Supplement for Lecture 21: Cross-Validation

Load Data

```
data("NCbirths")

NCB = NCbirths[,-which(names(NCbirths) %in% c("ID","BirthWeightGm","RaceMom","HispMom","Low","Premie"))]

str(NCB)
```

```
## 'data.frame':    1450 obs. of  9 variables:
##  $ Plural       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex          : int  1 2 1 1 1 1 2 2 2 2 ...
##  $ MomAge       : int  32 32 27 27 25 28 25 15 21 27 ...
##  $ Weeks        : int  40 37 39 39 39 43 39 42 39 40 ...
##  $ Marital      : int  1 1 1 1 1 1 1 2 1 2 ...
##  $ Gained       : int  38 34 12 15 32 32 75 25 28 37 ...
##  $ Smoke        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BirthWeightOz: int  111 116 138 136 121 117 143 113 120 124 ...
##  $ MomRace      : Factor w/ 4 levels "black","hispanic",..: 4 4 4 4 4 4 4 4 4 4 ...
```

Clean Data

```
NCB$Plural = factor(ifelse(NCB$Plural==1,"Single",ifelse(NCB$Plural==2,"Twin","Triplet")))
table(NCB$Plural)
```

```
##
##  Single Triplet    Twin
##    1401       4      45
```

```
NCB$Sex = factor(ifelse(NCB$Sex==1,"Male","Female"))
table(NCB$Sex)
```

```
##
## Female   Male
##    706    744
```

```
NCB$Marital = factor(ifelse(NCB$Marital==1,"Married","Not Married"))
table(NCB$Marital)
```

```
##
##     Married Not Married
##         950         500
```

```
NCB$Smoke = factor(ifelse(NCB$Smoke==1, "Yes","No"))
table(NCB$Smoke)
```

```
##
##   No  Yes
## 1236  209
```

```r
NCB=na.omit(NCB)

str(NCB)
```

```
## 'data.frame':    1409 obs. of  9 variables:
##  $ Plural       : Factor w/ 3 levels "Single","Triplet",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex          : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 2 1 1 1 1 ...
##  $ MomAge       : int  32 32 27 27 25 28 25 15 21 27 ...
##  $ Weeks        : int  40 37 39 39 39 43 39 42 39 40 ...
##  $ Marital      : Factor w/ 2 levels "Married","Not Married": 1 1 1 1 1 1 1 2 1 2 ...
##  $ Gained       : int  38 34 12 15 32 32 75 25 28 37 ...
##  $ Smoke        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BirthWeightOz: int  111 116 138 136 121 117 143 113 120 124 ...
##  $ MomRace      : Factor w/ 4 levels "black","hispanic",..: 4 4 4 4 4 4 4 4 4 4 ...
##  - attr(*, "na.action")= 'omit' Named int [1:41] 58 169 179 201 244 245 248 255 304 335 ...
##   ..- attr(*, "names")= chr [1:41] "58" "169" "179" "201" ...
```

Fit Full Model with Interactions and without Interactions

```r
#Fit Empty Model
empty=lm(BirthWeightOz~1,data=NCB)
summary(empty)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ 1, data = NCB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.441  -10.441    1.559   13.559   64.559
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 116.4407     0.5894   197.5 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 1408 degrees of freedom
```

```r
#Fit Full Model without Interactions
full = lm(BirthWeightOz~.,data=NCB)
summary(full)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ ., data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.925 -10.203  -0.528  10.315  53.461
##
## Coefficients:
##                     Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        -63.20152    7.27854  -8.683 < 0.0000000000000002 ***
```

```
## PluralTriplet      -35.07044    8.40430  -4.173         0.0000319283 ***
## PluralTwin         -25.90487    2.70451  -9.578 < 0.0000000000000002 ***
## SexMale              3.39238    0.88041   3.853             0.000122 ***
## MomAge               0.38579    0.08226   4.690         0.0000030041 ***
## Weeks                4.10762    0.17751  23.140 < 0.0000000000000002 ***
## MaritalNot Married  -1.83057    1.13739  -1.609             0.107745
## Gained               0.27735    0.03226   8.598 < 0.0000000000000002 ***
## SmokeYes            -7.29759    1.29603  -5.631         0.0000000217 ***
## MomRacehispanic      3.62009    1.63875   2.209             0.027332 *
## MomRaceother         0.94079    2.59643   0.362             0.717153
## MomRacewhite         3.97614    1.19060   3.340             0.000861 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.49 on 1397 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.4449
## F-statistic: 103.6 on 11 and 1397 DF,  p-value: < 0.00000000000000022
```

```
#plot(full,which=c(1,2)) #residuals look fine

#Fit Full Model with Interactions
fullinteract = lm(BirthWeightOz~.*.,data=NCB)
summary(fullinteract)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ . * ., data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.318 -10.170  -0.442  10.504  54.820
##
## Coefficients: (7 not defined because of singularities)
##                                      Estimate   Std. Error t value
## (Intercept)                       -145.4760123   41.8842103  -3.473
## PluralTriplet                     -126.1480732  210.5906029  -0.599
## PluralTwin                           0.1493945   28.3681659   0.005
## SexMale                             -5.2015824   15.4426034  -0.337
## MomAge                               0.2761758    1.2894832   0.214
## Weeks                                6.5122476    1.0687329   6.093
## MaritalNot Married                  -6.9768632   18.4587373  -0.378
## Gained                               2.7811908    0.5707663   4.873
## SmokeYes                            41.2468515   19.2885073   2.138
## MomRacehispanic                     96.0458626   31.2967202   3.069
## MomRaceother                        30.1042766   44.7697042   0.672
## MomRacewhite                         9.4723613   18.6932942   0.507
## PluralTriplet:SexMale               22.3189974   23.8062857   0.938
## PluralTwin:SexMale                   0.5443383    5.5819310   0.098
## PluralTriplet:MomAge                 2.8481053    4.9017044   0.581
## PluralTwin:MomAge                   -0.4036361    0.5716718  -0.706
## PluralTriplet:Weeks                 -0.7328360    2.6484242  -0.277
## PluralTwin:Weeks                     0.3173235    0.7453535   0.426
## PluralTriplet:MaritalNot Married            NA           NA      NA
## PluralTwin:MaritalNot Married       -7.3964752    8.1576639  -0.907
## PluralTriplet:Gained                        NA           NA      NA
```

```
## PluralTwin:Gained                    -0.3691367    0.1895640  -1.947
## PluralTriplet:SmokeYes                       NA           NA      NA
## PluralTwin:SmokeYes                   2.7159020   14.5823986   0.186
## PluralTriplet:MomRacehispanic                NA           NA      NA
## PluralTwin:MomRacehispanic          -18.3708485   14.5211544  -1.265
## PluralTriplet:MomRaceother                   NA           NA      NA
## PluralTwin:MomRaceother                      NA           NA      NA
## PluralTriplet:MomRacewhite                   NA           NA      NA
## PluralTwin:MomRacewhite             -14.6562017    8.8105122  -1.663
## SexMale:MomAge                       -0.0023346    0.1668647  -0.014
## SexMale:Weeks                         0.1880253    0.3765504   0.499
## SexMale:MaritalNot Married           -1.2937953    2.3201468  -0.558
## SexMale:Gained                        0.0366090    0.0655126   0.559
## SexMale:SmokeYes                     -0.6085578    2.6397098  -0.231
## SexMale:MomRacehispanic              -3.9278915    3.3260394  -1.181
## SexMale:MomRaceother                  0.0862587    5.2983121   0.016
## SexMale:MomRacewhite                  1.6412632    2.3884213   0.687
## MomAge:Weeks                         -0.0052575    0.0330533  -0.159
## MomAge:MaritalNot Married             0.1479168    0.1962651   0.754
## MomAge:Gained                        -0.0007407    0.0062411  -0.119
## MomAge:SmokeYes                      -0.4994723    0.2368523  -2.109
## MomAge:MomRacehispanic                0.3482324    0.3172115   1.098
## MomAge:MomRaceother                  -0.1999746    0.5447646  -0.367
## MomAge:MomRacewhite                   0.3950757    0.2268419   1.742
## Weeks:MaritalNot Married             -0.0165623    0.4508566  -0.037
## Weeks:Gained                         -0.0658038    0.0134289  -4.900
## Weeks:SmokeYes                       -0.7112224    0.4605171  -1.544
## Weeks:MomRacehispanic                -2.5973683    0.7536319  -3.446
## Weeks:MomRaceother                   -0.6612244    1.1230627  -0.589
## Weeks:MomRacewhite                   -0.5029288    0.4491329  -1.120
## MaritalNot Married:Gained             0.0261705    0.0793486   0.330
## MaritalNot Married:SmokeYes           1.2661802    3.1100218   0.407
## MaritalNot Married:MomRacehispanic    3.4203932    3.5470298   0.964
## MaritalNot Married:MomRaceother      -2.2329341    7.3838461  -0.302
## MaritalNot Married:MomRacewhite       1.5542637    2.9138718   0.533
## Gained:SmokeYes                      -0.1233010    0.0915112  -1.347
## Gained:MomRacehispanic                0.0214439    0.1264229   0.170
## Gained:MomRaceother                   0.1167465    0.2090528   0.558
## Gained:MomRacewhite                   0.0867782    0.0820437   1.058
## SmokeYes:MomRacehispanic            -27.7865067   17.2293561  -1.613
## SmokeYes:MomRaceother                -3.6803925    8.0956158  -0.455
## SmokeYes:MomRacewhite                -5.8226332    3.6049513  -1.615
##                                         Pr(>|t|)
## (Intercept)                             0.000530 ***
## PluralTriplet                           0.549260
## PluralTwin                              0.995799
## SexMale                                 0.736295
## MomAge                                  0.830442
## Weeks                             0.00000000144 ***
## MaritalNot Married                      0.705512
## Gained                            0.00000123052 ***
## SmokeYes                                0.032662 *
## MomRacehispanic                         0.002191 **
## MomRaceother                            0.501428
```

```
## MomRacewhite                            0.612430
## PluralTriplet:SexMale                   0.348656
## PluralTwin:SexMale                      0.922330
## PluralTriplet:MomAge                    0.561308
## PluralTwin:MomAge                       0.480271
## PluralTriplet:Weeks                     0.782048
## PluralTwin:Weeks                        0.670368
## PluralTriplet:MaritalNot Married              NA
## PluralTwin:MaritalNot Married           0.364732
## PluralTriplet:Gained                          NA
## PluralTwin:Gained                       0.051706 .
## PluralTriplet:SmokeYes                        NA
## PluralTwin:SmokeYes                     0.852280
## PluralTriplet:MomRacehispanic                 NA
## PluralTwin:MomRacehispanic              0.206050
## PluralTriplet:MomRaceother                    NA
## PluralTwin:MomRaceother                       NA
## PluralTriplet:MomRacewhite                    NA
## PluralTwin:MomRacewhite                 0.096446 .
## SexMale:MomAge                          0.988839
## SexMale:Weeks                           0.617624
## SexMale:MaritalNot Married              0.577186
## SexMale:Gained                          0.576385
## SexMale:SmokeYes                        0.817707
## SexMale:MomRacehispanic                 0.237830
## SexMale:MomRaceother                    0.987013
## SexMale:MomRacewhite                    0.492090
## MomAge:Weeks                            0.873644
## MomAge:MaritalNot Married               0.451186
## MomAge:Gained                           0.905545
## MomAge:SmokeYes                         0.035146 *
## MomAge:MomRacehispanic                  0.272490
## MomAge:MomRaceother                     0.713613
## MomAge:MomRacewhite                     0.081800 .
## Weeks:MaritalNot Married                0.970702
## Weeks:Gained                       0.00000107319 ***
## Weeks:SmokeYes                          0.122725
## Weeks:MomRacehispanic                   0.000585 ***
## Weeks:MomRaceother                      0.556115
## Weeks:MomRacewhite                      0.263007
## MaritalNot Married:Gained               0.741589
## MaritalNot Married:SmokeYes             0.683978
## MaritalNot Married:MomRacehispanic      0.335069
## MaritalNot Married:MomRaceother         0.762388
## MaritalNot Married:MomRacewhite         0.593843
## Gained:SmokeYes                         0.178081
## Gained:MomRacehispanic                  0.865334
## Gained:MomRaceother                     0.576627
## Gained:MomRacewhite                     0.290378
## SmokeYes:MomRacehispanic                0.107034
## SmokeYes:MomRaceother                   0.649459
## SmokeYes:MomRacewhite                   0.106506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 16.35 on 1353 degrees of freedom
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4537
## F-statistic: 22.26 on 55 and 1353 DF,  p-value: < 0.00000000000000022
```

```
#plot(fullinteract,which=c(1,2)) #residuals look fine

#Nested F-test to Compare Models (Increasing Complexity and Nested)
anova(empty,full, fullinteract)
```

```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ 1
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital + Gained +
##     Smoke + MomRace
## Model 3: BirthWeightOz ~ (Plural + Sex + MomAge + Weeks + Marital + Gained +
##     Smoke + MomRace) * (Plural + Sex + MomAge + Weeks + Marital +
##     Gained + Smoke + MomRace)
##   Res.Df    RSS Df Sum of Sq        F               Pr(>F)
## 1   1408 689271
## 2   1397 379647 11    309624 105.2599 < 0.0000000000000002 ***
## 3   1353 361807 44     17840   1.5162              0.01707 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Different Subsets of Variables

```
#Model without Mom's Information
nomom = lm(BirthWeightOz~.-MomAge-Marital-Smoke-MomRace,data=NCB)
summary(nomom)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ . - MomAge - Marital - Smoke - MomRace,
##     data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.581 -11.226  -0.128  11.205  50.877
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   -59.24424    7.12364  -8.317 < 0.0000000000000002 ***
## PluralTriplet -28.23082    8.68682  -3.250             0.001182 **
## PluralTwin    -23.13489    2.78823  -8.297 0.00000000000000247 ***
## SexMale         3.30266    0.91244   3.620             0.000306 ***
## Weeks           4.29905    0.18283  23.514 < 0.0000000000000002 ***
## Gained          0.28205    0.03314   8.511 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.1 on 1403 degrees of freedom
## Multiple R-squared:  0.4051, Adjusted R-squared:  0.403
## F-statistic: 191.1 on 5 and 1403 DF,  p-value: < 0.00000000000000022
```

```r
#ANOVA Tables of full and nomom Models
anova455(full)
```

```
## ANOVA Table
## Model: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital + Gained + Smoke + MomRace
##
##         Df Sum Sq Mean Sq F value                 P(>F)
## Model   11 309624 28147.7  103.58 < 0.00000000000000022 ***
## Error 1397 379647   271.8
## Total 1408 689271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova455(nomom)
```

```
## ANOVA Table
## Model: BirthWeightOz ~ (Plural + Sex + MomAge + Weeks + Marital + Gained + Smoke + MomRace) - MomAge
##
##         Df Sum Sq Mean Sq F value                 P(>F)
## Model    5 279217   55843  191.07 < 0.00000000000000022 ***
## Error 1403 410054     292
## Total 1408 689271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Hand Calculation of Nested F-Test Statistic
SSModelfull = 309624
SSModelreduced = 279217
nsubset = length(coef(full)) - length(coef(nomom))

Fstat = ((SSModelfull-SSModelreduced)/nsubset)/(summary(full)$sigma^2)
Fstat
```

```
## [1] 18.64827
```

```r
#Is any of Mom's information useful?
anova(nomom,full)
```

```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ (Plural + Sex + MomAge + Weeks + Marital + Gained +
##     Smoke + MomRace) - MomAge - Marital - Smoke - MomRace
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital + Gained +
##     Smoke + MomRace
##   Res.Df    RSS Df Sum of Sq      F                 Pr(>F)
## 1   1403 410054
## 2   1397 379647  6     30407 18.648 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which Model is Best?

```r
#empty Model RMSE
sqrt(sum((NCB$BirthWeightOz - fitted(empty))^2)/(nrow(NCB)-1))
```

```
## [1] 22.12553
```

```
#nomom Model RMSE
summary(nomom)$sigma
```

```
## [1] 17.09589
```

```
#full Model RMSE
summary(full)$sigma
```

```
## [1] 16.48511
```

```
#fullinteract Model RMSE
summary(fullinteract)$sigma
```

```
## [1] 16.3527
```

Use Cross Validation to Determine Which Model is Best

```
#Randomly Choose Rows to Be in Training Data
set.seed(216)
train.select = sample(1:1409,size = floor(0.8*1409),replace=FALSE)

#Split Data Up into Train and Test Sets
TRAIN = NCB[train.select,]
TEST = NCB[-train.select,]

#Check Representation
table(TRAIN$Plural)
```

```
##
##  Single Triplet    Twin
##    1091       4      32
```

```
table(TRAIN$Sex)
```

```
##
## Female   Male
##    565    562
```

```
table(TRAIN$Marital)
```

```
##
##     Married Not Married
##         728         399
```

```
table(TRAIN$Smoke)
```

```
##
##  No Yes
## 958 169
```

```
table(TRAIN$MomRace)
```

```
##
##    black hispanic    other    white
##      264      116       38      709
```

```
#Refit All Models to TRAIN
empty=lm(BirthWeightOz~1,data=TRAIN)
nomom = lm(BirthWeightOz~.-MomAge-Marital-Smoke-MomRace,data=TRAIN)
```

```
full=lm(BirthWeightOz~.,data=TRAIN)
fullinteract=lm(BirthWeightOz~.*.,data=TRAIN)

#Calculate In-Sample RMSE for all 4 models
sqrt(mean((TRAIN$BirthWeightOz - predict(empty))^2))
```

```
## [1] 22.09607
```

```
sqrt(mean((TRAIN$BirthWeightOz - predict(nomom))^2))
```

```
## [1] 17.11258
```

```
sqrt(mean((TRAIN$BirthWeightOz - predict(full))^2))
```

```
## [1] 16.44581
```

```
sqrt(mean((TRAIN$BirthWeightOz - predict(fullinteract))^2))              #BEST
```

```
## [1] 15.96207
```

```
#Calculate Out-of-Sample RMSE for all 4 models
sqrt(mean((TEST$BirthWeightOz - predict(empty,newdata=TEST))^2))
```

```
## [1] 22.21974
```

```
sqrt(mean((TEST$BirthWeightOz - predict(nomom,newdata=TEST))^2))
```

```
## [1] 16.89467
```

```
sqrt(mean((TEST$BirthWeightOz - predict(full,newdata=TEST))^2))          #BEST
```

```
## [1] 16.35943
```

```
sqrt(mean((TEST$BirthWeightOz - predict(fullinteract,newdata=TEST))^2))
```

```
## Warning in predict.lm(fullinteract, newdata = TEST): prediction from a
## rank-deficient fit may be misleading
```

```
## [1] 16.53506
```