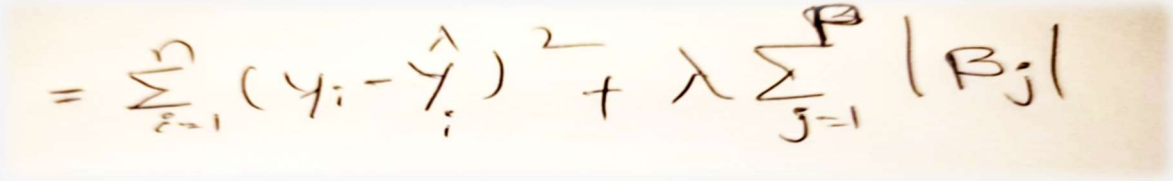
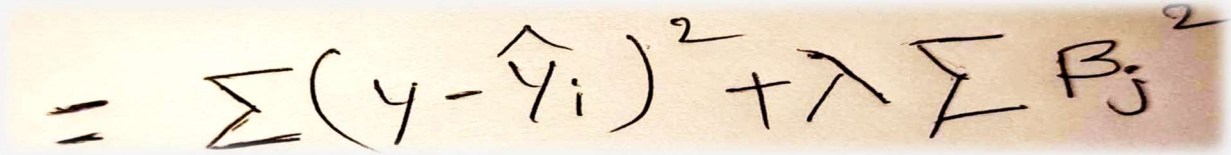
Question 1:

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?

Lasso Regression is one of the methods called Regularized regression method to deal with model overfit case of fewer datapoints are available than the number of parameters that we need to estimate. Lasso regression can also be used for feature selection since it shrinks the coefficient to zero.
Cost function for Losso regression is



Ridge Regression is also regularization technique which is used to analyze multiple regression data which is suffering from multicollinear. Ridge regression also use penalty to minimize residual sum of squares. In Ridge regression the model coefficient can shrink very near to zero but cannot be zero. Because of this there is no feature selection in Ridge regression.
Cost function for Ridge Regression is



Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced.

Optimal Value of alpha for Ridge

- Ridge – 2.0
- Lasso – 6e-05

Top 5 most significant variables in Ridge are:

```
OverallQual_Excellent =0.201510
Neighborhood_NridgHt  =0.137584
OverallQual_Very Good =0.128839
Neighborhood_NoRidg   =0.126702
Neighborhood_Crawfor  =0.117898
```

Top 5 most significant variables in Lasso are:

```
SaleCondition_Partial =0.231
SaleCondition_Others  =0.161
SaleCondition_Normal  =0.157
GarageFinish_Unf      =0.144
GarageFinish_RFn      =0.140
```

Double the values of alpha

- Ridge – 4.0
- Lasso - 0.00012

Top 5 most significant variables in Ridge are:

```
OverallQual_Fair   0.180887
Neighborhood_OldTown      0.121967
OverallQual_Very Poor     0.121187
Neighborhood_Edwards      0.108935
Neighborhood_StoneBr      0.108012
```

Top 5 most significant variables in Lasso are:

```
SaleCondition_Partial =0.226
SaleCondition_Others  =0.142
SaleCondition_Normal  =0.157
GarageFinish_Unf      =0.139
GarageFinish_RFn      =0.134
```

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?

I have observed decent results in both Losso and Ridge regression. But Losso provides feature selection option , so I will apply Lasso regression.

Question 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?

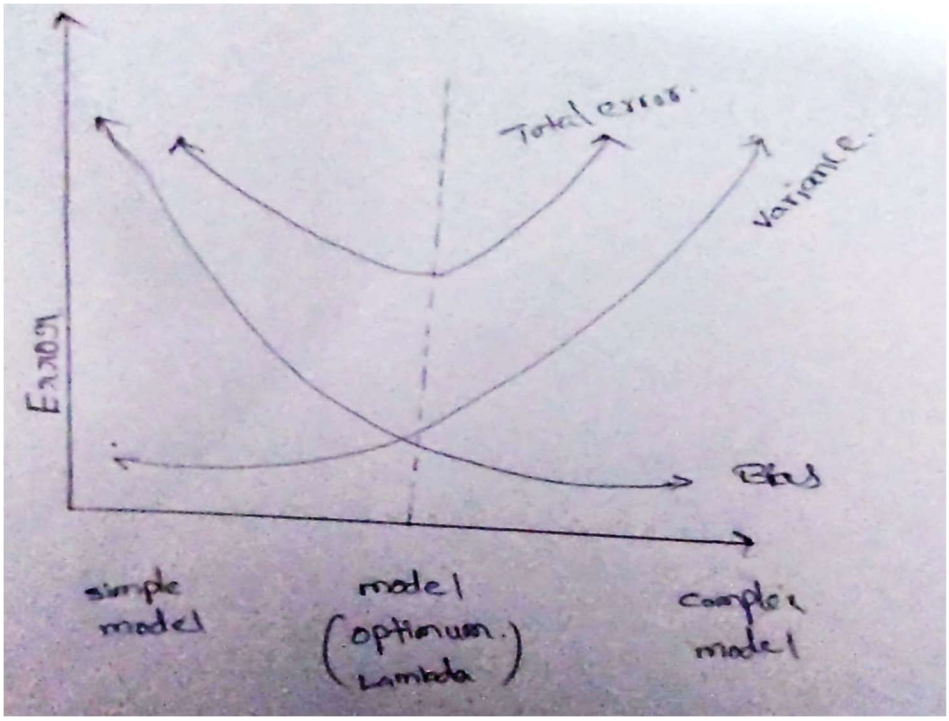
After removing the top 5 predictor variable in lasso model below are the new top 5 variable identified in new model where the 5 are excluded.

```
GarageFinish_No Garage 0.219
GarageType_Others', 0.158
GarageType_No Garage', 0.154
GarageType_Detchd', 0.145
GarageType_BuiltIn', 0.137
```

Question 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?

A model should be as simple as possible to make it robust and generalize. A model complexity depends on magnitude of coefficient and number of coefficients.



Regularization helps to manage the model complexity by essentially shrinking the model coefficient estimates towards zero. This discourages the model from becoming model overfit.

Extreme value of coefficient makes the model more complex then higher the overfitting.

It is trade of between variance and Bias. The higher lambda the simpler the model and lower the lambda the model becomes complex.

The implication of accuracy in terms is that a robust model should perform equally on both data sets that is seen(train) and unseen data set (test).