Final Course Project of Bike Sharing Demand Data Set

Yihao Zhao

Abstract

This data analysis project is based on the Seoul Bike Sharing Demand dataset, it is an observational study of the number of rented bikes with various weather information. By using the linear regression and some necessary applied statistical skills (statistical tests, model selection, etc.), the major finding of this project is the response variable of the number of rented bikes is most relevant and impacted by the predictor variables of Hour (renting hour of the day), Temperature (temperature in Celsius of the day), Humidity (%), Solar radiation (MJ/m2), Rainfall (mm), and Seasons (Winter, Spring, Summer, and Autumn).

In addition, I have discovered some additional variables (interaction terms, polynomial terms, non-linear transformation terms) such as the interaction terms of Humidity&Rainfall that make my model more precise and practical. I have also performed a Box-Cox transformation on my response variable of rented bike counts and found the optimal lambda value of 0.3 as the corresponding transformation on the response variable. Besides, by analyzing and modifying the linear model above, I have inspected the collinearity and model complexity of my optimal model and data; showing that there is no need to be concerned about the collinearity and size in the linear regression model. Overall, I would create an optimal linear model predicting the number of rented bikes by using this Bike Sharing Demand data set to better answer and explore the questions regarding my interests and the potential growth of the business of rental bikes.

Introduction

In order to reduce the waiting time and improve the mobility comfort of urban citizens, the rental bike also known as sharing bike can be a convenient transaction tool in modern and big cities. However, the major concern of this rental bike business is whether a stable supply of rental bikes can be

implemented within the city. Therefore, this data analysis project is primarily focusing on the prediction of rented bike counts required at each hour for the stable supply of rental bikes based on the Seoul Bike Sharing Demand dataset.

The Seoul Bike Sharing Demand dataset has 544 observations (size) and 14 attributes (available variables), including Date (year-month-day), Rented Bike Count (Count of bikes at each hour), Hour (Hour of the day), Temperature (Temperature in Celsius), Humidity (%), Windspeed (m/s), Visibility (10m), Dew point temperature (Celsius), Solar radiation (MJ/m2), Rainfall (mm), Snowfall (cm), Seasons (Winter, Spring, Summer, Autumn), Holiday (Holiday/No holiday), and Functional Day (Non Functional hours, Functional hours).

As for analyzing this dataset, I have come out with several questions that can assist me in better understanding the relationship between the response variable of rented bike counts and other predictor variables. First, what are some strong relationship predictor variables for predicting the rented bike counts? Second, what are some unnecessary variables that can take out of the linear model to make our model more accreted? Most importantly, what is the best optimal model for predicting the rented bike counts in this data? Besides, I am very interested in predicting the supply of rental bikes since I would like to know if the linear model of this dataset could apply somewhere else, and I would like to acquire more information about how people would ride the bike in different weather conditions.

Results

By applying some statistical tests, including the ANOVA test and the F-statistic test. My optimal linear model for predicting rented bike counts is showing a much more significant statistical level of importance; the optimal linear model appears to be "(y^0.3 - 1) / 3 ~ Hour + Temperature.C. + Humidity + Solar Radiation + Rainfall + Seasons + Hour: Seasons + Humidity: Rainfall."

<u>Data importing & cleaning</u>

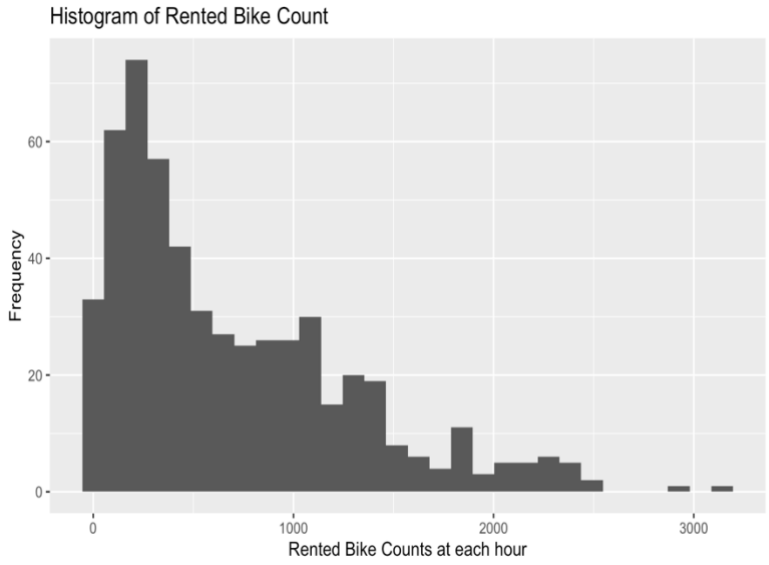
This is the beginning of the data analysis, by importing the dataset of "bike" as "bike" and mutating a new data variable as Date combining the information of the existing variable of Day, Month,

and Year. Below is a figure that shows the codes and the output of the first 6 rows of the new subset of the dataset “bike”.

Solar.Radiation..MJ.m2. <dbl>	Rainfall.mm. <dbl>	Snowfall..cm. <dbl>	Seasons <chr>	Holiday <chr>	Date <date>
0.00	0	3	Autumn	No Holiday	2020-11-25
0.00	0	0	Winter	No Holiday	2020-12-29
1.74	0	0	Autumn	No Holiday	2020-10-22
1.68	0	0	Autumn	No Holiday	2020-10-16
0.00	0	0	Autumn	No Holiday	2020-10-13
1.33	0	0	Spring	No Holiday	2020-03-26

Interpretation of the response variable of rented bike count

The histogram on the right indicates the center of the distribution of the rented bike count per hour falls around 1500 bikes per hour with a range of around 0 to 3500 bikes per hour. The peak of this histogram is roughly around 200 to 300, and there are also several potential outliers above 3000 bikes per hour.

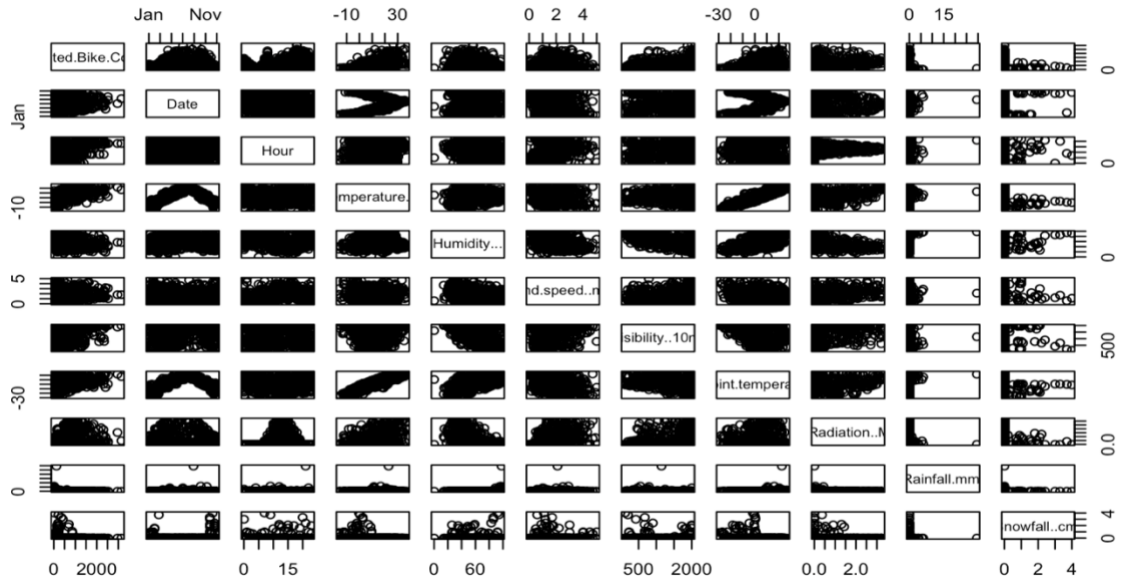


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.0	210.8	498.0	692.3	1044.8	3146.0

By looking at the summary table on the left, the response variable of Rented Bike Count has a precise range from 4 to 3146 bikes at each hour, it has a median of 498 bikes per hour and a mean of approximately 692 bikes per hour.

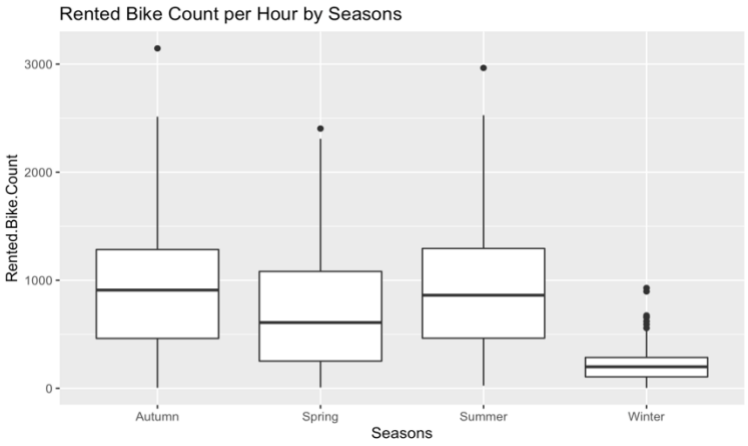
By creating a scatterplot matrix for Rented Bike Count and its quantitative predictors below, the matrix seems to indicate that the predictor variables of Month, Hour, Temperature.C., Humidity, Wind speed, Visibility, Dew.point.temperature.C., Solar. Radiation, Rainfall, and Snowfall have a possible stronger relationship with the response variable Rented.Bike.Count. Within the predictor variables just mentioned, it seems like the variables of Month, Hour, Temperature.C., Visibility,

Dew.point.temperature.c. have a positive linear relationship with the response variable, and the variables of Solar. Radiation, Rainfall, and Snowfall may have a negative linear relationship with the response variable.



Boxplots of Rented Bike Counts with the categorical variable of Seasons

From the side-by-side boxplot on the right, it seems to indicate that the center the number of the rented bike is around 900 bikes when it is the Autumn season as well as the season Summer, the center the number of the rented bike is about 600 bikes during Spring, the center of the rented bike counts is about 100 during Winter. Overall, it seems to indicate that the Winter season has the strongest negative relationship with the number of rented bikes.



The first model that includes all first-order terms for predictors

Based on the summary of the model below, we can see that the first model has a statistically significant level since its p-value is very small. The R-square value of this model is about 0.576, which indicates that about 57.6% of the variation in rented bike count in Seoul bike sharing demand can be explained by this linear relationship with all first-order terms of the predictors.

```
model1 = lm(Rented.Bike.Count ~ . - X, bike)
summary(model1)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ . - X, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -957.40 -249.83  -45.07  189.65 1688.16
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1518.67835 3756.49547   0.404 0.686170
## Hour                         27.65150    2.68486  10.299  < 2e-16 ***
## Temperature.C.               27.53405   12.09735   2.276 0.023243 *
## Humidity...                  -7.24152    3.41598  -2.120 0.034480 *
## Wind.speed..m.s.              4.97136   19.45269   0.256 0.798389
## Visibility..10m.              0.01961    0.03742   0.524 0.600535
## Dew.point.temperature.C.     -4.66192   12.86580  -0.362 0.717235
## Solar.Radiation..MJ.m2.     -66.07296   27.62813  -2.392 0.017128 *
## Rainfall.mm.                -49.07993   12.84564  -3.821 0.000149 ***
## Snowfall..cm.                17.17933   43.05807   0.399 0.690068
## SeasonsSpring              -192.86614   63.64829  -3.030 0.002564 **
## SeasonsSummer              -142.04810   69.30741  -2.050 0.040902 *
## SeasonsWinter              -458.59540   73.27380  -6.259 8.03e-10 ***
## HolidayNo Holiday            75.18538   69.96831   1.075 0.283060
## Date                         -0.04864    0.20225  -0.241 0.810029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.4 on 529 degrees of freedom
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5648
## F-statistic: 51.33 on 14 and 529 DF,  p-value: < 2.2e-16
```

The second model with additional variables based on model 1

On the second linear model, the interaction terms of Hour & Seasons and Humidity &Rainfall are added to model 2. By printing the summary table of model 2 and performing the ANOVA F-statistic test on model 1 and model 2. The result of the test indicates that model 2 is preferable since the p-value of the test is about 6.036e-12, which is below 0.05. As the null hypothesis is model 1 and the alternative hypothesis is model 2, there is enough evidence to reject the null hypothesis, model 1.

```
## Call:
## lm(formula = Rented.Bike.Count ~ . - X + Hour:Seasons + Humidity...:Rainfall.mm.,
##     data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1006.20  -201.90   -42.61   140.18  1652.00
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -6.062e+02  3.583e+03  -0.169 0.865721
## Hour                        3.330e+01  5.175e+00   6.435 2.79e-10 ***
## Temperature.C.              3.202e+01  1.151e+01   2.781 0.005619 **
## Humidity...                -5.086e+00  3.266e+00  -1.557 0.120030
## Wind.speed..m.s.           -4.560e+00  1.876e+01  -0.243 0.808016
## Visibility..10m.            2.432e-02  3.561e-02   0.683 0.494922
## Dew.point.temperature.C.   -9.722e+00  1.225e+01  -0.794 0.427798
## Solar.Radiation..MJ.m2.    -6.810e+01  2.639e+01  -2.580 0.010147 *
## Rainfall.mm.               -3.119e+03  7.974e+02  -3.912 0.000104 ***
## Snowfall..cm.               4.471e+00  4.094e+01   0.109 0.913065
## SeasonsSpring              -1.586e+02  1.054e+02  -1.504 0.133209
## SeasonsSummer              -2.581e+02  1.011e+02  -2.554 0.010927 *
## SeasonsWinter              -8.079e+01  1.100e+02  -0.735 0.462966
## HolidayNo Holiday           1.034e+02  6.692e+01   1.545 0.122884
## Date                        5.320e-02  1.927e-01   0.276 0.782614
Analysis of Variance Table

Model 1: Rented.Bike.Count ~ (X + Hour + Temperature.C. + Humidity... +
    Wind.speed..m.s. + Visibility..10m. + Dew.point.temperature.C. +
    Solar.Radiation..MJ.m2. + Rainfall.mm. + Snowfall..cm. +
    Seasons + Holiday + Date) - X
Model 2: Rented.Bike.Count ~ (X + Hour + Temperature.C. + Humidity... +
    Wind.speed..m.s. + Visibility..10m. + Dew.point.temperature.C. +
    Solar.Radiation..MJ.m2. + Rainfall.mm. + Snowfall..cm. +
    Seasons + Holiday + Date) - X + Hour:Seasons + Humidity...:Rainfall.mm.
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    529 82279888
2    525 73621413  4   8658475 15.436 6.036e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Furthermore, the R-square value of model 1 is about 0.58 as mentioned, and the R-square value of model 2 increases to about 0.62. This indicates that there is more variation in the response variable of rented bike count, 62% of the variation, which can be explained by its linear relationship with all first-order terms plus some additional interaction terms predictors.

Model selection on model 2

By performing a model selection of stepwise search focusing on the metric of BIC on model 2, the final selected predictors are "Hour", "Temperature.C.", "Humidity...", "Rainfall.mm.", "Seasons",

"Hour:Seaons", "Humidity...:Rainfall.mm". Below is a figure showing the code and its output of the final step of the stepwise search on the metric of BIC.

```
##
## Step:  AIC=6513.38
## Rented.Bike.Count ~ Hour + Temperature.C. + Humidity... + Solar.Radiation..MJ.m2. +
##     Rainfall.mm. + Seasons + Hour:Seasons + Humidity...:Rainfall.mm.
##
##                             Df Sum of Sq      RSS    AIC
## <none>                                    74146433 6513.4
## - Solar.Radiation..MJ.m2.    1   1160492 75306926 6515.5
## + Holiday                    1    352705 73793728 6517.1
## + Dew.point.temperature.C.   1    125358 74021075 6518.8
## + Visibility..10m.           1     41217 74105216 6519.4
## + Snowfall..cm.              1      8444 74137989 6519.6
## + Date                       1      8048 74138385 6519.6
## + Wind.speed..m.s.           1      4267 74142166 6519.7
## - Humidity...:Rainfall.mm.   1   2113467 76259901 6522.4
## - Hour:Seasons               3   6313446 80459879 6538.9
## - Temperature.C.             1   7977864 82124297 6562.7
```

Interpretation of the quantitative and categorical predictors on model 3

Based on the summary table of model 3, the quantitative predictors include Hour, Temperature, Humidity, Solar Radiation, and Rainfall; the categorical predictor includes only Seasons. The quantitative predictor of Temperature can be interpreted as each 1 degree of Celsius increase in Temperature on the bike renting day, I would expect the estimated number of the rented bike to increase by 23, holding the other predictor variables constant. For the categorical predictor of Seasons, the baseline level of the predictor Seasons is Autumn. For a season of Autumn, I would predict the average rented bike counts to increase by about 754, if the season is Spring, I would predict the average rented bike counts to decrease by around 180, if the season is Summer, I would predict the average rented bike counts to decrease by approximately 13, and if the season is Winter, I would predict the average rented bike counts to decrease by about 121, with the fixing values of the other predictors.

R-square value comparison between model 1 and model 3

Based on the summary tables of model 1 and model 3, the R-square value of model 1 is about 0.56, and the R-square value of model 3 is about 0.62. As a result, I prefer model 3 since its R-square value is greater than model1. In other words, the R-square value of model 1 indicates that there are about

56% of the variation in Rented Bike Counts can be explained by its linear relationship with the predictors of "Hour", "Temperature.C.", "Humidity...", "Wind. speed..m.s.", "Visibility..10m.", "Dew.point.temperature.C.", "Solar.Radiation..MJ.m2.", "Rainfall.mm.", "Snowfall..cm." ,"Seasons", "Holiday", "Date". The R-square value of model 3 indicates that approximately 62% of the variation in Rented Bike Counts can be explained by its linear relationship with the predictors of "Hour", "Temperature.C.", "Humidity...", "Solar. Radiation..MJ.m2.", "Rainfall.mm.", "Seasons", "Hour: Season", "Humidity...: Rainfall.mm."

Analysis of linear model 3

According to the previous model selection and additional variables added, the most preferable linear model will be model 3, which has predictor variables of Hour, Temperature, Humidity, Solar Radiation, Rainfall, Seasons, and the interaction terms of Hour & Seasons and Humidity &Rainfall.

Checking the collinearity of model 3

By calculating the Variance Inflation factors of each quantitative predictor of model 3, the VIF of Hour is about 1.08, the VIF of Temperature is roughly 1.30, the VIF of Humidity is 1.65, the VIF of Solar Radiation is about 1.58, and the VIF of Rainfall is about 1.05. Based on these VIF values, none of them is greater than 5, so there is no need to be concerned about the collinearity of model 3.

Is collinearity something to be concerned about in linear regression?

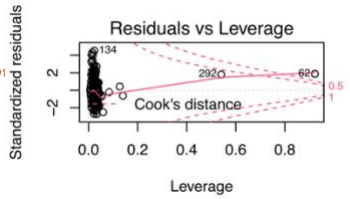
From what I have learned and understand so far, I think collinearity is something to be concerned about in linear regression for several reasons. First, Collinearity is when quantitative variables are highly correlated with each other, and it is sometimes called multicollinearity when it occurs in multiple linear regression. Second, collinearity is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. Besides, collinearity will undermine the statistical significance of each predictor variable. In other words, collinearity will reduce the precision of the estimated coefficients, which leads to the untrusted p-values and low significance level of the model.

Unusual observation

By looking at the figures below, the observations with high leverage are 11, 62, 142, 184, 229, 292, 365, and 476. The observation 62 has the largest leverage; the potential outliers are 11, 22, 28, 32, 42, 66, 83, 98,134, 151, 185, 188, 193, 235, 249, 277, 295, 317, 319, 350, 359, 377, 390, 402, 418, 442, 446, 489, 493, 512, 519, 534. The observations of 42, 185, 317, 319, 377, and 442 have the larger standardized residuals. Among these observations, observation 62 is the influence point also known as unusual observation since the Cook's distance is about 3.21 based on the Cook's distance levels of 0.5 and 1.0.

```
n = length(resid(fitted_model)) # number of rows of data
p = length(coef(fitted_model)) # number of coefficients
hatvalues(fitted_model)[which(hatvalues(fitted_model) > 2 * p / n)] # high leverage observatio
```

```
##          11          62         142         184         229         292         365
## 0.05953005 0.92223685 0.05182101 0.08371034 0.05352634 0.54041467 0.12708219
##         476
## 0.14026731
```




```
cooks.distance(fitted_model)[which(cooks.distance(fitted_model) > 4 / n)] # influence points
```

```
##          11          19          22          27          28          32
## 0.030855830 0.007395713 0.012733194 0.007971450 0.014479353 0.007586329
##          42          62          83         134         151         185
## 0.015153350 3.211128874 0.019168509 0.037817097 0.011382694 0.009687380
##         188         193         235         240         249         277
## 0.008397147 0.014224688 0.013316271 0.009081822 0.011915954 0.009027929
##         292         295         317         319         333         344
## 0.293195281 0.012472717 0.021419194 0.009751081 0.007760849 0.007441780
##         350         359         377         390         418         440
## 0.008928846 0.009996490 0.022365594 0.007897952 0.008601159 0.010852282
##         442         446         472         489         500         512
## 0.014604979 0.013425081 0.008651826 0.008990246 0.007816556 0.008549679
##         534
## 0.019762601
```

```
rstandard(fitted_model)[which(abs(rstandard(fitted_model)) > 2)] # Outliers
```

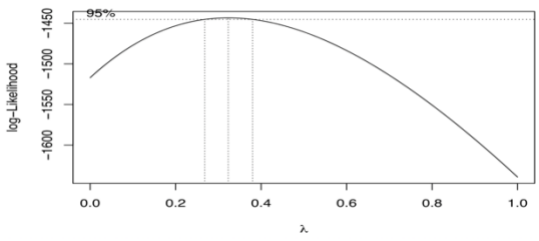
```
##         11         22         28         32         42         66         83         98
## -2.517356  2.592407  2.921721  2.031423  4.075498  2.179669  2.713813  2.231083
##        134        151        185        188        193        235        249        277
##  4.525283  2.174437  3.281614 -2.183142 -2.325854 -2.385643  2.651862  2.430295
##        295        317        319        350        359        377        390        402
## -2.699459  3.972614  3.161356  2.098271  2.417514  4.296980  2.336947 -2.087215
##        418        442        446        489        493        512        519        534
## -2.322072  3.399371  2.730018  2.024416  2.157456 -2.127500  2.162249 -2.722639
```

Model complexity vs. Dataset size

By applying the model complexity rule of thumb, 5 times the number of coefficients lesser or equal to the number of rows of data, there is no need to be concerned about the size of my fitted model. Since the bike data of model 3 has 544 rows and 13 coefficients, the calculation of the model complexity rule of thumb shows that 5 times 13 is 65 which is smaller than 544. Therefore, the bike dataset with model 3 passes the model complexity rule.

Box-Cox Transformation of Response Variable

Based on the Box-Cox plot on the right, an approximate 95% confidence interval for lambda is (0.28, 0.39). The center of this confidence interval appears to be a lambda of 0.3, so the most optimal



lambda would then correspond to a transformation of Rented.Bike.Count^{0.3}-1 / (0.3) . Alternatively, I could simply raise the response variable of rented bike count to the 0.3th power, as this would approximate the full Box-Cox transformation.

Report of the Statistical test

By performing the one-way ANOVA test on the new model 3 with transformed response variable of Rented.Bike.Count^{0.3}-1 / (0.3) on the new bike dataset that filtered out the unusual observation of 62. Since the p-value of each corresponding coefficient is all below the level of 0.05, except the interaction term of Hour & Seasons. This indicates that there is enough evidence to reject the null hypothesis of each coefficient except for the interaction term of Hour & Seasons; the null hypothesis of each coefficient is the estimated coefficient equal to zero. Therefore, the interaction term of Hour $ Seasons could be removed from model 3 since it does not have a statistically significant.

```
## Analysis of Variance Table
##
## Response: (((Rented.Bike.Count^0.3) - 1)/3)
##                            Df Sum Sq Mean Sq  F value  Pr(>F)
## Hour                        1 32.732  32.732 239.8116 < 2e-16 ***
## Temperature.C.              1 78.600  78.600 575.8586 < 2e-16 ***
## Humidity...                 1 21.295  21.295 156.0187 < 2e-16 ***
## Solar.Radiation..MJ.m2.     1  0.542   0.542   3.9727 0.04676 *
## Rainfall.mm.                1 13.598  13.598  99.6218 < 2e-16 ***
## Seasons                     3 16.741   5.580  40.8827 < 2e-16 ***
## Hour:Seasons                3  0.749   0.250   1.8291 0.14079
## Humidity...:Rainfall.mm.    1  0.782   0.782   5.7271 0.01705 *
## Residuals                 530 72.341   0.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion

According to the data analysis above, I have known that there are generally 6 first-order predictor variables that are the most relevant variables to the response variable of rented bike count, including Hour (renting hour of the day), Temperature (temperature in Celsius of the day), Humidity (%), Solar radiation (MJ/m2), Rainfall (mm), and Seasons (Winter, Spring, Summer, and Autumn). Besides, I have also discovered a few interaction terms, Humidity $ Rainfall, as the additional predictor to the linear model for predicting the number of rented bikes. Overall, my optimal fitted model for predicting the rented bike count is "y^ {0.3}-1 / (0.3) ~ Hour + Temperature + Humidity + Solar Radiation + Rainfall + Seasons + Humidity: Rainfall."

In addition, I would like to talk about the limitation of this dataset; I think it is also necessary to have some other available variables in the dataset such as the type of the rented bike, the location where

the bikes are rented, the group of people who rent the rental bike, the cost of the rental bike, etc. If I had unlimited time on this data analysis project, I would try to find out any potentially additional variables (interaction terms, polynomial terms, non-linear transformation terms) that could be added to my fitted model, and I would also like to change my response variable; for example, predicting the renting hour of each bike to see if this model will work. Besides, I would also provide the appropriate interpretations of each coefficient in my fitted model to express my model with readable and understandable content.

References

Provided by Canvas

[1] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020
[2] Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020