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: Optimal value of Alpha for ridge and lasso regression is that minimizes prediction error MSE on a chosen validation set. It's determined using cross validation by trying different alpha values.

In the case of Ridge regression Alpha controls the amount of regularization applied on model. Increase in alpha increases regularization by penalizing the size of coefficients. In the case of lasso regression Alpha controls amount of regularizations including additional effect of performing feature selection by driving some coefficients to exactly 0.

If we double the value of alpha for both ridge and lasso, in case of ridge, the model will further penalize the coefficients but will not drive them to zero .so model will continue to use all predictors variables with little change in performance. Whereas in case of lasso penalty term on coefficients will be doubled resulting more coefficients driven to exactly 0 and so the model will have lesser predictors variables.

Most important predictors will be the ones with nonzero coefficients in lasso regression model.

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?

For this assignment I found Optimum lambda value for Ridge is 5 and Lasso is 0.0001

Ridge: MSE(test) 0.011000494608608283

r2_train_ridge: 0.91632583459692

r2_test_ridge: 0.9138335990775496

Lasso:MSE(test) 0.011195981595568857

r2_train_lasso: 0.9167457542585093

r2_test_lasso: 0.9123023579204124

AS per result obtained from both regression

Mean square error in case of Lasso is slightly higher than Ridge. And R square values(test) is higher in case of Ridge than Lasso.

Lasso helps in feature reduction as the coefficient value of one of the feature became 0. So based on Lasso features that can affect selecting the Slaes price are Zoning classifications, Living

area square feet, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accommodated in the garage, Total basement area in square feet and the Basement finished square feet area.Therefore, the variables predicted by Lasso in the above bar chart as significant variables for predicting the price of a house.

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?

In such case we have to perform Lassp regression again on dataset excluding those five predictors that are not in incoming data. This will help identify the most important predictors among the remaining variables. After rebuilding the model, the five most important predictor variables will be the ones with the highest coefficients in the Lasso model. These variables will have the most significant impact on the predictions, considering the revised dataset.

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?

Model is said to be more robust if it's able to identify and generalize the patterns in the data than memorizing them, variation in data does not affect its performance. Robustness can be tested by multiple validation methods on unseen data.

A generalizable model that avoids overfitting which can be achieved by regularization. A overfitted model has high variance and change in data can affect unseen test data, failing to identify all patterns of unseen test data. The model shouldn’t be too complex to be robust and generalizable.

A complex model can have high accuracy but to get the model robust and generalizable we have to tradeoff between complexity and maintaining the balance between bias and variance of error by applying ridge and lasso regression method of regularization and selecting the predictors that are significant for target.