Introduction to Regularisation: Ridge and LASSO Regression

Models should be as simple as possible, but not more so.
- Albert Einstein

Outline

Learning Objectives

In this video, you will learn to:

- Understand Regularisation can solve the Multicollinearity problem.
- Understand Regularisation can solve the Overfitting problem.
- Understand LASSO Regression is good for interpretability.

Introduction to the Multicollinearity and Overfitting Problems

Multicollinearity Problem

- Multicollinearity: Some predictor variables are strongly correlated.
- Multicollinearity can cause the following problems:
  1. Create inaccurate estimates of the regression coefficients, e.g., it may produce a wrong sign.
  2. Give false, or non-significant p-values.
  3. Degrade the interpretability, and the predictability of the model.

Overfitting Problem

- Overfitting may occur if the model is overly trained on the training dataset, and it becomes too complex.
- The overly trained model may learn the ”noise” of the training dataset.
- As a result, it performs poorly against the test dataset, and it cannot generalise well to any unseen data.
- If the model has a low error rate on the training dataset, but a high error rate on the test dataset, it signals Overfitting.

Solution: Regularisation

Solution: Regularisation

- Regularisation helps to solve the Multicollinearity and the Overfitting problems.
- Regularisation reduces the model complexity by penalizing the large coefficients of the predictors.
  - Ridge Regression solves the Multicollinearity, by shrinking the coefficients of the correlated predictors to some small numbers.
  - LASSO Regression solves the Multicollinearity, by reducing the coefficients of some correlated predictors to exactly zero.
- Only LASSO Regression, but not Ridge Regression, performs the ***Variable Selection***.
- Regularised models tend to have a slightly higher error rate on the training dataset, but in return, they have a lower error rate on the test dataset.

A Case Study

Case Study: Car Sales Dataset

Story

Mr. Yap is a chief manager of a car sales company that specialises in selling 2nd hand cars.

Focus Question

To identify the key factors that impact the car sale price.

Source: `https://www.freepik.com/`

Inspect the Dataset

- Load the dataset, and check the first few observations.

```
df.carprice <- read_excel("data/carprice_sample.xlsx") %>%
                    as.data.frame()
head(df.carprice)
```

```
  Number_of_Doors Highway_MPG City_MPG Popularity  Price
1               3          17       12       5657  33196
2               4          24       16       1385  29903
3               2          22       15        640  33677
4               4          28       18       1624  33582
5               4          24       17        210  32006
6               4          23       16        190  35663
```

- Predictor variables: Number of Doors, Highway MPG, City MPG, Popularity.
- Dependent variable: Price.

Inspect the Dataset

- Check the structure of the data frame.

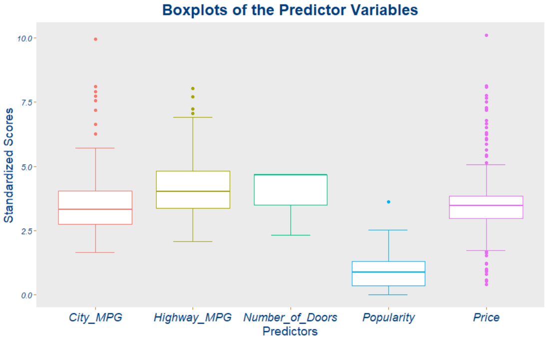
```
str(df.carprice)
```

```
'data.frame': 600 obs. of  5 variables:
 $ Number_of_Doors: num  3 4 2 4 4 4 4 4 4 4 ...
 $ Highway_MPG    : num  17 24 22 28 24 23 24 22 32 50 ...
 $ City_MPG       : num  12 16 15 18 17 16 16 16 23 54 ...
 $ Popularity     : num  5657 1385 640 1624 210 ...
 $ Price          : num  33196 29903 33677 33582 32006 ...
```

- There are 600 observations, and all the 5 variables are numerical.

Standardisation

- For Ridge and LASSO Regression, it is compulsory to standardise all the numerical variables, such that they have a constant standard deviation, which is 1.
- We will elaborate it more in another video.
- Suppose the standardisation is performed.

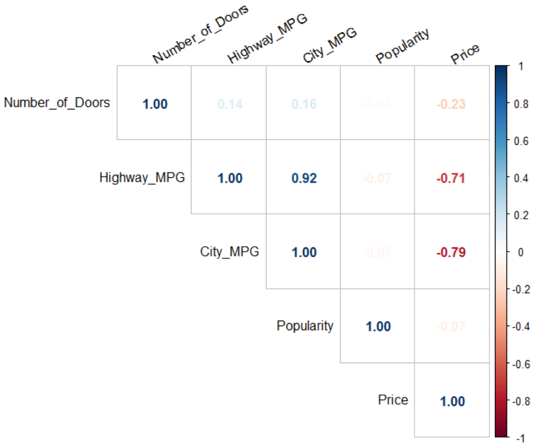


- From the chart, all the numerical variables have been standardised properly.

Correlation Matrix

- Let us analyse the relationship between the 5 variables.

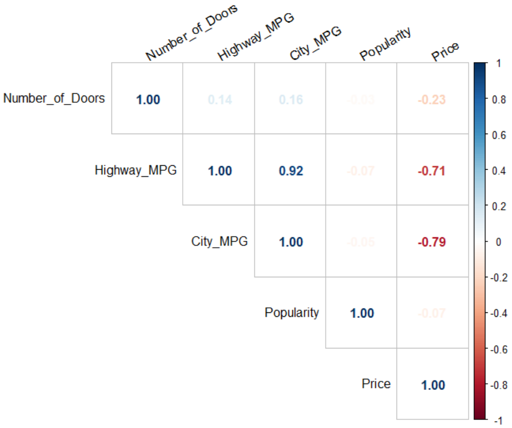
```
corrplot(cor(df.carprice), method = "number", type = "upper",
         tl.col = "black", tl.srt = 30)
```



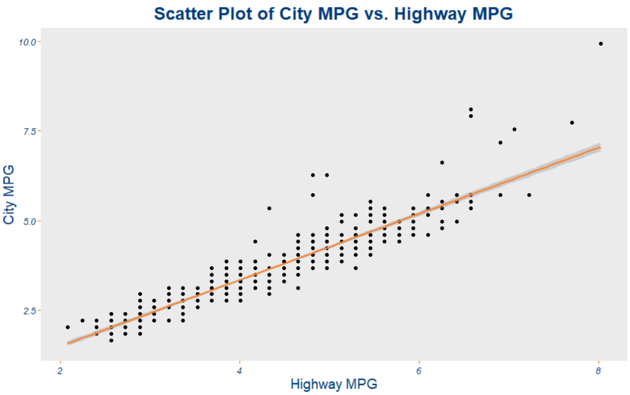
Correlation Matrix

From the chart, we notice the following facts:

- Highway MPG and City MPG are strongly and positively correlated, with r = 0.92.
- Both Highway MPG and City MPG are strongly and negatively correlated with Price, with r = -0.71 and -0.79, respectively.
- The correlations between other pairs of factors are generally weak.

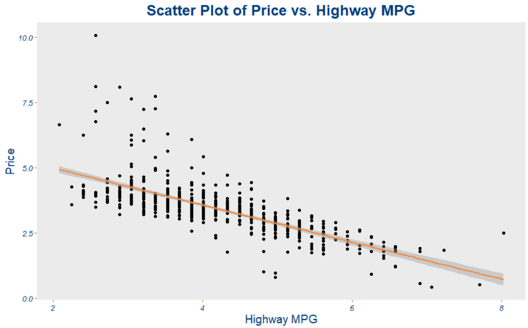


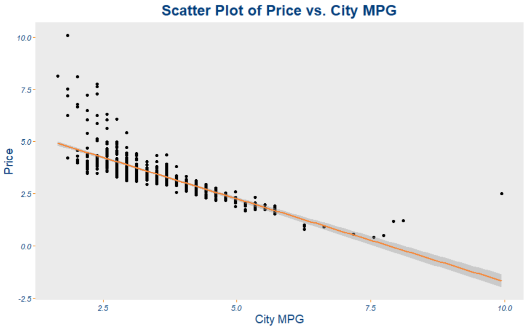
City MPG vs. Highway MPG



- From the scatterplot, Highway MPG and City MPG are strongly correlated.
- As Highway MPG increases, City MPG will also increase. Vice versa.

Highway MPG, City MPG and Price





- These plots tally with the correlation values above (-0.71; -0.79), indicating the strong negative correlation between both MPG and price,
- The higher the Highway MPG (or the City MPG), the less the sales price.

Build the 1st Multiple Linear Regression Model

- Let us fit the MLR model to the car price data.

```
model1 <- lm(Price ~., data = df.carprice)
summary(model1)
```

From the summary table, we notice two issues here:

- The coefficients of Highway MPG is positive, which is not consistent with our earlier observation that Highway MPG is negatively correlated with price (r = -0.71).
- P-value for Highway MPG is not significant. This is a bit unexpected and not consistent with the fact that the Highway MPG is strongly correlated with price (r = -0.71).

```
Call:
lm(formula = Price ~ ., data = df.carprice)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7753 -0.2162 -0.0916  0.0589  5.1232 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.64162    0.14128  47.009  < 2e-16 ***
Number_of_Doors -0.11065    0.02440  -4.535 6.98e-06 ***
Highway_MPG      0.11398    0.06266   1.819   0.0694 .  
City_MPG        -0.88575    0.06272 -14.121  < 2e-16 ***
Popularity      -0.10803    0.02420  -4.463 9.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5897 on 595 degrees of freedom
Multiple R-squared:  0.6546,	Adjusted R-squared:  0.6523 
F-statistic: 281.9 on 4 and 595 DF,  p-value: < 2.2e-16
```

How Ridge and LASSO Regression Solve the Problems

Case 1: Multicollinearity

- The above issues are due to Multicollinearity, as Highway MPG and City MPG are strongly correlated.

- Multicollinearity is one of the common problems in data science.
  - Multicollinearity makes it hard to interpret the coefficients of the regression models.
  - It also reduces the power of the linear regression models to identify the key predictors that are statistically significant.

Case 1: Multicollinearity

- We can use VIF scores to detect the Multicollinearity.

```
vif(model1)
```

```
Number_of_Doors      Highway_MPG           City_MPG       Popularity
       1.025777         6.763217           6.777843         1.009109
```

- Multicollinearity exists, as the VIF scores of Highway MPG and City MPG are above 5.
- There is one key difference between correlation matrix and VIF scores.
- Correlation matrix shows the bivariate relationship between any two variables.
- VIF score of any predictor variable represents how well the variable is explained by all other predictor variables.

Build the 2nd Multiple Linear Regression Model

- Let us build the 2nd MLR Model, by removing City MPG.

```
model2 <- lm(Price ~.-City_MPG, data = df.carprice)
summary(model2)
```

It is interesting to note that:

- Highway MPG becomes statistically significant, as its p-value is below 0.05.
- All the coefficients, in Model 2, are consistent with the correlation matrix and scatter plots.
- The coefficients of “Number of Doors” and “Popularity” in Model 2 have minimal changes, compared with that of Model 1.

```
Call:
lm(formula = Price ~ . - City_MPG, data = df.carprice)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8872 -0.3212 -0.0775  0.2276  5.2348 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.06795    0.15935  44.356  < 2e-16 ***
Number_of_Doors -0.13660    0.02809  -4.863 1.48e-06 ***
Highway_MPG     -0.70104    0.02816 -24.898  < 2e-16 ***
Popularity      -0.12774    0.02790  -4.579 5.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6808 on 596 degrees of freedom
Multiple R-squared:  0.5389,	Adjusted R-squared:  0.5365 
F-statistic: 232.2 on 3 and 596 DF,  p-value: < 2.2e-16
```

Solving Case 1: Multicollinearity

Ridge Regression

- Recall the 2nd MLR model, denoted as "***MLR adjusted***", is as follows:

$$\text{Price} = 7.068 - 0.137 * \text{Number of Doors} - 0.701 * \text{Highway MPG} - 0.128 * \text{Popularity}.$$

- The following table summarises the coefficients of the ***Ridge Regression*** model.

	(Intercept)	Number_of_Doors	Highway_MPG	City_MPG	Popularity
s0	6.69086	-0.1107518	-0.1238088	-0.617263	-0.1047929

$$\text{Price} = 6.691 - 0.111 * \text{Number of Doors} - 0.124 * \text{Highway MPG} - 0.617 * \text{City MPG} - 0.105 * \text{Popularity}.$$

  - The coefficients of the predictors, Highway MPG and City MPG, are all negative, as expected.
  - The Ridge Regression model has solved the Multicollinearity problem.

Solving Case 1: Multicollinearity

LASSO Regression

- The following table summarises the coefficients of the ***LASSO Regression*** model.

	(Intercept)	Number_of_Doors	Highway_MPG	City_MPG	Popularity
s0	6.629672	-0.09721505	.	-0.7661423	-0.09455915

$$\text{Price} = 6.630 - 0.097 * \text{Number of Doors} + 0 * \text{Highway MPG} - 0.766 * \text{City MPG} - 0.095 * \text{Popularity}.$$

- In the LASSO Regression model, the coefficient of Highway MPG is 0.
- LASSO Regression has performed variable selection by setting some coefficient to be zero.
- LASSO Regression has successfully solved the Multicollinearity problem.

Compare the Performance of MLR, Ridge and LASSO Regression Models

	MSE	MAE	RMSE	MAPE	R^2
MLR adjusted	0.460	0.431	0.679	0.137	0.539
Ridge model	0.356	0.333	0.596	0.095	0.644
LASSO model	0.347	0.318	0.589	0.088	0.652

- The adjusted MLR model performs the worst.
  - The MSE of the adjusted MLR model is 0.460, which is the highest MSE among the three models.
  - The R^2 of the adjusted MLR model is 0.539, which is the lowest R^2 among the three models.
- The Ridge and LASSO Regression models have a better performance than the adjusted MLR model.
- In a nutshell, both Ridge and LASSO Regression can effectively solve the multicollinearity problem without compromising the accuracy.

Case 2: Small Training Dataset

- By mentioning ***small***, we actually mean that the ratio of the training dataset size to the number of predictors is small.
- Suppose the training dataset has **32** observations, and the number of predictors is **4**.
- In such case, the ratio of the training dataset size to the number of predictors is **8**.
- The common rule of thumb is that for every one predictor variable, it is recommended to have at least **100** observations.

Train a MLR Model using a Small Training Dataset

- Let us split the entire dataset (**600**) into the training dataset and the test dataset of size **32** and **568**, respectively.

```
set.seed(6674)
sample <- sample(nrow(df.carprice), 32)
training <- df.carprice[sample, ]
test <- df.carprice[-sample, ]
```

- Next, we use the small training dataset to train the 3rd MLR model, which will be denoted as "***MLR baseline***".

```
model3 <- lm(Price ~.-Highway_MPG, data = training)
summary(model3)
```

Build the Baseline MLR Model

```
Call:
lm(formula = Price ~ . - Highway_MPG, data = training)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7480 -0.2743 -0.0849  0.1436  1.8213 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.65850    0.57416  13.339 1.18e-13 ***
Number_of_Doors -0.19706    0.09685  -2.035   0.0514 .  
City_MPG        -0.90431    0.10776  -8.392 3.96e-09 ***
Popularity      -0.18903    0.09016  -2.097   0.0452 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5232 on 28 degrees of freedom
Multiple R-squared:  0.7319,	Adjusted R-squared:  0.7031 
F-statistic: 25.48 on 3 and 28 DF,  p-value: 3.725e-08
```

- Note that we do not include the predictor, “Highway MPG”, in order to solve the multicollinearity problem.
- From the coefficients and p values, we can conclude that the Multicollinearity problem has been resolved.

Performance of the Baseline MLR Model

	MSE	MAE	RMSE	MAPE
Baseline MLR Train	0.240	0.321	0.489	0.083
Baseline MLR Test	0.387	0.379	0.622	0.116

- All the error rates of the baseline MLR model on the test dataset are consistently higher than those on the training dataset.
- For example, the MSE of the model on the training dataset is 0.240, while the MSE on the test dataset is 0.387.
- The above problem is commonly referred to as ***Overfitting***.

Case 2: Overfitting

- In our case, Overfitting is due to the small size of the training dataset.
- The small training dataset may not well represent the test dataset, or any other unseen data.
- The model may have been overfitted to the small training dataset, such that it may lose the ability to generalise well to any unseen data.
- One solution is to train the model with more data.
- Another solution is Ridge and LASSO Regression.

Solving Case 2: Overfitting

- Let us compare the error metrics of the baseline MLR, Ridge and LASSO Regression models, on both the training and the test datasets.

	MSE	MAE	RMSE	MAPE
Baseline MLR Train	0.240	0.321	0.489	0.083
Ridge Model Train	0.251	0.321	0.501	0.082
LASSO Model Train	0.241	0.310	0.491	0.078

	MSE	MAE	RMSE	MAPE
Baseline MLR Test	0.387	0.379	0.622	0.116
Ridge Model Test	0.369	0.366	0.608	0.107
LASSO Model Test	0.370	0.360	0.609	0.107

Solving Case 2: Overfitting

	MSE	MAE	RMSE	MAPE
Baseline MLR Train	0.240	0.321	0.489	0.083
Ridge Model Train	0.251	0.321	0.501	0.082
LASSO Model Train	0.241	0.310	0.491	0.078

- The accuracy levels of the three models, on the training dataset, are similar.

	MSE	MAE	RMSE	MAPE
Baseline MLR Test	0.387	0.379	0.622	0.116
Ridge Model Test	0.369	0.366	0.608	0.107
LASSO Model Test	0.370	0.360	0.609	0.107

- The accuracy levels of the Ridge and LASSO Regression models, on the test dataset, are higher than that of the Multiple Linear Regression model.

Solving Case 2: Overfitting

- Both the Ridge and LASSO Regression models can reduce the error for the test dataset, without compromising the accuracy on the training dataset.
- The Overfitting problem exists, as the errors are consistently higher on the test dataset, compared with that on the training dataset.
- Nevertheless, both the Ridge and LASSO Regression models have minimised the differences of error metrics between the training and test datasets, to some extent.

Solving Case 2: Overfitting

	MSE	MAE	RMSE	MAPE
Baseline MLR Train	0.240	0.321	0.489	0.083
Baseline MLR Test	0.387	0.379	0.622	0.116

	MSE	MAE	RMSE	MAPE
Ridge Model Train	0.251	0.321	0.501	0.082
Ridge Model Test	0.369	0.366	0.608	0.107

- In summary, both Ridge and LASSO Regression can minimise the Overfitting problem to a certain extent.

Case 3: A Large Number of Predictors

- It is difficult to interpret the MLR model with too many coefficients.
- It is helpful to simplify the model by retaining a smaller set of important predictors.

- LASSO Regression can achieve this goal, by shrinking some predictors' coefficients to 0.
- This is also called "***Variable Selection***", which can
  - Prevent Overfitting;
  - Improve the model interpretability;
  - Make it easier to execute the business solution in practice.

Solving Case 3: A Large Number of Predictors

- Recall the coefficients of the LASSO Regression model, when we solve the case 1: Multicollinearity.

```
   (Intercept) Number_of_Doors Highway_MPG   City_MPG  Popularity
s0    6.629672     -0.09721505           . -0.7661423 -0.09455915
```

- Note that the coefficient of the predictor, Highway MPG, has been reduced to zero.
- In general, when the number of predictors is large, LASSO Regression can shrink the coefficients of some predictors to exactly zero.
- By performing Variable Selection, LASSO Regression can reduce the model complexity, and improve the model interpretability.
- Ridge Regression cannot perform variable selection directly.
- If you value the business interpretability, and want a simple model with fewer parameters, LASSO Regression is a better choice.

Summary

Summary

We have learnt to:

- Understand the impact of the Multicollinearity and the Overfitting problems.
- Understand how Regularisation: Ridge and LASSO Regression, have successfully solved, or minimised the Multicollinearity and the Overfitting problems.

In the next video,

We will learn more about Bias, Variance and their Trade-off.

References

Wessel N. van Wieringen (2021)
Lecture notes on ridge regression

Dataset: Car Features and MSRP
https://www.kaggle.com/CooperUnion/cardataset