Breakout Session 7: Model Selection

We will go back to using the bike data set (refer to Breakout Session 3 for information on the variables in the bike data set). We will work through several different concepts and possibilities within this data set (the data is located on the GitHub). First, let's naively try an automatic selection with the variables in the data set (remove the dteday variable for any analysis). We are trying to model the cnt of total rental bikes. First, we need to create a training and a test data set. We will use the seed of 18954 and partition the data into a 70/30 split.

```
bike=read.csv("https://raw.githubusercontent.com/IAA-Faculty/statistical_foundations/master/bike.csv")
bike<- bike %>% dplyr::select (-dteday) %>% mutate(id = row_number())
set.seed(18954)
train=bike %>% sample_frac(0.7)
test= anti_join(bike,train,by='id')
```

1. Using the AIC criterion, do a forward selection on the training data set (with the data set as is…without the id variable). What happens?

```
lm.0=lm(cnt~1,data=bike)
lm.full=lm(cnt~.-id,data=bike)
step(lm.0,scope=list(lower=lm.0,upper=lm.full),direction = "forward")
```

This will give an error because you have two variables that perfectly describe the response (casual and registered…these two variables added together give the cnt variable).

2. Now also remove the casual and registered variables and try forward, backward and stepwise with AIC and BIC on the training data set. Comment on the similarities and differences you observe.

```
lm.full=lm(cnt~.-id-casual-registered,data=bike)
```

AIC Forward

```
step(lm.0,scope=list(lower=lm.0,upper=lm.full),direction="forward")
```

cnt ~ temp + hr + yr + hum + season + atemp + holiday + windspeed + weekday + weathersit + workingday

Backward

cnt ~ season + yr + hr + holiday + weekday + workingday + weathersit + temp + atemp + hum + windspeed

Stepwise

cnt ~ temp + hr + yr + hum + season + atemp + holiday + windspeed + weekday + weathersit + workingday

Fairly consistent variables across all models. Notice that forward and backwards both have temp and atemp in them (this is NOT good since we KNOW that these two variables are correlated!!). Another interesting tidbit is all of them have workingday and weekday (which are also both VERY correlated!!).

BIC Forward
cnt ~ temp + hr + yr + hum + season + atemp + holiday + windspeed + weekday

Backward
cnt ~ season + yr + hr + holiday + weekday + atemp + hum + windspeed

Stepwise
cnt ~ hr + yr + hum + season + atemp + holiday + windspeed + weekday

Using BIC has LESS variables in all directions (which we knew that it would). Only forward has both temp and atemp. Working day does not show up here. There are some issues with types of variables and multicollinearity that we have not gotten to yet…building up to it.

3.