Chapter 7 section 3

Dr. Akram Almohalwas

Cross Validation in R

We will randomly divide the data set into two parts (the training sample and the testing sample: 67% and 33% respectively or 70% 30%).

. We will create a regression model based on the training sample.

. We will use the regression model based on the training sample to compute $\widehat{Y}$ in the test sample.

. We will calculate the coefficient of correlation between the y-hat calculated for the testing sample using the model created from the training data and the actual Y in the testing sample.

. We will draw the plot of $\widehat{Y}$ and Y and decided

. Based on the coefficient of correlation between Y and Y-hat and the scatterplot, we will decide how good of a model we have for prediction.

North Carolina Births:

Variables in the data set

. fage father's age in years.

. mage mother's age in years.

. meduc mother's education

. weeks length of pregnancy in weeks.

. premie whether the birth was classified as premature (premie) or fullterm.

. visits number of hospital visits during pregnancy.

. marital whether mother is married or not married at birth.

. gained weight gained by mother during pregnancy in pounds.

. weight weight of the baby at birth in pounds.

. lowbirthweight whether baby was classified as low birthweight (low) or not (not low).

. gender gender of the baby, female or male.

. habit status of the mother as a nonsmoker or a smoker.

. whitemom whether mom is white or not white. Possible situations with analysis of covariance.

Case one. Coincident regression line. The simplest model is

when. . The dummy variable has no effect on Y. . The regression line is exactly the same for both values of the dummy variable.

```
options(warn=-1)
births <- read.delim("~/STAT 101A/Data Sets/ncbirths.txt")
attach(births)
head(births)

##   fage mage      mature weeks    premie visits marital racemom hispmom
## 1   NA   13 younger mom    39 full term     10       2       2       N
## 2   NA   14 younger mom    42 full term     15       2       2       N
## 3   19   15 younger mom    37 full term     11       2       1       M
## 4   21   15 younger mom    41 full term      6       2       1       M
## 5   NA   15 younger mom    39 full term      9       2       2       N
## 6   NA   15 younger mom    38 full term     19       2       2       N
##   gained weight lowbirthweight sexbaby     habit
## 1     38   7.63        not low    male nonsmoker
## 2     20   7.88        not low    male nonsmoker
## 3     38   6.63        not low  female nonsmoker
## 4     34   8.00        not low    male nonsmoker
## 5     27   6.38        not low  female nonsmoker
## 6     22   5.38            low    male nonsmoker

dim(births)

## [1] 1000   14
```

Now we want to split our data into 70 Training-30 Testing data sets.

```
# 70% of the sample size
smp_size <- floor(0.70 * nrow(births))
smp_size

## [1] 700
## set the seed to make your partition reproductible
set.seed(123456)

train_ind <- sample(seq_len(nrow(births)), size = smp_size)

train <- births[train_ind, ]
test <-  births[-train_ind, ]

write.table(train, "~/STAT 101A/Data Sets/birthsTrain.txt", sep="\t")
write.table(test, "~/STAT 101A/Data Sets/birthsTest.txt", sep="\t")

head(train)

##     fage mage      mature weeks    premie visits marital racemom hispmom
## 798   35   33 younger mom    37 full term     13       1       1       M
## 753   34   32 younger mom    37 full term     11       1       1       M
## 391   NA   24 younger mom    34    premie      7       2       1       N
## 341   21   24 younger mom    40 full term     18       1       1       M
## 360   NA   24 younger mom    38 full term     16       2       1       N
## 198   21   21 younger mom    38 full term     10       1       1       M
##     gained weight lowbirthweight sexbaby     habit
## 798     21   9.63        not low    male nonsmoker
## 753     28   5.56        not low    male nonsmoker
## 391     18   5.06            low    male nonsmoker
## 341     35   7.13        not low  female nonsmoker
## 360     15   6.31        not low  female nonsmoker
## 198     35   6.50        not low  female nonsmoker

trm1<-lm(weight~weeks+mage+fage+visits+gained,data=train)
summary(trm1)

##
## Call:
## lm(formula = weight ~ weeks + mage + fage + visits + gained,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5301 -0.7200 -0.0727  0.7555  3.5749
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.873434   0.712114  -8.248 1.16e-15 ***
## weeks        0.320044   0.017512  18.276  < 2e-16 ***
## mage         0.014622   0.012214   1.197 0.231767
## fage         0.004379   0.010836   0.404 0.686311
## visits      -0.006819   0.012580  -0.542 0.587972
## gained       0.011070   0.003309   3.346 0.000876 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.117 on 560 degrees of freedom
##   (134 observations deleted due to missingness)
## Multiple R-squared:  0.3926, Adjusted R-squared:  0.3871
## F-statistic: 72.38 on 5 and 560 DF,  p-value: < 2.2e-16
```

```
anova(trm1)
```

```
## Analysis of Variance Table
##
## Response: weight
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## weeks       1 431.43  431.43 346.0122 < 2.2e-16 ***
## mage        1   5.57    5.57   4.4669 0.0349993 *
## fage        1   0.15    0.15   0.1206 0.7285116
## visits      1   0.15    0.15   0.1225 0.7264376
## gained      1  13.96   13.96  11.1929 0.0008761 ***
## Residuals 560 698.24    1.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(car)
vif(trm1)
```

```
##    weeks     mage     fage   visits   gained
## 1.035316 2.537696 2.498718 1.059334 1.010175
```

```
# trbackAIC <- step(trm1,direction="backward", data=train)

trm2<-lm(weight~weeks+gained,data=train)
summary(trm2)
```

```
##
## Call:
## lm(formula = weight ~ weeks + gained, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5074 -0.7202 -0.0626  0.7359  4.3035
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.537333   0.598831  -9.247  < 2e-16 ***
## weeks        0.322118   0.015463  20.831  < 2e-16 ***
## gained       0.011086   0.002992   3.705 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.119 on 678 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.4017, Adjusted R-squared:  0.3999
## F-statistic: 227.6 on 2 and 678 DF,  p-value: < 2.2e-16
```

```
anova(trm2)
```

```
## Analysis of Variance Table
##
## Response: weight
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## weeks       1 553.26  553.26 441.486 < 2.2e-16 ***
## gained      1  17.20   17.20  13.728 0.0002285 ***
## Residuals 678 849.65    1.25                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(trm2)
```

```
##    weeks   gained 
## 1.001877 1.001877
```

```
trm2
```

```
## 
## Call:
## lm(formula = weight ~ weeks + gained, data = train)
## 
## Coefficients:
## (Intercept)        weeks       gained  
##    -5.53733      0.32212      0.01109
```

```
# par(mfrow=c(1,1))
# y_hat<-trm2$fitted.values
# length(y_hat)
# length(train$weight)
# plot(y_hat,train$weight,xlab="Fitted Values")
# abline(lsfit(trm2$fitted.values,weight))


Test <- read.delim("~/STAT 101A/Data Sets/birthsTest.txt")
head(Test)
```

```
##    fage mage      mature weeks    premie visits marital racemom hispmom
## 11   30   16 younger mom    45 full term      9       2       1       M
## 13   NA   16 younger mom    40 full term      4       2       2       N
## 18   16   16 younger mom    24    premie      5       2       2       N
## 20   18   17 younger mom    37 full term     10       2       2       N
## 25   26   17 younger mom    38 full term     11       2       1       M
## 27   NA   17 younger mom    39 full term     12       2       2       N
##    gained weight lowbirthweight sexbaby     habit
## 11     28   7.44        not low    male nonsmoker
## 13     12   6.00        not low  female nonsmoker
## 18     12   1.50            low    male nonsmoker
## 20     39   6.19        not low  female nonsmoker
## 25     30   9.50        not low  female nonsmoker
## 27     50   7.50        not low    male nonsmoker
```

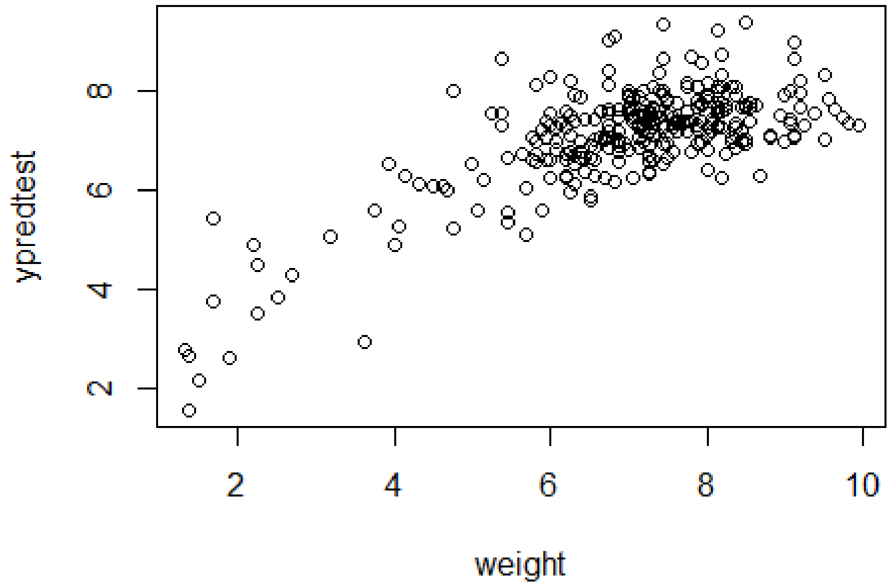
```
attach(Test)

## The following objects are masked from births:
##
##     fage, gained, habit, hispmom, lowbirthweight, mage, marital,
##     mature, premie, racemom, sexbaby, visits, weeks, weight

ypredtest<-  -5.962291  + 0.334483* weeks + 0.008246*gained
summary(ypredtest)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.561   6.746   7.289   7.072   7.652   9.370       9

plot(weight,ypredtest,data=Test)
```



```
mtest<-lm(ypredtest~weight,data=Test)
summary(mtest)

##
## Call:
## lm(formula = ypredtest ~ weight, data = Test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69577 -0.43907 -0.00595  0.41997  2.37471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.56309    0.19488   18.28   <2e-16 ***
## weight       0.50285    0.02723   18.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7369 on 289 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.5412, Adjusted R-squared:  0.5397
## F-statistic:   341 on 1 and 289 DF,  p-value: < 2.2e-16
```