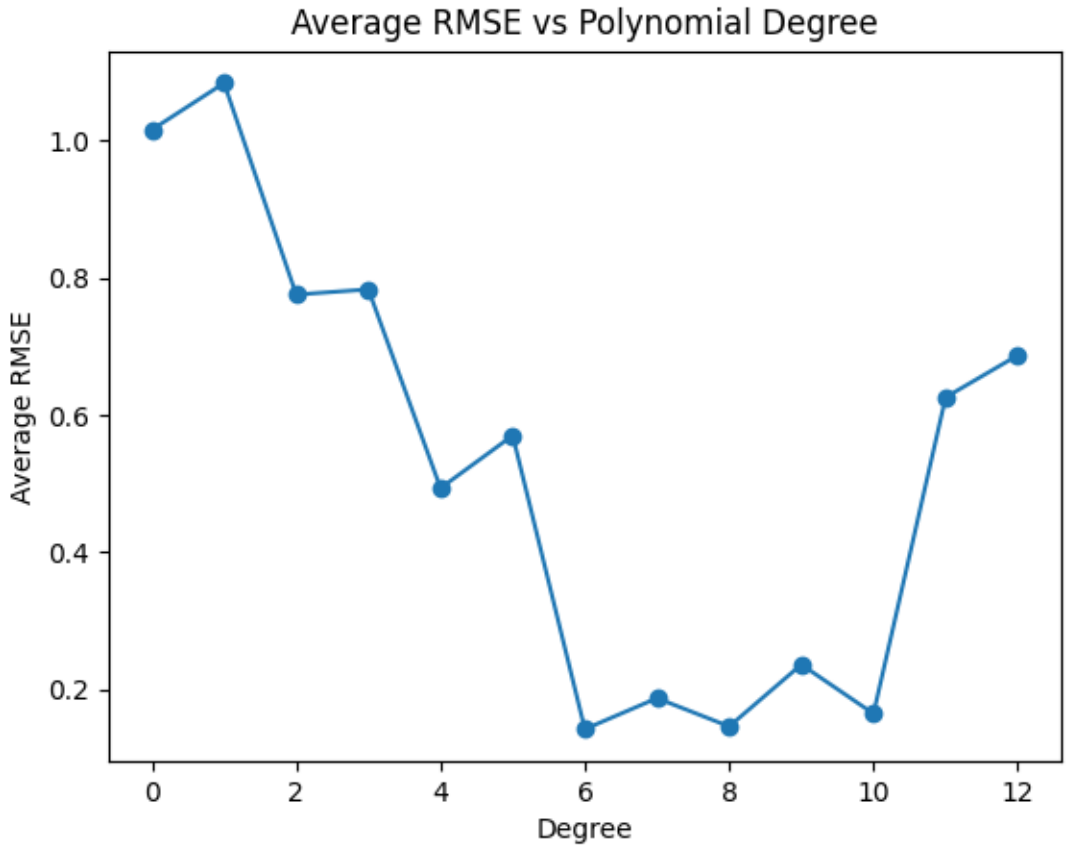
Erika Valle-Baird
University of Michigan, Dearborn
CIS 481
Project 1
Generalized Linear Regression, Model Selection via Cross Validation, and Model Evaluation; Application: Polynomial Curve-Fitting for Working-Age Data
*UPDATED

AVERAGES OF RMSE VALUES OBTAINED DURING 6-FOLD CV FOR EACH CASE:

Degree 0: 1.0155605589711303
Degree 1: 1.0835561617376988
Degree 2: 0.7754293123514141
Degree 3: 0.7830016853814804
Degree 4: 0.4934818442564504
Degree 5: 0.5701350133901353
Degree 6: 0.1423526253934837
Degree 7: 0.18750847543439889
Degree 8: 0.14631007436526883
Degree 9: 0.23647291096685455
Degree 10: 0.16488080003181632
Degree 11: 0.6257835760349929
Degree 12: 0.6869783309076073



OPTIMAL DEGREE D OBTAINED VIA 6-FOLD CV:

The optimal degree, d, obtained via 6-fold cross-validation was found to be degree 6.

COEFFICIENT-WEIGHTS OF D-DEGREE POLYNOMIAL LEARNED ON ALL TRAINING DATA:

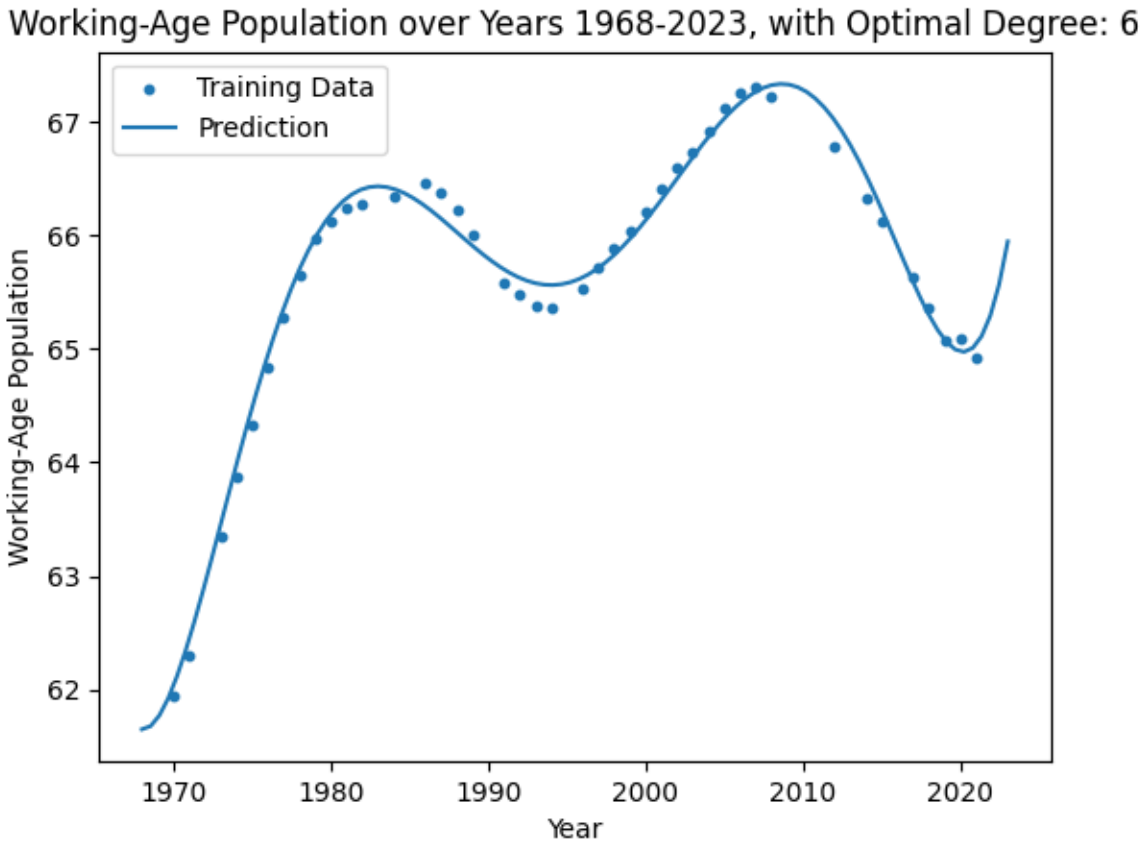
$f(x) = 0.35654293x + 3.37042403x^2 + 0.1362716x^3 - 2.94644226x^4 + 0.00841331x^5 + 0.53773951x^6$

For degree 6, the coefficients for 6-degree polynomial learned on all training data:
[0, 0.35654293, 3.37042403, 0.1362716, -2.94644226, 0.00841331, 0.53773951]

TRAINING AND TEST RMSE OF THAT FINAL, LEARNED POLYNOMIAL:

Train RMSE of Learned Polynomial of Degree 6: 0.1054
Test RMSE of Learned Polynomial of Degree 6: 0.1143

PLOT CONTAINING ALL THE TRAINING DATA ALONG WITH RESULTING POLYNOMIAL CURVES FOR THAT FINAL, LEARNED POLYNOMIAL WITH DEGREE D, FOR RANGE OF YEARS 1968-2023 AS INPUT:


Working-Age Population over Years 1968-2023, with Optimal Degree: 6
Training Data
Prediction
Working-Age Population
Year
67
66
65
64
63
62
1970
1980
1990
2000
2010
2020

BRIEF DISCUSSION OF YOUR FINDINGS AND OBSERVATIONS:

The averages of RMSE values obtained during the 6-folds of cross-validation for each degree include:
Degree 0: 1.0155605589711303
Degree 1: 1.0835561617376988
Degree 2: 0.7754293123514141
Degree 3: 0.7830016853814804
Degree 4: 0.4934818442564504
Degree 5: 0.5701350133901353
Degree 6: 0.1423526253934837 (Lowest Average RMSE)
Degree 7: 0.18750847543439889
Degree 8: 0.14631007436526883
Degree 9: 0.23647291096685455
Degree 10: 0.16488080003181632
Degree 11: 0.6257835760349929
Degree 12: 0.6869783309076073

Visually, as viewed in the graph, we see the minimum point again the average RMSE's occur at degree 6.

The optimal degree d obtained via 6-fold cross-validation, as seen above is the lowest average root mean squared error. This is degree 6, with a value of 0.1423526253934837.

For degree 6 of the learned polynomial, learned on all training data, the coefficient-weights include [0, 0.35654293, 3.37042403, 0.1362716, -2.94644226, 0.00841331, 0.53773951]. As a function, this would look like: $f(x) = 0.35654293x + 3.37042403x^2 + 0.1362716x^3 - 2.94644226x^4 + 0.00841331x^5 + 0.53773951x^6$.

For the final learned polynomial, the training root mean squared error for degree 6 equated to 0.10540106673270382. After training was complete, the testing root mean squared error for degree 6 was equivalent to 0.0299653279101365.
***Previously had training data included when performing the model fit and prediction. Fixed my mistake, as testing must be done separately from the training data.

The average root mean squared error, i.e RMSE, begin to decrease as the complexity of the polynomial increases until we reach degree 6. Here we are at the lowest average RMSE found. From here, the average RMSE bounces up and down, maintaining degree 6 as its minimum point. When viewed in the final graph, against the training data, we see that the learned polynomial of degree 6, is

similar in shape to the training data. This similarity in shape displays the fitting for the line of best fit.

To run the File:
The file can be run in the command prompt! :) Changing the directory to where ever you place the file, for example, my file was on my desktop so I used the command cd C:\Users\airik\Desktop. After executing this statement, I followed up with the command: python mine.py. With “mine.py” being the file to execute. The python file has been updated with necessary changes for testing RMSE.