Project: Population Regression

Instructions: Please work in your preassigned groups to complete and submit your work to the appropriate folder in Canvas by the due date.

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

(i) Powerpoint slides named as Group-X-Project.pptx (15 slides max)
(ii) Completed Word file named as Group-X-Project.docx (with all results)
(iii) Print preview of ipynb file named as Group-X-Project.pdf (with all results)
(iv) Your working ipynb file named as Group-X-Project.ipynb
(v) Your data files (either csv or excel).

1. 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/

2. 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 2022 from: https://www.singstat.gov.sg/

a) Graph the total population vs year.
b) Use linear regression to build an estimator of the total population of Singapore in the future. Use the data for years 2019 and earlier as training data.
c) Performance metrics:

i. What are the slope and y-intercept of the best fit line? Plot the best fit line over the empirical data.
ii. What is the R^2 coefficient and mean squared error (MSE) of the estimator on the training data?
iii. Use years greater than 2020 as test data and predict the population for those years.
iv. What is the MSE of the estimator on the test data?

d) What is your estimate of Singapore’s population in 2023, 2030 and 2050? Do you think these estimates are reasonable? Explain your answer.
e) What pattern do you expect for human population growth in Singapore?
f) How could you improve your estimates of the future population?

3. Data Source: World Bank

Repeat Question 2 for your own country’s population from 1950 to 2022. You may download the data from the World Bank: https://data.worldbank.org/indicator/SP.POP.TOTL

4. Project Presentation – Each group is to do a 10 minute project presentation. You can use PowerPoint or any other tools. Please focus on the following points:

1. Introduction – Why population forecasting is an important problem worth working on
2. Problem Description – What is the problem that you are solving, what is the context, etc.
3. Experimental Setup – Describe any assumptions you make, experiment setup , etc.
4. Results Demonstration – Visualize your results in a clear and easy to understand manner
5. Results Analysis – What insights do you get from your results, key reflections, etc.
6. Reflections & Conclusions – What is the takeaway message, what did you learn, etc.

5. Appendix: On the R^2 Coefficient

The coefficient of determination, or R^2, 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 R^2 is shown below.

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

The better the linear regression fits the data in comparison to the simple average, the closer the value of R^2 is to 1.

See the WikiPedia entry on R^2: https://en.wikipedia.org/wiki/Coefficient_of_determination
See video on R^2 from Khan Academy: https://youtu.be/lng4ZgConCM