

Q1.1



Train Error: 6.47473

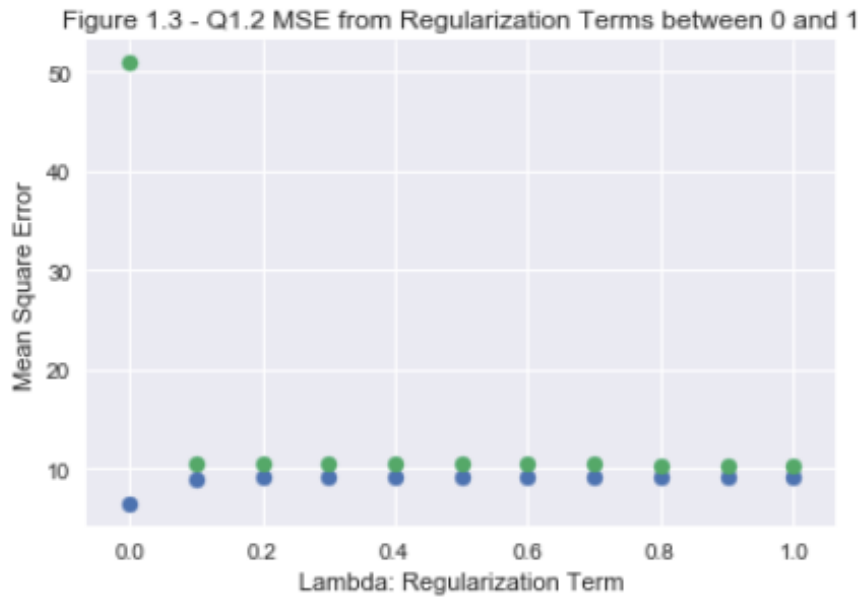


Test Error: 50.88464

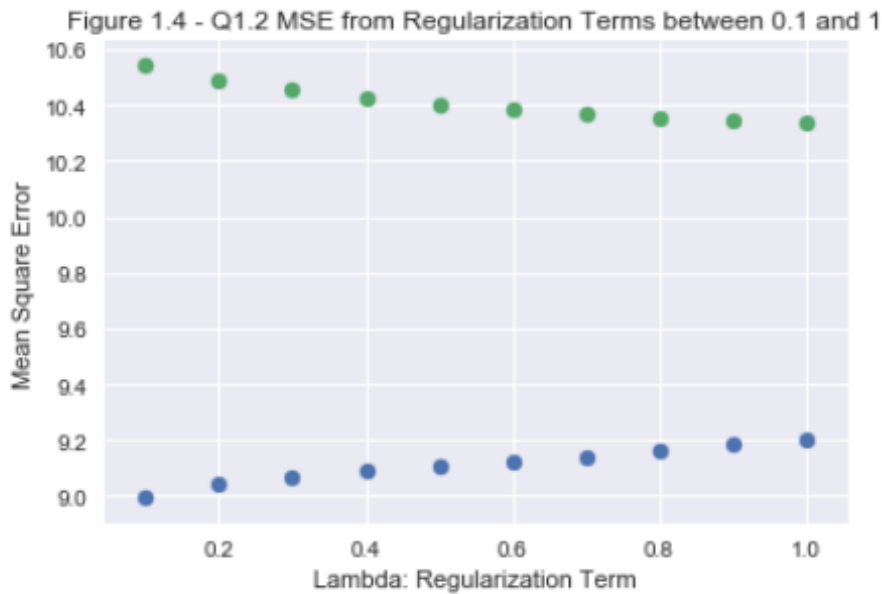
Figure 1.0 shows that visually a 20 degree polynomial fit explains the training data well from $x = -1.00$ to $x = 1.00$. The fit on the test data (Fig 1.1) illustrates that the model does not generalize well to new points. This is an example of over fitting. Over fitting is also reinforced by the values of the train and test error. The test error is 7.86 times the training error, again supporting that a $n=20$ degree model is over fitting the data.

Q1.2

By varying the regularization term, we get a significant improvement in the training and test set results. However it's difficult to visualize since the original training error is so high. The first data point was thus discarded to be able to visualize the regularization terms greater than 0 to 1.



The best value of lambda was determined by looking at the percentage improvement in generalization (i.e. our performance on the test set) as we increase the value of lambda.



The first introduction of a lambda turn led to an almost 80% improvement in test accuracy as shown in Figure 1.5. However after this improvement, subsequent variations in lambda only lead to incremental improvements in test accuracy. To see the incremental gains in smaller values, the first value in this graph was discarded and the data again visualized in Figure 1.6.

The percentage improvement in the test MSE was always a positive number so a lambda of 1.0 was selected. A lambda of 1.0 had a corresponding training error of 9.20 and a test error of 10.33. Using this lambda the fit was visualized to see if it was possible to guess the degree of the polynomial .

Figure 1.5 - Q1.3 Percentage Improvement in Generalization with Various Values of Lambda 0.0 - 1.0

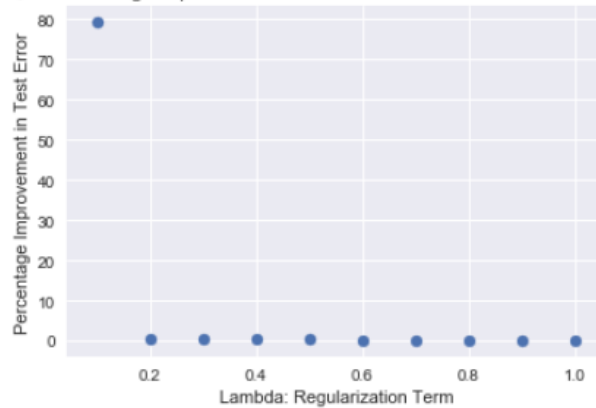
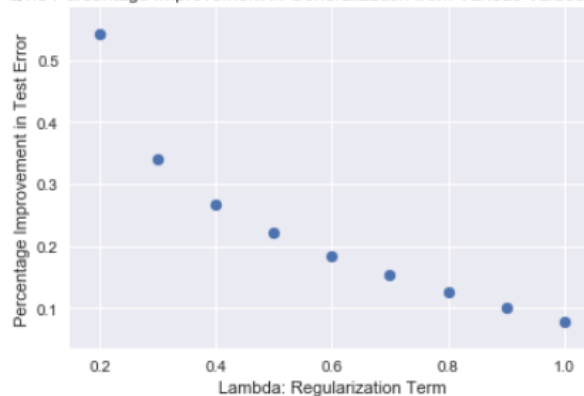
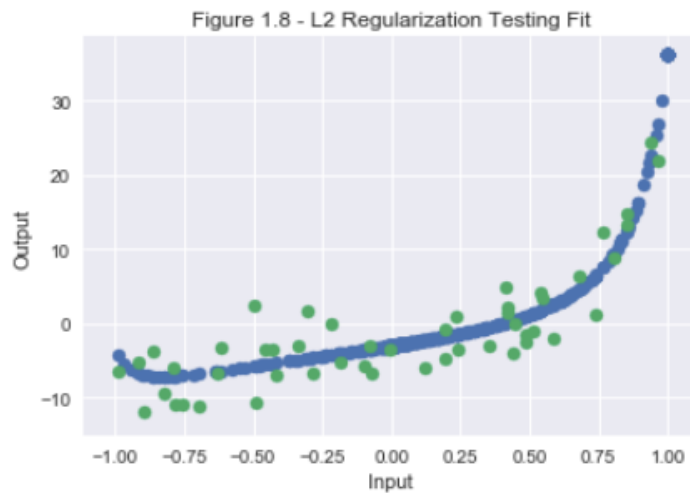


Figure 1.6 - Q1.3 Percentage Improvement in Generalization from Various Values of Lambda: 0.1 - 1.0



Q1.3



I think the polynomial used to generate this model has a degree of 2 or 3. It is difficult to infer the exact degree from the visualization because there appears to be multiple degrees of freedom in the graph. A cross-validation of the polynomial degree may help achieve better clarity.

2.1

Figure 2.0 - Stochastic Gradient Descent Fit - Training Data

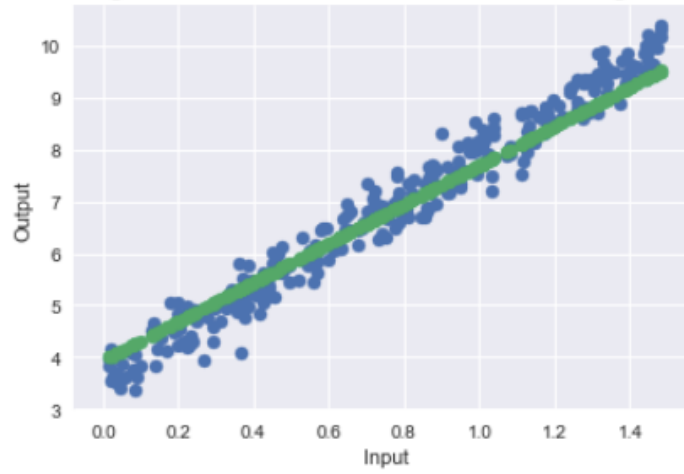


Figure 2.1 - Stochastic Gradient Descent Fit - Test Data

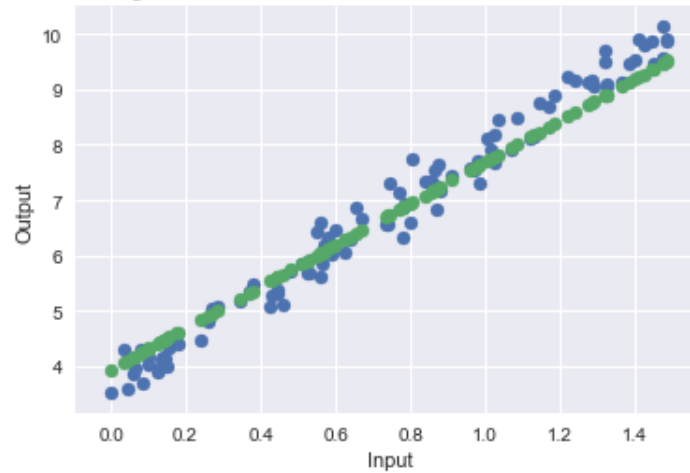


Figure 2.2 - MSE Error for Q2 - Training Set

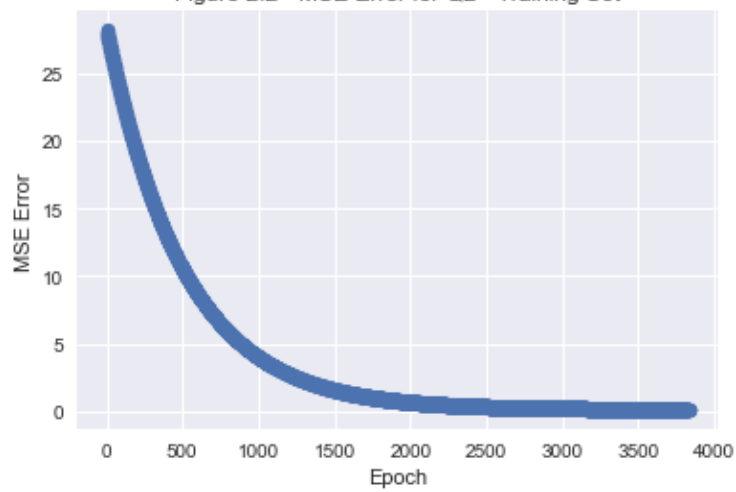


Figure 2.3 - MSE Error for Q2 - Test Set

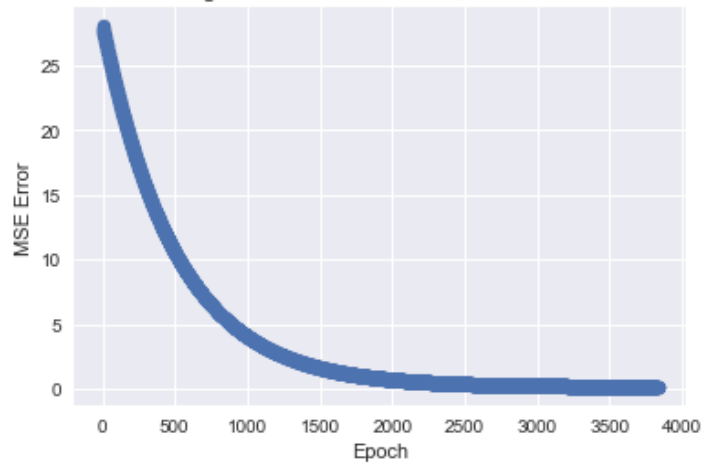
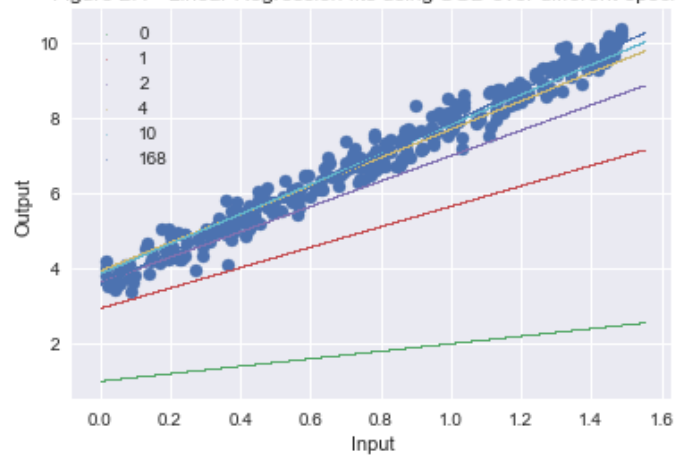


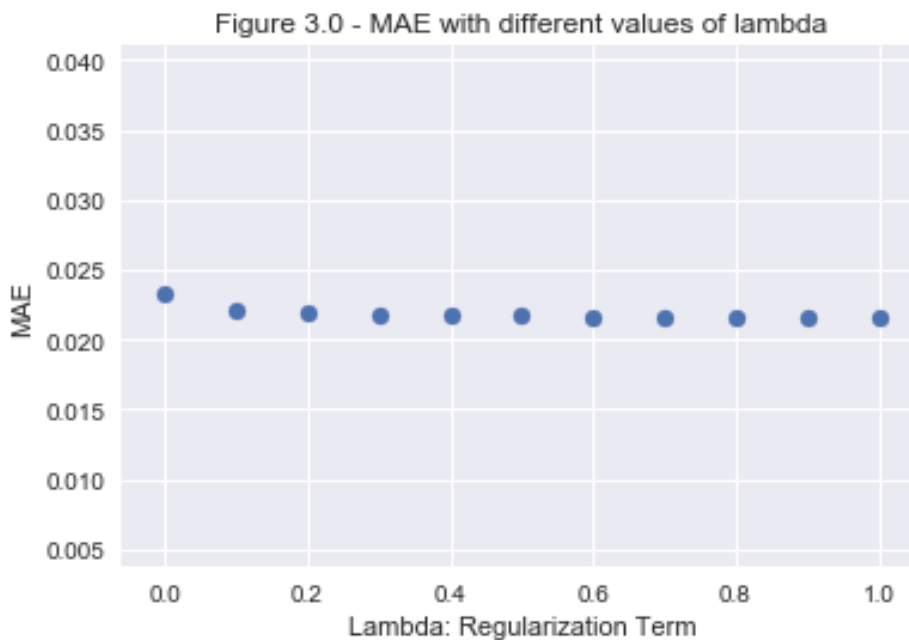
Figure 2.4 - Linear Regression fits using SGD over different epochs



3.1

The process of replacing a missing value with the mean is preferred over an alternative like deleting the columns with missing data as it does not reduce the effective sample size. The samples with missing data points may follow a pattern that is not random and deleting these would cause the new sample set to not be representative of the original. Furthermore the mean is fairly simple to calculate and substitute in but reduces the variability and variance of the data, thus introducing some bias. [1] That being said I do not have a better method and left it as is.

3.2

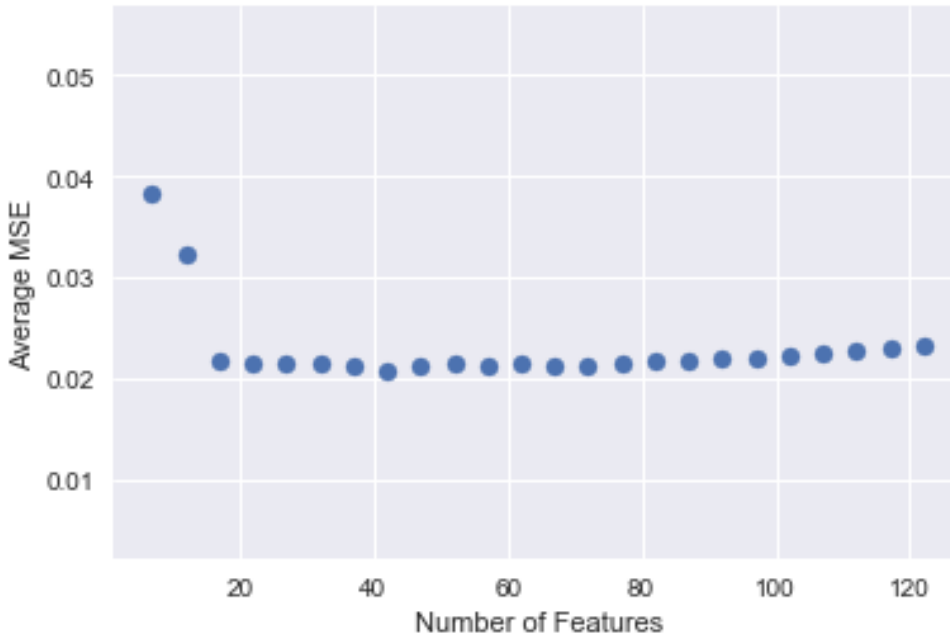


The MSE does not depreciate significantly with L2 regularization. It seems to plateau around 0.035 so a regularization constant lambda of 0.2 was chosen since there isn't significant improvement after this point in the cross validation error. From this regularization, the coefficients of the new model were looked at to help with feature selection.

3.3

Starting at the full set of 122 features, the 5 with the least absolute value were removed progressively. Then a new model was retrained and again the coefficients were removed. For each training, the average MSE was computed.

Figure 3.1 - Feature Set Reduction using Coefficients of Regularied Linear Model



The best performance of the model appeared to be around 40 features. This model was extracted from the list and the feature set along with the weights is shown in the Jupyter notebook.

References

[1] Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402.

Olivier Simard-Morissette

260563480

January 26, 2018