LASSO Regression

The aim of science is to seek the simplest explanation of complex facts...
Seek simplicity and distrust it.
- A. N. Whitehead

Outline

Learning Objectives

In this video, you will learn to:

- Understand the model, the cost function and the regularisation parameter λ of LASSO Regression.
- Learn to train and evaluate a LASSO Regression model in R.
- Learn to use the Cross Validation method to pick the optimal λ value.

Introduction to LASSO Regression

Cost Function for LASSO Regression

LASSO Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_n X_n$$

- As we apply LASSO Regression to regularise some MLR model, the LASSO Regression model shares the same type as MLR.
- The coefficients of the LASSO Regression model are chosen as the ones that minimise the following cost function:

$$\begin{aligned}\text{Cost Function} &= \sum_i \text{Residual}_i^2 + \lambda \sum_{j=1}^{n} |\text{Coefficients}| \\ &= \sum_i \text{Residual}_i^2 + \lambda \sum_{j=1}^{n} |\beta_j| \qquad \text{where } \lambda \geq 0\end{aligned}$$

Common Features of Ridge and LASSO Regression

- Both models need an input of λ.
- λ can be zero or any positive value.
- When λ is zero, there is no penalty. Both Ridge and LASSO Regression models will produce the same coefficients as the MLR model.
- When λ is a positive number, the penalty term has an effect of shrinking the coefficients. Both Ridge and LASSO Regression models tend to have smaller coefficients, compared with MLR models.
- In general, when λ increases, it enforces stronger regularisation on the model, and the coefficients of the model will approach zero.

Difference between Ridge and LASSO Regression

L1 and L2 Norm (ℓ^1 and ℓ^2 Norm)

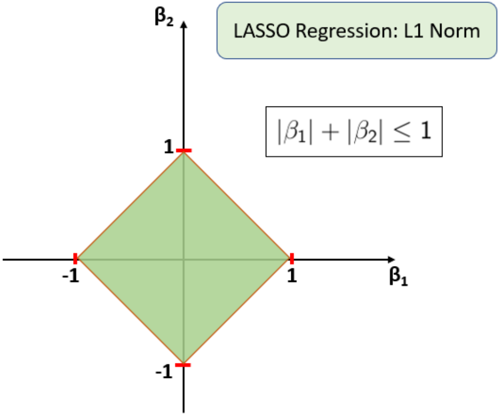
- For a Linear Regression model, $Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_n X_n$,
- The collection of coefficients of all predictors, $(\beta_1, \beta_2, \cdots, \beta_n)$, is denoted as the ***coefficients vector***, β.
- A norm is a function measuring the distance of a vector from the origin.
- L1 Norm of β is defined as: $\|\beta\|_1 = \sum_{j=1}^{n} |\beta_j|$.
- L2 Norm of β is defined as: $\|\beta\|_2 = \sqrt{\sum_{j=1}^{n} \beta_j^2}$.

$$\text{Cost Function of Ridge} = \text{RSS} + \lambda \sum_{j=1}^{n} \beta_j^2 = \text{RSS} + \lambda \|\beta\|_2^2$$

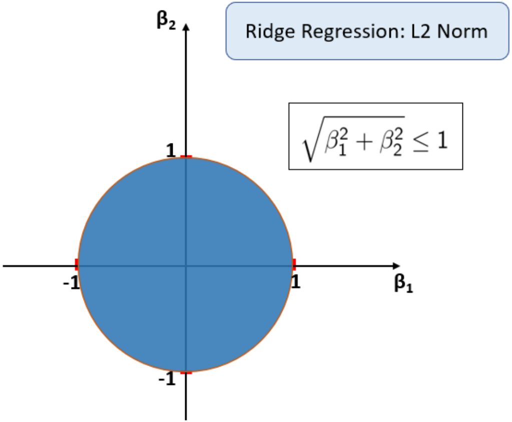
$$\text{Cost Function of LASSO} = \text{RSS} + \lambda \sum_{j=1}^{n} |\beta_j| = \text{RSS} + \lambda \|\beta\|_1$$

L1 Norm vs. L2 Norm

- LASSO Regression: $\|\beta\|_1$



- Ridge Regression: $\|\beta\|_2$



Recap on the glmnet() Function

```
glmnet(x, y, alpha = 1, lambda = K)
```

The inputs include:

- x is a data matrix of predictor variables, and y is the dependent variable.
- Alpha is the mixing parameter, that determines the type of the Regression model. Here, we choose alpha $= 1$, for LASSO Regression.
- Lambda is the regularisation parameter.

Assumptions of LASSO Regression Models

Assumptions of LASSO Regression Models

1. ***Independence***: Each observation is independent from the others.
2. ***Linearity***: The relationship, between the predictors Xs and the dependent variable Y, is linear.
3. ***Constant Variance*** The residuals are evenly scattered around the center line of zero.

Case Study: Predicting Housing Price

Mr. Tan's Focus Question

What is the expected selling price of houses from one neighbourhood, given the conditions and relevant factors of the area?

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

Analyse: Model Building ($\lambda = 0.1$)

- Let us first try $\lambda = 0.1$.

```
model_LASSO_trial1 <- glmnet(train.x,train.y, alpha = 1, lambda =
    0.1)
t(coef(model_LASSO_trial1))
```

```
   (Intercept) Crime_rate Industry Number_of_rooms Access_to_highways   Tax_rate
s0   -3.523875 -0.1242237        .       0.7116442                  . -0.1377232
```

- The above list gives the (standardised) coefficients of the LASSO Regression model.
- For example, the coefficient of “Crime rate” is -0.124, and the coefficient of “Industry” is 0.
- LASSO Regression performs ***Variable Selection*** by setting the coefficients of two predictors, “Industry” and “Access to highways”, to zero.

Analyse: Model Building ($\lambda = 0.5, 1$)

- Next, try $\lambda = 0.5$.

```
model_LASSO_trial2 <- glmnet(train.x,train.y, alpha = 1, lambda =
    0.5)
t(coef(model_LASSO_trial2))
```

```
   (Intercept) Crime_rate Industry Number_of_rooms Access_to_highways Tax_rate
s0   -0.849903          .        .       0.3663317                  .        .
```

- Finally, try $\lambda = 1$.

```
model_LASSO_trial3 <- glmnet(train.x,train.y, alpha = 1, lambda = 1)
t(coef(model_LASSO_trial3))
```

```
   (Intercept) Crime_rate Industry Number_of_rooms Access_to_highways Tax_rate
s0    2.432427          0        .               .                  .        .
```

Compare the Coefficients of three Models ($\lambda = 0.1, 0.5, 1$)

	lambda = 0.1	lambda = 0.5	lambda = 1
(Intercept)	-3.5238752	-0.8499030	2.432427
Crime_rate	-0.1242237	.	0.000000
Industry	.	.	.
Number_of_rooms	0.7116442	0.3663317	.
Access_to_highways	.	.	.
Tax_rate	-0.1377232	.	.

Look at the coefficient of the predictor, "Number of rooms".

- When λ increases from 0.1 to 0.5, the coefficient decreases from 0.712 to 0.366.
- If λ further increases to 1, the coefficient changes to 0.
- In general, a larger λ value imposes a higher degree of regularisation.
- Consequently, the absolute values of the predictors' coefficients tend to approach 0.
- If λ is large enough, the coefficients eventually become 0.

Train 100 LASSO Regression Models

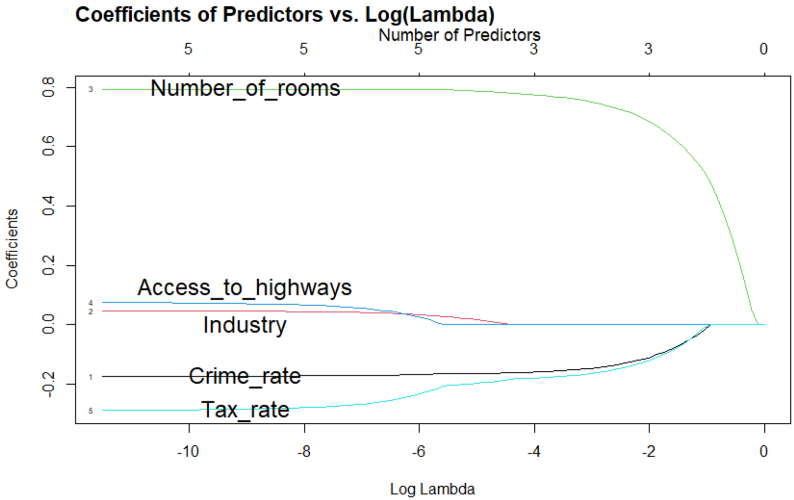
- Let us first train 100 LASSO Regression models using a sequence of lambda values, from 10^{-5} to 1.

```
lambda <- 10^seq(-5, 0, length = 100)
LASSO_model <- glmnet(train.x,train.y, alpha = 1, lambda = lambda)
```

- Then we use the following code chunk to generate the plot for the Coefficients vs. Log(Lambda):

```
add_lbs <- function(fit, offset_x=2.5) {
  L <- length(fit$lambda)
  x <- log(fit$lambda[L])+ offset_x
  y <- fit$beta[, L]
  labs <- names(y)
  text(x, y, labels=labs, cex = 1.5)
}
plot(LASSO_model, xvar = "lambda", label = TRUE)
add_lbs(LASSO_model)
legend("topright", lwd = 1, col = 1:6, legend = colnames(train.x),
    cex = .7)
```

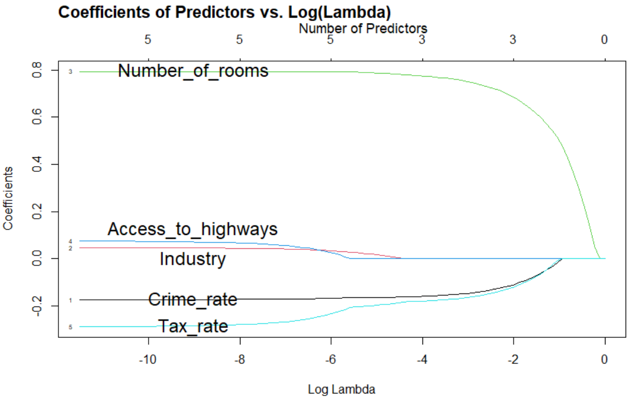
Coefficients vs. Log(λ)



Coefficients vs. Log(λ)

The plot shows:

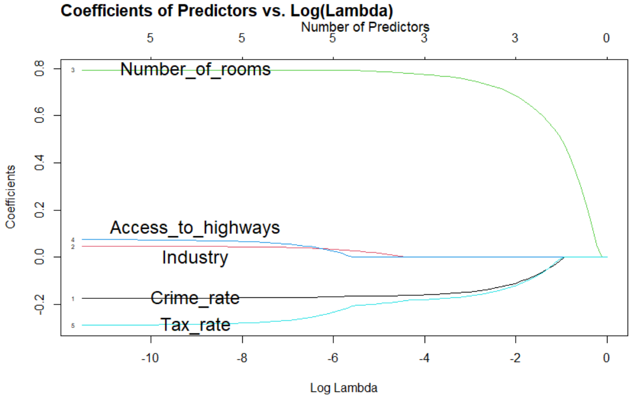
- X axis is the Logarithm of the regularisation parameter λ.
- Y axis is the standardised coefficients for each predictor.
- As λ increases, the predictors' coefficients will approach zero, and stabilize at zero from some point onwards.



Coefficients vs. Log(λ)

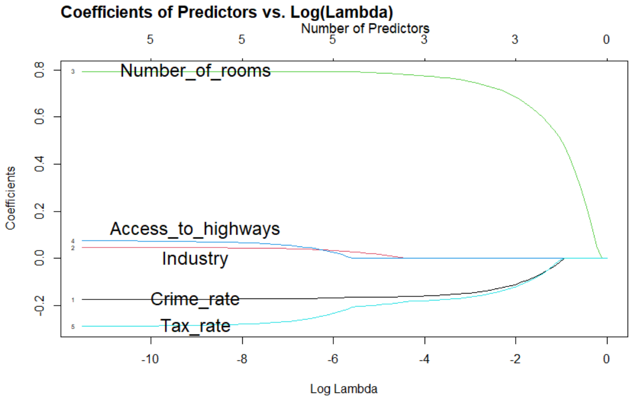
The plot shows:

- The numbers on the top axis, say, 5, 5, 3, 3, 0, indicates how many coefficients are non-zero.
- When $\log(\lambda)$ equals -6, namely, λ is approximately, 0.0025, the LASSO model contains all the 5 predictors.
- When $\log(\lambda)$ equals -4, namely, λ is around 0.018, the LASSO model retains 3 predictors out of 5.
- When $\log(\lambda)$ equals 0, namely, λ is 1, the LASSO model has deselected all the five predictors.



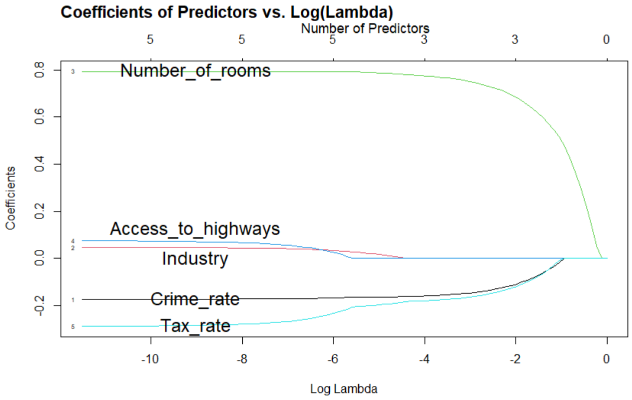
Coefficients vs. Log(λ)

- Recall that Multicollinearity exists, and the sign of the coefficient of "Access to highways", in the MLR model, is positive, which is problematic.
- For the coefficient of "Access to highways":
  - When Log(λ) increases from -10 to -5.5, it gradually decreases to 0.
  - When Log(λ) further increases from -5.5, it remains as 0.
- The coefficient of "Tax rate" remains negative, and it only starts to approach 0, when Log(λ) is more than -2.



Coefficients vs. Log(λ)

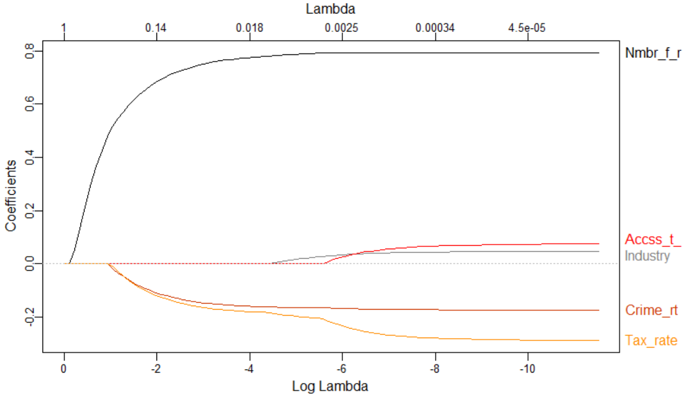
- In general, as λ increases, only one of the strongly correlated predictors has a positive or negative coefficient, while the rest are linked to zero coefficients.
- This may explain a bit on how LASSO Regression copes with Multicollinearity.



Coefficients vs. Log(λ)

- We can also use the "plot_glmnet()" function from the "plotmo" package, to generate a similar plot.

```
plot_glmnet(LASSO_model)
```



Cross Validation: cv.glmnet()

```
set.seed(123)
cv_LASSO <- cv.glmnet(train.x, train.y, alpha = 1, type.measure = "
   mse")
cv_LASSO
```

```
Call:  cv.glmnet(x = train.x, y = train.y, type.measure = "mse", alpha = 1)
Measure: Mean-Squared Error
     Lambda Index Measure     SE Nonzero
min 0.02241    40  0.4233 0.1153       3
1se 0.25175    14  0.5287 0.1179       3
```

```
cv_LASSO$lambda.min
```

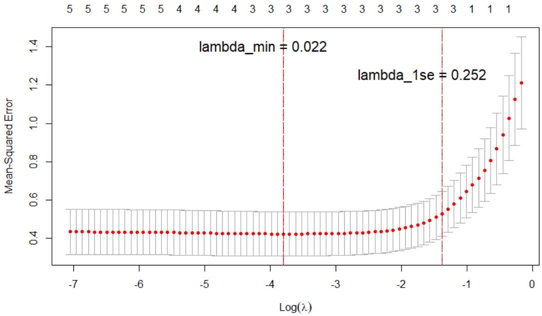
```
[1] 0.02241138
```

```
cv_LASSO$lambda.1se
```

```
[1] 0.2517524
```

Cross Validation: lambda.min and lambda.1se

- `plot(cv_ridge)`



- When Log(λ) ranges between -7 and -2, the Cross Validation MSE rates are similar.
- If Log(λ) increases from -2 and onwards, the Cross Validation error increases dramatically.
- Here, the left red vertical line indicates where Log(lambda.min) lies, and the right red vertical line indicates where Log(lambda.1se) lies.

Build the Final LASSO Regression Model

Analyse: Build the Final LASSO Regression Model

```
glm_LASSO <- glmnet(train.x, train.y, alpha = 1, lambda =
    cv_LASSO$lambda.min)
t(coef(glm_LASSO))
```

```
   (Intercept) Crime_rate Industry Number_of_rooms Access_to_highways   Tax_rate
s0   -3.944037 -0.1574161        .       0.7718826                  . -0.1773956
```

- "Number of rooms" is the most important predictor, since its standardised coefficient, namely, 0.772, has the highest absolute value among all.
- "Industry" and "Access to highways" are the least important predictors, as their standardised coefficients are equal to zero.
- Only LASSO Regression, but not Ridge Regression, can perform Variable Selection.

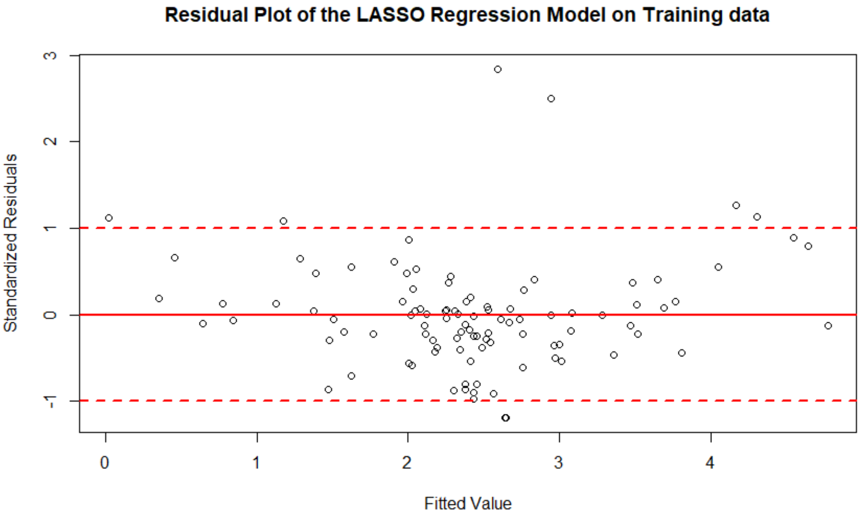
Analyse: Evaluating the Final LASSO Regression Model

	MSE	MAE	RMSE	MAPE
LASSO Train	0.388	0.424	0.623	0.210
LASSO Test	0.472	0.441	0.687	0.224

From the summary table, we can see:

- The error metrics are consistently higher on the test dataset, compared with those on the training dataset.
- In practice, we use these error metrics to compare different models, and perform the model selection.
- We will show in the next video that the LASSO Regression model performs better than the MLR model.

Analyse: Residual Plots



Apply: Make Predictions

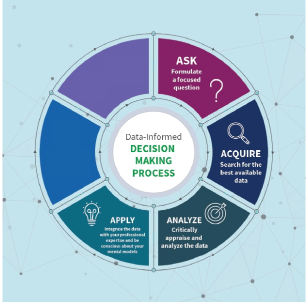
```
new_data
```

```
  Crime_rate Industry Number_of_rooms Access_to_highways Tax_rate
1    0.00632     2.31           6.575                  1      296
```

```
new_x <- data.matrix(new_data/
    scaler)
```

```
predict(glm_LASSO, new_x ) *
    scaler[6]
```

```
        s0
1 27.29209
```



Source: `https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions`

Features of the LASSO Regression Models

Features of the LASSO Regression Models

Let us summarise some features of the LASSO Regression models:

1. Just like Ridge Regression, the LASSO Regression model has the effect of shrinking the coefficients of predictors towards zero.
2. LASSO Regression can perform ***Variable Selection***.
3. By Variable Selection, LASSO Regression helps to ***solve Multicollinearity***, and ***improve the model interpretability***.
4. The regularisation parameter, λ, controls the amount of regularisation, and regularisation controls the amount of bias and variance.
5. With the Bias-Variance trade-off, the optimal LASSO Regression model can ***minimise Overfitting***.

Summary

Summary

We have learned to:

- Understand how LASSO Regression works, and compare it with Ridge Regression.
- Understand how regularisation parameter, λ, affects the LASSO Regression model coefficients.
- Can use the "glmnet()" function to train a LASSO Regression model with the optimal λ, that is obtained from the "cv.glmnet()" function.

In the next video,

We will introduce Elastic Net Regression, and learn to implement it in R.

References

- Wessel N. van Wieringen (2021), *Lecture notes on ridge regression*
- Hastie, Qian, and Tay (2021), *An Introduction to glmnet*
  https://glmnet.stanford.edu/articles/glmnet.html
- Dataset: the Boston Housing Dataset
  https://www.cs.toronto.edu/ delve/data/boston/bostonDetail.html
- Shubham.jain Jain (2017), *A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R*
  https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/