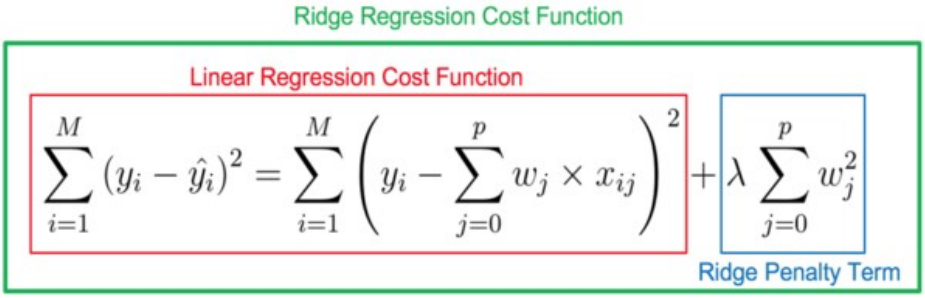
Assignment - Advanced Regression - Problem Statement - Part II

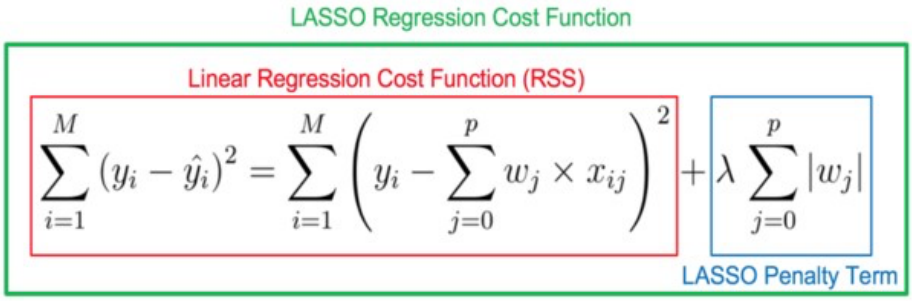
Question 1:
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

1. The optimal alpha value for ridge regression is 20, and lasso regression is 0.0003
2. If we choose to double the value of alpha for both Ridge and Lasso, then we know that for cost function for Ridge is



- And if the lambda value is increased by double, the value of the Ridge penalty term will be doubled



- And same goes for Lasso
- Model change increased in Root mean squared error value o Decreased R2 values

3. Most important predictor variables after the change is implemented
    - Ridge
        i From OverallQual (0.219164) to OverallQual (0.223053) o Lasso
        ii From Condition2_PosN (-3.164559) to Condition2_PosN (-3.164559)

Question 2:
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:
For given lambda value, I will go with **Ridge** regression when there is multicollinearity in the data and use **Lasso**, when we have to perform variable selection. And there is below given statics in this support of my selection:

	Alpha	R-squared Train	R-squared Test	RMSE Train	RMSE Test
Ridge	20	0.911	0.894	0.299	0.322
Lasso	0.0003	0.945	0.857	0.235	0.375

We have found that Lasso has a higher R2 for training but has less R2 for the test, the gap between it is around 9%. On the other hand, we have Ridge, which has optimal R2 for both train and test. That's why I choose Ridge. To make it more optimal, I can use RFE on Ridge to make it better.

Question 3:
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
After building the model, there are the top 5 most essential predictor variables after removing the top 5 most important predictor variables in the **Lasso**.

Condition2_Norm	0.325719
Neighborhood_NridgHt	0.316924
Neighborhood_Crawfor	0.29781
Neighborhood_StoneBr	0.295042
GrLivArea	0.27871

Question 4:
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:
Robustness is a property that **checks how good our algorithm** is when being tested on the **new dataset(test).**

Basically, it should show:

1. **Low test error**
2. **Test error should be near training error**

So we have a model with low training and test error, and both the error should be closed and which can be achieved by using **Regularization**.

If the accuracy is **not maintained**, then the model can be **underfitted** or **overfitted**.