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)
```

Clean Data

```
NCB$Plural = COMPLETE
table(NCB$Plural)

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

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

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

NCB=na.omit(NCB)

str(NCB)
```

Fit Full Model with Interactions and without Interactions

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

#Fit Full Model without Interactions
full = lm(BirthWeightOz~.,data=NCB)
summary(full)
#plot(full,which=c(1,2)) #residuals look fine

#Fit Full Model with Interactions
fullinteract = lm(BirthWeightOz~.*.,data=NCB)
summary(fullinteract)
#plot(fullinteract,which=c(1,2)) #residuals look fine

#Nested F-test to Compare Models (Increasing Complexity and Nested)
```

```
COMPLETE
```

Test Different Subsets of Variables

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

#ANOVA Tables of full and nomom Models
COMPLETE


#Hand Calculation of Nested F-Test Statistic
SSModelfull = COMPLETE
SSModelreduced = COMPLETE
nsubset = COMPLETE

Fstat = COMPLETE
Fstat

#Is any of Mom's information useful?
COMPLETE
```

Which Model is Best?

```
#empty Model RMSE
COMPLETE

#nomom Model RMSE
summary(nomom)$sigma

#full Model RMSE
summary(full)$sigma

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

Use Cross Validation to Determine Which Model is Best

```
#Randomly Choose Rows to Be in Training Data
set.seed(216)
train.select = COMPLETE

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

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

```
table(TRAIN$Marital)
table(TRAIN$Smoke)
table(TRAIN$MomRace)

#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
COMPLETE

#Calculate Out-of-Sample RMSE for all 4 models
COMPLETE
```