# Module 3: Assignment 1

## Maliszewski, Angela

## Model Validation

Deliverable: All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

Libraries: For this assignment you will need the following libraries: tidyverse, lubridate, and tidymodels.

Before beginning the assignment tasks, read-in the “bike\_cleaned.csv” file into a data frame called “bike”. This is the same data that you used in the Module 2 Multiple Linear Regression and Special Issues assignment.

As we did in that assignment you should convert “dteday” from a character variable to a date variable. Convert the remaining character variables to factors. You can do this one variable at a time or use a “mutate\_if”.

Finally, convert the “hr” variable into a factor.

bike <- read\_csv("bike\_cleaned.csv")

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

str(bike)

## tibble [17,379 x 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ instant : num [1:17379] 1 2 3 4 5 6 7 8 9 10 ...  
## $ dteday : chr [1:17379] "1/1/2011" "1/1/2011" "1/1/2011" "1/1/2011" ...  
## $ season : chr [1:17379] "Winter" "Winter" "Winter" "Winter" ...  
## $ mnth : chr [1:17379] "Jan" "Jan" "Jan" "Jan" ...  
## $ hr : num [1:17379] 0 1 2 3 4 5 6 7 8 9 ...  
## $ holiday : chr [1:17379] "NotHoliday" "NotHoliday" "NotHoliday" "NotHoliday" ...  
## $ weekday : chr [1:17379] "Saturday" "Saturday" "Saturday" "Saturday" ...  
## $ workingday: chr [1:17379] "NotWorkingDay" "NotWorkingDay" "NotWorkingDay" "NotWorkingDay" ...  
## $ weathersit: chr [1:17379] "NoPrecip" "NoPrecip" "NoPrecip" "NoPrecip" ...  
## $ temp : num [1:17379] 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...  
## $ atemp : num [1:17379] 0.288 0.273 0.273 0.288 0.288 ...  
## $ hum : num [1:17379] 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...  
## $ windspeed : num [1:17379] 0 0 0 0 0 0.0896 0 0 0 0 ...  
## $ casual : num [1:17379] 3 8 5 3 0 0 2 1 1 8 ...  
## $ registered: num [1:17379] 13 32 27 10 1 1 0 2 7 6 ...  
## $ count : num [1:17379] 16 40 32 13 1 1 2 3 8 14 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. instant = col\_double(),  
## .. dteday = col\_character(),  
## .. season = col\_character(),  
## .. mnth = col\_character(),  
## .. hr = col\_double(),  
## .. holiday = col\_character(),  
## .. weekday = col\_character(),  
## .. workingday = col\_character(),  
## .. weathersit = col\_character(),  
## .. temp = col\_double(),  
## .. atemp = col\_double(),  
## .. hum = col\_double(),  
## .. windspeed = col\_double(),  
## .. casual = col\_double(),  
## .. registered = col\_double(),  
## .. count = col\_double()  
## .. )

bike = bike %>% mutate(dteday = mdy(dteday))

bike = bike %>% mutate\_if(is.character,as.factor)

bike$hr<-as.factor(bike$hr)  
str(bike)

## tibble [17,379 x 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ instant : num [1:17379] 1 2 3 4 5 6 7 8 9 10 ...  
## $ dteday : Date[1:17379], format: "2011-01-01" "2011-01-01" ...  
## $ season : Factor w/ 4 levels "Fall","Spring",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ mnth : Factor w/ 12 levels "Apr","Aug","Dec",..: 5 5 5 5 5 5 5 5 5 5 ...  
## $ hr : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ holiday : Factor w/ 2 levels "Holiday","NotHoliday": 2 2 2 2 2 2 2 2 2 2 ...  
## $ weekday : Factor w/ 7 levels "Friday","Monday",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ workingday: Factor w/ 2 levels "NotWorkingDay",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ weathersit: Factor w/ 4 levels "HeavyPrecip",..: 4 4 4 4 4 3 4 4 4 4 ...  
## $ temp : num [1:17379] 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...  
## $ atemp : num [1:17379] 0.288 0.273 0.273 0.288 0.288 ...  
## $ hum : num [1:17379] 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...  
## $ windspeed : num [1:17379] 0 0 0 0 0 0.0896 0 0 0 0 ...  
## $ casual : num [1:17379] 3 8 5 3 0 0 2 1 1 8 ...  
## $ registered: num [1:17379] 13 32 27 10 1 1 0 2 7 6 ...  
## $ count : num [1:17379] 16 40 32 13 1 1 2 3 8 14 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. instant = col\_double(),  
## .. dteday = col\_character(),  
## .. season = col\_character(),  
## .. mnth = col\_character(),  
## .. hr = col\_double(),  
## .. holiday = col\_character(),  
## .. weekday = col\_character(),  
## .. workingday = col\_character(),  
## .. weathersit = col\_character(),  
## .. temp = col\_double(),  
## .. atemp = col\_double(),  
## .. hum = col\_double(),  
## .. windspeed = col\_double(),  
## .. casual = col\_double(),  
## .. registered = col\_double(),  
## .. count = col\_double()  
## .. )

#Task 1: Split the data into training and testing sets. Your training set should have 70% of the data. Use a random number (set.seed) of 1234. Your split should be stratified by the “count” variable.

set.seed(1234)  
bike\_split = initial\_split(bike, prob = 0.70, strata = count)  
train = training(bike\_split)  
test = testing(bike\_split)

#Task 2: How many rows of data are in each set (training and testing)?   
ANSWER: 13,036 rows in the training set and 4,343 rows in the testing set.

#Task 3: Build a linear regression model (using the training set) to predict “count” using the variables “season”, “mnth”, “hr”, “holiday”, and “weekday”, “temp”, and “weathersit”.

Comment on the quality of the model. Be sure to note the Adjusted R-squared value. ANSWER: Adjusted R-Squared is 0.6229 which is decent. The p-value for slope coefficients are significant for the seasons, summer months and all hours of the day indicating they are significant predictors of count. The p-value for slope coefficients are not significant for weather indicating weather is not a significant predictor of count.

bike\_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, train) %>%  
 step\_dummy(all\_nominal()) %>% #makes categorical  
 step\_center(all\_predictors()) %>% #centers the predictors  
 step\_scale(all\_predictors()) #scales the predictors

lm\_model =   
 linear\_reg() %>%   
 set\_engine("lm")

lm\_wflow =  
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(bike\_recipe)

lm\_fit = fit(lm\_wflow, train)

summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -413.11 -61.65 -10.20 52.16 493.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 188.95743 0.97281 194.238 < 2e-16 \*\*\*  
## temp 55.83290 2.27360 24.557 < 2e-16 \*\*\*  
## season\_Spring -11.22585 3.12246 -3.595 0.000325 \*\*\*  
## season\_Summer -13.62697 2.81452 -4.842 1.30e-06 \*\*\*  
## season\_Winter -26.69916 2.62085 -10.187 < 2e-16 \*\*\*  
## mnth\_Aug -5.55531 2.28035 -2.436 0.014857 \*   
## mnth\_Dec -1.28590 2.25701 -0.570 0.568865   
## mnth\_Feb 2.45670 2.16432 1.135 0.256358   
## mnth\_Jan 2.19256 2.26073 0.970 0.332141   
## mnth\_Jul -10.82710 2.31302 -4.681 2.88e-06 \*\*\*  
## mnth\_Jun -4.42159 1.57131 -2.814 0.004901 \*\*   
## mnth\_Mar 3.95464 1.73653 2.277 0.022783 \*   
## mnth\_May 0.03385 1.39465 0.024 0.980638   
## mnth\_Nov -2.73418 2.42310 -1.128 0.259180   
## mnth\_Oct 1.43289 2.39321 0.599 0.549363   
## mnth\_Sep 2.31759 2.11255 1.097 0.272635   
## hr\_X1 -3.51511 1.35168 -2.601 0.009318 \*\*   
## hr\_X2 -4.92587 1.35108 -3.646 0.000268 \*\*\*  
## hr\_X3 -7.23713 1.35222 -5.352 8.85e-08 \*\*\*  
## hr\_X4 -7.75210 1.33785 -5.794 7.01e-09 \*\*\*  
## hr\_X5 -4.63960 1.34701 -3.444 0.000574 \*\*\*  
## hr\_X6 7.07217 1.36506 5.181 2.24e-07 \*\*\*  
## hr\_X7 34.28937 1.35968 25.219 < 2e-16 \*\*\*  
## hr\_X8 61.76220 1.35137 45.703 < 2e-16 \*\*\*  
## hr\_X9 34.32084 1.37026 25.047 < 2e-16 \*\*\*  
## hr\_X10 22.49373 1.35709 16.575 < 2e-16 \*\*\*  
## hr\_X11 28.04106 1.36144 20.597 < 2e-16 \*\*\*  
## hr\_X12 35.48147 1.35297 26.225 < 2e-16 \*\*\*  
## hr\_X13 36.46909 1.36774 26.664 < 2e-16 \*\*\*  
## hr\_X14 33.04659 1.38001 23.947 < 2e-16 \*\*\*  
## hr\_X15 33.47360 1.36802 24.469 < 2e-16 \*\*\*  
## hr\_X16 45.47455 1.37024 33.187 < 2e-16 \*\*\*  
## hr\_X17 77.04941 1.37746 55.936 < 2e-16 \*\*\*  
## hr\_X18 70.45233 1.35922 51.833 < 2e-16 \*\*\*  
## hr\_X19 49.09129 1.36249 36.031 < 2e-16 \*\*\*  
## hr\_X20 31.88322 1.35057 23.607 < 2e-16 \*\*\*  
## hr\_X21 22.32101 1.36261 16.381 < 2e-16 \*\*\*  
## hr\_X22 14.85926 1.36249 10.906 < 2e-16 \*\*\*  
## hr\_X23 6.97465 1.35440 5.150 2.65e-07 \*\*\*  
## holiday\_NotHoliday 4.63306 1.02585 4.516 6.35e-06 \*\*\*  
## weekday\_Monday -3.02126 1.29875 -2.326 0.020018 \*   
## weekday\_Saturday -0.45549 1.27739 -0.357 0.721414   
## weekday\_Sunday -7.04302 1.27548 -5.522 3.42e-08 \*\*\*  
## weekday\_Thursday -1.59975 1.27486 -1.255 0.209557   
## weekday\_Tuesday -2.67683 1.27097 -2.106 0.035212 \*   
## weekday\_Wednesday -1.54779 1.27570 -1.213 0.225043   
## weathersit\_LightPrecip -14.03077 21.72138 -0.646 0.518328   
## weathersit\_Misty 8.99013 34.46960 0.261 0.794241   
## weathersit\_NoPrecip 19.55991 37.32871 0.524 0.600294   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.1 on 12987 degrees of freedom  
## Multiple R-squared: 0.6243, Adjusted R-squared: 0.6229   
## F-statistic: 449.6 on 48 and 12987 DF, p-value: < 2.2e-16

#Task 4: Use the predict functions to make predictions (using your model from Task 3) on the training set. Hint: Be sure to store the predictions in an object, perhaps named “predict\_train” or similar.

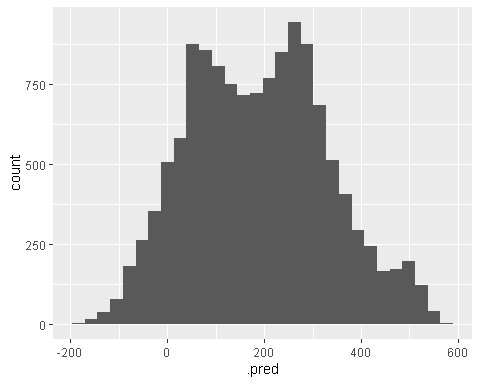
Develop a histogram of the predictions (Hint: The predictions are likely stored in a variable called “.pred” in your predictions object).

Comment on the distribution of the predictions.   
ANSWER: The histrogram shows the predictions have a bimodal distribution. This makes sense since people tend to rent bikes most frequently during morning and evening commute hours.

predict\_train = predict(lm\_fit,train)

ggplot(predict\_train, aes(x = .pred)) + geom\_histogram()

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



#Task 5: Determine the R-squared value of the model on the testing set.

Comment on how this value compares to the model’s performance on the training set. ANSWER: Adjusted R-squared is 0.623 on the test data. Adjusted R-Squared is 0.6229 on the train data. The performance between the two data sets is similar indicating the model is not overfitting the data and just as good at predicting outcomes with data not used to build the model as the original model itself.

bike\_recipe\_test = recipe(count ~ season + mnth + hr + holiday + weekday + temp + weathersit, test) %>%  
 step\_dummy(all\_nominal()) %>% #makes categorical  
 step\_center(all\_predictors()) %>% #centers the predictors  
 step\_scale(all\_predictors()) #scales the predictors

lm\_model\_test =   
 linear\_reg() %>%   
 set\_engine("lm")

lm\_wflow\_test =  
 workflow() %>%  
 add\_model(lm\_model\_test) %>%  
 add\_recipe(bike\_recipe\_test)

lm\_fit\_test = fit(lm\_wflow\_test, test)

summary(lm\_fit\_test$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -376.81 -64.23 -8.34 51.71 508.12   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 190.98089 1.70420 112.065 < 2e-16 \*\*\*  
## temp 54.48544 3.95112 13.790 < 2e-16 \*\*\*  
## season\_Spring -18.28142 5.62905 -3.248 0.001172 \*\*   
## season\_Summer -27.34322 5.20492 -5.253 1.57e-07 \*\*\*  
## season\_Winter -32.42941 4.72608 -6.862 7.77e-12 \*\*\*  
## mnth\_Aug 0.17197 3.83346 0.045 0.964221   
## mnth\_Dec -6.45880 3.92993 -1.643 0.100355   
## mnth\_Feb 0.77228 3.73741 0.207 0.836304   
## mnth\_Jan 0.17878 4.00267 0.045 0.964376   
## mnth\_Jul -6.67344 4.02536 -1.658 0.097421 .   
## mnth\_Jun 0.05325 2.74128 0.019 0.984503   
## mnth\_Mar 0.63626 3.09457 0.206 0.837109   
## mnth\_May 0.18229 2.41789 0.075 0.939907   
## mnth\_Nov -5.67133 4.30469 -1.317 0.187749   
## mnth\_Oct -2.40625 4.20927 -0.572 0.567585   
## mnth\_Sep 5.06223 3.61027 1.402 0.160935   
## hr\_X1 -3.83540 2.33969 -1.639 0.101229   
## hr\_X2 -6.34968 2.31543 -2.742 0.006126 \*\*   
## hr\_X3 -7.97725 2.26126 -3.528 0.000423 \*\*\*  
## hr\_X4 -8.66114 2.33709 -3.706 0.000213 \*\*\*  
## hr\_X5 -5.96418 2.35752 -2.530 0.011447 \*   
## hr\_X6 5.65950 2.29169 2.470 0.013566 \*   
## hr\_X7 33.16919 2.31811 14.309 < 2e-16 \*\*\*  
## hr\_X8 63.58697 2.34770 27.085 < 2e-16 \*\*\*  
## hr\_X9 28.82286 2.25516 12.781 < 2e-16 \*\*\*  
## hr\_X10 21.87474 2.33422 9.371 < 2e-16 \*\*\*  
## hr\_X11 27.51270 2.34077 11.754 < 2e-16 \*\*\*  
## hr\_X12 37.78983 2.41535 15.646 < 2e-16 \*\*\*  
## hr\_X13 31.82093 2.37982 13.371 < 2e-16 \*\*\*  
## hr\_X14 29.48873 2.34136 12.595 < 2e-16 \*\*\*  
## hr\_X15 35.73592 2.41609 14.791 < 2e-16 \*\*\*  
## hr\_X16 49.15862 2.39667 20.511 < 2e-16 \*\*\*  
## hr\_X17 77.49375 2.33521 33.185 < 2e-16 \*\*\*  
## hr\_X18 70.88053 2.39164 29.637 < 2e-16 \*\*\*  
## hr\_X19 46.34644 2.34130 19.795 < 2e-16 \*\*\*  
## hr\_X20 33.53552 2.38135 14.083 < 2e-16 \*\*\*  
## hr\_X21 21.70748 2.30525 9.417 < 2e-16 \*\*\*  
## hr\_X22 13.64239 2.29628 5.941 3.06e-09 \*\*\*  
## hr\_X23 5.89831 2.33705 2.524 0.011644 \*   
## holiday\_NotHoliday 3.61604 1.81337 1.994 0.046205 \*   
## weekday\_Monday -4.18376 2.30403 -1.816 0.069464 .   
## weekday\_Saturday -1.20260 2.24526 -0.536 0.592250   
## weekday\_Sunday -3.50347 2.25699 -1.552 0.120669   
## weekday\_Thursday -1.01554 2.23587 -0.454 0.649706   
## weekday\_Tuesday -3.49570 2.24818 -1.555 0.120044   
## weekday\_Wednesday -3.04838 2.24442 -1.358 0.174471   
## weathersit\_LightPrecip 14.55577 30.27304 0.481 0.630672   
## weathersit\_Misty 55.29464 50.12550 1.103 0.270035   
## weathersit\_NoPrecip 66.97556 53.76453 1.246 0.212935   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 112.3 on 4294 degrees of freedom  
## Multiple R-squared: 0.6271, Adjusted R-squared: 0.623   
## F-statistic: 150.5 on 48 and 4294 DF, p-value: < 2.2e-16