Homework1

Mert ŞEN

Boston Housing

I chosed the "Boston Housing" dataset for my homework. It is available on MASS package. Attributes and explanations of Boston Housing dataset are in the below:

```
library(MASS)
help("Boston")
```

```
httpd yardım sunucusu başlatılıyor ... tamamlandı
```

Lets take a 80% of sampling of our dataset and observe. We have 506 samples and 80% of this is equals 408 samples.

```
trainingDataIndex<-sample(1:nrow(Boston),408)
traingingData<-Boston[trainingDataIndex,]#it 408 row and all columns
testDataIndex<-setdiff(1:nrow(Boston),trainingDataIndex)#we are going to use this when we
testData<-Boston[testDataIndex,]
```

Let's produce three different fitted models and compare which model is best for predicting Boston housing prices.

For this example, we will call it Fitted Model 1. We will use all the attributes (columns) and check if they have an effect on the median value of owner-occupied homes, which is a measure of the housing market value in Boston

```
fittedModel1<-lm(medv ~ ., data = traingingData )
summary(fittedModel1)
```

```
Call:
```

```
lm(formula = medv ~ ., data = traingingData)

Residuals:
    Min      1Q  Median      3Q     Max
-10.501  -2.805  -0.605   1.852  26.727

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.883711   5.724969   5.744 1.85e-08 ***
crim         -0.130012   0.036183  -3.593 0.000368 ***
zn            0.048280   0.015040   3.210 0.001435 **
indus         0.025035   0.069267   0.361 0.717978
chas          3.381729   0.891847   3.792 0.000173 ***
nox         -16.079841   4.190502  -3.837 0.000145 ***
rm            4.132351   0.459139   9.000  < 2e-16 ***
age          -0.005279   0.014354  -0.368 0.713232
dis          -1.455596   0.224408  -6.486 2.64e-10 ***
rad           0.268497   0.074127   3.622 0.000330 ***
tax          -0.010508   0.004280  -2.455 0.014518 *
ptratio      -0.896280   0.147344  -6.083 2.80e-09 ***
black         0.008117   0.003134   2.590 0.009955 **
lstat        -0.533841   0.055401  -9.636  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.721 on 394 degrees of freedom
Multiple R-squared:  0.7515,	Adjusted R-squared:  0.7433
F-statistic: 91.66 on 13 and 394 DF,  p-value: < 2.2e-16
```

For this example, let's call it Fitted model 2. Now we are going to use the age of the building, proportion of residential area, and crime rate to observe if they have an effect on

```r
fittedModel2<-lm(medv~age+zn+crim, data = traingingData)
summary(fittedModel2)
```

```
Call:
lm(formula = medv ~ age + zn + crim, data = traingingData)

Residuals:
    Min      1Q  Median      3Q     Max
-14.237  -4.736  -1.870   2.046  30.753
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.59992    1.38417  19.217  < 2e-16 ***
age         -0.05399    0.01781  -3.032  0.00258 ** 
zn           0.08622    0.02030   4.248 2.68e-05 ***
crim        -0.30427    0.05106  -5.959 5.54e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.11 on 404 degrees of freedom
Multiple R-squared:  0.2482,    Adjusted R-squared:  0.2426 
F-statistic: 44.45 on 3 and 404 DF,  p-value: < 2.2e-16
```

For this example, let's call it Fitted model 3. We are going to use nitrogen oxide concentration, accessibility to radial highways, and pupil-teacher ratio to observe if they have an effect on

```r
fittedModel3<-lm(medv~ptratio+nox+rad+dis,data = traingingData)
summary(fittedModel3)
```

```
Call:
lm(formula = medv ~ ptratio + nox + rad + dis, data = traingingData)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.753  -4.834  -1.029   3.345  31.798 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  93.42981    5.28689  17.672  < 2e-16 ***
ptratio      -2.15542    0.18924 -11.390  < 2e-16 ***
nox         -48.59177    5.28925  -9.187  < 2e-16 ***
rad           0.09077    0.05792   1.567    0.118    
dis          -1.24579    0.26935  -4.625 5.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.166 on 403 degrees of freedom
Multiple R-squared:  0.4145,    Adjusted R-squared:  0.4087 
F-statistic: 71.33 on 4 and 403 DF,  p-value: < 2.2e-16
```

When we observe all three Fitted models, Fitted model 1 is the most effective one among them. This is because it has a higher Adjusted R-squared value of 0.5328 and all of its coefficients are statistically significant with a code of 3 "*" which means we can be 100% sure that these properties have a major effect on the housing market value in Boston. After Fitted model 1, Fitted model 3 follows it, and Fitted model 2 is the least effective Fitted model.

Now, let's check our predictions and see how much they deviated from the values in our test dataset. Since it can be difficult to interpret only the predicted numbers, we can use the "accuracy()" function to determine the error percentage.

```r
fittedModel1Predict<-predict(fittedModel1,newdata = testData)
fittedModel2Predict<-predict(fittedModel2,newdata = testData)
fittedModel3Predict<-predict(fittedModel3,newdata = testData)
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```r
accuracy(fittedModel1Predict,testData$medv)
```

```
                 ME     RMSE      MAE       MPE     MAPE
Test set 0.09945238 4.928077 3.284952 -0.845866 18.23931
```

```r
accuracy(fittedModel2Predict,testData$medv)
```

```
                ME     RMSE      MAE       MPE    MAPE
Test set -1.265359 7.551472 5.391201 -16.98777 28.6975
```

```r
accuracy(fittedModel3Predict,testData$medv)
```

```
                ME     RMSE      MAE       MPE     MAPE
Test set -1.564599 6.865225 5.079023 -18.18524 29.11139
```

To determine the prediction accuracy, we can use the MAPE (Mean Absolute Percentage Error) metric. Upon calculation, we can see that Fitted model 1 has the most acceptable error rate of 17%, while Fitted model 2 and Fitted model 3 have error rates of 29% and 27%, respectively. Therefore, we can conclude that Fitted model 1 is the most suitable model for our prediction modelling.