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?

Answer 1
Optimal value of Alpha for Ridge - 4
Optimal value of Alpha for Lasso - 0.0001

The regularization strength in ridge and lasso regression is controlled by the parameter alpha. Increasing the value of alpha would result in a stronger regularization, leading to a reduction in the magnitude of the coefficients for each feature, i.e. more feature coefficients would be shrunk towards zero.

So if you double the value of alpha for both ridge and lasso, the most important predictor variables after the change is implemented would be those with non-zero coefficients in the lasso model. These variables would be considered the most significant in terms of their effect on the target variable because their coefficients would have a stronger regularization, which would make them more robust and less susceptible to overfitting.

After doubling the value of alpha in a lasso regression model, the most important predictor variables would be those with non-zero coefficients. The lasso regression not only reduces the magnitude of the coefficients for each predictor but also performs feature selection by setting some coefficients to zero. So, the variables with non-zero coefficients in the lasso model would be considered the most significant in terms of their effect on the target variable.

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?

Answer 2
Since, the relationship between the predictors and the target variable is approximately linear, ridge regression may be a better choice because it shrinks the coefficients towards zero but never sets them exactly to zero.

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?

Answer 3

If we exclude the five most important predictor variables as determined by the lasso model, the new most important predictor variables would be those that remain in the model. To determine the new most important predictor variables, you would have to refit the model without the excluded variables and evaluate the coefficients for the remaining variables.

It's important to note that excluding predictor variables can impact the performance of the model, and it may be necessary to try different models with different subsets of the predictors to determine the best possible model. Additionally, it may be possible to use feature engineering techniques to create new predictors that are related to the excluded variables, which could help to mitigate the impact of excluding those variables from the model.

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?

Answer 4

To make sure that a model is robust and generalizable, it's important to follow a number of best practices in model development and evaluation:

- *Split the data into training and testing sets:* It's important to evaluate the model's performance on a holdout set of data that was not used for training to get a sense of its generalization performance. This can be done by splitting the data into a training set, used for model development, and a testing set, used for evaluation.

- *Use cross-validation:* Cross-validation is a technique that involves repeatedly splitting the data into training and validation sets, fitting the model on the training set and evaluating its performance on the validation set. This provides a more robust estimate of the model's generalization performance, as it takes into account multiple splits of the data.

- *Regularize the model:* Regularization is a technique used to prevent overfitting, which is when a model fits the training data too closely and has poor generalization performance. Lasso and ridge regression are two common forms of regularization used in linear models.

- *Use multiple metrics to evaluate model performance:* It's important to use multiple metrics to evaluate the model's performance, rather than relying on a single metric such

as accuracy. For example, precision, recall, F1 score, and ROC AUC are common metrics used in classification problems.

- *Avoid overfitting:* Overfitting occurs when a model fits the training data too closely, leading to poor generalization performance. To avoid overfitting, it's important to choose a model that is not too complex for the data and to use techniques such as regularization and early stopping.

The implications of making sure a model is robust and generalizable for its accuracy are that a model that is robust and generalizable will have higher accuracy on new, unseen data. On the other hand, a model that is not robust and generalizable may have high accuracy on the training data but low accuracy on new data, which is a hallmark of overfitting. By following the best practices outlined above, you can increase the chances that your model will be robust and generalizable and will have high accuracy on new data.