In this project I performed a polynomial model and ridge regression, while finding the optimal degree and regularization parameter to accurately predict the population. I performed k-fold cross-validation on Ridge regression models with polynomial features of varying degrees(6) and different regularization parameters for each fold using GridSearchCV. I then calculated the root mean squared error for each fold and stored them in a dataframe. For the ridge regression, I set the param grid to have all the necessary alpha values and degree and used the solver for the ridge as ‘cholesky’ and fit\_intercept to True as discussed in class.

After conducting extensive analysis, it has been determined that the optimal degree for the regularized model is 6, and the optimal regularization parameter is 4.98e-02. This was done using the grid search CV which on the training set across the different cross validation steps. After the params were found, I picked the best params based on the least RMSE score to use it in the final model.

The coefficients for the d∗-degree polynomial are [0, 1.41027678, 4.28238228, -1.92793174, -4.38398966, 0.97135127, 1.32894803, -0.13404705, -0.14013278], indicating the importance of each variable in the model.

The RMSE on the training set is 0.12, while the RMSE on the test set is 0.10, indicating that the model is accurately predicting values and is not overfitting to the training data. The training RMSE for the optimal degree of 6 is 0.12, while the test RMSE is 0.10, demonstrating that the model is performing well on both the training and test data. Likewise, the training RMSE for the optimal regularization parameter of 4.98e-02 is 1.02, while the test RMSE is 0.96, indicating that regularization has helped the model generalize better on the test data.

The optimal degree and regularization parameter for this polynomial regression model have been identified, and the model is accurately predicting values with low RMSE on both training and test data. The RMSE for k-fold cross-validation with kF=5 was the lowest of any cross-validation fold, indicating that the model performed well at that kF value.

For the non-regularized polynomial regression model, Kfold CV was used to test the model using cross-validation with k-fold values ranging from 2 – 6. The best value of kF was found to be 5, whereas best degree of the polynomial was discovered to be 8. The RMSE on the training set was 0.076, indicating good performance on the training data, while the RMSE on the test set was 0.087, demonstrating that the model will generalize well on unseen data.