Advanced Linear Regression Model

Prashant Ramdas Patil

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?

The optimal value of alpha for ridge and lasso regression that I found was 100, 0.05 respectively. As the value of Alpha increases :-

1. **The RSS steadily increases**
2. **The variance decreases**
3. **Bias of the model increases**
4. **Test error will high**
5. **Some of the coefficient in LASSO regression will become 0.**

I find the following 5 important predictors with original Alpha :

For Ridge :

```
betas.Ridge.nlargest(n=5)
```

```
OverallQual              0.978797
GrLivArea                0.828889
GarageArea               0.655927
1stFlrSF                 0.575225
Neighborhood_NridgHt     0.573199
```

For Lasso :

```
betas.Lasso.nlargest(n=5)
```

```
GrLivArea                 2.063303
OverallQual               1.441807
YearBuilt                 0.696347
GarageArea                0.631497
Neighborhood_NridgHt      0.630043
```

By doubling the value of Alpha , the model generates following top 5 parameters :

```
betas.Ridge.nlargest(n=5)
```

```
OverallQual               0.893527
GrLivArea                 0.726013
GarageArea                0.621882
1stFlrSF                  0.538231
Neighborhood_NridgHt      0.517016
Name: Ridge, dtype: float64
```

```
betas.Lasso.nlargest(n=5)
```

```
GrLivArea                 2.015418
OverallQual               1.645453
GarageArea                0.665562
BsmtFinSF1                0.585149
Neighborhood_NridgHt      0.583721
Name: Lasso, dtype: float64
```

I observed that that coefficients have changed and Garage area, overall quality are in top five of predictor values.

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?

Given the management wants to understand how Sale price of the houses vary with independent variables , it will be easier to explain with Lasso regression. The Lasso model pushes coefficients to zero and hence it be comes relatively easier to explain the model with fewer variables.

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?

I took out the top 5-predictor variables from training and test data set. On building the model following top 5 variables are seen:

```
betas1.Lasso.nlargest(n=5)
```

```
2ndFlrSF        1.883267
1stFlrSF        1.782655
YearBuilt       1.195983
TotalBsmtSF     0.932699
OverallCond     0.560943
Name: Lasso, dtype: float64
```

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?

To make the model robust and generalizable , I would think of following –

1. **Carefully inspecting every independent variable data. Handle missing values carefully based on nature of that variable and selecting appropriate imputing methods or corrections in the**
2. **Addressing outliers in the target variable. Also if some independent variable occurrences is negligible small, I think of removing it before generating the mode that has large number of independent variable.**
3. **Keep optimal number of**
4. **Keep checking VIF scores to address multicollinearity in the model.**

5. **Regularization - Choosing the correct value of alpha while applying Ridge /Lasso model. In case where accuracy is more desired over explanation, I would think of choosing Ridge as it tends to keep all variable where as Lasso where model needs to be explained with respect to independent variables.**
6. **In addition while coming up with the model check and ensure for error terms –**
    a. **Residuals should be randomly scattered around 0.**
    b. **The spread of the residuals should be constant.**
    c. **There should be no outliers in the data**

A robust and generalizable model then helps addresses following issues –

1. **Non-constant variance**
2. **Autocorrelation and time series issue**
3. **Multicollinearity**
4. **Overfitting**
5. **Extrapolation**