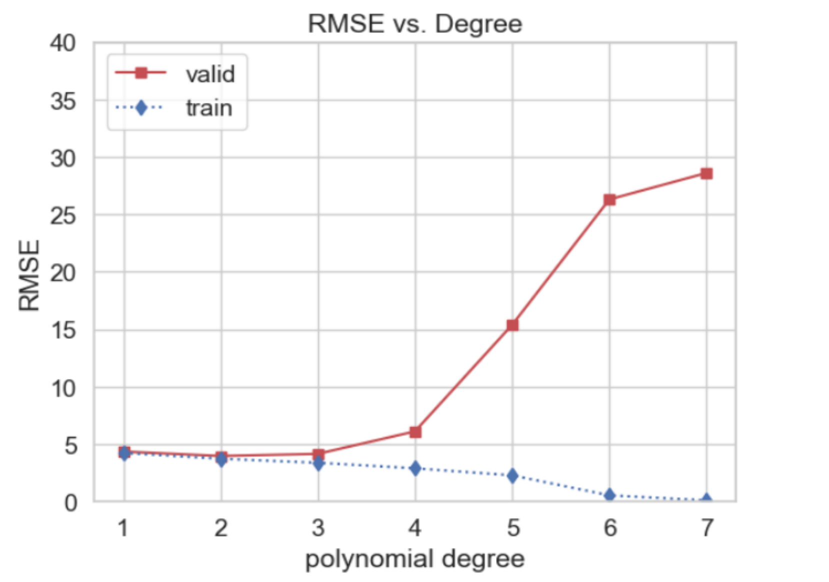
Erin Sarlak

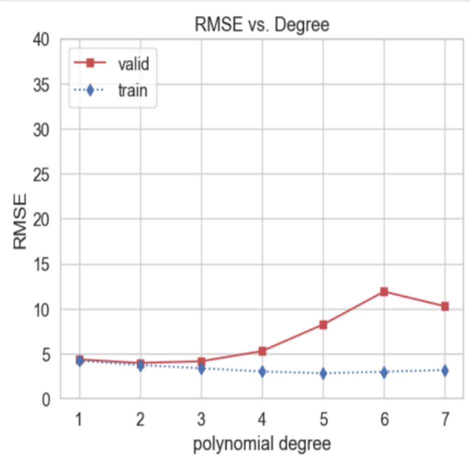
Problem 1: Polynomial Regression – Selecting Degree on a Fixed Validation

Figure 1: Error vs. Degree

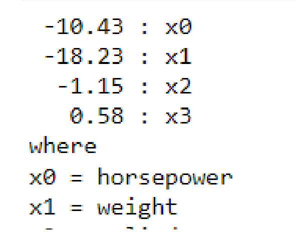


Caption on Figure 1: The plot looks like I expected because the training error decreases with increasing polynomial degree, and the validation error skyrockets after a certain polynomial. The reason for this is that higher polynomial degrees create more complex models that can overfit the data. This leads to fitting the training data quite well, but failing to fit any other new data points such as ones in validation set in this case. I recommend degree 2 because it has the least error for validation set, which implies that it generalizes the data the best.

Short Answer 1a: Preprocessing by scaling the data is important for this data set because some features have very high values, and numpy.linalg.solve() function used to find the coefficients of the parameters produce inaccurate results solutions for calculations involving large numbers. Models with higher polynomials fit the data better due to flexibility, and therefore have decreasing training errors. However, in this case, more complex models produce about the same training error compared to models with lower polynomial degrees. This is an indication of the error in calculations due to numpy.linalg.solve().



Short Answer 1b :



```
  -10.43 : x0
  -18.23 : x1
   -1.15 : x2
    0.58 : x3
where
x0 = horsepower
x1 = weight
```

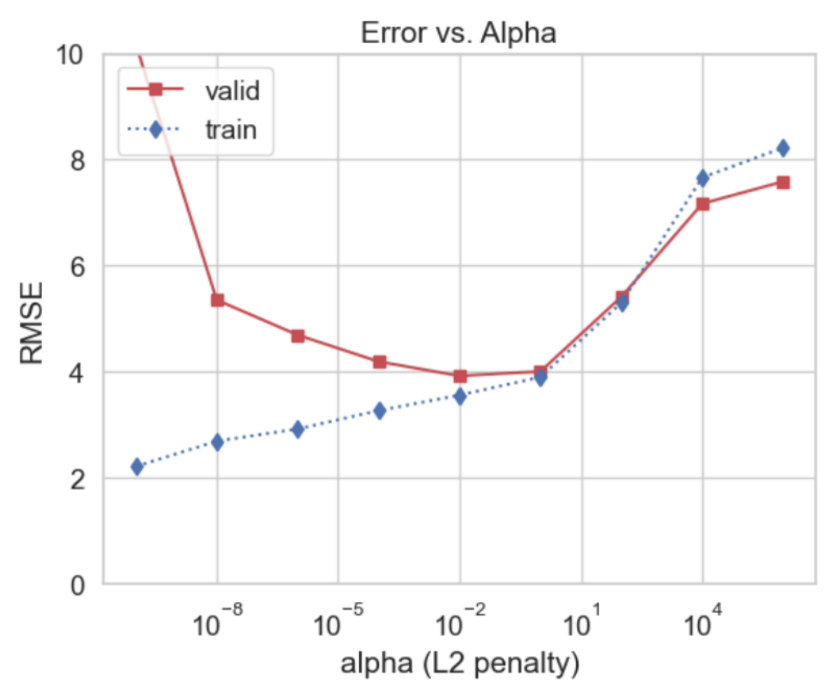
Figure 1: Weight Parameters for Model with Degree 1

Based on the coefficients, increasing engine weight leads to decreasing mpg. This makes sense because vehicles greater in weight tend to have less fuel efficiency. However, there is a positive correlation between displacement and the mpg, which is not logical. Usually, larger engines with higher displacements use up more fuel which lead to lower fuel efficiency.

Short Answer 1c: The models with degree 3 and 4 have more parameters and include more complex high-order polynomial terms compared to models with degree 1 and 2. Also, the parameter coefficients of the models with degree 3 and 4 are much larger in magnitude (positive and negative) than those of models with degree 1 and 2. This increased complexity leads to overfitting, which means the models with degree 3 and 4 fit the training data well, but don't generalize the trend well. Therefore, these more complex models have a lower training error, and higher a validation error.

Problem 2: Penalized Polynomial Regression – Selecting Alpha on a Fixed Validation Set

Figure 1: Error vs. Alpha



Caption of Figure 2: The graph looks like what I expected. The training error increases with increasing alpha because the increasing penalty pushes the coefficients closer to zero, and this leads to the model becoming less flexible and not fit the data well. However, the validation curve takes the shape of an irregular. For very small alpha values, the model overfits data due to regularization effect being negligible, and for very high values, the regularization effect becomes very strong, and causes the model to become less flexible. The model should generalize for unknown points good, and therefore we pick an alpha that gives the least error for validation set, which is alpha = 0.01.

Short Answer 2a: The coefficients of the weight parameters for the chosen degree-4 model are much lower than those in 1c. This is due to using penalized linear regression which encourages the coefficients to approach zero.

Erin Sarlak

Short Answer 2b:

I would pick the lowest alpha from the given set, which is 1.e-10 to minimize the error of the training set. However, this is problematic because a very low alpha value would make the effect 'ridge' penalty negligible, and therefore, the model would become more flexible with coefficients higher in magnitude. Such a model overfits any other data, in other words doesn't generalize well, which is also seen on the graph in the first data point for validation error.

Problem 3: Penalized Polynomial Regression + Model Selection with Cross-Validation

Table 2: Model vs. RMSE (on test set)

	Properties	**RMSE on Test Set**
Baseline Model	Horizontal Line	7.131
Model 1b	Degree: 2	3.992
Model 2b	Degree: 4, Alpha: 0.01	3.879
Model 3b	Degree: 7, Alpha: 0.1	3.817

Caption of Table 3:

The table shows that the RMSE value for the test set is the least in model 3b. This indicated that model 3b fits unseen data the best, and should be used for future predictions. The rankings of the methods match what I expected based on the class content.

- First, the baseline model performs poorly because using mean value for mpg is not provide a relationship between features and mpg values, and therefore is an ineffective method for prediction.
- Model 2b performs much better than baseline because it implements a linear regression model that leverages a polynomial function of degree-2 to better fit the data, and specifically degree-2 does a better job than the others.
- Model 2b fits the set data slightly better than Model 1b because it utilizes ridge penalty in calculating its root mean squared error avoids overfitting for the polynomial degree of 4.
- Lastly, model 3b does the best among all because it uses a high polynomial degree which makes the model more precise, with a high alpha value that prevents overfitting.