# Advanced Regression Assignment Subjective Questions

**Q.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?**

**Ans.** Optimal value of alpha for Ridge Regression comes out to be 30 and optimal value of alpha for Lasso Regression comes out to be 200.

If we consider the above values of alpha for building the model using Ridge and Lasso regression then the accuracies are as follows:

- Accuracy on Train set using Ridge Regression: 87.25%
- Accuracy on Test set using Ridge Regression: 86.25%
- Accuracy on Train set using Lasso Regression: 87.53%
- Accuracy on Test set using Lasso Regression: 86.03%

The number of coefficients present in the final model using Lasso regression are 76.

If we consider the values of alpha for building the model using Ridge and Lasso regression double the previous values that is alpha = 60 for Ridge regression and alpha = 400 for Lasso Regression, then the accuracies are as follows:

- Accuracy on Train set using Ridge Regression: 86.10%
- Accuracy on Test set using Ridge Regression: 85.83%
- Accuracy on Train set using Lasso Regression: 86.16%
- Accuracy on Test set using Lasso Regression: 85.17%

The number of coefficients present in the final model using Lasso regression are 54.

**We can observe that as we increase the value of alpha more than optimal value, the accuracy of the model starts decreasing gradually on the Train set and Test set.**

**Predictors:** We have considered the final model using Lasso regression here.

| **Predictors with alpha = 200 (Lasso)** | **Predictors with alpha = 400 (Lasso)** |
|---|---|
| MSSubClass | MSSubClass |
| Neighborhood_NridgHt | Neighborhood_NridgHt |
| BsmtFullBath | BsmtFullBath |
| Neighborhood_OldTown | OverallCond |
| OverallCond | Neighborhood_OldTown |

**We can observe that when we change the value of alpha, then some coefficients that contribute most to model building  also change.**

**Q.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?**

**Ans:** Once the Data is cleaned and Scaled, firstly the model is built using Ridge Regression. After performing the k fold cross validation using different values of alpha, the graph is plotted which denotes alpha value vs negative absolute mean error. From the graph we can infer that the optimal value of alpha for Ridge regression is 30. We get good accuracy on both Train and Test sets using alpha = 30 in case of Ridge regression.

Similarly for Lasso regression we perform k fold cross validation using different values of alpha. When the graph is plotted for alpha value vs negative absolute mean error, we can infer that the optimal value of alpha for Lasso regression is 200. Here also the accuracy is good for both Train and Test data.

Both the Models are giving good accuracy on the Train and Test set. We decided to keep Model built using Lasso Regression because it gave few extra benefits like dimensionality reduction which keeps the valuable features intact and eliminates less important features.

¶

**Q.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?**

**Ans:** If the five most important predictor variables in the lasso model are not available in the incoming data, then we have to delete the five most predictors from the train set and rebuild the model with remaining features.

Previous list of five most important predictors:

| | Featuere | Coef |
|---|---|---|
| **0** | MSSubClass | 190436.444697 |
| **59** | Neighborhood_NridgHt | 45571.176323 |
| **13** | BsmtFullBath | 30970.862227 |
| **60** | Neighborhood_OldTown | 28835.251836 |
| **4** | OverallCond | 23340.602926 |

We removed the five most important predictor variables in the lasso model and got the optimal value of alpha as 30 for Ridge regression and 200 for Lasso Regression for the new model.

We choose to go with the Lasso Regression model which has accuracy of 86.63% on Train set and 85% on Test set.

After rebuilding the model we get following new list of most important predictors:

| | Featuere | Coef |
|---|---|---|
| **0** | LotFrontage | 205114.681501 |
| **56** | Neighborhood_SWISU | 39557.374823 |
| **11** | BsmtHalfBath | 30135.476641 |
| **3** | MasVnrArea | 26680.066703 |
| **47** | Neighborhood_Edwards | 17454.198547 |

**Q.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?**

**Ans:** Ideally the model is robust and generalisable when it works well on unseen data. We should not let the model learn the entire dataset and get complex unnecessarily.

If the accuracy on the Train set and Test set is almost the same then the model can be said as generalisable. But when the accuracy on the Train set is high and on the Test set it is very low then there is definitely an issue of overfitting which has to be overcome.

In order to make the model robust, apply proper regularisation technique on the dataset, perform hyperparameter tuning  and perform outlier treatment as and when required.

Even if accuracy is moderate on Train set and the model is working well for unseen data then the model can be termed as robust and generalisable.