## Prediction with Census Data

In this project, you’ll explore predicting a population pyramid from the population pyramid of the prior year using census data.

I’ve downloaded CSVs from data.census.gov for the American Community Survey (ACS) 1-year estimate for Age and Sex. There is a CSV file for years 2010 through 2023, though 2020 is missing. As example links, I downloaded the 2010 CSV from <https://data.census.gov/table/ACSST1Y2010.S0101> and the 2023 CSV from <https://data.census.gov/table/ACSST1Y2023.S0101>. The downloaded CSV can be found in the “data” folder in OneDrive.

For this project, you can ignore the margin of error and simply use point estimates. Although the data has information for both age and sex, for this project, we will only look at age.

Each data point will be one of the 50 US states and a pair of adjacent years.

Below is a table of how you should split the years.

|  |  |  |
| --- | --- | --- |
| Input year (x) | Target year (y) | Split |
| 2010 | 2011 | train |
| 2011 | 2012 | train |
| ... | ... | train |
| 2017 | 2018 | train |
| 2018 | 2019 | train |
| 2021 | 2022 | test |
| 2022 | 2023 | test |

Thus, you should have 9\*50 = 450 training points and 2\*50=100 test points.

Each point has 18 input dimensions (for the 18 age groups) and 18 target dimensions.

Note the columns are slightly different between 2010-2016 and 2017-2023. When no population estimate for an age group is available, multiple the total population by the proportion in the age group.

Use [sklearn.linear\_model.Ridge](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html) to train an L2-regularized linear regression model. Evaluate the test mean relative error (relative error = abs(prediction-true)/true) for alpha (the regularization parameter) as powers of 10. What is the optimal value of alpha (restricted to be a power of 10)? What is the test mean relative error for the optimal value of alpha?

Use [matplotlib.pyplot.matshow](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.matshow.html) to plot an 18x18 matrix of the regression coefficients (with a colorbar legend). Do you notice a pattern in the coefficients? Can you quantify this pattern and give a plausible explanation?

**To submit this report, include:**

* The optimal alpha and test mean relative error
* The regression coefficient visualization (as an image)
* A paragraph describing the pattern, it’s quantification, and a plausible explanation
* Your code used for processing the CSVs, training the model, and visualizing/analyzing the results.