MODELING OF BOSTON HOUSE PRICES DATA

Rajat Agarwal

7/18/2020

```
rm(list=ls())
getwd()        # It  will print  the current directory whre we are working
```

```
## [1] "/cloud/project"
```

```
library(MASS) # Importing the library containing the Boston Housing Data
data<-Boston
dim(data)
```

```
## [1] 506  14
```

```
n=nrow(data)
p=ncol(data) #p is Number of unknown coefficient including intercept to be estimated for regression lin
n;p
```

```
## [1] 506
```

```
## [1] 14
```

```
Y=(data$medv)    # Setting the Response
X=as.matrix(cbind(rep(1,n),data[,-which(colnames(data)=='medv')]))    #Design Matrix or Predictor matri
colnames(X)[1]='intercept'
det(t(X)%*%X)
```

```
## [1] 1.140194e+53
```

```
#determinant of (t(X)%*%X) is not zero . Hence we conclude the matrix is full column rank matrix
beta_hat=solve(t(X)%*%X)%*%t(X)%*%Y          #Least square solution
beta_hat  # estimated the regression coefficient
```

```
##                     [,1]
## intercept  3.645949e+01
## crim      -1.080114e-01
## zn         4.642046e-02
## indus      2.055863e-02
## chas       2.686734e+00
## nox       -1.776661e+01
## rm         3.809865e+00
## age        6.922246e-04
## dis       -1.475567e+00
## rad        3.060495e-01
## tax       -1.233459e-02
## ptratio   -9.527472e-01
## black      9.311683e-03
## lstat     -5.247584e-01
```

```r
#----------------------------Testing Significance of Regressors for predicting mean hosing price------

# #Ho=beta(i)=0 for all i=1(1)13 vs H1:- beta(i) not equalto 0 for atleast one  i
# FULL MODEL  is
# Y=beta0 + beta1*CRIM + beta2*ZN + beta3*INDUS + beta4*CHAS + beta5*NOX + beta6*rm + beta7*AGE + beta8
SSreg=sum((X%*%beta_hat-mean(Y))^2) #Residual sum of square under null hypthesis
SSreg
```

```
## [1] 31637.51
```

```r
SSres= sum((Y-(X%*%beta_hat))^2)#Residual sum of square for full model
SSres
```

```
## [1] 11078.78
```

```r
Fcal=(SSreg*(n-p))/((p-1)*SSres)   #Test statistics
Fcal
```

```
## [1] 108.0767
```

```r
alpha=0.05
Ftab=qf(1-alpha,p-1,n-p)
Ftab
```

```
## [1] 1.740074
```

```r
# (Decision Rule reject Ho if Fcal > Ftab   otherwise accept it)
# (Since Fcal=108.0767  is greater than Ftab=1.740074 under 5% level of significance so we reject Ho)
# Thus we know there exist atleast one regressor which is significant

#---------------------------------'Performing Individual regressor test'-----------------------------

C=solve(t(X)%*%X)
ftab=qf(1-alpha,1,n-p)  #tabulatted value of f distribution with (1,n-p)dof at 5% level of significance
fcal=rep(0,p-1)
#checking the significance of individual regressor
for (i in 1:(p-1))
  {
  fcal[i]=((beta_hat[i+1])^2)/((SSres/(n-p))*C[i+1,i+1])
  if(fcal[i]>ftab) print(" regressor is significant")
  else print(paste(colnames(X)[i+1],"regressor is not significant"))
}
```

```
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] "indus regressor is not significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] "age regressor is not significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
## [1] " regressor is significant"
```

```r
# As we see regressor corresponding to INDUS and AGE is not signifiacnt. Now , we consider new model wi
# all regressor except the regressor corrresponding to INDUS and AGE i.e. beta3 and beta7 and including
# intercept term


#----------------------------------Model M1 is True or not ------------------------------------------
# Our New model is
# Y=beta0 + beta1*CRIM + beta2*ZN + beta4*CHAS + beta5*NOX + beta6*rm + beta8*DIS + beta9*RAD + beta10*
# #Hypothesis of interest is Ho:-"M1 is true" at 5% level of significance
# #Testing Ho:M1 is true  BETA2=0  where BETA=(BETA1,BETA2)'

Z=X[,c(-which(colnames(X)=='indus'),-which(colnames(X)=='age'))]
W=X[,c(which(colnames(X)=='indus'),which(colnames(X)=='age'))]
H=X%*%solve(t(X)%*%X)%*%t(X)
H1=Z%*%solve(t(Z)%*%Z)%*%t(Z)
SSregg=t(Y)%*%(H-H1)%*%Y    # Extra Sum of square due to BETA2
FcalN=(SSregg)*(n-p)/(SSres*2)
FtabN=qf(1-alpha,2,n-p)
if (FcalN<FtabN) print("M1 is TRUE MODEL")  else print(" M1 is FALSE MODEL ")
```

```
## [1] "M1 is TRUE MODEL"
```

```r
#----------------------CONFIDENCE INTERVAL OF NOX FROM FULL MODEL AND SLRM-----------------------------
cnox=which(colnames(X)=='nox')
cl=beta_hat[cnox]-sqrt(qf(1-alpha,1,n-p)*C[cnox,cnox]*(SSres/(n-p)))
cu=beta_hat[cnox]+sqrt(qf(1-alpha,1,n-p)*C[cnox,cnox]*(SSres/(n-p)))
CONFIDENCE_INTERVAL=c(cl,cu)
CONFIDENCE_INTERVAL
```

```
## [1] -25.27163 -10.26159
```

```r
#Considering SLRM with predictor as NOX variable and response MEDV
#model Y=b0 + b1*nox

cintercept=which(colnames(X)=='intercept')
X1=as.matrix(X[,c(cintercept,cnox)])
b_hat_nox=solve(t(X1)%*%X1)%*%t(X1)%*%Y
b_hat_nox[2]
```

```
## [1] -33.91606
```

```r
c1=solve(t(X1)%*%X1)
RSS1=t(Y-X1%*%b_hat_nox)%*%(Y-X1%*%b_hat_nox)
c1l=b_hat_nox[2]-(qt(0.975,n-2)*sqrt(c1[2,2]*(RSS1/(n-2))))
c1u=b_hat_nox[2]+(qt(0.975,n-2)*sqrt(c1[2,2]*(RSS1/(n-2))))
CON._INT=c(c1l,c1u)
CON._INT
```

```
## [1] -40.19584 -27.63627
```

```r
#Since estimate of regressor lie in confidence interval so we say NOX is a significant estimator of pre


#-------------------------Testing the equality of regressor of CRIM and AGE --------------------------
# "Considering Full Model"
# "Test Hypothesis Ho: beta(CRIRM)=beta(AGE) "
cld=(beta_hat[2]-beta_hat[8])-(qt(0.998,n-p)*(sqrt((SSres/(n-p))*(C[2,2]+C[8,8]-2*C[2,8]))))
```

```r
cud=(beta_hat[2]-beta_hat[8])+(qt(0.998,n-p)*(sqrt((SSres/(n-p))*(C[2,2]+C[8,8]-2*C[2,8]))))
CON.INTd=c(cld,cud)
CON.INTd
```

```
## [1] -0.211250650 -0.006156515
```

```r
# " This interval does not contain zero. So, there is  significant difference in the
# regressor coefficient  of CRIM AND AGE at 5% level of significance "

# "Considering model with standardized value of predictors"
Xd=matrix(0,n,p-1)
for(i in 1:p-1)
  {
    Xd[,i]=(X[,i+1]-mean(X[,i+1]))/sd(X[,i+1])
    }
Xd=cbind(rep(1,n),Xd)
beta_hatd=solve(t(Xd)%*%Xd)%*%t(Xd)%*%Y
Cd=solve(t(Xd)%*%Xd)
RSSd= t(Y-Xd%*%beta_hatd)%*%(Y-Xd%*%beta_hatd)
clsd=(beta_hatd[2]-beta_hatd[8])-(qt(0.975,n-p)*(sqrt((RSSd/(n-p))*(Cd[2,2]+Cd[8,8]-2*Cd[2,8]))))
cusd=(beta_hatd[2]-beta_hatd[8])+(qt(0.975,n-p)*(sqrt((RSSd/(n-p))*(Cd[2,2]+Cd[8,8]-2*Cd[2,8]))))
CON.INTsd=c(clsd,cusd)
CON.INTsd
```

```
## [1] -1.86777495 -0.02932485
```

```r
# "Confidence interval does not contain value 0 .So, there is no change in decision also
# after standardizing the model that is there is  significant difference in the
# regressor coefficient  of CRIM AND AGE at 5% level of significance"
```