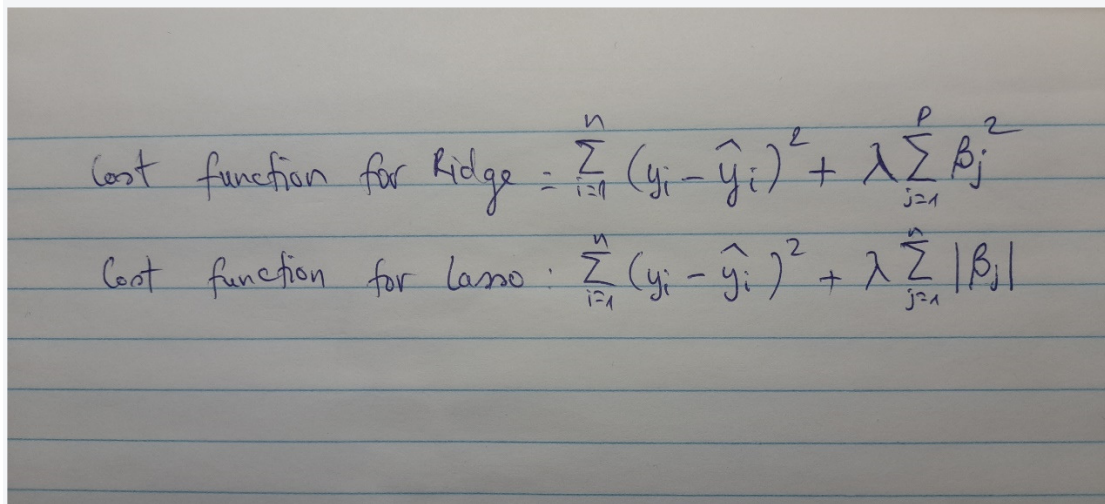


Question 1

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?

Ans:

- + The optimal value of alpha for ridge regression is 10.
- + The optimal value of alpha for lasso regression is 0.05



Cost function for Ridge = $\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$

Cost function for Lasso = $\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$

- Doubling the value of the optimal alpha in Ridge regression would increase the amount of shrinkage applied to the coefficients, resulting in smaller coefficient values. This can help reduce the impact of individual features and mitigate the effects of multicollinearity.

Top 10 predictor variables after the change in Ridge model:

RoofMatl_CompShg	0.241392
RoofMatl_Tar&Grv	0.197429
MSZoning_RL	0.194236
GrLivArea	0.175307
PoolArea	0.171047
PoolQC_No Pool	0.150960
MSZoning_RM	0.150102

OverallQual	0.139877
YearBuilt	0.124751
1stFlrSF	0.116565

So the most important predictor in Ridge model after the change is RoofMatl_CompShg.

- Doubling the value of the optimal alpha in Lasso regression would increase the penalty for non-zero coefficients, leading to a more sparse solution. This means that more features would be shrunk to zero, resulting in a model with fewer features and potentially improved interpretability

Top 10 predictor variables after the change in Lasso model:

OverallQual	0.348582
GrLivArea	0.208030
GarageCars	0.142838
YearRemodAdd	0.076889
YearBuilt	0.060193
TotalBsmtSF	0.048129
MSZoning_RM	-0.042853
1stFlrSF	0.034942
CentralAir	0.024102
Fireplaces	0.022590

So the most important predictor in Ridge model after the change is OverallQual .

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?

Ans:

Below is the summary comparison between the 2 models:

- Ridge Regression model on test dataset: r^2 score= 0.784, MAE= 0.256, RMSE= 0.4584
- Lasso Regression model on test dataset: r^2 score= 0.8472, MAE= 0.2696, RMSE= 0.385

Lasso model has almost the same r^2 between Training & test datasets, whereas Ridge has higher r^2 value in Training dataset than in Test dataset.

Lasso has fewer features after eliminating the features as 33, in comparison to Ridge's number of features 270, thus the Lasso model is simpler and this implies better performance.

In conclusion, I'd choose the Lasso model.

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?

Ans:

Top 5 most important predictor variables in Lasso model.

OverallQual	0.348582
GrLivArea	0.208030
GarageCars	0.142838
YearRemodAdd	0.076889
YearBuilt	0.060193

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?

Ans:

- How to make sure a model is robust and generalizable:
 - Use sufficient and dataset: use a large enough training dataset and well-diverse would help the model learn the pattern better
 - Perform Train-Test splitting practice on dataset: Separate into 2 dataset so we can evaluate the performance of the model on the unseen data in Test dataset
 - Regularization: apply regularization to prevent overfitting and improve model's generalization.
 - Hyper-parameter tuning: This help find the optimal configuration for the model, and help keep underfitting and overfitting balanced.
- Implications of the same for the accuracy of the model:
 - Accuracy in training may not reach 100%: when we prioritize generalization, it's more important for the model to be able to recognize the underlying pattern and perform well on unseen data, rather than perfectly fit the training data but poorly perform on unseen data.
 - Reduce the risk of overfitting: leveraging regularization techniques would help reduce the complexity of the model as well as avoid overfitting.