimates of the future population?

Immigration and population aging are both factors affecting the total population of Singapore.

However, we cannot treat them separately. These two factors can influence each other. Migrant immigrants will greatly alleviate the process of population aging, and population aging will attract a large number of young migrant workers to work here. Therefore, the population data model we have established needs to be further revised: reduce the weight of the adverse factor of population aging, To better predict the future population.

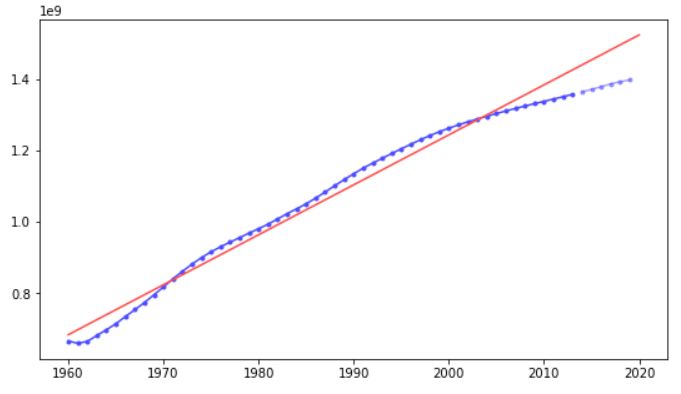
1. **Data Source: The World Bank (https://www.worldbank.org/)**

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.  
       
     Answer:Data matches
  2. Use the total population data for China. Graph the total population vs year.  
       
     Answer:The figure is shown below  
     
  3. 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.
  4. Performance metrics:

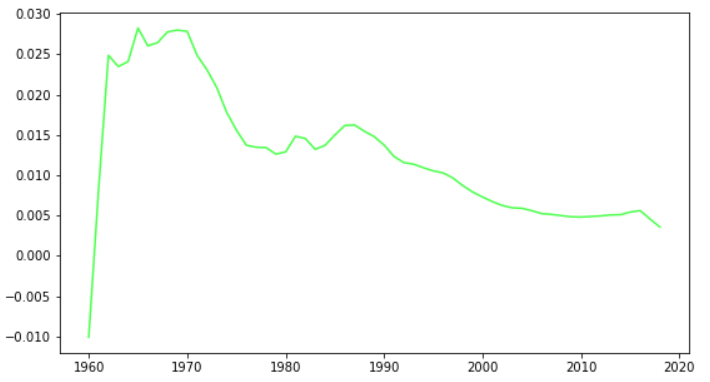
What are the slope and y-intercept of the best fit line? Plot the best fit line over the empirical data.  
  
 Answer:  
 Slope: 0.010006665423016379

Intercept: -19.595724663432584



* + 1. What is the *R2* coefficient for the best fit line?  
         
       Answer:The value of R2 is 0.9827150095054414
    2. What is the mean squared error (MSE) of the estimator on the training data?  
         
       Answer:The MSE of training data is 0.0004278362975097043
    3. Use years greater than 2013 as test data and predict the population for those years.   
         
       Answer:

|  |  |
| --- | --- |
| Year | Prediction |
| 2014 | 1439834950 |
| 2015 | 1453821420 |
| 2016 | 1467807890 |
| 2017 | 1481794350 |
| 2018 | 1495780820 |
| 2019 | 1509767290 |

* + 1. 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.  
         
       Answer:The MSE of the forecast data is 0.0045007114664355995
  1. What is your estimate of China’s population in 2030 and 2050? Do you think these estimates are reasonable? Explain your answer.  
       
     Answer:  
     China's population forecast in 2030 is 1663618416  
     China's population forecast in 2050 is 1943347743  
       
     Explain:  
       
     *The figure above is a graph of China's population growth rate*

I think the results predicted by the model are accurate for the following reasons:

* According to the figure, it is not difficult to find that after 1974, the population growth rate showed a downward trend. However, the downward trend has eased in the past two years, and the overall population is still in a steady upward state.
* According to the fertility measures successively implemented after China's reform and opening up. For example, the family planning policy implemented by China in 1970 gradually led to the decline of the growth rate curve. At the same time, the separate two child policy and three child policy in recent years have further controlled the continuous decline of growth rate.
* To sum up, through calculation, it is not difficult to find that the growth from 2010 to 2030 will exceed that from 2030 to 2050. Therefore, we can further verify the accuracy of the model prediction.  
  1. What pattern do you expect for human population growth in China?  
       
     Answer:I believe that due to the implementation of the strict family planning policy, China's population growth model has entered a modern model with low birth rate, low mortality and low natural growth rate.
  2. How could you improve your estimates of the future population?  
       
     Answer:An important factor affecting China's population change is China's legal system. Therefore, if we want to better understand the trend of China's future population development, we should add China's legal control to the model, so that our model can get better analysis results.

1. **Appendix: On the *R2* Coefficient**

The coefficient of determination, or *R2*, is a measure that provides some information about the goodness of fit of a model. In the context of regression, it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses. The most widely used expression for *R2* is shown below.

A picture containing object, clock, drawing

Description automatically generated

A drawing of a person

Description automatically generated

A close up of a logo

Description automatically generated

The better the linear regression fits the data in comparison to the simple average, the closer the value of *R2* is to 1.

See the WikiPedia entry on *R2*: <https://en.wikipedia.org/wiki/Coefficient_of_determination>

See video on *R2* from Khan Academy: <https://youtu.be/lng4ZgConCM>