Supplement for Lecture 16: Techniques for Choosing Predictors

Load Data

```
data("BodyFat") # Load Data

bf = BodyFat

head(bf)
```

```
##   Bodyfat Age Weight Height Neck Chest Abdomen Ankle Biceps Wrist
## 1    32.3  41 247.25  73.50 42.1 117.0   115.6  26.3   37.3  19.7
## 2    22.5  31 177.25  71.50 36.2 101.1    92.4  24.6   30.1  18.2
## 3    22.0  42 156.25  69.00 35.5  97.8    86.0  24.0   31.2  17.4
## 4    12.3  23 154.25  67.75 36.2  93.1    85.2  21.9   32.0  17.1
## 5    20.5  46 177.00  70.00 37.2  99.7    95.6  22.5   29.1  17.7
## 6    22.6  54 198.00  72.00 39.9 107.6   100.0  22.0   35.9  18.9
```

Check for Multicollinearity

```
# Correlation Matrix from Base R


# Tile Plot of Correlation Matrix (Correlogram) from corrplot package


#Scatterplots of Bodyfat Variable with Each Other Predictor


# Cool Visual from PerformanceAnalytics package
```

Variance Inflation Factor

```
mod.full = lm(COMPLETE,data=bf)
summary(mod.full)
plot(mod.full)
vif(mod.full) # From car package
```

```
mod.noWeight = lm(COMPLETE,data=bf)
vif(mod.noWeight)

mod.noWeightChest = lm(COMPLETE,data=bf)
vif(mod.noWeightChest)
```

Fit All Subsets

The `regsubsets()` function fits all subset models up to a maximum number of variables. Notice the `nvmax` argument. The asterisk indicates which variables are included in the best model for each possible choice for k.

```
all =  #From leaps package
summary(all)

all2 =  #From leaps package
summary(all2)
```

Now we identify the "best" model based off the criteria R-Squared, adjusted R-Squared, and Mallow's Cp. We can use the `ShowSubsets()` function created by Dr. McLean.

```
# Best Model According to R-Squared
out2[COMPLETE,]

# Best Model According to Adjusted R-Squared
out2[COMPLETE,]

# Best Model According to Mallows Cp
out2[COMPLETE,]
```

We can also tell the function using `nbest` the number of top models for each choice of k that we want to see in the output. We can also calculate the BIC for each of the models and identify the best model according to BIC.

```
all3 = COMPLETE
out3 = ShowSubsets(all3)
out3

#Get BIC for each of the models
summ.all3 = COMPLETE
summ.all3$bic

#Find Best Model According to BIC
out3[which.min(summ.all3$bic),]

#Calculated adjusted R-squared by hand
1-(1-0.7471)*((100-1)/(100-3-1)) #Notice these equals the adjusted R-Squared in the BIC Model
```

Now we fit our "best" models according to adjusted R-squared/Mallow's Cp and BIC.

```
mod.rsqmallow = lm(Bodyfat ~ Age + Weight + Abdomen + Wrist, data=bf)
plot(mod.rsqmallow)
vif(mod.rsqmallow)

mod.bic = lm(Bodyfat ~ Weight + Abdomen + Wrist,data=bf)
plot(mod.bic)
vif(mod.bic)
```

Backwards, Forwards, and Stepwise Algorithms

Built-in `step()` function doesn't use p-values to determine what variables to remove or keep. It uses the AIC measurement which is similar to BIC. Let's first look at backward's elimination.

```
summary(back.out)
vif(back.out)
```

Now let's look at forward selection. In this case, we need to start by initiating the empty model and telling the `step()` function to consider all models up to possibly the full model.

```
summary(forward.out)
vif(forward.out)
```

Now, we specify the `direction="both"` to conduct stepwise regression where variables can be both added and removed.

```
summary(step.out)
vif(step.out)
```

All Models Summarized

```
mod.full
mod.noWeight
mod.noWeightChest
mod.rsqmallow
mod.bic
back.out
forward.out
step.out
```