**1. OUTPUT OF MULTIPLE LINEAR REGRESSION**

> setwd("~/Desktop/Data Mining/Project3")

> redWine <- read.csv("~/Desktop/Data Mining/Project3/winequality-red.csv", sep=";")

> View(redWine)

> require(rpart)

> null=lm(quality~1, data=redWine)

> full=lm(quality~., data=redWine)

> step(null, scope=list(lower=null, upper=full), direction="forward")

Start: AIC=-682.5

quality ~ 1

Df Sum of Sq RSS AIC

+ alcohol 1 236.295 805.87 -1091.65

+ volatile.acidity 1 158.967 883.20 -945.14

+ sulphates 1 65.865 976.30 -784.89

+ citric.acid 1 53.405 988.76 -764.61

+ total.sulfur.dioxide 1 35.707 1006.46 -736.24

+ density 1 31.887 1010.28 -730.19

+ chlorides 1 17.318 1024.85 -707.29

+ fixed.acidity 1 16.038 1026.13 -705.29

+ pH 1 3.473 1038.69 -685.84

+ free.sulfur.dioxide 1 2.674 1039.49 -684.61

<none> 1042.17 -682.50

+ residual.sugar 1 0.197 1041.97 -680.80

Step: AIC=-1091.65

quality ~ alcohol

Df Sum of Sq RSS AIC

+ volatile.acidity 1 94.074 711.80 -1288.1

+ sulphates 1 44.977 760.89 -1181.5

+ citric.acid 1 31.953 773.92 -1154.3

+ pH 1 26.362 779.51 -1142.8

+ fixed.acidity 1 24.623 781.25 -1139.3

+ total.sulfur.dioxide 1 8.270 797.60 -1106.2

+ density 1 5.203 800.67 -1100.0

<none> 805.87 -1091.7

+ chlorides 1 0.611 805.26 -1090.9

+ free.sulfur.dioxide 1 0.325 805.55 -1090.3

+ residual.sugar 1 0.041 805.83 -1089.7

Step: AIC=-1288.14

quality ~ alcohol + volatile.acidity

Df Sum of Sq RSS AIC

+ sulphates 1 19.6916 692.10 -1331.0

+ total.sulfur.dioxide 1 6.3730 705.42 -1300.5

+ pH 1 5.9515 705.84 -1299.6

+ fixed.acidity 1 5.7061 706.09 -1299.0

+ density 1 1.9410 709.86 -1290.5

<none> 711.80 -1288.1

+ free.sulfur.dioxide 1 0.6621 711.13 -1287.6

+ chlorides 1 0.3762 711.42 -1287.0

+ citric.acid 1 0.1936 711.60 -1286.6

+ residual.sugar 1 0.0101 711.79 -1286.2

Step: AIC=-1331

quality ~ alcohol + volatile.acidity + sulphates

Df Sum of Sq RSS AIC

+ total.sulfur.dioxide 1 8.2176 683.89 -1348.1

+ chlorides 1 7.4925 684.61 -1346.4

+ fixed.acidity 1 3.3282 688.78 -1336.7

+ pH 1 3.0454 689.06 -1336.0

+ free.sulfur.dioxide 1 1.1129 690.99 -1331.6

<none> 692.10 -1331.0

+ citric.acid 1 0.2522 691.85 -1329.6

+ density 1 0.2222 691.88 -1329.5

+ residual.sugar 1 0.0143 692.09 -1329.0

Step: AIC=-1348.1

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide

Df Sum of Sq RSS AIC

+ chlorides 1 8.0370 675.85 -1365.0

+ pH 1 3.3094 680.58 -1353.8

+ fixed.acidity 1 2.1037 681.78 -1351.0

+ free.sulfur.dioxide 1 1.3557 682.53 -1349.3

<none> 683.89 -1348.1

+ residual.sugar 1 0.2634 683.62 -1346.7

+ density 1 0.1077 683.78 -1346.3

+ citric.acid 1 0.0730 683.81 -1346.3

Step: AIC=-1365

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides

Df Sum of Sq RSS AIC

+ pH 1 5.9189 669.93 -1377.1

+ fixed.acidity 1 2.4065 673.44 -1368.7

+ free.sulfur.dioxide 1 1.2403 674.61 -1365.9

<none> 675.85 -1365.0

+ residual.sugar 1 0.5531 675.30 -1364.3

+ citric.acid 1 0.1615 675.69 -1363.4

+ density 1 0.1526 675.70 -1363.4

Step: AIC=-1377.06

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides + pH

Df Sum of Sq RSS AIC

+ free.sulfur.dioxide 1 2.39413 667.54 -1380.8

<none> 669.93 -1377.1

+ citric.acid 1 0.80525 669.13 -1377.0

+ residual.sugar 1 0.28390 669.65 -1375.7

+ density 1 0.04468 669.89 -1375.2

+ fixed.acidity 1 0.01040 669.92 -1375.1

Step: AIC=-1380.79

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides + pH + free.sulfur.dioxide

Df Sum of Sq RSS AIC

<none> 667.54 -1380.8

+ citric.acid 1 0.47480 667.06 -1379.9

+ residual.sugar 1 0.16673 667.37 -1379.2

+ density 1 0.03079 667.51 -1378.9

+ fixed.acidity 1 0.00663 667.53 -1378.8

Call:

lm(formula = quality ~ alcohol + volatile.acidity + sulphates +

total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,

data = redWine)

Coefficients:

(Intercept) alcohol

4.430099 0.289303

volatile.acidity sulphates

-1.012753 0.882665

total.sulfur.dioxide chlorides

-0.003482 -2.017814

pH free.sulfur.dioxide

-0.482661 0.005077

> data = redWine[,c(1,5,6,7,9,10,11,12)]

> require(caTools)

> y= data[,8]

> train\_rows = sample.split(y, SplitRatio=0.8)

> head(train\_rows)

[1] TRUE TRUE TRUE TRUE TRUE TRUE

> train = data[ train\_rows,]

> test=data[!train\_rows,]

> lm.fit = lm(quality~.,data=train)

> testdata = data.frame(test)

> pred\_data = predict(lm.fit, testdata)

> test\_quality = test[,8]

> error <- sqrt((sum((test\_quality-pred\_data)^2))/length(test\_quality))

> error

**[1] 0.6609043**

> pdf("Proj3\_Linear\_Predictions.pdf")

> plot(test\_quality,pred\_data)

> dev.off()

RStudioGD

2

**2. OUTPUT OF RIDGE REGRESSION WITH 8 SELECTED VARIABLES**

> redWine <- read.csv("~/Desktop/Data Mining/Project3/winequality-red.csv", sep=";")

> View(redWine)

> require(rpart)

> require(MASS)

> require(caTools)

> null=lm(quality~1, data=redWine)

> full=lm(quality~., data=redWine)

> step(null, scope=list(lower=null, upper=full), direction="forward")

Start: AIC=-682.5

quality ~ 1

Df Sum of Sq RSS AIC

+ alcohol 1 236.295 805.87 -1091.65

+ volatile.acidity 1 158.967 883.20 -945.14

+ sulphates 1 65.865 976.30 -784.89

+ citric.acid 1 53.405 988.76 -764.61

+ total.sulfur.dioxide 1 35.707 1006.46 -736.24

+ density 1 31.887 1010.28 -730.19

+ chlorides 1 17.318 1024.85 -707.29

+ fixed.acidity 1 16.038 1026.13 -705.29

+ pH 1 3.473 1038.69 -685.84

+ free.sulfur.dioxide 1 2.674 1039.49 -684.61

<none> 1042.17 -682.50

+ residual.sugar 1 0.197 1041.97 -680.80

Step: AIC=-1091.65

quality ~ alcohol

Df Sum of Sq RSS AIC

+ volatile.acidity 1 94.074 711.80 -1288.1

+ sulphates 1 44.977 760.89 -1181.5

+ citric.acid 1 31.953 773.92 -1154.3

+ pH 1 26.362 779.51 -1142.8

+ fixed.acidity 1 24.623 781.25 -1139.3

+ total.sulfur.dioxide 1 8.270 797.60 -1106.2

+ density 1 5.203 800.67 -1100.0

<none> 805.87 -1091.7

+ chlorides 1 0.611 805.26 -1090.9

+ free.sulfur.dioxide 1 0.325 805.55 -1090.3

+ residual.sugar 1 0.041 805.83 -1089.7

Step: AIC=-1288.14

quality ~ alcohol + volatile.acidity

Df Sum of Sq RSS AIC

+ sulphates 1 19.6916 692.10 -1331.0

+ total.sulfur.dioxide 1 6.3730 705.42 -1300.5

+ pH 1 5.9515 705.84 -1299.6

+ fixed.acidity 1 5.7061 706.09 -1299.0

+ density 1 1.9410 709.86 -1290.5

<none> 711.80 -1288.1

+ free.sulfur.dioxide 1 0.6621 711.13 -1287.6

+ chlorides 1 0.3762 711.42 -1287.0

+ citric.acid 1 0.1936 711.60 -1286.6

+ residual.sugar 1 0.0101 711.79 -1286.2

Step: AIC=-1331

quality ~ alcohol + volatile.acidity + sulphates

Df Sum of Sq RSS AIC

+ total.sulfur.dioxide 1 8.2176 683.89 -1348.1

+ chlorides 1 7.4925 684.61 -1346.4

+ fixed.acidity 1 3.3282 688.78 -1336.7

+ pH 1 3.0454 689.06 -1336.0

+ free.sulfur.dioxide 1 1.1129 690.99 -1331.6

<none> 692.10 -1331.0

+ citric.acid 1 0.2522 691.85 -1329.6

+ density 1 0.2222 691.88 -1329.5

+ residual.sugar 1 0.0143 692.09 -1329.0

Step: AIC=-1348.1

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide

Df Sum of Sq RSS AIC

+ chlorides 1 8.0370 675.85 -1365.0

+ pH 1 3.3094 680.58 -1353.8

+ fixed.acidity 1 2.1037 681.78 -1351.0

+ free.sulfur.dioxide 1 1.3557 682.53 -1349.3

<none> 683.89 -1348.1

+ residual.sugar 1 0.2634 683.62 -1346.7

+ density 1 0.1077 683.78 -1346.3

+ citric.acid 1 0.0730 683.81 -1346.3

Step: AIC=-1365

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides

Df Sum of Sq RSS AIC

+ pH 1 5.9189 669.93 -1377.1

+ fixed.acidity 1 2.4065 673.44 -1368.7

+ free.sulfur.dioxide 1 1.2403 674.61 -1365.9

<none> 675.85 -1365.0

+ residual.sugar 1 0.5531 675.30 -1364.3

+ citric.acid 1 0.1615 675.69 -1363.4

+ density 1 0.1526 675.70 -1363.4

Step: AIC=-1377.06

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides + pH

Df Sum of Sq RSS AIC

+ free.sulfur.dioxide 1 2.39413 667.54 -1380.8

<none> 669.93 -1377.1

+ citric.acid 1 0.80525 669.13 -1377.0

+ residual.sugar 1 0.28390 669.65 -1375.7

+ density 1 0.04468 669.89 -1375.2

+ fixed.acidity 1 0.01040 669.92 -1375.1

Step: AIC=-1380.79

quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +

chlorides + pH + free.sulfur.dioxide

Df Sum of Sq RSS AIC

<none> 667.54 -1380.8

+ citric.acid 1 0.47480 667.06 -1379.9

+ residual.sugar 1 0.16673 667.37 -1379.2

+ density 1 0.03079 667.51 -1378.9

+ fixed.acidity 1 0.00663 667.53 -1378.8

Call:

lm(formula = quality ~ alcohol + volatile.acidity + sulphates +

total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,

data = redWine)

Coefficients:

(Intercept) alcohol

4.430099 0.289303

volatile.acidity sulphates

-1.012753 0.882665

total.sulfur.dioxide chlorides

-0.003482 -2.017814

pH free.sulfur.dioxide

-0.482661 0.005077

> data = redWine[,c(1,5,6,7,9,10,11,12)]

> y= data[,8]

> train\_rows = sample.split(y, SplitRatio=0.8)

> head(train\_rows)

[1] FALSE TRUE FALSE TRUE TRUE TRUE

> train = data[ train\_rows,]

> test=data[!train\_rows,]

> train2 = train[,-8]

> testdata = data.frame(test)

> select(lm.ridge(train$quality~.,data=train2,lambda=seq(0,10,0.001)))

modified HKB estimator is 10.67167

modified L-W estimator is 10.80418

smallest value of GCV at 10

> ridgereg = lm.ridge(train$quality~.,data=train2,lambda=10)

> lm.ridge(train$quality~.,data=train2,lambda=10)

fixed.acidity

4.181635510 0.004511302

chlorides free.sulfur.dioxide

-2.620187254 0.007984390

total.sulfur.dioxide pH

-0.004316709 -0.703921621

sulphates alcohol

1.117235497 0.317763734

> pred\_data <-scale(testdata[,1:7],center=F,scale=ridgereg$scales)%\*% ridgereg$coef

> sumh =0;

> #Add the intercept value obtained from lm.ridge function.

> tt <-nrow(testdata)

> for(i in 1:tt){

+ sumh = sumh+(pred\_data[i]+4.18-testdata[i,8])^2

+ }

> sumg = sumh/tt

> sumg

**[1] 0.4479991**

> head(pred\_data)

[,1]

1 0.9171757

3 1.2262977

15 1.0504644

20 1.0262733

21 1.0037606

28 1.4820333

> pred\_data1 = pred\_data + 4.18

> head(pred\_data1)

[,1]

1 5.097176

3 5.406298

15 5.230464

20 5.206273

21 5.183761

28 5.662033

> test2 = testdata[,8]

> pdf("part3Ridge.pdf")

> plot(test2,pred\_data1)

> dev.off()

RStudioGD

2

**3. OUTPUT OF RIDGE REGRESSION WITH ALL VARIABLES**

> redWine <- read.csv("~/Desktop/Data Mining/Project3/winequality-red.csv", sep=";")

> View(redWine)

> require(rpart)

> require(caTools)

> require(MASS)

> y = redWine[,12]

> train\_rows = sample.split(y, SplitRatio = 0.8)

> train = redWine[train\_rows,]

> test = redWine[!train\_rows,]

> train2= train[,-12]

> select(lm.ridge(train$quality~.,data=train2,lambda = 10))

modified HKB estimator is 20.80607

modified L-W estimator is 16.29286

smallest value of GCV at 10

> ridgereg = (lm.ridge(train$quality~.,data=train2,lambda = 10))

> lm.ridge(train$quality~.,data=train2,lambda = 10)

fixed.acidity

6.1413276709 0.0001536191

volatile.acidity citric.acid

-1.1172519488 -0.0460245300

residual.sugar chlorides

0.0143567260 -2.0733471046

free.sulfur.dioxide total.sulfur.dioxide

0.0032093894 -0.0031584565

density pH

-1.4531349116 -0.4817268083

sulphates alcohol

0.9236025177 0.2653664959

> pred\_data <- scale(test[,1:11],center = F,scale = ridgereg$scales)%\*%ridgereg$coef

> sumh<-0

> tt <- nrow(test)

> #Add the intercept value obtained from lm.ridge function.

> for(i in 1:tt){

+ sumh=sumh+(pred\_data[i]+6.14-test[i,12])^2

+ }

> error = sumh/tt

> error

**[1] 0.421975**

> test2= test[,12]

> pred\_data1 = pred\_data + 6.14

> pdf("Part3\_ridge1.pdf")

> plot(test2,pred\_data1)

> dev.off()

RStudioGD

2