# Model Validation Assignment

## BAN502 Predictive Analytics

### Susan Wiggins

The libraries needed:

library(tidyverse)

## -- Attaching packages ------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.1 v dplyr 0.8.1  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ---------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

Import of dataset:

library(readr)  
hour <- read\_csv("hour.csv")

## Parsed with column specification:  
## cols(  
## instant = col\_double(),  
## dteday = col\_date(format = ""),  
## season = col\_double(),  
## yr = col\_double(),  
## mnth = col\_double(),  
## hr = col\_double(),  
## holiday = col\_double(),  
## weekday = col\_double(),  
## workingday = col\_double(),  
## weathersit = col\_double(),  
## temp = col\_double(),  
## atemp = col\_double(),  
## hum = col\_double(),  
## windspeed = col\_double(),  
## casual = col\_double(),  
## registered = col\_double(),  
## count = col\_double()  
## )

View(hour)

Formation of tibble data frame bike:

bike = as\_tibble(hour)

Using the same code in past assignment:

bike = bike %>% mutate(season = as\_factor(as.character(season))) %>% mutate(season = fct\_recode(season, "Spring" = "1", "Summer" = "2", "Fall" = "3", "Winter" = "4"))  
  
bike = bike %>% mutate(yr = as\_factor(as.character(yr)))  
bike = bike %>% mutate(mnth = as\_factor(as.character(mnth)))  
bike = bike %>% mutate(hr = as\_factor(as.character(hr)))  
  
bike = bike %>% mutate(holiday = as\_factor(as.character(holiday))) %>% mutate(holiday = fct\_recode(holiday, "NotHoliday" = "0", "Holiday" = "1"))  
  
bike = bike %>% mutate(workingday = as\_factor(as.character(workingday))) %>% mutate(workingday = fct\_recode(workingday, "NotWorkingDay" = "0", "WorkingDay" = "1"))  
  
bike = bike %>% mutate(weathersit = as\_factor(as.character(weathersit))) %>% mutate(weathersit = fct\_recode(weathersit, "NoPrecip" = "1", "Misty" = "2", "LightPrecip" = "3", "HeavyPrecip" = "4"))  
  
  
bike = bike %>% mutate(weekday = as\_factor(as.character(weekday))) %>% mutate(weekday = fct\_recode(weekday, "Sunday" = "0", "Monday" = "1", "Tuesday" = "2", "Wednesday" = "3", "Thursday" = "4", "Friday" = "5", "Saturday" = "6"))  
  
  
  
glimpse(bike)

## Observations: 17,379  
## Variables: 17  
## $ instant <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...  
## $ dteday <date> 2011-01-01, 2011-01-01, 2011-01-01, 2011-01-01, 20...  
## $ season <fct> Spring, Spring, Spring, Spring, Spring, Spring, Spr...  
## $ yr <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ mnth <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ hr <fct> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ holiday <fct> NotHoliday, NotHoliday, NotHoliday, NotHoliday, Not...  
## $ weekday <fct> Saturday, Saturday, Saturday, Saturday, Saturday, S...  
## $ workingday <fct> NotWorkingDay, NotWorkingDay, NotWorkingDay, NotWor...  
## $ weathersit <fct> NoPrecip, NoPrecip, NoPrecip, NoPrecip, NoPrecip, M...  
## $ temp <dbl> 0.24, 0.22, 0.22, 0.24, 0.24, 0.24, 0.22, 0.20, 0.2...  
## $ atemp <dbl> 0.2879, 0.2727, 0.2727, 0.2879, 0.2879, 0.2576, 0.2...  
## $ hum <dbl> 0.81, 0.80, 0.80, 0.75, 0.75, 0.75, 0.80, 0.86, 0.7...  
## $ windspeed <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0896, 0.0...  
## $ casual <dbl> 3, 8, 5, 3, 0, 0, 2, 1, 1, 8, 12, 26, 29, 47, 35, 4...  
## $ registered <dbl> 13, 32, 27, 10, 1, 1, 0, 2, 7, 6, 24, 30, 55, 47, 7...  
## $ count <dbl> 16, 40, 32, 13, 1, 1, 2, 3, 8, 14, 36, 56, 84, 94, ...

### Task 1

Split the data into training and testing sets.

set.seed(1234)  
train.rows = createDataPartition(y = bike$count, p=0.7, list = FALSE) #70% in training  
train = bike[train.rows,]   
test = bike[-train.rows,]

### Task 2

How many rows of data are in each set ( training and testing)?

The training set has 12167 rows and 17 variables(columns) of data. The testing set has 5212 rows abd 17 variables(columns) of data.

bike = bike %>% dplyr::select("season", "mnth", "hr", "holiday", "weekday", "temp","weathersit", "count")  
  
#set.seed(1234)  
#train.rows = createDataPartition(y = bike2$count, p=0.7, list = FALSE) #70% in training  
#train2 = bike2[train.rows,]   
#test2 = bike2[-train.rows,]  
  
summary(bike)

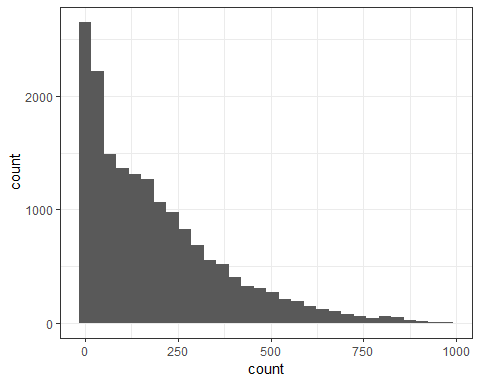
## season mnth hr holiday   
## Spring:4242 5 :1488 16 : 730 NotHoliday:16879   
## Summer:4409 7 :1488 17 : 730 Holiday : 500   
## Fall :4496 12 :1483 13 : 729   
## Winter:4232 8 :1475 14 : 729   
## 3 :1473 15 : 729   
## 10 :1451 12 : 728   
## (Other):8521 (Other):13004   
## weekday temp weathersit count   
## Saturday :2512 Min. :0.020 NoPrecip :11413 Min. : 1.0   
## Sunday :2502 1st Qu.:0.340 Misty : 4544 1st Qu.: 40.0   
## Monday :2479 Median :0.500 LightPrecip: 1419 Median :142.0   
## Tuesday :2453 Mean :0.497 HeavyPrecip: 3 Mean :189.5   
## Wednesday:2475 3rd Qu.:0.660 3rd Qu.:281.0   
## Thursday :2471 Max. :1.000 Max. :977.0   
## Friday :2487

glimpse(bike)

## Observations: 17,379  
## Variables: 8  
## $ season <fct> Spring, Spring, Spring, Spring, Spring, Spring, Spr...  
## $ mnth <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ hr <fct> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ holiday <fct> NotHoliday, NotHoliday, NotHoliday, NotHoliday, Not...  
## $ weekday <fct> Saturday, Saturday, Saturday, Saturday, Saturday, S...  
## $ temp <dbl> 0.24, 0.22, 0.22, 0.24, 0.24, 0.24, 0.22, 0.20, 0.2...  
## $ weathersit <fct> NoPrecip, NoPrecip, NoPrecip, NoPrecip, NoPrecip, M...  
## $ count <dbl> 16, 40, 32, 13, 1, 1, 2, 3, 8, 14, 36, 56, 84, 94, ...

ggplot(bike, aes(x=count)) + geom\_histogram() + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Task 3

AFTER you split, then do visualization and modeling with the **training set**.

Our Y (response) variable in this dataset is “Count”.

model1 = lm(count ~.,train)  
summary(model1)

##   
## Call:  
## lm(formula = count ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.136e-11 -4.300e-14 0.000e+00 5.000e-14 3.396e-11   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.593e-09 2.973e-10 -5.360e+00 8.48e-08 \*\*\*  
## instant -4.337e-15 8.321e-16 -5.213e+00 1.89e-07 \*\*\*  
## dteday 1.064e-13 1.985e-14 5.358e+00 8.58e-08 \*\*\*  
## seasonSummer -4.868e-13 5.835e-14 -8.343e+00 < 2e-16 \*\*\*  
## seasonFall 3.451e-13 6.886e-14 5.011e+00 5.48e-07 \*\*\*  
## seasonWinter 6.781e-14 5.922e-14 1.145e+00 0.252186   
## yr1 -1.741e-13 3.861e-13 -4.510e-01 0.651994   
## mnth2 -3.830e-13 5.788e-14 -6.617e+00 3.82e-11 \*\*\*  
## mnth3 1.592e-13 8.319e-14 1.914e+00 0.055607 .   
## mnth4 6.575e-15 1.222e-13 5.400e-02 0.957092   
## mnth5 -1.202e-13 1.501e-13 -8.010e-01 0.423079   
## mnth6 -2.407e-13 1.781e-13 -1.352e+00 0.176484   
## mnth7 -4.120e-13 2.107e-13 -1.956e+00 0.050537 .   
## mnth8 -2.651e-13 2.394e-13 -1.108e+00 0.268048   
## mnth9 2.022e-14 2.670e-13 7.600e-02 0.939623   
## mnth10 -1.834e-13 2.980e-13 -6.150e-01 0.538289   
## mnth11 -6.587e-14 3.295e-13 -2.000e-01 0.841563   
## mnth12 -1.814e-13 3.575e-13 -5.070e-01 0.611838   
## hr1 6.358e-13 6.389e-14 9.951e+00 < 2e-16 \*\*\*  
## hr2 -6.934e-13 6.427e-14 -1.079e+01 < 2e-16 \*\*\*  
## hr3 1.002e-12 6.505e-14 1.540e+01 < 2e-16 \*\*\*  
## hr4 -9.678e-13 6.562e-14 -1.475e+01 < 2e-16 \*\*\*  
## hr5 3.161e-13 6.502e-14 4.861e+00 1.18e-06 \*\*\*  
## hr6 -5.534e-13 6.522e-14 -8.485e+00 < 2e-16 \*\*\*  
## hr7 -5.851e-13 6.683e-14 -8.755e+00 < 2e-16 \*\*\*  
## hr8 -1.053e-13 7.310e-14 -1.441e+00 0.149684   
## hr9 -1.395e-13 6.672e-14 -2.091e+00 0.036529 \*   
## hr10 -8.705e-14 6.610e-14 -1.317e+00 0.187888   
## hr11 1.389e-13 6.724e-14 2.066e+00 0.038811 \*   
## hr12 3.199e-13 6.839e-14 4.678e+00 2.93e-06 \*\*\*  
## hr13 2.723e-13 6.813e-14 3.997e+00 6.45e-05 \*\*\*  
## hr14 4.623e-13 6.953e-14 6.650e+00 3.06e-11 \*\*\*  
## hr15 2.874e-13 6.943e-14 4.139e+00 3.51e-05 \*\*\*  
## hr16 4.037e-13 7.108e-14 5.679e+00 1.38e-08 \*\*\*  
## hr17 1.857e-12 7.542e-14 2.462e+01 < 2e-16 \*\*\*  
## hr18 -1.327e-12 7.603e-14 -1.745e+01 < 2e-16 \*\*\*  
## hr19 -4.261e-13 7.008e-14 -6.080e+00 1.24e-09 \*\*\*  
## hr20 -5.617e-13 6.804e-14 -8.255e+00 < 2e-16 \*\*\*  
## hr21 -6.222e-13 6.708e-14 -9.276e+00 < 2e-16 \*\*\*  
## hr22 -2.222e-13 6.705e-14 -3.314e+00 0.000921 \*\*\*  
## hr23 -1.787e-13 6.691e-14 -2.670e+00 0.007593 \*\*   
## holidayHoliday 2.656e-13 6.040e-14 4.398e+00 1.10e-05 \*\*\*  
## weekdaySunday -2.780e-15 3.456e-14 -8.000e-02 0.935889   
## weekdayMonday -4.787e-13 3.806e-14 -1.258e+01 < 2e-16 \*\*\*  
## weekdayTuesday -5.330e-13 3.766e-14 -1.415e+01 < 2e-16 \*\*\*  
## weekdayWednesday -5.270e-13 3.765e-14 -1.400e+01 < 2e-16 \*\*\*  
## weekdayThursday -4.708e-13 3.759e-14 -1.252e+01 < 2e-16 \*\*\*  
## weekdayFriday -3.343e-13 3.657e-14 -9.142e+00 < 2e-16 \*\*\*  
## workingdayWorkingDay NA NA NA NA   
## weathersitMisty -6.353e-14 2.301e-14 -2.761e+00 0.005764 \*\*   
## weathersitLightPrecip -3.304e-13 3.885e-14 -8.503e+00 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -1.487e-13 1.020e-12 -1.460e-01 0.884111   
## temp 2.020e-12 3.436e-13 5.880e+00 4.21e-09 \*\*\*  
## atemp -9.756e-13 3.550e-13 -2.748e+00 0.006002 \*\*   
## hum -3.120e-13 6.714e-14 -4.647e+00 3.41e-06 \*\*\*  
## windspeed -3.788e-13 8.386e-14 -4.517e+00 6.32e-06 \*\*\*  
## casual 1.000e+00 3.130e-16 3.195e+15 < 2e-16 \*\*\*  
## registered 1.000e+00 1.174e-16 8.516e+15 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.018e-12 on 12110 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 6.831e+30 on 56 and 12110 DF, p-value: < 2.2e-16

model2 = lm(count~., test)  
summary(model2)

##   
## Call:  
## lm(formula = count ~ ., data = test)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.588e-11 -6.600e-14 -2.000e-15 6.500e-14 2.320e-11   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.188e-09 3.728e-10 3.187e+00 0.001449 \*\*   
## instant 3.328e-15 1.041e-15 3.196e+00 0.001402 \*\*   
## dteday -7.939e-14 2.489e-14 -3.190e+00 0.001433 \*\*   
## seasonSummer 4.414e-13 7.298e-14 6.048e+00 1.57e-09 \*\*\*  
## seasonFall 4.177e-13 8.666e-14 4.820e+00 1.47e-06 \*\*\*  
## seasonWinter 7.793e-13 7.315e-14 1.065e+01 < 2e-16 \*\*\*  
## yr1 2.208e-13 4.890e-13 4.520e-01 0.651612   
## mnth2 2.791e-13 7.481e-14 3.731e+00 0.000193 \*\*\*  
## mnth3 5.168e-13 1.066e-13 4.847e+00 1.29e-06 \*\*\*  
## mnth4 3.736e-13 1.567e-13 2.385e+00 0.017126 \*   
## mnth5 6.930e-13 1.906e-13 3.635e+00 0.000280 \*\*\*  
## mnth6 7.121e-13 2.267e-13 3.141e+00 0.001692 \*\*   
## mnth7 8.587e-13 2.699e-13 3.181e+00 0.001475 \*\*   
## mnth8 8.256e-13 3.046e-13 2.710e+00 0.006752 \*\*   
## mnth9 6.764e-13 3.395e-13 1.992e+00 0.046374 \*   
## mnth10 2.820e-13 3.785e-13 7.450e-01 0.456281   
## mnth11 2.563e-13 4.189e-13 6.120e-01 0.540573   
## mnth12 3.328e-13 4.541e-13 7.330e-01 0.463661   
## hr1 2.962e-13 8.095e-14 3.659e+00 0.000255 \*\*\*  
## hr2 5.693e-14 8.073e-14 7.050e-01 0.480670   
## hr3 6.247e-13 8.050e-14 7.761e+00 1.01e-14 \*\*\*  
## hr4 4.712e-13 7.953e-14 5.924e+00 3.34e-09 \*\*\*  
## hr5 7.850e-13 7.931e-14 9.897e+00 < 2e-16 \*\*\*  
## hr6 4.151e-13 7.871e-14 5.274e+00 1.39e-07 \*\*\*  
## hr7 6.911e-13 8.464e-14 8.165e+00 3.99e-16 \*\*\*  
## hr8 1.476e-12 8.869e-14 1.665e+01 < 2e-16 \*\*\*  
## hr9 4.563e-13 8.154e-14 5.596e+00 2.31e-08 \*\*\*  
## hr10 8.990e-14 8.075e-14 1.113e+00 0.265596   
## hr11 8.126e-14 8.192e-14 9.920e-01 0.321269   
## hr12 3.559e-14 8.394e-14 4.240e-01 0.671592   
## hr13 4.590e-14 8.720e-14 5.260e-01 0.598687   
## hr14 -8.143e-14 8.463e-14 -9.620e-01 0.336025   
## hr15 1.463e-14 8.625e-14 1.700e-01 0.865266   
## hr16 3.541e-13 8.573e-14 4.130e+00 3.69e-05 \*\*\*  
## hr17 1.328e-13 9.557e-14 1.389e+00 0.164820   
## hr18 7.527e-13 8.885e-14 8.472e+00 < 2e-16 \*\*\*  
## hr19 9.864e-13 8.752e-14 1.127e+01 < 2e-16 \*\*\*  
## hr20 6.962e-13 8.472e-14 8.218e+00 2.60e-16 \*\*\*  
## hr21 4.491e-13 8.394e-14 5.350e+00 9.17e-08 \*\*\*  
## hr22 2.968e-13 8.282e-14 3.584e+00 0.000341 \*\*\*  
## hr23 4.374e-13 8.298e-14 5.271e+00 1.41e-07 \*\*\*  
## holidayHoliday -4.151e-13 7.161e-14 -5.797e+00 7.15e-09 \*\*\*  
## weekdaySunday 2.069e-13 4.317e-14 4.791e+00 1.70e-06 \*\*\*  
## weekdayMonday 7.040e-13 4.756e-14 1.480e+01 < 2e-16 \*\*\*  
## weekdayTuesday 7.830e-13 4.744e-14 1.650e+01 < 2e-16 \*\*\*  
## weekdayWednesday 8.776e-13 4.747e-14 1.849e+01 < 2e-16 \*\*\*  
## weekdayThursday 6.657e-13 4.746e-14 1.403e+01 < 2e-16 \*\*\*  
## weekdayFriday 6.253e-13 4.603e-14 1.359e+01 < 2e-16 \*\*\*  
## workingdayWorkingDay NA NA NA NA   
## weathersitMisty 1.074e-13 2.889e-14 3.718e+00 0.000203 \*\*\*  
## weathersitLightPrecip 6.108e-14 5.034e-14 1.213e+00 0.225121   
## weathersitHeavyPrecip -1.510e-12 5.952e-13 -2.536e+00 0.011226 \*   
## temp -1.256e-12 4.828e-13 -2.602e+00 0.009304 \*\*   
## atemp -5.803e-13 5.023e-13 -1.155e+00 0.248031   
## hum -4.866e-13 8.474e-14 -5.742e+00 9.89e-09 \*\*\*  
## windspeed 1.398e-13 1.083e-13 1.290e+00 0.197011   
## casual 1.000e+00 4.085e-16 2.448e+15 < 2e-16 \*\*\*  
## registered 1.000e+00 1.466e-16 6.821e+15 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.354e-13 on 5155 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 4.49e+30 on 56 and 5155 DF, p-value: < 2.2e-16

### Task 3

Commit on the quality of the model. Be sure to note the Adjusted R-Squared value.

The linear regression model on the training data implies that the variables for month and weekday do not have significance when looking at the p-value due to being greater than .05, Sunday is the only value in weekday that shows a significance for bike rentals. As for the other variables tested in the training method the weathersit, temp, hr, and season all show significance due to having p-value less than .05. The Adjusted R-squared value is good at .6214 indicating that our model is sufficient.

### Task 4

Use the predict functions to make predictions:

predict\_train = predict(model1, newdata = train)

## Warning in predict.lm(model1, newdata = train): prediction from a rank-  
## deficient fit may be misleading

head(predict\_train, n=6)

## 1 2 3 4 5 6   
## 40 32 13 1 1 2

Comment on the predictions:

The training data tends to show lower values in prediction, with negative tendencies. The lowest value for prediction is -57.81925 and the highest value listed is 13.80902. The lower and upper prediction is a fairly large range in data points between the two at around the -270’s to +160’s.

### Task 5

predict\_test = predict(model2, newdata = test)

## Warning in predict.lm(model2, newdata = test): prediction from a rank-  
## deficient fit may be misleading

head(predict\_test, n=6)

## 1 2 3 4 5 6   
## 16 14 56 84 94 17

Comment on the predictions:

The testing data tends to show one lower values in prediction, and the other values are much greater in value than the training prediction. The lowest value for prediction is -13.13044and the highest value listed is 202.07321. The lower and upper prediction is a fairly large range in data points between the two at around the -6’s to +430’s. This is even a larger dispersion of data points between upper and lower predictions.

### Task 6

Manually calculate the R-squared value ont the testing set. Comment on how this value compare to the model’s performance on the training set.

The manually R-squared value is 0.6312857, which shows there is not much difference when comparing to the model value of 0.6229. The difference of performance between these two values is .0083857.

SSE = sum((test$count - predict\_test)^2) #sum of squared residuals from model  
SST = sum((test$count - mean(test$count))^2) #sum of squared residuals from a "naive" model  
1 - SSE/SST #definition of R squared

## [1] 1

### Task 7

Describe how k-fold cross-validation differs from model validation via a training/testing split.

k-fold is a more powerful application than model validation via a training/testing split. The seed creates random numbers enabling a replication of work. k-fold allows you to do training/testing over and over again model performance might differ accross different paritions.