# A Study on Population Prediction based on Linear Regression

## I. Introduction

Population issues have gained public attention in recent decades. This concept refers to the challenges and concerns related to the size and growth of human populations. It encompasses a wide range of issues such as overpopulation, population aging, migration, urbanization, and declining birth rates. Overpopulation refers to the condition where the number of people in a given area exceeds the available resources and environmental carrying capacity, leading to the depletion of resources, environmental degradation, and an increased risk of conflict. Population aging is another issue that has become increasingly relevant in recent years, as the median age of populations in many countries has increased due to declining birth rates and improved longevity. This shift has implications for healthcare, pensions, and the workforce, among other things. These issues have far-reaching consequences for individuals, communities, and the planet as a whole.

Linear regression models have been widely used in population forecasting since they are simple and easy to implement. They are based on the idea that the relationship between two variables can be represented by a straight line, with the population as the dependent variable and one or multiple predictor variables as the independent variables. The predictor variables can be demographic factors such as age, gender, and migration, or socioeconomic factors such as income, education, and employment.

Several studies have shown that linear regression models can provide accurate predictions in sociology areas. For example, a study on risk assessment of crowd gathering in public spaces in Shanghai (Yang, 2022) applies linear regression to predict the risk level of different regions. Another study on the estimation of the multi-level spatial distribution of the interregional migrant population (Liu, et al., 2020) precisely predicted the population variation tendency during the epidemic period in Wuhan, China. Therefore, it is feasible to predict the social characteristics of population size by linear regression.

Despite the simplicity and accuracy of linear regression models, they have some limitations. These models assume a linear relationship between the predictor variables and the population, which may not always be the case in reality. Additionally, they do not account for non-linear relationships or unexpected events that may impact population trends. To overcome these limitations, we ascended the dimension of the dataset by increasing related features to make multivariate linear regression.

The remainder of the essay is organized as follows. In the next section, the pre-processing, fitting, and result enhancement methods are briefly reviewed; sections 3 and 4 explain the prediction results and enhancement results of known Singapore and Chinese population statistical data; and section 5 summarizes the conclusions of this study.

## II. Methodology

### 2.1 Linear Regression

In this case, we used year as the explanatory variable and population number as the response variable for the prediction. In such cases, we often use simple linear regression methods. Meanwhile, to increase the accuracy of the regression, we use a high-dimensional dataset to train the model. The general mathematical expression of both two algorithms is

where refers to the response variables (population numbers), refers to the explanatory variables (years, population number of ethnic groups, etc.), and refers to the random errors that match the normal distribution .

### 2.2 Data Collection

The Department of Statistics Singapore provides a detailed population structure record from 1957 to 2022 on its official website. The brief statistical data of the Chinese population is published from the dataset of the World Bank. The corresponding sources are listed in Table 1.

**Table 1**

*Source of Datasets*

|  |  |
| --- | --- |
| **Dataset** | **Source** |
| *Population and Population Structure of Singapore* | https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data |
| *Population, total* | https://data.worldbank.org/indicator/SP.POP.TOTL |

### 2.3 Datasets Preparation

We selected the following attributes from the Singapore population structure dataset as the final dataset. For the convenience of training, we transpose the corresponding data matrix and rename each column, shown in Table 2.

**Table 2**

*Description of Final Dataset for Singapore Population Prediction*

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Column Name** |
| Data Series | Recorded year (from 1957-2022). | Year |
| Total Residents | Population number in the corresponding year. | TR |
| Total Male Residents | Male population number in the corresponding year. | TMR |
| Total Female Residents | Female population number in the corresponding year. | TFR |
| Sex Ratio (Males Per Thousand Females) | Sex ratio represented by the value of male / (1000 female). | SR |
| Median Age of Resident Population (Years) | Median age of residents in the corresponding year. | MAR |

Due to the small amount of information contained in the data set, we only carried out one-dimensional linear regression. The attributes are shown in the following table.

**Table 3**

*Attributes in the Chinese Population Dataset*

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Column Name** |
| Year | Population statistics recorded year (from 1960-2021). | Year |
| Population, total | Population number in the corresponding year. | PPL |

### 2.4 Data Preprocessing

For the required simple linear regression, we did not implement any data-preprocessing phase because the coefficient and the y-interception are needed. For further multivariate linear regression, we applied normalization to enhance the effects of the regression. Meanwhile, we applied correlation analysis to determine the influences between the attributes.

### 2.5 Model Evaluation

To evaluate the fitting effect of the model, we applied 2 indicators: coefficient and mean squared error (MSE). refers to the number% of dependent variables derived from the regression equation that can be explained by independent variables. Thus, it is used to judge the degree of data fitting. The calculations of these indicators are

, and

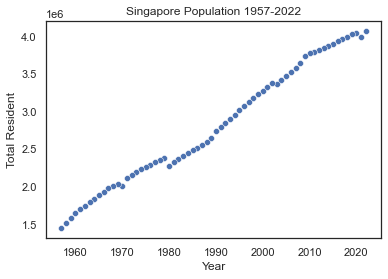
.

## III. Experiment Results for Singapore Population Data Prediction and Enhancement

### 3.1 Dataset Visualization

**Figure 1**

*Year-Population Plot of Singapore population statistics*

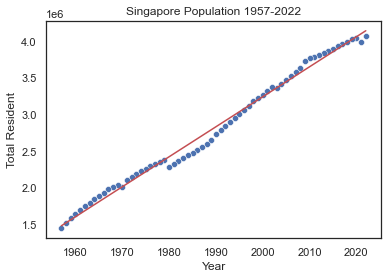


### 3.2 Linear Regression Results

The coefficient gained from this linear regression model is Slope = 40975.96590614, y-Intercept = -78707770.39387262.

**Figure** 2

*The best-fit line for Singapore population statistics*



### 3.3 Metrics

coefficient: 0.9857883020995728

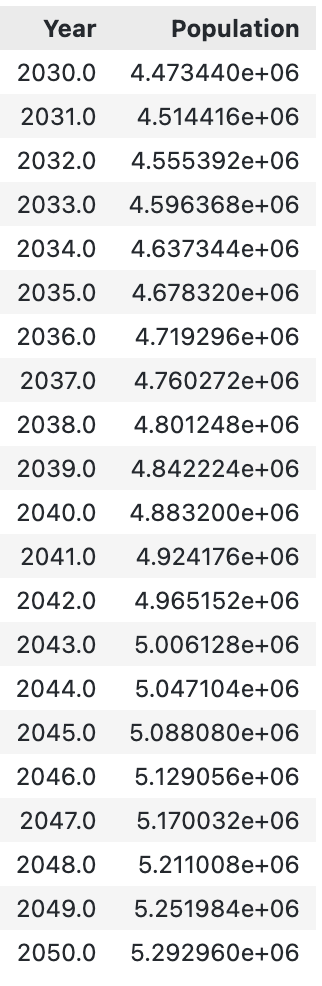
Training MSE = 6551722294.557682 =

Testing MSE = 2907187388.7105937 =

### 3.4 Prediction for 2030-2050 and Analysis

**Figure 4**

*The predicted population of Singapore*



This result is partly reasonable. It is because, in natural population growth principles, there should be competition between samples and the growth rate might naturally decrease to 0. However, in the linear regression model, the growth rate is a constant value. Perhaps it fits the situation in a short term, but it might not precisely reflect the population growth tendency in long term (e.g., >100 years).

Meanwhile, we suppose the population growth model of Singapore is in the middle of the logistic pattern.

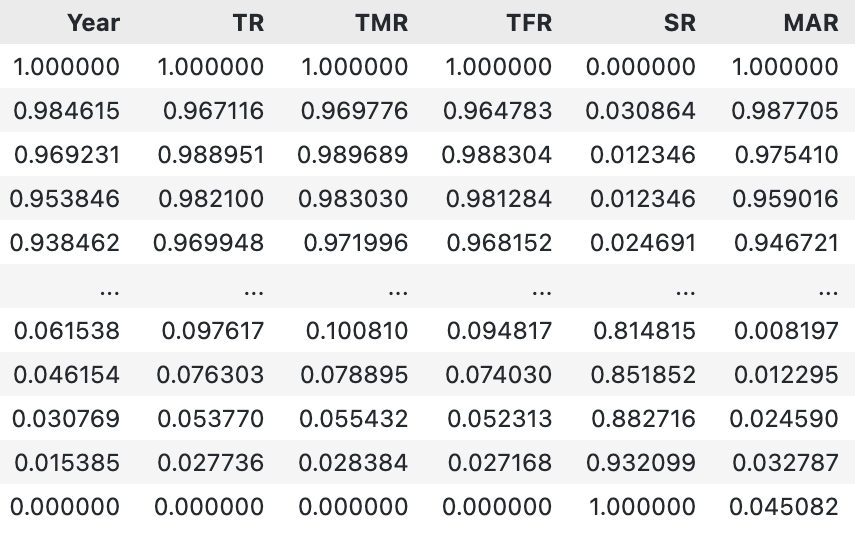
### 3.5 Future Work for Model Enhancement

In this case, we apply a multivariate linear regression. We selected some attributes from the Singapore official statistics data to increase the dimensions of the feature set. Before the formal multivariate linear regression, we conducted a correlation analysis to select the most influencing ones from these additional attributes.

To prevent the imbalanced influences of data from different ranges, a normalization phase should be implemented.

**Figure 5**

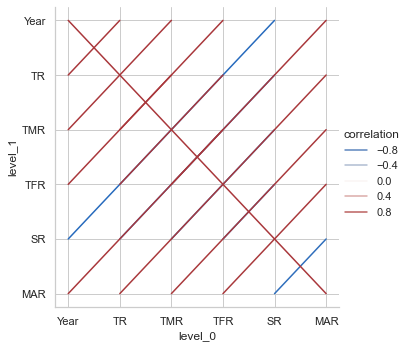
*Part of the normalized dataset*



Based on the correlation matrix, we plotted a correlation plot, shown in Figure 6.

**Figure 6**

*Correlation plot of the attributes*



From the correlation plot above, we can conclude that TMR and SR have the greatest positive influence on TR because the correlation coefficients between these attributes and TR are equal to 1. Therefore, we can add these two attributes to our X matrix to improve the accuracy of the prediction.

The testing result of multivariate regression has a 0-like MSE and its R^2 coefficient equals to 1, which is much more precise than the previous simple linear regression with [Testing MSE, R^2] = [2907187388.7105937, 0.9857883020995728].

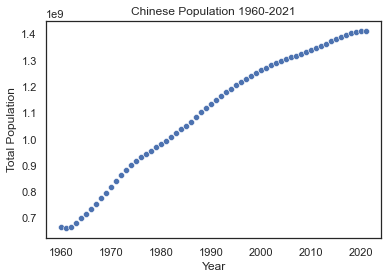
Therefore, multivariate linear regression can be considered an effective enhancement method.

## IV. Experiment Results for Chinese Population Data Prediction

### 4.1 Dataset Visualization

**Figure 7**

*Year-Population Plot of Chinese population statistics*

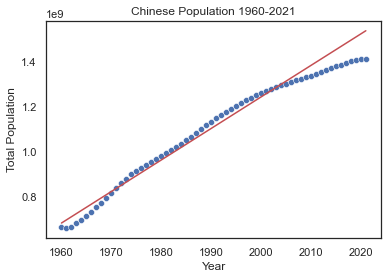


### 4.2 Linear Regression Results

The coefficient gained from this linear regression model is Slope = 14006789.21288355, y-Intercept = -2.67690896e+10.

**Figure** 8

*The best-fit line for Chinese population statistics*



### 4.3 Metrics

coefficient: 0.9832476129933141

Training MSE = 811985540736408.5=

Testing MSE = 9065457889543732.0=

### 4.4 Prediction for 2030-2050 and Analysis

**Figure 9**

*The predicted population of China*



This prediction is not reasonable. From the figure plotted in the previous stage, it can be easily found that the growth rate of the Chinese population is descending. The predicted growth rate (slope) is a large constant because the dataset used includes the phases of acceleration of Chinese population growth. It does not match the current situation in China.

We suppose the population growth pattern of China is the logistic pattern.

### 4.5 Future Work for Model Enhancement

As mentioned earlier, multivariate linear regression can be used to improve the precision of the prediction. However, due to the lack of further information, we did not verify this enhancement method by predicting the Chinese population.

## V. Conclusion

In this study, we analyzed the sample population dataset structures and used simple and multivariate linear regression to predict the population. Our practice and research are aimed at getting better accuracy of population prediction. Grounded in model evaluation methods such as score and mean squared error, we have discovered that multivariate linear regression has a great improvement on prediction. Thus, we recommend multi-dimensional population datasets as training references.

In conclusion, a linear regression-based population prediction study provides valuable insights into the dynamics of population growth and change. By examining the relationship between key attributes in population structure and population size, this method enables us to make predictions about future population trends and to identify areas that may be at risk of overpopulation or population decline. Nevertheless, population growth is a complex and dynamic phenomenon that is influenced by a wide range of factors, including economic, political, social, and environmental factors. Therefore, it is crucial to continuously monitor and update population projections as new data becomes available and to consider the potential impact of unanticipated events and emerging trends.

# Bibliography

Liu, Z., Qian, J., Du, Y., Wang, N., Yi, J., Sun, Y., . . . Zhou, C. (2020). A multi-level spatial distribution estimation model for interregional migrants based on multi-source spatio-temporal big data: A case study of people migrating out of Wuhan during COVID-19. *Geo-Information Science (Geo-Inf Sci)*, 147-160.

Yang, Y. (2022). Research on risk assessment method of crowd gathering in urban open public Space supported by spatiotemporal big data. *Shanghai Normal University*.