# ADVANCED REGRESION ASSIGNMENT – SUBJECTIVE QUESTIONS

**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**

Using Cross Validation with 5 folds, I was able to determine the optimal values for alpha as given below :-

1. Ridge Regression: Alpha = **0.035**
2. Lasso Regression: Alpha = **0.0001**

Upon doubling the optimal values, the following were observed in each of the regression approaches :-

1. Ridge Regression: Alpha = **0.070**
2. Lasso Regression: Alpha = **0.0002**

**Ridge Regression Coefficients– Alpha: 0.035 vs 0.070**

| | | |
|---|---|---|
| 15 | Exterior1st_BrkComm | -2.418645 |
| 14 | Exterior1st_AsphShn | -0.822021 |
| 11 | HouseStyle_2.5Unf | -0.580483 |
| 5 | LotShape_IR3 | -0.487257 |
| 3 | House_Age_Current | -0.245971 |
| 10 | HouseStyle_1Story | 0.181452 |
| 0 | LotArea | 0.181856 |
| 13 | HouseStyle_SLvl | 0.198458 |
| 8 | Neighborhood_NridgHt | 0.247829 |
| 16 | Exterior1st_BrkFace | 0.268833 |
| 1 | OverallQual | 0.316883 |
| 20 | GarageType_Detchd | 0.316907 |
| 17 | Exterior2nd_Brk Cmn | 0.332469 |
| 7 | Neighborhood_NoRidge | 0.348419 |
| 19 | GarageType_BuiltIn | 0.358170 |
| 18 | GarageType_Attchd | 0.368368 |
| 2 | GrLivArea | 0.401316 |
| 12 | HouseStyle_SFoyer | 0.406895 |
| 9 | Neighborhood_StoneBr | 0.413361 |
| 4 | MSSubClass_75 | 0.452729 |
| 6 | Neighborhood_Crawfor | 0.452932 |

| | | |
|---|---|---|
| 15 | Exterior1st_BrkComm | -2.321792 |
| 14 | Exterior1st_AsphShn | -0.794963 |
| 11 | HouseStyle_2.5Unf | -0.574118 |
| 5 | LotShape_IR3 | -0.485061 |
| 3 | House_Age_Current | -0.245942 |
| 10 | HouseStyle_1Story | 0.180902 |
| 0 | LotArea | 0.181598 |
| 13 | HouseStyle_SLvl | 0.198064 |
| 8 | Neighborhood_NridgHt | 0.247565 |
| 16 | Exterior1st_BrkFace | 0.268718 |
| 17 | Exterior2nd_Brk Cmn | 0.313777 |
| 20 | GarageType_Detchd | 0.316613 |
| 1 | OverallQual | 0.317089 |
| 7 | Neighborhood_NoRidge | 0.347754 |
| 19 | GarageType_BuiltIn | 0.357486 |
| 18 | GarageType_Attchd | 0.368192 |
| 2 | GrLivArea | 0.401252 |
| 12 | HouseStyle_SFoyer | 0.405859 |
| 9 | Neighborhood_StoneBr | 0.412309 |
| 4 | MSSubClass_75 | 0.447056 |
| 6 | Neighborhood_Crawfor | 0.452351 |

## Lasso Regression Coefficients – Alpha: 0.0001 vs 0.0002

| | | |
|---|---|---|
| 15 | Exterior1st_BrkComm | -2.376864 |
| 14 | Exterior1st_AsphShn | -0.750829 |
| 11 | HouseStyle_2.5Unf | -0.550493 |
| 5 | LotShape_IR3 | -0.478013 |
| 3 | House_Age_Current | -0.245659 |
| 10 | HouseStyle_1Story | 0.179876 |
| 0 | LotArea | 0.181524 |
| 13 | HouseStyle_SLvl | 0.194946 |
| 8 | Neighborhood_NridgHt | 0.245173 |
| 16 | Exterior1st_BrkFace | 0.266106 |
| 17 | Exterior2nd_Brk Cmn | 0.308480 |
| 20 | GarageType_Detchd | 0.313383 |
| 1 | OverallQual | 0.317845 |
| 7 | Neighborhood_NoRidge | 0.343691 |
| 19 | GarageType_BuiltIn | 0.353357 |
| 18 | GarageType_Attchd | 0.365609 |
| 2 | GrLivArea | 0.401047 |
| 12 | HouseStyle_SFoyer | 0.401663 |
| 9 | Neighborhood_StoneBr | 0.407029 |
| 4 | MSSubClass_75 | 0.424086 |
| 6 | Neighborhood_Crawfor | 0.449790 |

| | | |
|---|---|---|
| 15 | Exterior1st_BrkComm | -2.229725 |
| 14 | Exterior1st_AsphShn | -0.650620 |
| 11 | HouseStyle_2.5Unf | -0.514008 |
| 5 | LotShape_IR3 | -0.466516 |
| 3 | House_Age_Current | -0.245316 |
| 10 | HouseStyle_1Story | 0.177711 |
| 0 | LotArea | 0.180910 |
| 13 | HouseStyle_SLvl | 0.191026 |
| 8 | Neighborhood_NridgHt | 0.242252 |
| 16 | Exterior1st_BrkFace | 0.263276 |
| 17 | Exterior2nd_Brk Cmn | 0.264090 |
| 20 | GarageType_Detchd | 0.309591 |
| 1 | OverallQual | 0.319021 |
| 7 | Neighborhood_NoRidge | 0.338289 |
| 19 | GarageType_BuiltIn | 0.347882 |
| 18 | GarageType_Attchd | 0.362717 |
| 4 | MSSubClass_75 | 0.389645 |
| 12 | HouseStyle_SFoyer | 0.395363 |
| 9 | Neighborhood_StoneBr | 0.399631 |
| 2 | GrLivArea | 0.400711 |
| 6 | Neighborhood_Crawfor | 0.446060 |

In both cases, it was observed that that the r2_score for train and test marginally increased but the difference could be considered negligible. The most important predictor variables were found to be the same **21** variables. In a few cases however, the coefficients of the predictor variables slightly changed indicating a change in their influence on the dependent variable, Sale Price. Lasso Regression did not eliminate any new features in the process. The above images can be used to compare and arrive at these conclusions.

### 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

I would choose Ridge Regression over Lasso. One of the reasons is that the r2_score for Ridge is only marginally lesser than Lasso's in test dataset to the point that it is negligible. But the primary reason is when considering computational resources. It is a known fact that Lasso uses

an iterative process in arriving at the final model that is computationally expensive when compared to Ridge. The advantage of Lasso is that it is able to eliminate unnecessary features by bringing their coefficients down to zero. Since in our case, Lasso did not eliminate any features from the final **21**, it loses its primary advantage. Therefore if I were to consider scaling up my model, I would prefer Ridge Regression in my case.

**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**

The original 5 most important variables with their corresponding coefficients as per the Lasso Regression model are the following:-

1. Exterior1st_BrkComm (-2.376864)
2. Exterior1st_AsphShn (-0.750829 )
3. HouseStyle_2.5Unf (-0.550493)
4. LotShape_IR3 (0.449790)
5. Neighborhood_Crawfor (0.424086)

Upon removing the above from the incoming data, the following are the new top predictor variables with the corresponding coefficients within brackets:-

1. GrLivArea (0.409238)
2. HouseStyle_SFoyer (0.405678)
3. Neighborhood_StoneBr (0.391068)
4. GarageType_Attchd (0.388291)
5. GarageType_BuiltIn (0.344850)

**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**

For a model to be generalizable, we need to consider the bias variance tradeoff. Too much variance results in a model that makes little mistakes in training but is way off in predicting unseen data because of overfitting. Too much bias can be a result of an overly generalized or simplistic model which reduces accuracy and that isn't useful either. Our goal to a robust and generalizable model can be achieved only if we strike a balance between variance and bias. This is where the role of a regularizer comes in. By adding a regularizer next to the error term in the model's equation, we are able to control the complexity of the model by penalizing and preventing complex coefficients or too much slope from creeping into the model. However, we need to be careful to set the right amount of penalty as the regularizer could be the difference between overfitting or an overly generalized model, both resulting in poor accuracy in real world scenarios. By finding the optimal regularizer for our model (through approaches like cross validation), we are able to balance variance and bias and in the process deliver a model that can generalize just enough to hit the right accuracy with unseen data.