Analytics Foundations: Breakout Session 11

Today's dataset comes from a bike sharing company (Capital Bike Share). Each *hour*, the number of riders (**cnt**) is given, along with various other attributes as shown in the table below:

cnt	Count of total rental bikes including both casual and registered
dteday	Date
instant	Record index (ID)
season	Season (1:winter, 2:spring, 3:summer, 4:fall)
yr	Year (0:2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	Whether day is holiday or not
weekday	Day of the week
workingday	If day is neither weekend nor holiday is 1, otherwise is 0
weathersit	1: Clear, few clouds, partly cloudy, partly cloudy 2: Mist + cloudy, mist + broken clouds, Mist + few clouds, Mist 3: Light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds 4: Heavy rain + ice pallets + thunderstorm + mist, snow + fog
temp	Normalized temperature in Celsius. Values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. Values are divided to 50 (max)
hum	Normalized humidity. Values are divided to 100 (max)
windspeed	Normalized wind speed. Values are divided to 67 (max)
casual	Count of casual users
registered	Count of registered users

Source: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

Data:

You can obtain the dataset by running the following code:

```
bike <- read.csv('https://raw.githubusercontent.com/IAA-Faculty/statistical_foundations/master/bike.csv')
```

Questions:

1. You want to build a couple of different models and see which one is better. We will learn in later lectures how to do this with test datasets, but for right now we will only do this with training data. First, we need to split the data into training and test. Run the following code to get the training and test split:

```
set.seed(123)

bike <- bike %>% mutate(id = row_number())

train <- bike %>% sample_frac(0.7)

test <- anti_join(bike, train, by = 'id')
```

   What is the observation count in each dataset (train and test)?

2. You can't decide which variable is better to predict number of users (**cnt**), actual temperature (**temp**) or what the temperature feels like (**atemp**). You know they are highly correlated (correlation of approximately 0.99), but you feel that each might still provide some valuable information. Maybe the temperature provides (ever so slightly) different information than the feeling temperature. To keep them both in the model, build a ridge regression that has four variables: actual temperature (**temp**), feeling temperature (**atemp**), humidity (**hum**) and wind speed (**windspeed**). Use CV to find the penalty that minimizes the MSE. What is that lambda value?

3. Get predictions for this ridge regression on the test dataset. What are the first 5 predictions from the test dataset from this model? What are the respective actual values of **cnt** from this dataset?

4. A MLR with only feeling temperature, humidity, and windspeed had a test dataset MAPE of 478%. What was the MAPE from your ridge regression? From this comparison, which model would you choose?