# Advanced Regression Assignment – Part II

**Student:** Jheser Guzman Illanes
**Email:** dicotips@gmail.com

## Assignment-based Subjective Questions

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?**

   a. **What is the optimal value of alpha for ridge and lasso regression?**

   Based on the results from "AdvancedRegressionAssignment" this repository Jupyter Notebook (Section "Final Conclusion and Observations"), the optimal value for alpha Lasso and Ridge Regression are:

   - *Ridge Regression = 10*
   - *Lasso Regression = 0.0002.*

   b. **What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?**

   If we change the Alpha for Ridge and Lasso Regression to 20 and 0.0004, respectively, we will obtain:

   **Ridge Regression:**

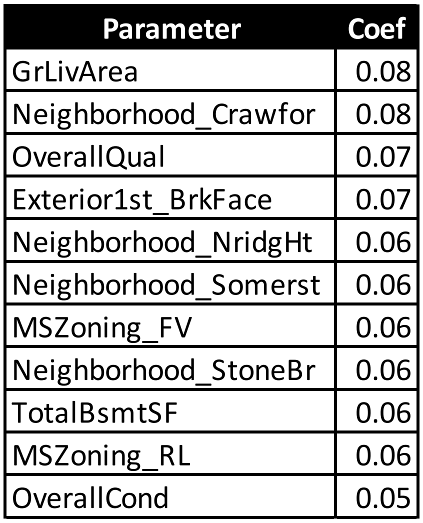
   - **R2 Score Train Change:** from 0.922 to 0.919 (decrease)
   - **R2 Score Test Change:** from 0.892 to 0.891 (decrease)

   **Lasso Regression:**

   - **R2 Score Train Change:** from 0.926 to 0.922 (decrease)
   - **R2 Score Test Change:** from 0.894 to 0.893 (decrease)

   c. **What will be the most important predictor variables after the change is implemented?**

   Based on the results from "AdvancedRegressionAssignment" this repository Jupyter Notebook (Section "Final Conclusion"), the most important predictors are:



| Parameter | Coef |
|---|---|
| GrLivArea | 0.08 |
| Neighborhood_Crawfor | 0.08 |
| OverallQual | 0.07 |
| Exterior1st_BrkFace | 0.07 |
| Neighborhood_NridgHt | 0.06 |
| Neighborhood_Somerst | 0.06 |
| MSZoning_FV | 0.06 |
| Neighborhood_StoneBr | 0.06 |
| TotalBsmtSF | 0.06 |
| MSZoning_RL | 0.06 |
| OverallCond | 0.05 |

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?**

   GridSearchCv and KFold Cross Validation allow us to determine the optimal value for alpha in Ridge and Lasso Regression. We set up the HyperParameters in order to extract the optimal alpha.

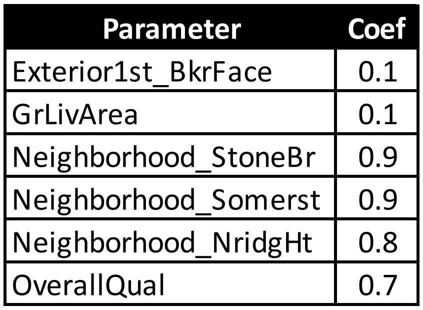
   All regression models used in the analysis However all the models are showing close value of R2_score. *Lasso Regression* is slightly better with a difference of +0.01

3. **After building the model, you realized 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?**

   The top 5 most important predictors in the initial analysis are:

   - MSZoning_FV
   - MSZoning_RL
   - MSZoning_RH
   - MSZoning_RM
   - Neighborhood_Crawfor

   If we rerun the analysis removing those variables from the initial dataset, the new predictors with Lasso Regression are:



| Parameter | Coef |
|---|---|
| Exterior1st_BkrFace | 0.1 |
| GrLivArea | 0.1 |
| Neighborhood_StoneBr | 0.9 |
| Neighborhood_Somerst | 0.9 |
| Neighborhood_NridgHt | 0.8 |
| OverallQual | 0.7 |

4. **How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?**

Overfitting and Underfitting concept allow us to determine if a model is robust or not. To verify if the model is robust and generalizable for unknown data, we need to compare the Accuracy and performance metrics between the Training and the Test datasets.

A sign of overfitting is when the Accuracy is higher in the Training data compared against the Test data. This means that the model learned the datapoints from the training dataset including its noise.

In order to avoid overfitting in our models, I use Regularization, which will reduce the variance by compromising bias. This will make the model more robust and generalizable when the accuracy based on the Test data increase.