STAT537 Project Report

Sylvia Hu(1728935), Enxi Lin(1537565), Ting Wang(1438592)

1. Introduction

Traffic congestion has already been a serious problem in many urban cities, which increases the demand for public transportation. Bike renting is remarkable with its low renting rate, convenience, and environmental friendliness. Thus, bike-sharing is becoming more and more popular in the world, especially in areas with high population density. Therefore, it might be valuable to analyze what factor will have a significant effect on bike renting demand and to fit a model to predict the demand.

We used the dataset of bike renting amount of the capital city of Korea — Seoul, from a published journal [E, S. V., Park, J., & Cho, Y. (2020)]. It recorded the count of the rented bikes every hour every day for 365 days from December 1st, 2017 to November 30th, 2018. It also includes information on Hour, Temperature, Humidity, Windspeed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, Snowfall, Seasons, Holiday/NoHoliday, Functioning Day (Yes/No). The original dataset has 8760 samples, but we were only interested in the renting demand at the peak hour of 6 pm. Also, the renting amounts are all 0 when Functioning Day equals 0, which is useless. After filtering the dataset, we got 353 samples in total. Then we added a variable WeekStatus indicating whether it was a weekday or not. The count of rented bikes is our response variable and we have 11 explanatory variables (5 categorical ones and 6 continuous ones).

2. Review of Methodology

This is an observational study design, the dataset was split into a training and a testing dataset with a ratio of 7:3. We trained the models on the training dataset and analyzed the prediction performance on the testing dataset using r squared, adjusted r squared, and 5-fold cross-validated MSE (CV-MSE).

4 methods were used to select variables:

2.1. Backward Stepwise Selection

Backward Stepwise is a stepwise regression approach by starting with all candidate variables in the model. Then remove the predictor with the highest p-value greater than the model fit criterion, until all p-values are less than the criterion [Faraway (2002)]. The p-values come from the t-test of variables' coefficients.

2.2. Conscientious Approach

Manually remove candidate variables with p-values larger than the model fit criterion from the full model until all p-values are less than the criterion [Faraway (2002)].

2.3. Ridge Regression

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \le t$$

The goal of this algorithm is to minimize the least-squares with the L2 penalty, which is the sum of squares of the magnitude of the coefficients [Bhattacharyya

(2020)]. Ridge Regression controls the multicollinearity between variables by a biasing constant k. It proceeds by adding the k to the diagonal elements of the correlation matrix [NCSS (2021)]. It shrinks the coefficients towards zero.

2.4. Lasso Regression

$$\sum_{i=1}^{n} (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

The goal of this algorithm is to minimize the least-squares with the L1 penalty, which is the sum of the absolute coefficients. The penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero. λ controls the strength of the L1 penalty. As λ increases, bias increases, and more coefficients are set to zero and eliminated from the model [Beyer (1990)]

3. Data Analysis

We draw the scatterplot matrix (Appendix A) for all continuous variables. The variables visibility and solar are highly skewed. The variable count is bimodal and we take a square root transformation of it. However, its distribution in the histogram plot (Appendix B) is still not normal. We will check the normality and residual plot later to see if the transformation is acceptable.

In addition to all the main effects, we also consider the interactions between categorical variables. There is no data for some level of combination of Snow vs Season, Snow vs Holiday (see interaction plots in Appendix C). For example, there is no snow in spring and summer. And there is no interaction between Weekstatus vs Holiday, Rain vs Weekstatus, and Seasons and Weekstatus (see interaction plots in Appendix D). We do not include those five interactions mentioned above in the full model. There are interactions between Holiday vs Seasons, Rain vs Holiday, Snow vs Weekstatus, and Rain vs Season. We include those four interactions (Appendix E) in our full models. And the combination of non-holiday weekdays in summer with no rain and no snow will give the highest rented bike count.

As a result, our full model would be:
Sqrt(Count) ~ Temp + Hum + Wind + Visb + Dew + Solar + Rain + Snow + Seasons + Holiday + WeekStatus + Rain:Seasons + Snow:Weekstatus + Rain:Holiday + Holiday:Seasons.

We also calculated Condition Index and VIF to check if there are multicollinearity issues. Condition Index equals 12141.6, which is much larger than 30.
And the VIF value table for the main effect variable is as below (Note: $GVIF^2$=VIF). The high VIF value of Seasons, which is a variable with four levels, can be safely ignored [ALLISON (2012)]. The high VIF values of Tem and Dew indicate that there might be collinearity between them, which makes sense in basic science. From a physical angle, the higher the temperature, the greater the capacity of the air to hold water vapor which leads to higher dew point temperature.

Variable	Temp	Hum	Wind	Visb	Dew	Solar	Rain	Snow	Seasons	Holiday
GVIF	137.87	31.25	1.35	1.84	177.23	4.54	41.61	2	32501	6.47

Table 1

Due to possible multicollinearity issues and to make the project more interesting, we focused on comparing the variable selection methods. We keep all the variables at the beginning and let the selection method choose the variables.

3.1. Ridge Regression

One of the methods we are comparing is ridge regression. First, we examine the ridge trace plot (Figure 1) and find a suitable smallest biasing constant k. We choose 0.045 since the following plot shows that it begins to stabilize at the moment about k = 0.045. Afterward, we do a regression with this value of k. From the regression result shown in Table 2, we can see that the coefficient estimate of Visibility is 0.003, and this close-to-zero coefficient means that Visibility is insignificant and should be removed from the model. Then we run the linear model with the rest and clean up large p-value predictors one by one to finalize the model. The final model for Ridge Regression is the following formula:

Sqrt(Count) ~ Hum + Dew + Solar + Rain + Holiday + Seasons + WeekStatus + Rain*Seasons

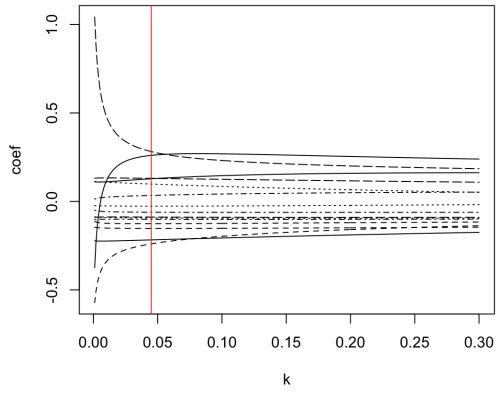


Figure 1

```
Residual Standard Error=6.0095
R-Square=0.977
F-statistic (df=16, 247)=656.2258
p-value=0
```

	Estimate	Std.Err	t-value	Pr(>|t|)
	24.2191	23.3177	1.0387	0.3000
Tmep	-0.6113	0.3356	-1.8216	0.0697
Hum	-0.5060	0.1064	-4.7550	0.0000
Wind	-0.3311	0.4188	-0.7904	0.4300
Visb	0.0003	0.0009	0.3853	0.7004
Dew	1.3446	0.3560	3.7773	0.0002
Solar	3.9887	1.6830	2.3700	0.0186
Rain	18.2778	24.4552	0.7474	0.4555
Snow	20.4227	20.5047	0.9960	0.3202
Seasons	15.4080	22.1825	0.6946	0.4880
Holiday	12.1259	20.5919	0.5889	0.5565
weekstatus	-0.8480	10.2941	-0.0824	0.9344
Rain*Seasons	-8.3570	10.2084	-0.8186	0.4138
Snow*weekstatus	-8.4407	10.2694	-0.8219	0.4119
Rain:Holiday	0.2434	18.0995	0.0135	0.9893
Seasons*Holiday	-5.4962	10.9927	-0.5000	0.6175

Table 2

3.2. Lasso Regression

Another method we used is Lasso Regression. First, we need to select the best lambda for the shrinkage penalty that produces the lowest possible MSE, which is at the left dash line in Figure 2. And the corresponding best lambda is about 0.04. From the Lasso Regression result in Table 3, we can see that the insignificant variables such as Temp and Seasons have zero coefficients now, and they were removed. But the intersection of Rain:Seasons is still in the model, so we could

not eliminate predictor Seasons. Then we run the linear model with the rest and use the backward selection method to get the final model. We get the same final model as that of Ridge Regression.

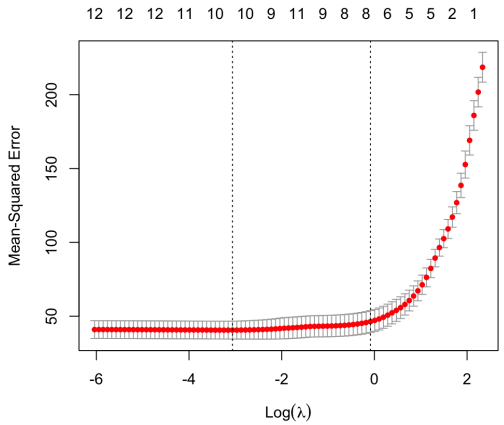


Figure 2

```
16 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)      55.9137908422
Temp              .
Hum              -0.3216256982
Wind             -0.2753999030
Visb              0.0004440715
Dew               0.6957501989
Solar             3.4671540033
Rain             -0.2126203164
Snow              3.0075425144
Seasons           .
Holiday           6.5603061297
weekstatus       -9.4042766566
Rain:Seasons     -3.8955601217
Snow:weekstatus   .
Rain:Holiday      .
Seasons:Holiday   .
```

Table 3

3.3. Other methods

We also did the backward elimination and conscientious selection. All the methods returned the same model (see regression result in Table 4):

```
Call:
lm(formula = Count ~ Hum + Dew + Solar + Rain + Holiday + weekstatus +
    Seasons + Rain:Seasons, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-25.8722  -2.9550   0.3703   3.6130  14.9003

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          52.39790    2.48912  21.051  < 2e-16 ***
Hum                  -0.30242    0.03810  -7.937 8.59e-14 ***
Dew                   0.65860    0.07989   8.244 1.20e-14 ***
Solar                 5.22549    1.78683   2.924  0.00379 **
Rain1               -31.92951    4.37813  -7.293 4.66e-12 ***
HolidayNo Holiday     6.74767    1.70399   3.960 9.95e-05 ***
weekstatusweekend    -9.57326    0.82892 -11.549  < 2e-16 ***
SeasonsSpring        -7.44269    1.32736  -5.607 5.78e-08 ***
SeasonsSummer        -9.98128    1.66554  -5.993 7.74e-09 ***
SeasonsWinter       -13.65486    1.64827  -8.284 9.21e-15 ***
Rain1:SeasonsSpring  13.81790    5.50651   2.509  0.01277 *
Rain1:SeasonsSummer  16.07013    4.96955   3.234  0.00140 **
Rain1:SeasonsWinter  25.85449    4.99474   5.176 4.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.819 on 234 degrees of freedom
Multiple R-squared:  0.8532,    Adjusted R-squared:  0.8456
F-statistic: 113.3 on 12 and 234 DF,  p-value: < 2.2e-16
```

Table 4

4. Interpretation and Predictions

We checked model prediction accuracy with the test dataset by R Squared, Adjusted R Squared, and 5-fold Cross-Validated MSE. Our model is predicted precisely since it has a large adjusted r squared being 0.8532 and a small CV-MSE being 47.2577, which means

an average error of less than 7. Also, it is without the collinearity between temperature and dew point temperature, which is shown in the basic science surrounding our data. The best model (Table 4) is shown below and it seems that Rain is the most contributing factor if uninfluenced by other variables.

$\sqrt{Count}$=52.3979-0.3024*Hum+0.6586*Dew+ 5.2255*Solar -31.9295*Rain+6.7477*No Holiday-9.5733*Weekend-7.4427*Spring-9.9813*Summer-13.6549*Winter+13.8179*Rain:Spring+ 16.0701*Rain:Summer+25.8545*Rain:Winter

We check the assumptions for the reduced model. In the Normal QQ plot (Appendix F), most of the points are located on the straight line and there is no clear pattern or curve. The residuals (Appendix G) are approximately even random distributed around y=0. The normality and equal variance assumptions are valid. In addition to this, we can also confirm that the square root transformation of variable count is acceptable.
The crPlots (Appendix H) let us see visually that all the continuous variables Hum, Dew, and Solar have a linear relationship with the response variable.
The influence Plot (Appendix I) shows that index 3019 and 4795 data are outliers, index 6523,2155,8227 and 7411 are leverages. The Cook's distance plot (Appendix I) shows us the data 2155, 6523 and 3091 are the most influential points. Then we will do the sensitivity test. We use the backward stepwise selection again without those data. It gives the model including the same predictor variables as the model before with only slightly different coefficients. And its prediction accuracy decreased from 0.8532 to 0.8084. The results of with and without those data points are consistent. As a result, we keep those data points.

5. Limitations and Future Direction

There are some limitations to our project. Firstly, we assumed the samples are independent, but there might be some correlations between the data of consecutive days. Also, the data was collected in the Asian area, which may not fit the North American market. The dataset has been analyzed by some researchers while most of them used Machine Learning methods.
In the future, we will focus on district-wise rental bike demand prediction. Because the demand for bike rental is closely related to regional factors such as population and topography. In addition, we might analyze the rented bike count on specific districts (with massive crowds) instead of the total amount of the whole city.

6. Project experience

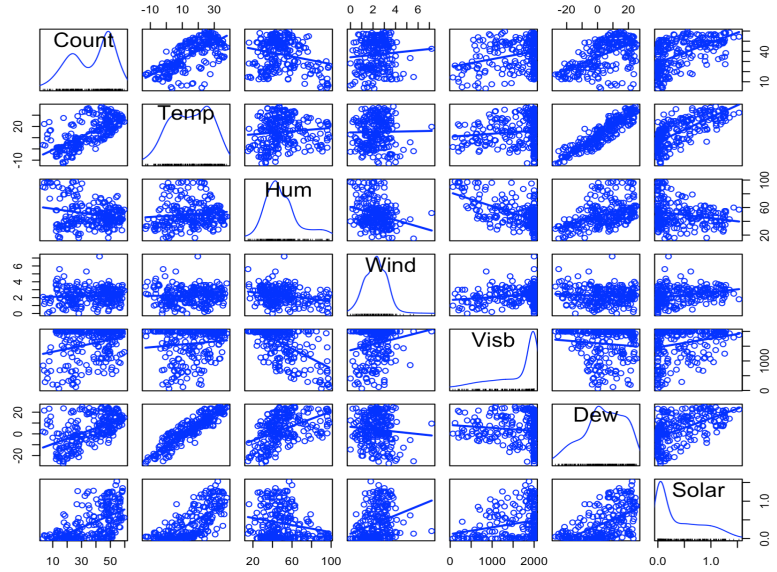
Working in a group increases productivity and improves our performance. All of my group members are so active and reliable. We selected data and set the goal of our projects together. And we cooperated with each other to write code, powerpoint and final report. The most beneficial part for me is learning how to deal with those practical problems using what we learned in class when processing the actual data. And I also have an improvement in fixing code and looking for resources to get a deeper understanding of some concepts.

Reference:

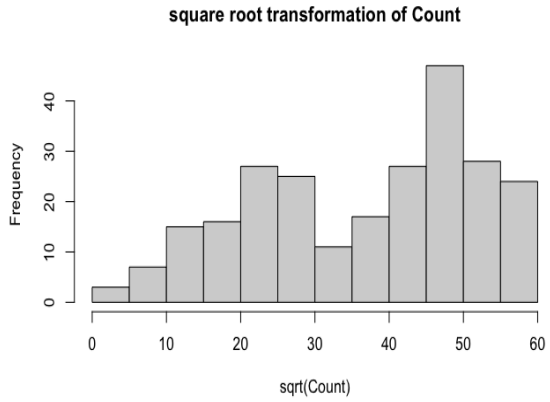
1. Dataset link: dataset link:https://data.mendeley.com/datasets/zbdtzxcxvg/2
2. E, S. V., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in Metropolitan City. *Computer Communications*, *153*, 353–366. doi:10.1016/j.comcom.2020.02.007
3. ALLISON, P. A. U. L. (2012, September 10). When can you safely ignore multicollinearity? *Statistical Horizons | Statistics training that makes sense*. https://statisticalhorizons.com/multicollinearity. Accessed 10 December 2021
4. Beyer, W. H. (1990). In CRC standard mathematical tables (pp. 536,571 and 2002). essay, Boca Raton: CRC Press.
5. Bhattacharyya, S. (2020, September 28). Ridge and lasso regression: L1 and L2 regularization. *Medium*. Towards Data Science. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b. Accessed 10 December 2021
6. NCSS. (2021) Chapter 335 Ridge Regression. *NCSS Statistical Software*. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf. Accessed 10 December 2021

Appendix:

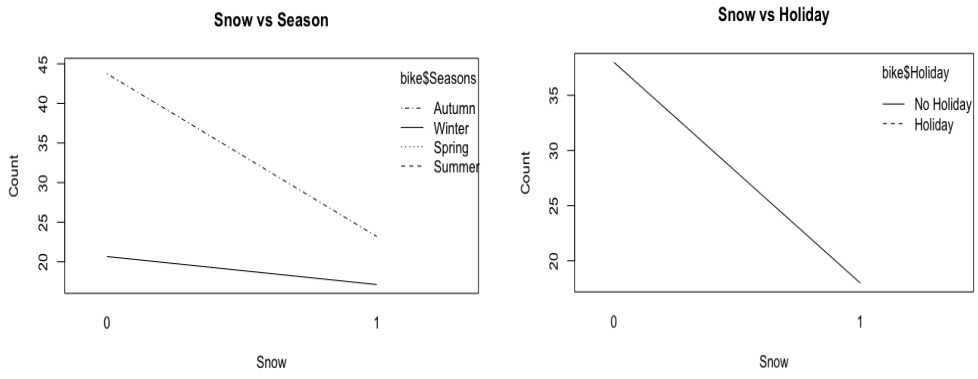
A. The scatterplot matrix for all continuous variables:



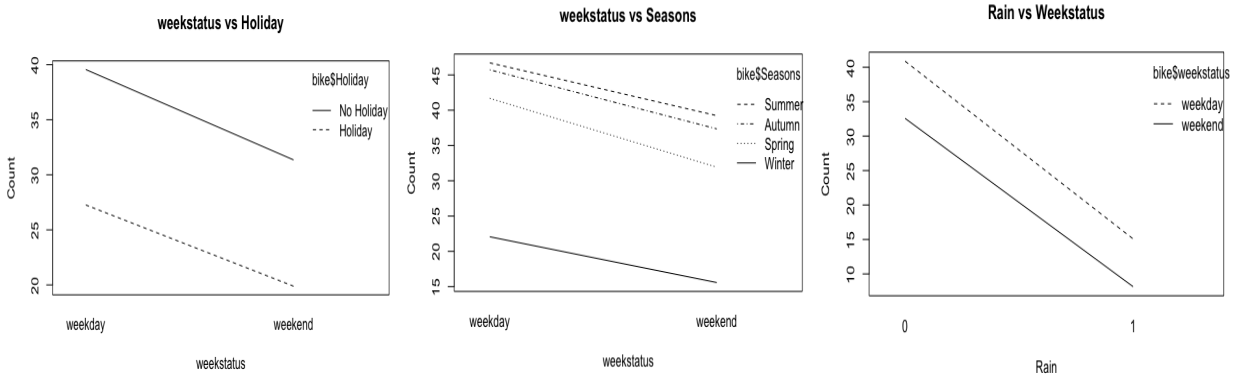
B. The histogram of Count variable after square root transformation:



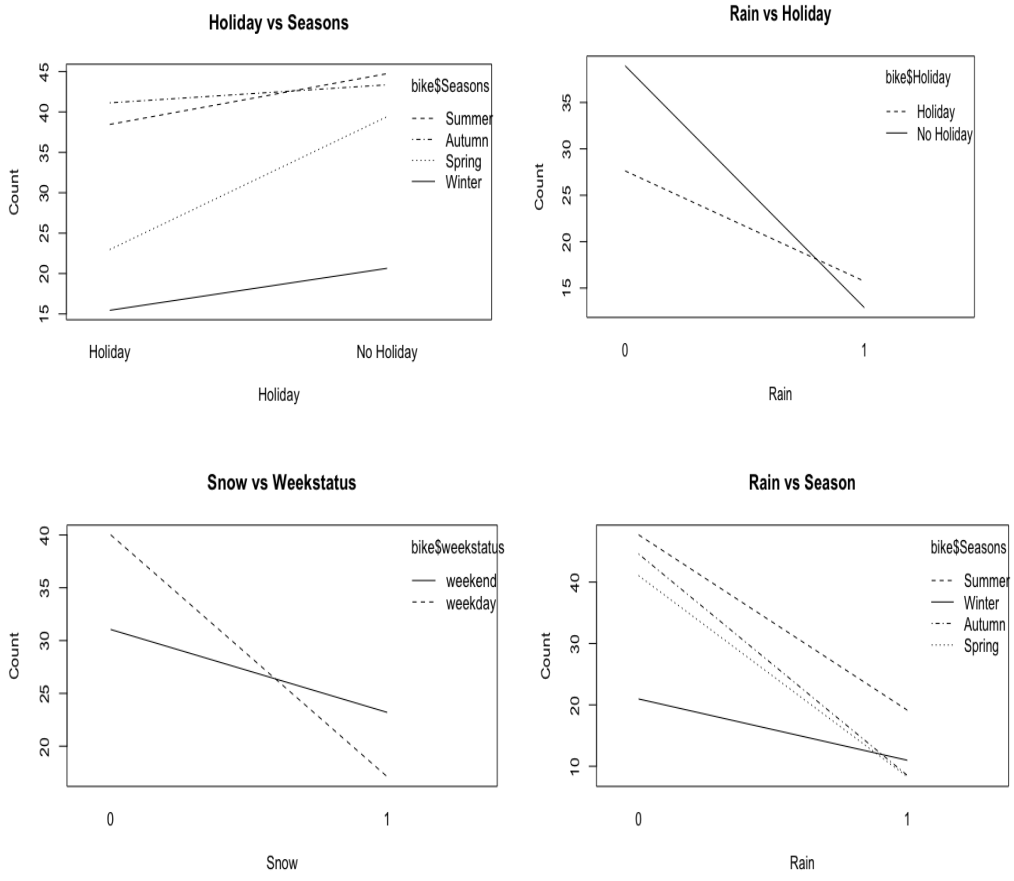
C. The interaction plots of Snow vs Season, Snow vs Holiday:



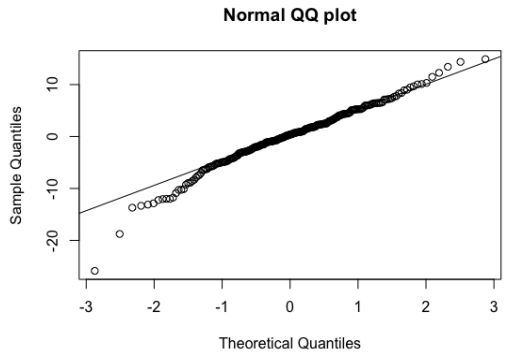
D. The interaction plots of Weekstatus vs Holiday, Rain vs Weekstatus, and Seasons and Weekstatus:



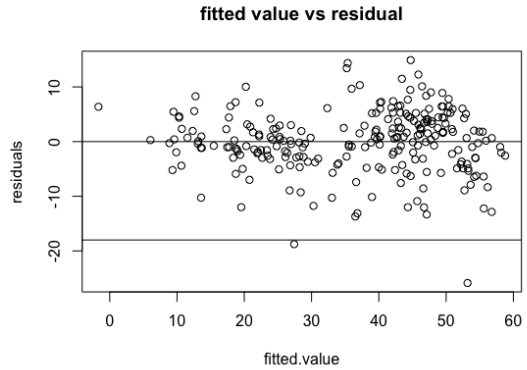
E. The interaction plots between Holiday vs Seasons, Rain vs Holiday, Snow vs Weekstatus, and Rain vs Season:



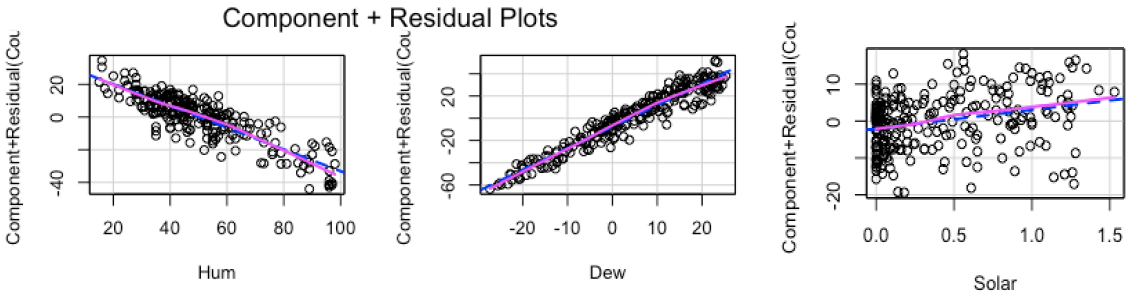
F. Normal QQ plot:



G. Residuals plot:



H. crPlots for all continuous variables in reduced model



I. Cook's distance and Influence Plot