TIME SERIES ANALYSIS BY USING BICYCLE RENTAL DATA

2021150416 통계학과 전예린

Abstract

During my recent trip to Japan, I have never seen a place where bicycles did not pass by. Bicycles are widely used since they make transportation convenient for many people. In Seoul, an unmanned bicycle rental system called '따릉이' was began in 2015, and usage is increasing every year. Currently, efforts are being made to improve the '따릉이' system, including the creation of a bicycle rental app and improvements to improve the convenience of bicycle rental and return. Among the factors that affect bicycle rental, time and temperature play a greater role than other variables. In addition, the importance of variables that affect the amount of bicycle rentals can be determined. Among models such as random forest regression and Holt and Winter's model for predicting rental count, the exponential smoothing model has been evaluated as having high performance, so it can be used in the future when selecting a model for predicting the amount of rentals.

Introduction

According to Seoul City's public bicycle rental history information data from the '서울시 열린 데이터 광장' in 2017, Mapogu was analyzed as the place with the highest amount of public bicycle rentals in Seoul. Accordingly, the analysis was conducted using data from Daycon, which has refined data on '따릉이' rental count by time and weather conditions in Mapo-gu. Among the variables in the data, precipitation is a binary variable and is 97% biased toward the value of 0 (no rain), so it is not to be given much significance in the interpretation. Additionally, unlike other variables, ozone and visibility are not information that people consider or can easily access when deciding whether to engage in outdoor activities, so they were excluded from direct analysis.

I'll analyze the rental count of bicycle by time of day and weather condition, and evaluate the accuracy of rental count prediction using tslm regression and random forest regression with count as the target variable. Also, I'll analyze the seasonal, trend, and remainder effects over time of the rental amount using time series deocomposition, and apply exponential smoothing and holt and winter models to compare model accuracy to select a model with good performance. Through this, we can identify weather conditions and come up with a plan for managing '따릉이' rentals by time zone.

Application of Statistical method and Results

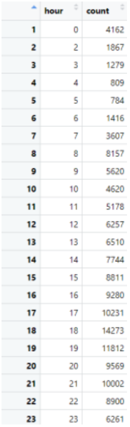
Before comprehensively considering the environmental factors that affect the rental amount of bicycle, conduct a one-to-one correspondence analysis between each factor and rental amount. To conduct the analysis, all missing values were removed and the variable names were kept as id, hour, hour_bef_temperature, hour_bef_precipitation, hour_bef_windspeed, hour_bef_humidity, hour_bef_visibility, hour_bef_ozone, hour_bef_pm2.5, hour_bef_10. When exploratory analysis of the data was performed, all variables except hour_bef_precipitation were continuous variables, had no major outliers, and generally satisfied normality, so it was proceeded as no more preprocessing.

Ⅰ. *Exploratory data analysis between bicycle rental amount and environmental variables*

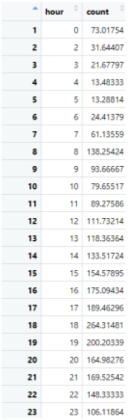
Ⅰ. *Bicycle rental by hour*

The first picture is the number of bicycle amount by hour, and the second is the average number of bicycle rentals by hour. The average number of bicycle rentals by hour is the highest at 18:00 with 264.3 units, and the lowest at 13 between 4 and 5 o'clock. On average, the amount of rentals is lower in the afternoon than in the morning, and is higher at 8 o'clock in the morning. If you check this with a bar plot, it is as follows.

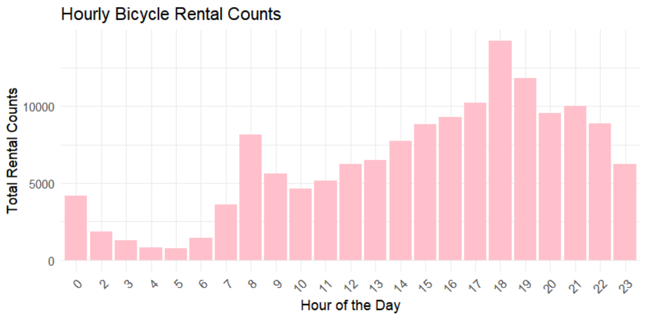
If we analyze the cause of the pattern of bicycle rental amount according to time of day, it can be expected that rental amount is high at 8 o'clock, as it is a time when activity levels are high during the morning hours, such as commuting to work and going to school. The rental number peaks at 8 o'clock in the morning and decreases, and continues to increase from 11 o'clock in the afternoon, peaking again at 18 o'clock and decreasing, which is when the rental amount is concentrated on returning to daily life after activities or going home.



	hour	count
1	0	4162
2	2	1867
3	3	1279
4	4	809
5	5	784
6	6	1416
7	7	3607
8	8	8157
9	9	5620
10	10	4620
11	11	5178
12	12	6257
13	13	6510
14	14	7744
15	15	8811
16	16	9280
17	17	10231
18	18	14273
19	19	11812
20	20	9569
21	21	10002
22	22	8900
23	23	6261

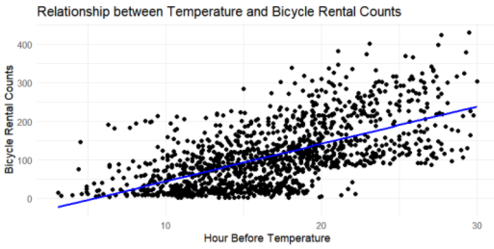


	hour	count
1	0	73.01754
2	2	31.64407
3	3	21.67797
4	4	13.48333
5	5	13.28814
6	6	24.41379
7	7	61.13559
8	8	138.25424
9	9	93.66667
10	10	79.65517
11	11	89.27586
12	12	111.73214
13	13	118.36364
14	14	133.51724
15	15	154.57895
16	16	175.09434
17	17	189.46296
18	18	264.31481
19	19	200.20339
20	20	164.98276
21	21	169.52542
22	22	148.33333
23	23	106.11864



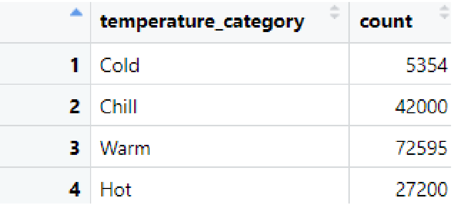
Ⅱ. *Bicycle rental by temperature*

To understand the number of bicycle rentals according to temperature, a scatterplot was drawn with temperature on the horizontal axis and rental volume on the vertical axis. The correlation coefficient between temperature and rental amount is 0.604439, and looking at the scatter plot, there appears to be a positive correlation to some extent.

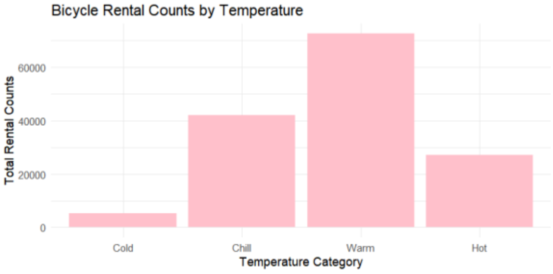


Ⅲ. Bicycle rental by percipitation

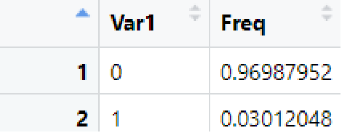
However, in order to understand whether bicycle rental number is high or low in a specific temperature range, converting hour_bef_temperature into a categorical variable is need. The temperature is based on Korea's average temperature of 17 degrees, and considering that the hour_bef_temperature category is [3.1, 30.0], the temperature is [3,10], [10,17], [17, 24], [17, 30]. It was divided into four categories and each was defined as 'cold', 'chill', 'warm', and 'hot'. As a result of calculating the number of users for each temperature category, the rental amount is calculated to be 5354, 42000, 72595, and 27200 for the temperature categories defined as cold, chill, warm, and hot, respectively, and for [17, 24], which corresponds to warm, the bicycle rental amount is calculated as [17, 24]. It can be seen that the rental amount is relatively large. Considering there is a lot of chill, the amount of bicycle rental is low when it is cold or hot (i.e., when the temperature is extremely low or too high). Below is a picture shown as a bar plot.



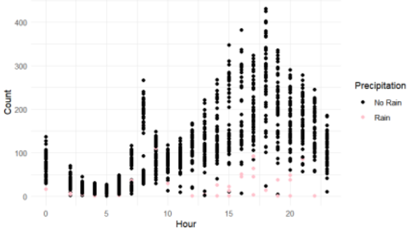
	temperature_category	count
1	Cold	5354
2	Chill	42000
3	Warm	72595
4	Hot	27200



In order to determine the rental amount according to whether or not it rained, we calculated the ratio between rain and non-rain, and found no rain (marked as 0) corresponds to about 97%, and rain (marked as 1) corresponds to about 3%. This means that 97% of bicycle rentals are on days when it is not raining, and 3% of bicycle rentals are on days when it rains. The figure below shows whether or not it will rain in the scatterplot of bicycle rental amount over time.



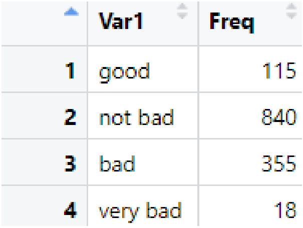
	Var1	Freq
1	0	0.96987952
2	1	0.03012048



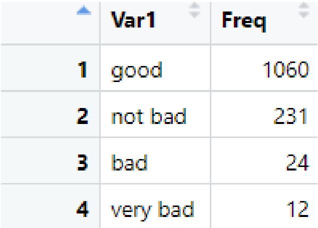
IV. Bicycle rental by fine dust

In fact, it is difficult to say among the variables, hour_bef_ozone or hour_bef_visibility are factors that cause people to refrain from outdoor activities as they can easily recognize that they may have a negative impact on health. However, fine dust concentration is located at the top along with temperature in weather apps, so it is one of the weather factors that people check easily and frequently. Therefore, there may be a fairly direct correlation with the amount of bicycle rental, so

I analyzed the bicycle rental amount according to the level of fine dust. The hour_bef_pm10, which corresponds to fine dust of 1/5 to 1/7 of a hair size, has a range of [9,269], and hour_bef_pm2.5, which corresponds to fine dust of 1/20 to 1/30 of a hair size, has a range of [8,90]. As with the temperature analysis, categorization was performed for each fine dust, and the poor level of air quality was divided according to the concentration of dust and the corresponding bicycle rental amount was obtained.



	Var1	Freq
1	good	115
2	not bad	840
3	bad	355
4	very bad	18

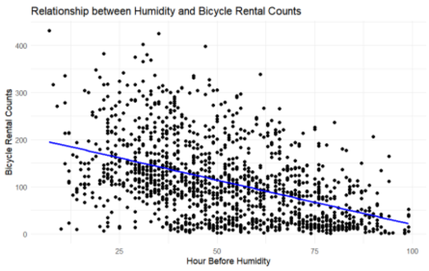


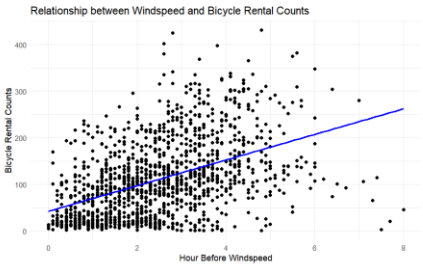
	Var1	Freq
1	good	1060
2	not bad	231
3	bad	24
4	very bad	12

Looking at the two tables above, when both pm10 and pm2.5 are in the very bad category, the bicycle rental amount is low. In addition, the amount of bicycle rentals is higher when PM10 is good than PM2.5. This may be because the category of fine dust is broader, but people search for information more frequently for fine dust(here, pm10) than ultrafine dust(here, pm2.5) and choose outdoor activities based on fine dust. Through a future analysis of the explanatory degree of each variable, we plan to find out which of the two contributes more to explaining the amount of bicycle rental.

V. Bicycle rental by Humidity and Windspeed

Among the remaining environmental variables, humidity has a negative correlation with the count variable with a correlation coefficient of -0.4591494, and windspeed has a positive correlation with a correlation coefficient of 0.4580833. The amount of bicycle rental by hour appears to be large when humidity is relatively low and windspeed is high.

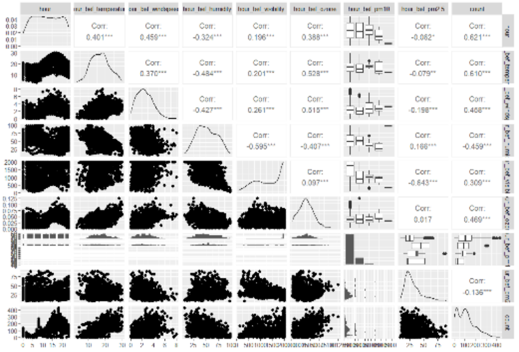




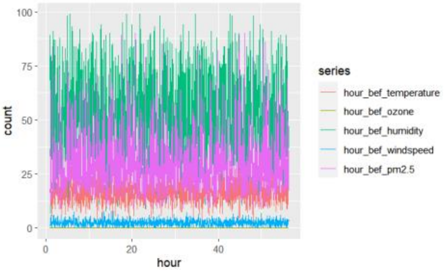
Ⅱ. Regression on time series data

Ⅰ. Regression with target variable as bicycle count

To do regression using the bicycle rental amount as the target variable, the correlation between variables was first confirmed. Since the precipitation variable is a binary, it was removed along with id. Although there is a lot of data, most variables do not have a clear correlation, but there are many variables that have a linear correlation with the count variable.

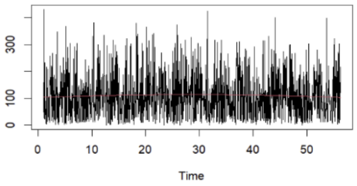


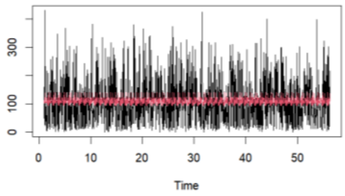
The autoplot function in the fpp2 library was used to determine how the number of bicycle rentals appears over time depending on temperature, ozone, humidity, and wind, which showed high explanatory power. In order to apply it to the autoplot function, bicycle data was transformed into a time series object with a period of 24.



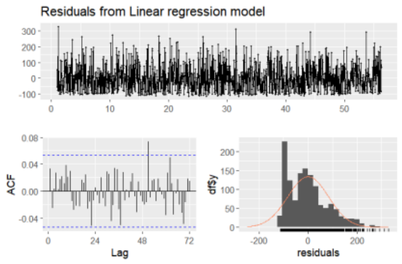
Looking at the plot, it appears that the count number has changed during conversion to time series data. However, when looking at changes over time in factors such as temperature, ozone, humidity, wind, and fine dust, the changes over time in temperature, ozone, and humidity appear similar.

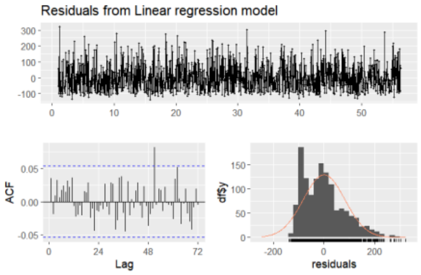
We converted the data into ts objects over time and plotted a ts.plot of rental volume over time using a naive regression approach. There is no trend of increasing or decreasing, and this is expected to be the result of analyzing rental volume over time on a daily basis, not on a yearly basis. However, naive regression approach cannot afford to explain oscillation by heterogeneity of variance. Therefore, cycle should be considered additionally. Before that, we will test the residuals of the fit using a naive regression approach.





Analyzing the residual test results for each regression fit, the naive regression approach does not show serial correlation except for residuals with lag 51, and the p-value in the Breusch-Godfrey test is very large at 0.8792, so there is no serial correlation. In addition, the results of the residual test fitted by adding a cycle element also show no serial correlation in the residuals except for lag 51, consistent with the application of the naive regression approach, and the p-value in the Breusch-Godfrey test is very large at 0.8997, so there is no serial.



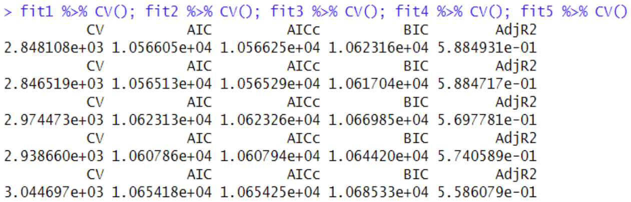


Comprehensively analyzing the results of the two regression models considering the naive approach and the cyclic component, it can be seen that the characteristic appearance of a large or small number of bicycle rentals is better represented by the aggregation of the number of bicycle rentals by hour rather than a time series model.

Ⅱ. Model selection and forest

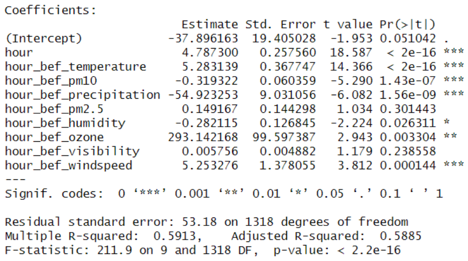
We'll find out the explanatory power by applying tslm to each variable. As a summary result of applying tslm with the count variable as the target, excluding the time variable (the time variable explains 38.52%), the variable with the lowest explanatory power among environmental factors is fine dust, and the highest explanatory power is temperature, which explains 37.22%. Temperature, ozone, humidity, and wind have the highest explanatory power in that order. As a result of data exploratory analysis, the precipitation variable is a dummy variable that is 1 if it rains, and 0 otherwise. Since 97% of days do not rain, it is thought that it would not be appropriate to explain the number of bicycle rentals even if the R^2 value appears to some extent. do. In addition, it is expected that the variable importance will be low.

Next, we create several regression models by excluding variables that had low explanatory power when considering only a single variable, and select the best model through the CV function. The figure below shows the model considering all variables in order from fit1 to fit5, removing hour_bef_pm2.5, removing hour_bef_pm10 and hour_bef_humidity, removing hour_bef_ _pm2.5, hour_bef_precipitation, hour_bef_visibility, hour_bef_windspeed, hour_bef_pm10, hour_bef_pm2.5, hour_bef_precipitation, hour_bef_vis ibility, hour_bef_windspeed.

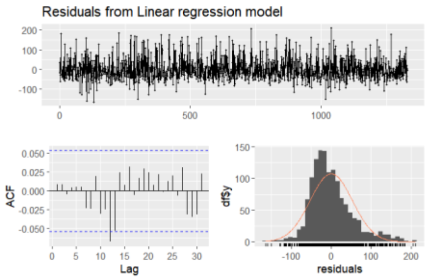


```
> fit1 %>% CV(); fit2 %>% CV(); fit3 %>% CV(); fit4 %>% CV(); fit5 %>% CV()
          CV          AIC         AICc          BIC        AdjR2
2.848108e+03 1.056605e+04 1.056625e+04 1.062316e+04 5.884931e-01
          CV          AIC         AICc          BIC        AdjR2
2.846519e+03 1.056513e+04 1.056529e+04 1.061704e+04 5.884717e-01
          CV          AIC         AICc          BIC        AdjR2
2.974473e+03 1.062313e+04 1.062326e+04 1.066985e+04 5.697781e-01
          CV          AIC         AICc          BIC        AdjR2
2.938660e+03 1.060786e+04 1.060794e+04 1.064420e+04 5.740589e-01
          CV          AIC         AICc          BIC        AdjR2
3.044697e+03 1.065418e+04 1.065425e+04 1.068533e+04 5.586079e-01
```

The fit5 has the largest adjusted R^2, but fit1, which has the next largest adjusted R^2, has the smallest CV, AIC, corrected AIC, and BIC, so the model that considers all variables is the best.



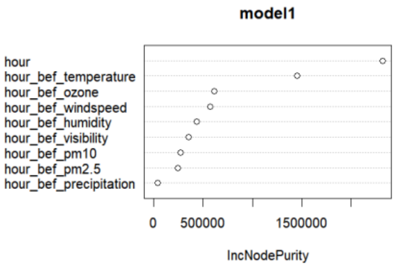
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -37.896163  19.405028  -1.953 0.051042 .  
hour                      4.787300   0.257560  18.587  < 2e-16 ***
hour_bef_temperature      5.283139   0.367747  14.366  < 2e-16 ***
hour_bef_pm10            -0.319322   0.060359  -5.290 1.43e-07 ***
hour_bef_precipitation -54.923253   9.031056  -6.082 1.56e-09 ***
hour_bef_pm2.5            0.149167   0.144298   1.034 0.301443    
hour_bef_humidity        -0.282115   0.126845  -2.224 0.026311 *  
hour_bef_ozone          293.142168  99.597387   2.943 0.003304 ** 
hour_bef_visibility       0.005756   0.004882   1.179 0.238558    
hour_bef_windspeed        5.253276   1.378055   3.812 0.000144 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.18 on 1318 degrees of freedom
Multiple R-squared:  0.5913,	Adjusted R-squared:  0.5885 
F-statistic: 211.9 on 9 and 1318 DF,  p-value: < 2.2e-16
```



The explanatory power of fit1 is 48%, and since the p-value is large, there is no serial correlation between residuals, and it is random as it does not show a specific pattern. Additionally, although the residuals are slightly skewed to the right, they are normal. As a result of checkresiduals function, the p value is greater than 0.5, so the null hypothesis that there is serial correlation between residuals is rejected, so there is no correlation between residuals, and it is stationary.

Ⅲ. *Random Forest Regression*

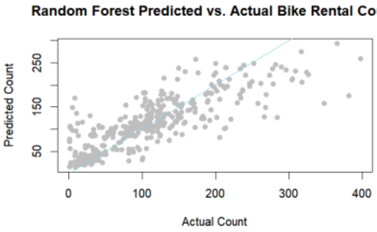
This time, regression was performed using randomforest. Random Forest is an ensemble learning method that builds a multitude of decision trees during training and outputs the average prediction (for regression) of the individual trees. It is based on the idea of bagging (bootstrap aggregating) and introduces randomness to improve the predictive performance and control overfitting. Also, Random Forest provides a feature importance score, allowing to identify which features contribute most to the model's predictions.



The dataset is divided and predicted using random forest regression using the count variable as the target variable. First, bicycle data is divided into train data and test data at a ratio of 7:3, which is the most commonly used ratio when dividing data. Next, a forecast is performed using the randomForest library with the count variable as the target variable. First, as a result of checking the results of the regression model fit with tslm using the CV function, we would like to train a model including all variables that showed high performance in the indices of CV, AIC, BIC, AICc, and R^2 using random forest regression.

To create a new model by reducing variables and compare performance, variables that appear to be highly important are identified in the random forest regression results including all variables. The importance appears to be high in the following order: hour, temperature, ozone, windspeed, humidity, visibility, pm10, pm2.5, and precipitation. Among these, we would like to create a new model using the four variables that were found to be highly important.

The model was fit again using four variables, and the model evaluation results for the train data and test data were as follows: MSE, MAE, and RMSE, respectively. The MSE, MAE, and RMSE for the train data are low, indicating that the error is less and the accuracy is higher.

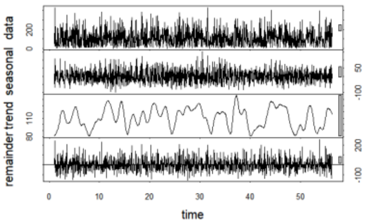


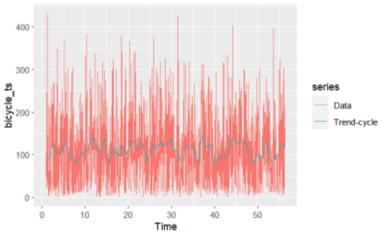
The actual number of bicycle rentals and the predicted number of rentals are consistent to some extent, and when it exceeds about 150 units, there is a difference between the predicted and actual number of rentals. In addition, it can be seen that the predicted number of rentals does not have a distribution of 300 to 400 units as much as the actual number of rentals.

Additionally, we would like to calculate the MSE and MAE for the test data through 10-fold cross validation. As a result of 10-fold cv, the test cv MSE was 1548.293 and the test CV MAE was 27.53347, which was lower than the MSE and MAE of 1990.798 and 30.9233 when 10-fold-cv was not performed, indicating that performance was improved.

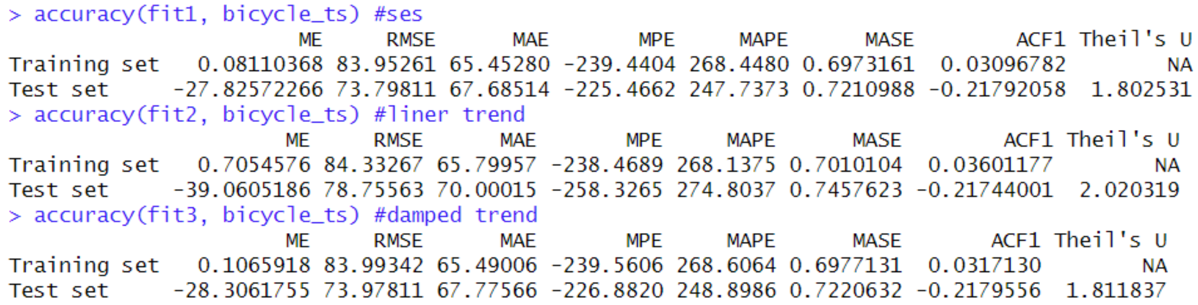
Ⅲ. Time series decomposition and various models

Next, we would like to proceed with time series decomposition to analyze the seasonal, trend, and remainder effects of bicycle rental over time. The count variable was decomposed into a time series object with a period of 24 and decomposition was performed.





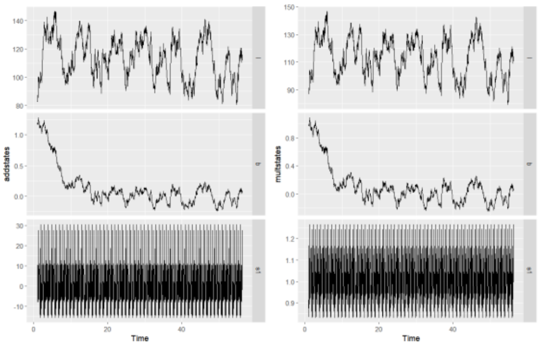
Next, we will compare the performance of simple exponential smoothing, Holt's linear trend, and Holt's damped trend models.



```
> accuracy(fit1, bicycle_ts) #ses
                      ME     RMSE      MAE       MPE     MAPE      MASE        ACF1 Theil's U
Training set   0.08110368 83.95261 65.45280 -239.4404 268.4480 0.6973161  0.03096782        NA
Test set     -27.82572266 73.79811 67.68514 -225.4662 247.7373 0.7210988 -0.21792058  1.802531
> accuracy(fit2, bicycle_ts) #liner trend
                     ME     RMSE      MAE       MPE     MAPE      MASE        ACF1 Theil's U
Training set   0.7054576 84.33267 65.79957 -238.4689 268.1375 0.7010104  0.03601177        NA
Test set     -39.0605186 78.75563 70.00015 -258.3265 274.8037 0.7457623 -0.21744001  2.020319
> accuracy(fit3, bicycle_ts) #damped trend
                     ME     RMSE      MAE       MPE     MAPE      MASE       ACF1 Theil's U
Training set   0.1065918 83.99342 65.49006 -239.5606 268.6064 0.6977131  0.0317130        NA
Test set     -28.3061755 73.97811 67.77566 -226.8820 248.8986 0.7220632 -0.2179556  1.811837
```

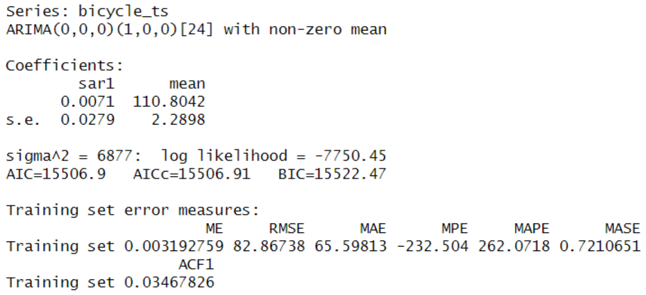
Comparing the RMSE, MAE, MASE, etc. of each model, the simple exponential smoothing model has the best performance. Additionally, when exponential smoothing is performed using the ets function, the model's accuracy is RMSE of 90.76602 and MAE of 68.71932.

Changes in bicycle rental amount over time were confirmed using the additive model and multiplicative model. Also, I performed Holt-Winters exponential smoothing on a time series on bicycle_ts. The seasonal component is specified as both "additive" and "multiplicative." After fitting the models, using autoplot visualized the states of the models.



Two models have similar aspects, but multiplicative winters exponential smoothing model has more oscillated seasonal effect in the first season.

For the Arima model's training set, RMSE, MAE, MASE, etc. appear less than ses, holt and winter's linear trend, and holt and winter's damped trend.



```
Series: bicycle_ts
ARIMA(0,0,0)(1,0,0)[24] with non-zero mean

Coefficients:
        sar1      mean
      0.0071  110.8042
s.e.  0.0279    2.2898

sigma^2 = 6877:  log likelihood = -7750.45
AIC=15506.9   AICc=15506.91   BIC=15522.47

Training set error measures:
                      ME     RMSE      MAE      MPE     MAPE      MASE
Training set 0.003192759 82.86738 65.59813 -232.504 262.0718 0.7210651
                   ACF1
Training set 0.03467826
```

As a result of applying various analysis techniques and analyzing variables, it appears that there is no significant volatility over time or any significant increase or decrease trend.

Discussion

As a result of examining the influence and relationship of variables that affect the amount of "따릉이" rental, it was high during frequent commuting times, when it is not raining, when fine dust is good, and when the temperature is warm, which seems to be consistent with our general logic. Surprisingly, there did not appear to be much use during times when the wind did not blow much. Through this, I realized that it would be better to use off-peak hours when there is something that needs to be done during the day, such as servicing the bicycle system, and to select areas that are highly affected by environmental factors that indicate high usage and relocated the bicycles. It will contribute to further supplementing the bicycle rental system by applying such methods.

To explain the amount of bicycle rental, we applied several models such as tslm, random forest regression, simple exponential smoothing, Holt and winter's liner trend and damped trend, and arima model. The model that explained the count well was a model that included all variables, and the temperature variable has the highest explanatory power. This means that the emperatures have a significant impact on whether or not to use a bicycle. In addition, as a result of comparing the seasonal, trend, and remainder effects of time series data by dividing them into additive and multiplicative models, the large oscillation in the seasonal effect shows that the seasonal effect almost does not exist. By dividing the data into train and test using random forest regression, the model was trained and the accuracy was analyzed, and the performance of model prediction through train data was found to be higher. In addition, the performance of the model predicted by applying the simple exponential smoothing technique and the model predicted by applying the arima model showed relatively high performance.

By categorizing temperature and fine dust, we calculated the amount of bicycle rental according to temperature, and air quality. In this process, the category value was divided by the number of classes based on the average temperature to determine the categories. However, it was not appropriate to use methods such as calculating the categories or dividing the total range of fine dust by the number of categories. I think that if the facts about temperature and fine dust had been further investigated and criteria for more specific and objective, different results might have been obtained.

In addition, if we analyze data containing differentiated data to select regions for bicycle reallocation, we expect to be able to identify the amount of bicycle deployment by region necessary to improve the efficiency of differentiated bicycle deployment in Seoul.