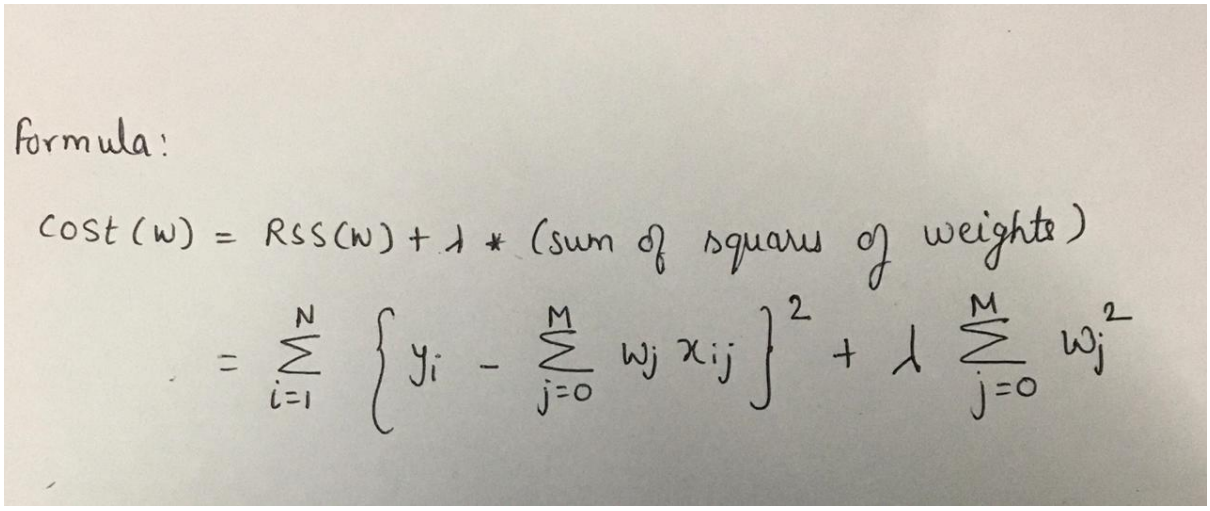
Question 1

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?

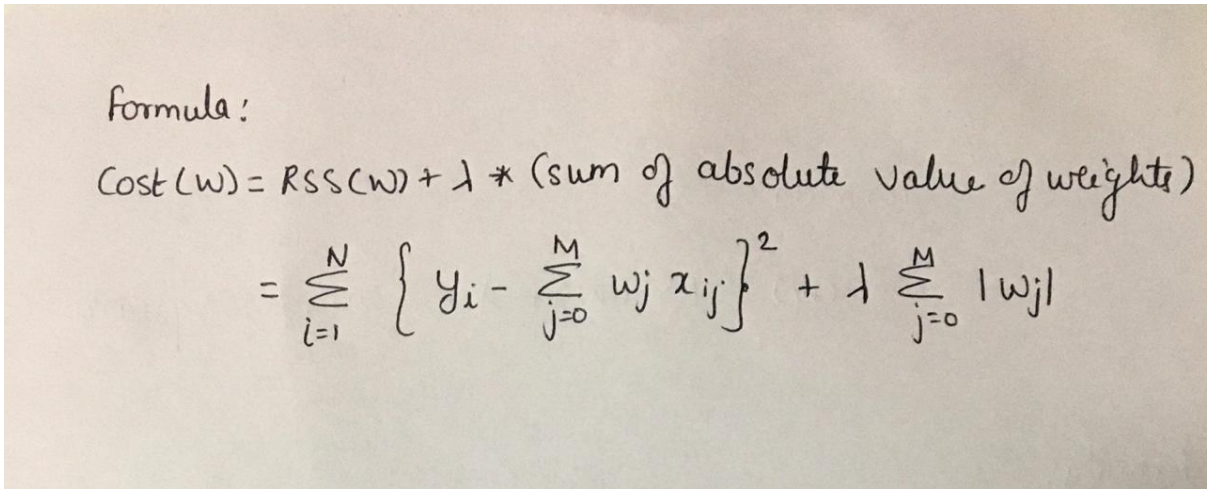
Answer:

Ridge regression:
Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.



For ridge regression, the optimal value of alpha is 20.

Lasso Regression :
Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).



In this case of Lasso regression, the optimal value for alpha is 1.

If we choose double the value of alpha for both ridge and lasso regression, model complexity will have a greater contribution to the cost. Because the minimum cost hypothesis is selected, this means that higher λ will bias the selection toward models with lower complexity.

After second model is build we compare the r square value of new model with the old one. The model which is having high r square of test and train dataset , we will select the features/variables from that model. And the variable is selected based on the high coefficient value.

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:
Lasso regression would be a better option it would help in feature elimination and the model will be more robust. Because

- In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed. In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output.
- Ridge regression can't zero out coefficients; thus, you either end up including all the coefficients in the model, or none of them. In contrast, the LASSO does both parameter shrinkage and variable selection automatically.

- Lasso regression can produce many solutions to the same problem.

- Ridge regression can only produce one solution to one problem.

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?

Answer:

Statistical measures can show the relative importance of the different predictor variables. However, these measures can't determine whether the variables are important in a practical sense. To determine practical importance, you'll need to use your subject area knowledge.

How you collect and measure your sample can bias the apparent importance of the variables in your sample compared to their true importance in the population.

If you randomly sample your observations, the variability of the predictor values in your sample likely reflects the variability in the population. In this case, the standardized coefficients and the change in R-squared values are likely to reflect their population values.

However, if you select a restricted range of predictor values for your sample, both statistics tend to underestimate the importance of that predictor. Conversely, if the sample variability for a predictor is greater than the variability in the population, the statistics tend to overestimate the importance of that predictor.

Also, consider the accuracy and precision of the measurements for your predictors because this can affect their apparent importance. For example, lower-quality measurements can cause a variable to appear less predictive than it truly is.

How you define "most important" often depends on your goals and subject area. While statistics can help you identify the most important variables in a regression model, applying subject area expertise to all aspects of statistical analysis is crucial. Real world issues are likely to influence which variable you identify as the most important in a regression model.

For example, if your goal is to change predictor values in order to change the response, use your expertise to determine which variables are the most feasible to change. There may be variables that are harder, or more expensive, to change. Some variables may be impossible to change. Sometimes a large change in one variable may be more practical than a small change in another variable.

"Most important" is a subjective, context sensitive characteristic. You can use statistics to help identify candidates for the most important variable in a regression model, but you'll likely need to use your subject area expertise as well.

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:

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used. This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.

The best accuracy is 100% indicating that all the predictions are correct. For an imbalanced dataset, accuracy is not a valid measure of model performance. For a dataset where the

default rate is 5%, even if all the records are predicted as 0, the model will still have an accuracy of 95%