# Module 3 - Assignment 1

## Model Validation

### Cooper, Sarah

library(tidyverse)

library(MASS)

library(caret)

library(readr)  
hour <- read\_csv("C:/Users/Sarah/Downloads/hour.csv")

bike <- hour  
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))) %>%  
mutate(yr = fct\_recode(yr,  
"2012" = "1",  
"2011" = "0",))  
bike = bike %>% mutate(mnth = as\_factor(as.character(mnth))) %>%  
mutate(mnth = fct\_recode(mnth,  
"Jan" = "1",  
"Feb" = "2",  
"Mar" = "3",  
"April" = "4",  
"May" = "5",  
"June" = "6",  
"July" = "7",  
"August" = "8",  
"Sept" = "9",  
"Oct" = "10",  
"Nov" = "11",  
"Dec" = "12"))  
bike = bike %>% mutate(hr = as\_factor(as.character(hr))) %>%  
mutate(hr = fct\_recode(hr,  
"12:00am" = "0",  
"1:00am" = "1",  
"2:00am" = "2",  
"3:00am" = "3",  
"4:00am" = "4",  
"5:00am" = "5",  
"6:00am" = "6",  
"7:00am" = "7",  
"8:00am" = "8",  
"9:00am" = "9",  
"10:00am" = "10",  
"11:00am" = "11",  
"12:00pm" = "12",  
"1:00pm" = "13",  
"2:00pm" = "14",  
"3:00pm" = "15",  
"4:00pm" = "16",  
"5:00pm" = "17",  
"6:00pm" = "18",  
"7:00pm" = "19",  
"8:00pm" = "20",  
"9:00pm" = "21",  
"10:00pm" = "22",  
"11:00pm" = "23",))  
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",))

#Task 1

train.rows = createDataPartition(y = bike$count, p=0.7, list = FALSE)  
train = bike[train.rows,]  
test = bike[-train.rows,]  
ctrl = trainControl(method = "cv",number = 10)  
set.seed(1234)

#Task 2

**There are 12167 rows in the training set, and 5212 rows in the testing set.**

#Task 3

mod1 = lm(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train)  
summary(mod1)

##   
## Call:  
## lm(formula = count ~ season + mnth + hr + holiday + weekday +   
## temp + weathersit, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -388.93 -61.73 -10.02 52.13 492.96   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -87.9303 6.9155 -12.715 < 2e-16 \*\*\*  
## seasonSummer 32.0886 6.2403 5.142 2.76e-07 \*\*\*  
## seasonFall 20.7843 7.4805 2.778 0.005470 \*\*   
## seasonWinter 58.4104 6.3352 9.220 < 2e-16 \*\*\*  
## mnthFeb -0.7478 5.1299 -0.146 0.884108   
## mnthMar 3.9829 5.7253 0.696 0.486646   
## mnthApril -2.6365 8.4834 -0.311 0.755971   
## mnthMay -1.7209 9.0430 -0.190 0.849074   
## mnthJune -13.5183 9.3133 -1.452 0.146662   
## mnthJuly -40.3884 10.4887 -3.851 0.000118 \*\*\*  
## mnthAugust -15.8140 10.2038 -1.550 0.121212   
## mnthSept 13.1266 9.0742 1.447 0.148040   
## mnthOct 0.9542 8.4297 0.113 0.909875   
## mnthNov -9.4750 8.1122 -1.168 0.242834   
## mnthDec -9.0177 6.4755 -1.393 0.163768   
## hr1:00am -17.0676 6.9403 -2.459 0.013938 \*   
## hr2:00am -26.2928 6.9898 -3.762 0.000170 \*\*\*  
## hr3:00am -38.4566 6.9984 -5.495 3.98e-08 \*\*\*  
## hr4:00am -39.9282 7.0487 -5.665 1.51e-08 \*\*\*  
## hr5:00am -25.3897 6.9755 -3.640 0.000274 \*\*\*  
## hr6:00am 35.1870 6.9522 5.061 4.22e-07 \*\*\*  
## hr7:00am 176.4017 7.0150 25.146 < 2e-16 \*\*\*  
## hr8:00am 304.3938 6.9632 43.714 < 2e-16 \*\*\*  
## hr9:00am 164.2820 6.9269 23.717 < 2e-16 \*\*\*  
## hr10:00am 113.9694 6.9600 16.375 < 2e-16 \*\*\*  
## hr11:00am 137.1642 6.9676 19.686 < 2e-16 \*\*\*  
## hr12:00pm 182.2887 7.0952 25.692 < 2e-16 \*\*\*  
## hr1:00pm 176.3672 6.9888 25.236 < 2e-16 \*\*\*  
## hr2:00pm 162.5949 7.0595 23.032 < 2e-16 \*\*\*  
## hr3:00pm 169.6027 7.0317 24.120 < 2e-16 \*\*\*  
## hr4:00pm 232.3041 7.0668 32.872 < 2e-16 \*\*\*  
## hr5:00pm 376.7537 7.0329 53.570 < 2e-16 \*\*\*  
## hr6:00pm 355.8418 6.9984 50.846 < 2e-16 \*\*\*  
## hr7:00pm 245.6580 7.0291 34.949 < 2e-16 \*\*\*  
## hr8:00pm 161.4756 6.9860 23.114 < 2e-16 \*\*\*  
## hr9:00pm 108.9130 6.9420 15.689 < 2e-16 \*\*\*  
## hr10:00pm 72.6050 6.9444 10.455 < 2e-16 \*\*\*  
## hr11:00pm 32.0728 6.9606 4.608 4.11e-06 \*\*\*  
## holidayHoliday -16.4410 6.3372 -2.594 0.009487 \*\*   
## weekdaySunday -14.9907 3.7306 -4.018 5.90e-05 \*\*\*  
## weekdayMonday -9.8897 3.8594 -2.562 0.010405 \*   
## weekdayTuesday -8.1606 3.7870 -2.155 0.031188 \*   
## weekdayWednesday -2.6587 3.7863 -0.702 0.482580   
## weekdayThursday -0.1411 3.7844 -0.037 0.970267   
## weekdayFriday 0.4656 3.7674 0.124 0.901635   
## temp 295.3296 12.1120 24.383 < 2e-16 \*\*\*  
## weathersitMisty -20.0258 2.3563 -8.499 < 2e-16 \*\*\*  
## weathersitLightPrecip -92.5001 3.7470 -24.687 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -170.6876 111.0271 -1.537 0.124233   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 110.8 on 12118 degrees of freedom  
## Multiple R-squared: 0.6277, Adjusted R-squared: 0.6262   
## F-statistic: 425.7 on 48 and 12118 DF, p-value: < 2.2e-16

**According to the linear regression model there are several strong predictors of count. The adjusted R-squared value is 0.63 which tells us the variability in factors can almost certainly be explained in the model.**

#Task 4

train\_preds = predict(mod1, newdata = train)  
 head(train\_preds)

## 1 2 3 4 5 6   
## -17.05120 -40.02543 -49.25059 -56.97942 -62.46672 12.22925

**All of the predictions in the training set, except for one, are negative.**

#Task 5

test\_preds = predict(mod1, newdata = test)  
 head(test\_preds)

## 1 2 3 4 5 6   
## -55.50784 147.53732 138.26430 204.26268 171.65129 59.96833

**Contrary to the training set, the testing set above provides more positive predictions.**

#Task 6

SSE = sum((test$count - test\_preds)^2)  
SST = sum((test$count - mean(test$count))^2)  
1 - SSE/SST

## [1] 0.6139696

**Manually calculating the R squared value produces a result that has no similarity to the predictions of the training dataset.**

#Task 7

**I think the k-fold and testing/training models are similiar in that they withdraw a small sample and test against the remainder. The difference is in the size of withholding each model withdraws. The K-fold cross validation approach will hold only 1 fold, evaluating the model multiple times. The train/test approach will hold a larger percentage, training usually being 80% and testing usually being 20%.**