Topic 4.3 Cross-Validation

Load needed packages.

```
library(Stat2Data)
```

EXAMPLE 4.3 Houses in NY: cross-validation

Load **Houses** data from Stat2Data package and look at the structure of the data.

```
data(HousesNY)
str(HousesNY)
```

```
## 'data.frame':    53 obs. of  5 variables:
##  $ Price: num  57.6 120 150 143 92.5 ...
##  $ Beds : int  3 6 4 3 3 2 2 4 4 3 ...
##  $ Baths: num  2 2 2 2 1 1 2 3 2.5 2 ...
##  $ Size : num  0.96 2.79 1.7 1.2 1.33 ...
##  $ Lot  : num  1.3 0.23 0.27 0.8 0.42 0.34 0.29 0.21 1 0.3 ...
```

Create Training (first 35 cases) and Holdout (last 18) samples.

```
train=HousesNY[1:35,]
holdout=HousesNY[36:53,]
```

Fit a model to predict Price based on Size for the training sample.

```
modelTrain=lm(Price~Size,data=train)
summary(modelTrain)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.911 -29.397  -2.135  25.980  67.093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.549     16.824   3.361  0.00197 **
## Size          33.611      9.354   3.593  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.89 on 33 degrees of freedom
## Multiple R-squared:  0.2812, Adjusted R-squared:  0.2594
## F-statistic: 12.91 on 1 and 33 DF,  p-value: 0.00105
```

```
anova(modelTrain)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Size       1  15714   15714  12.912 0.00105 **
## Residuals 33  40162    1217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 4.6a Price versus Size scatterplot with training line for training sample

```
plot(Price~Size,data=train, pch=16)
abline(modelTrain,col="blue")
```

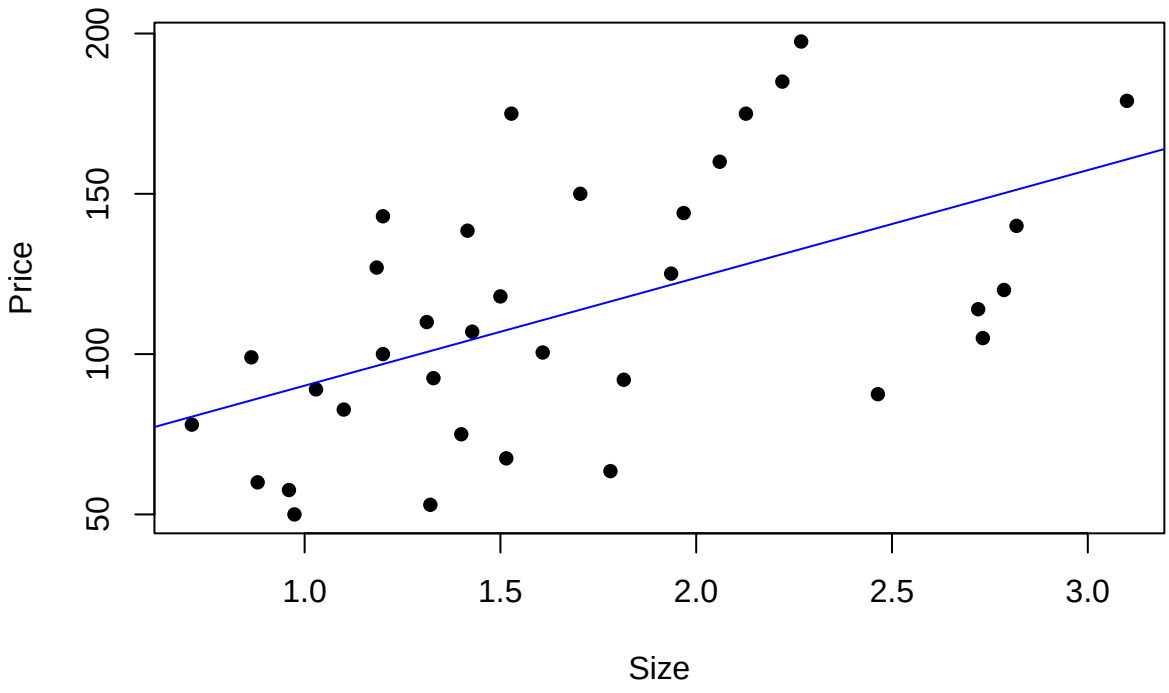
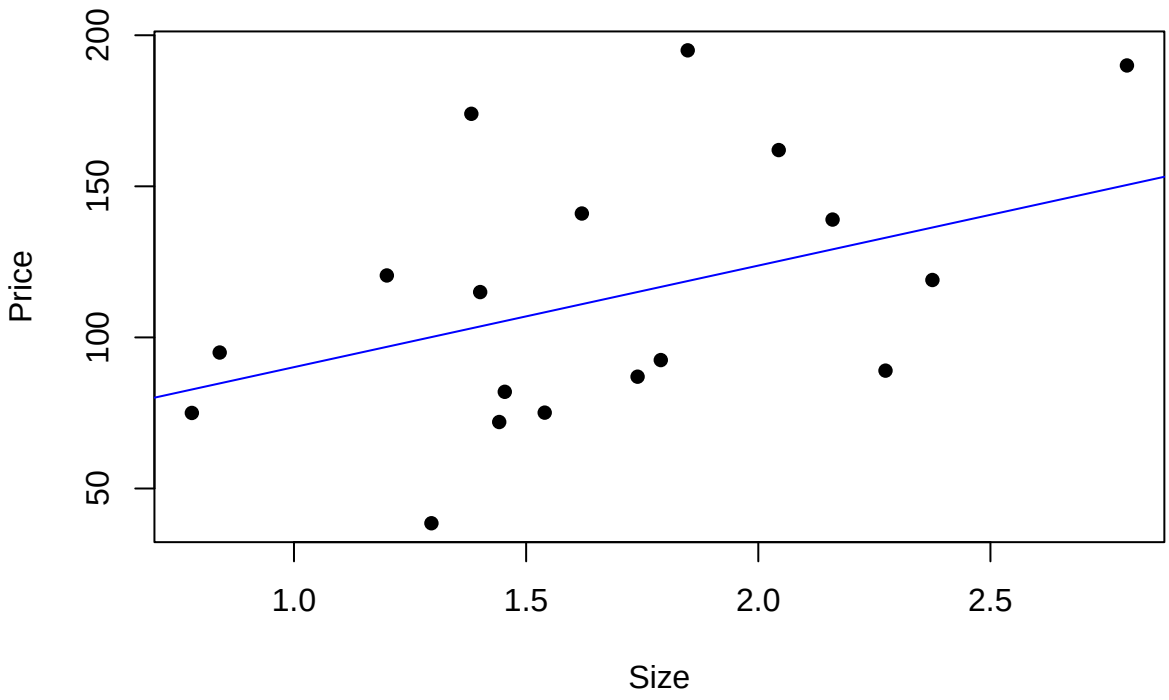


FIGURE 4.6b Price versus Size scatterplot with training line for holdout sample

```
plot(Price~Size,data=holdout, pch=16)
abline(modelTrain,col="blue")
```



Find predictions and prediction errors for the holdout sample, using the training model.

```
holdout$PriceHat=predict(modelTrain,holdout)
holdout$Residuals=holdout$Price-holdout$PriceHat
holdout[,6:7]
```

```
##     PriceHat  Residuals
## 36  96.88259  23.617406
## 37 105.01646 -33.016460
## 38  84.78263  10.217371
## 39 100.10925 -61.609252
## 40 129.14917   9.850831
## 41  82.76597  -7.765968
## 42 102.99980  71.000201
## 43 136.37554 -17.375537
## 44 132.98082 -43.980824
## 45 108.31034 -33.210339
## 46 116.71309 -24.213093
## 47 110.99922  30.000780
## 48 105.41979 -23.419792
## 49 125.25029  36.749709
## 50 118.66253  76.337468
## 51 150.45855  39.541448
## 52 103.63841  11.361592
## 53 115.03254 -28.032542
```

Compute mean, SSE, and MSE for the holdout residuals.

```
mean(holdout$Residuals)
```

```
## [1] 2.002944
```

```
SSE=sum(holdout$Residuals^2)
SSE
```

```
## [1] 25776.68
```

```
MSE=SSE/18
MSE
```

```
## [1] 1432.038
```

We compare MSE = 1432 to MSE = 1217 from the ANOVA table (see output above) for the initial training model.

EXAMPLE 4.4 Houses in NY: cross-validation correlation and shrinkage

Find the cross-validation correlation, square it, and compue shrinkage.

```
crossR=cor(holdout$Price,holdout$PriceHat)
crossRsq=crossR^2
shrinkage=summary(modelTrain)$r.squared-crossRsq
c(crossR,crossRsq,shrinkage)
```

```
## [1] 0.48729305 0.23745452 0.04377468
```

cross-validation correlation = `crossR`

cross-validation correlation squared = `crossRsq`

R^2 for training sample = summary(modelTrain)$r.squared