in class 9

Benjamin Smith

11/20/2019

```r
wages=read.table('~/R_work/stat_comp/wage.txt',header=T)
n=nrow(wages)
nTst=100
set.seed(195021)
tst=sample(1:n,size=nTst)
TRN.DATA=wages[-tst,]
TST.DATA=wages[tst,]

fm0=lm(Wage~1,data=TRN.DATA) # our 'baseline' model
fmA=lm(Wage~.,data=TRN.DATA) # note: Wage~. means regress Wage on all the other variables in 'data'
yHat0=predict(fm0,newdata=TST.DATA)
yHatA=predict(fmA,newdata=TST.DATA)

PRSS0=sum((TST.DATA$Wage-yHat0)^2)
PRSSA=sum((TST.DATA$Wage-yHatA)^2)
(R2.tst=(PRSS0-PRSSA)/PRSS0)
```

```
## [1] 0.1915918
```

```r
# R-sq. in the training sample
trnSS0=sum(residuals(fm0)^2)
trnSSA=sum(residuals(fmA)^2)
(R2.trn= (trnSS0-trnSSA)/trnSS0)
```

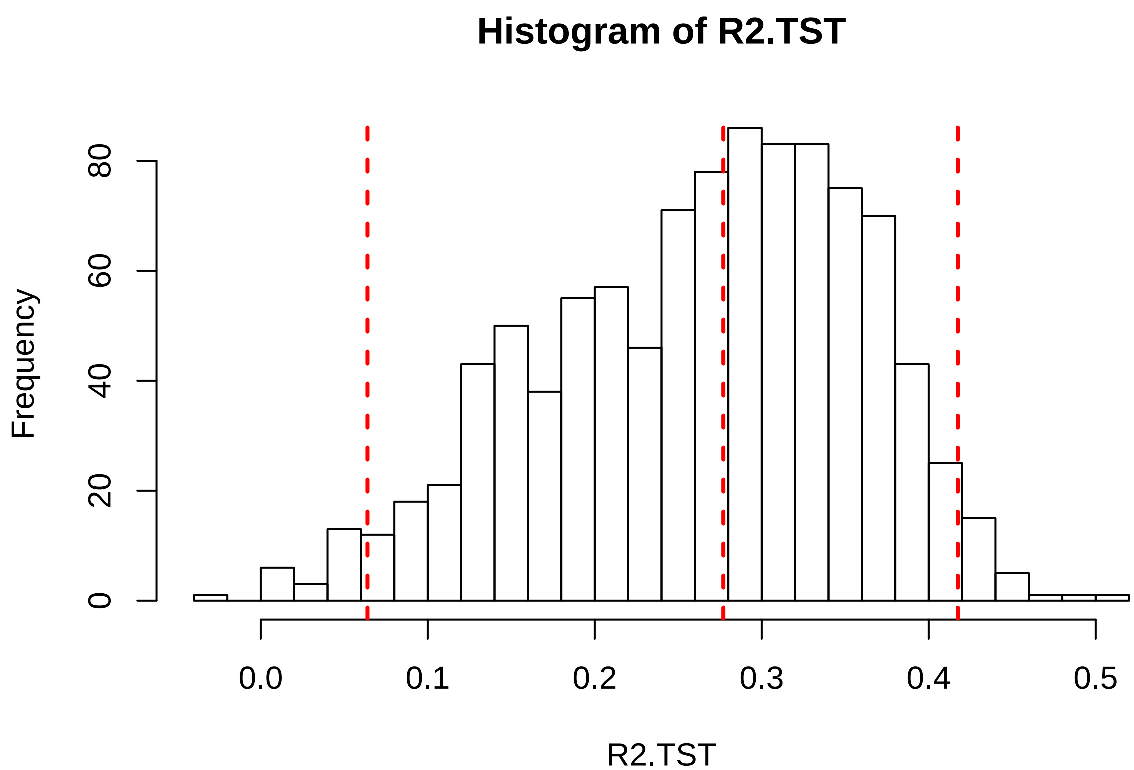
```
## [1] 0.2977964
```

```r
summary(fmA)
```

```
##
## Call:
## lm(formula = Wage ~ ., data = TRN.DATA)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.871 -2.627 -0.546  1.815 37.565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.12165    1.44459  -3.545 0.000436 ***
## Education    1.00506    0.09342  10.759  < 2e-16 ***
## South       -0.70230    0.47787  -1.470 0.142409
## Black       -0.71079    0.66163  -1.074 0.283308
## Hispanic    -0.74583    1.01123  -0.738 0.461203
## Sex         -2.11982    0.44245  -4.791 2.31e-06 ***
## Married      0.42074    0.47198   0.891 0.373207
## Experience   0.10548    0.01965   5.367 1.33e-07 ***
## Union        1.38867    0.58755   2.363 0.018559 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.441 on 419 degrees of freedom
## Multiple R-squared:  0.2978, Adjusted R-squared:  0.2844
## F-statistic: 22.21 on 8 and 419 DF,  p-value: < 2.2e-16
```

```r
n=nrow(wages)
nTst=100
nRep=1000
R2.TST=rep(NA,nRep)

for(i in 1:nRep){
  tst=sample(1:n,size=nTst)
  TRN.DATA=wages[-tst,]
  TST.DATA=wages[tst,]

  fm0=lm(Wage~1,data=TRN.DATA) # our 'baseline' model
  fmA=lm(Wage~.,data=TRN.DATA) # note: Wage~. means regress Wage on all the other variables in 'data'
  yHat0=predict(fm0,newdata=TST.DATA)
  yHatA=predict(fmA,newdata=TST.DATA)

  PRSS0=sum((TST.DATA$Wage-yHat0)^2)
  PRSSA=sum((TST.DATA$Wage-yHatA)^2)
  R2.TST[i]=(PRSS0-PRSSA)/PRSS0
}

hist(R2.TST,30);abline(v=quantile(R2.TST,prob=c(.025,.5,.975)),col=2,lwd=2,lty=2)
```




```r
n=nrow(wages)
nFolds=5
folds=rep(1:nFolds,ceiling(n/nFolds))[1:n] # this gives approximately balanced counts per fold
```

```
R2.TST=rep(NA,nFolds)

for(i in 1:nFolds){
  folds=sample(folds, size=n,replace=F) # randomizing the fold assigment
  tst=which(folds==i)
  TRN.DATA=wages[-tst,]
  TST.DATA=wages[tst,]

  fm0=lm(Wage~1,data=TRN.DATA) # our 'baseline' model
  fmA=lm(Wage~.,data=TRN.DATA) # note: Wage~. means regress Wage on all the other variables in 'data'
  yHat0=predict(fm0,newdata=TST.DATA)
  yHatA=predict(fmA,newdata=TST.DATA)

  PRSS0=sum((TST.DATA$Wage-yHat0)^2)
  PRSSA=sum((TST.DATA$Wage-yHatA)^2)
  R2.TST[i]=(PRSS0-PRSSA)/PRSS0
}
R2.TST
```

```
## [1] 0.3475282 0.1955655 0.3184900 0.2687620 0.3348280
```

Model Comparison using AIC/BIC/Adjusted R-2 and out-of-sample prediction R-sq.

Consider these two competing hypotheses: H1: `Wage~Sex+Education+Experience`, H2: `Wage~.`

- Fit the two models to the full data set, obtain R-sq., adjusted R-sq., AIC, BIC and a p-value from an F-test.

```
  wages=read.table('wage.txt',header=T)
  n=nrow(wages)
  nTst=100
  set.seed(195021)
  tst=sample(1:n,size=nTst)
  TRN.DATA=wages[-tst,]
  TST.DATA=wages[tst,]
  fm0=lm(Wage~1, data = TRN.DATA) # our 'baseline' model
  fm1=lm(Wage~Sex+Education+Experience,data=TRN.DATA)
  fm2=lm(Wage~.,data=TRN.DATA) # note: Wage~. means regress Wage on all the other variables in 'data'
  yHat0=predict(fm0,newdata=TST.DATA)
  yHat1=predict(fm1,newdata=TST.DATA)
  yHat2=predict(fm2,newdata=TST.DATA)

  PRSS0=sum((TST.DATA$Wage-yHat0)^2)
  PRSS1=sum((TST.DATA$Wage-yHat1)^2)
  PRSS2=sum((TST.DATA$Wage-yHat2)^2)
  (R2.tst1=(PRSS0-PRSS1)/PRSS0)
```

```
## [1] 0.1440046
```

```
  (R2.tst2=(PRSS0-PRSS2)/PRSS0)
```

```
## [1] 0.1915918
```

```
  # R-sq. in the training sample
  trnSS0=sum(residuals(fm0)^2)
  trnSS1=sum(residuals(fm1)^2)
  trnSS2=sum(residuals(fm2)^2)
```

```r
(R2.trn1= (trnSS0-trnSS1)/trnSS0)
```

```
## [1] 0.2789198
```

```r
(R2.trn2= (trnSS0-trnSS2)/trnSS0)
```

```
## [1] 0.2977964
```

```r
R2.trn1=summary(fm1)$r.squared
R2.adj.trn1=summary(fm1)$adj.r.squared
R2.trn2=summary(fm2)$r.squared
R2.adj.trn2=summary(fm2)$adj.r.squared

AIC(fm1)
```

```
## [1] 2503.027
```

```r
BIC(fm1)
```

```
## [1] 2523.323
```

```r
AIC(fm2)
```

```
## [1] 2501.673
```

```r
BIC(fm2)
```

```
## [1] 2542.265
```

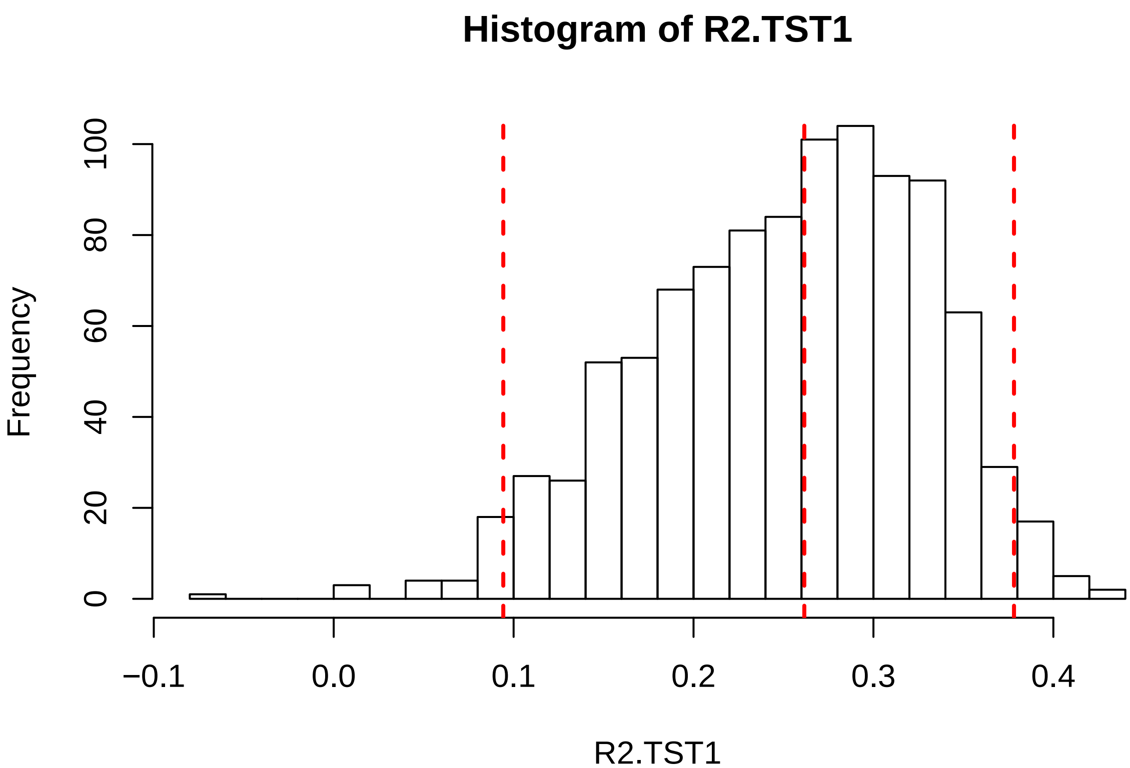
```r
anova(fm1,fm2)
```

```
## Analysis of Variance Table
##
## Model 1: Wage ~ Sex + Education + Experience
## Model 2: Wage ~ Education + South + Black + Hispanic + Sex + Married +
##     Experience + Union
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    424 8485.2
## 2    419 8263.0  5    222.13 2.2527 0.0484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

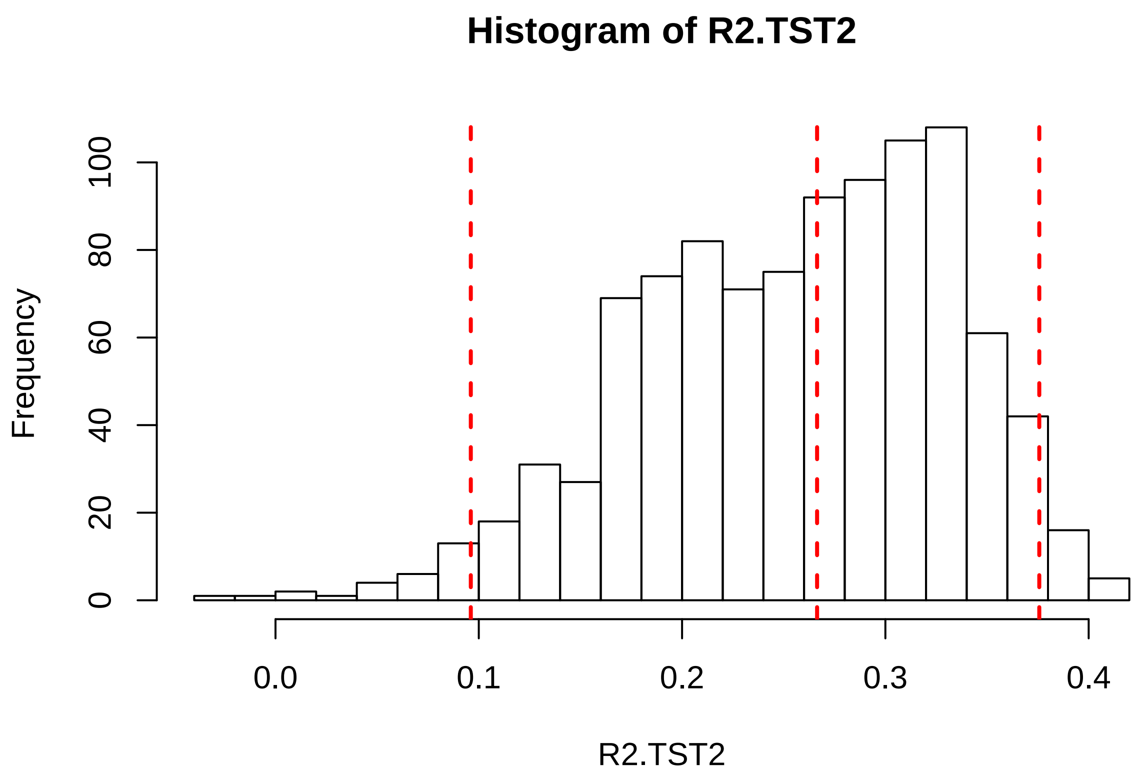
- Conduct 1000 training-testing evaluations (nTesting=150) to estimate prediciton R-sq. for H1 and H2.
- Report a table with AIC,BIC,Training R-sq., Training adj-Rsq. and prediction r-sq. for each of the models.
- Which model do you choose? Why?

```r
n=nrow(wages)
nTst=150
nRep=1000
R2.TST1=rep(NA,nRep)
R2.TST2=rep(NA,nRep)

for(i in 1:nRep){
  tst=sample(1:n,size=nTst)
  TRN.DATA=wages[-tst,]
  TST.DATA=wages[tst,]

  fm0=lm(Wage~1,data=TRN.DATA) # our 'baseline' model
  fmA=lm(Wage~Sex+Education+Experience,data=TRN.DATA) # note: Wage~. means regress Wage on all the othe
```

```
  yHat0=predict(fm0,newdata=TST.DATA)
  yHatA=predict(fmA,newdata=TST.DATA)

  PRSS0=sum((TST.DATA$Wage-yHat0)^2)
  PRSSA=sum((TST.DATA$Wage-yHatA)^2)
  R2.TST1[i]=(PRSS0-PRSSA)/PRSS0
}

hist(R2.TST1,30);abline(v=quantile(R2.TST1,prob=c(.025,.5,.975)),col=2,lwd=2,lty=2)
```




```
for(i in 1:nRep){
  tst=sample(1:n,size=nTst)
  TRN.DATA=wages[-tst,]
  TST.DATA=wages[tst,]

  fm0=lm(Wage~1,data=TRN.DATA) # our 'baseline' model
  fmA=lm(Wage~.,data=TRN.DATA) # note: Wage~. means regress Wage on all the other variables in 'data'
  yHat0=predict(fm0,newdata=TST.DATA)
  yHatA=predict(fmA,newdata=TST.DATA)

  PRSS0=sum((TST.DATA$Wage-yHat0)^2)
  PRSSA=sum((TST.DATA$Wage-yHatA)^2)
  R2.TST2[i]=(PRSS0-PRSSA)/PRSS0
}

hist(R2.TST2,30);abline(v=quantile(R2.TST2,prob=c(.025,.5,.975)),col=2,lwd=2,lty=2)
```




```
res <- cbind(c(mean(R2.TST1),mean(R2.TST2)), c(AIC(fm1), AIC(fm2)), c(BIC(fm1), BIC(fm2)), c(R2.trn1, R2

colnames(res) <- c("pred r-sq", "AIC", "BIC", "train r-sq", "train adj. r-sq")

res
```

```
##      pred r-sq      AIC      BIC train r-sq train adj. r-sq
## [1,] 0.2521552 2503.027 2523.323  0.2789198       0.2738178
## [2,] 0.2562663 2501.673 2542.265  0.2977964       0.2843891
```

```
mean(R2.TST1)
```

```
## [1] 0.2521552
```

```
mean(R2.TST2)
```

```
## [1] 0.2562663
```