STAT 330 Final Project

Ketherine Wang Sam Lee

Abstract

In this analysis we explore which factors primarily drive short-run fluctuations as they depend on Seasonality. We examine the changes in temperature as well as the demand for bike rentals in Seoul, South Korea. We conducted multiple linear regression using weather predictors from 2018 hourly data. We found that the the marginal effect of solar radiation on temperature change was most significant from 8:00 PM to 9:00 PM. Additionally, the marginal effects of solar radiation had a negative affect on bike demand; however, the interaction effects during peak demand times (rush hour times, for example), proved to be significantly positive.

1 Problem and Motivation

We are examining data from Seoul, Korea, to understand how weather and time influence bike sharing and temperature changes in the city. This data, which includes details about whether conditions and bike rentals over 8,456 days, is crucial for urban planning and environmental studies. Our goal is to identify which factors most affect bike usage and temperature fluctuations. This information is important because it helps city planners make better decisions about transportation and infrastructure, especially as cities around the world are focusing more on sustainable living and environment challenges. By understanding these patterns, we can help make cities like Seoul more efficient and responsive to both their climate and the needs of their residents.

1.1 Data Description

In this analysis we used data from UC Irvine's Machine Learning's Repository. We accessed Seoul Bike Sharing Data, which contains 8,760 rows of hourly data pertaining to the corresponding amount of rented bikes (bike demand) in Seoul, Korea. 8,465 rows are considered "functioning days," where bikes are able to be rented. In order to consider our problem of interest, we will only consider these 8,465 rows to model the bike demand in Seoul. This data set also contains corresponding data for the respective temperature (celcius), humidity (%), wind speed (m/s), visibility (10m), dew point temperature (celcius), solar radiation (MJ/m^2), rainfall precipitation (mm), snowfall precipitation (cm), seasons (Winter, Spring, Summer, Autumn), and an indicator determining whether the corresponding day is a holiday.

We will use these variables along with a subset of interaction terms which will be determined by elastic net regression to include in our models.

1.2 Questions of Interest

1. Can we use factors in weather and time to determine the best predictors of short-run for fluctuations in temperature for the climate in Seoul, Korea?
2. Can we determine which weather and time factors affect the bike sharing demand the most in Seoul, Korea?

If there are significant predictive factors, which one has the largest influence?

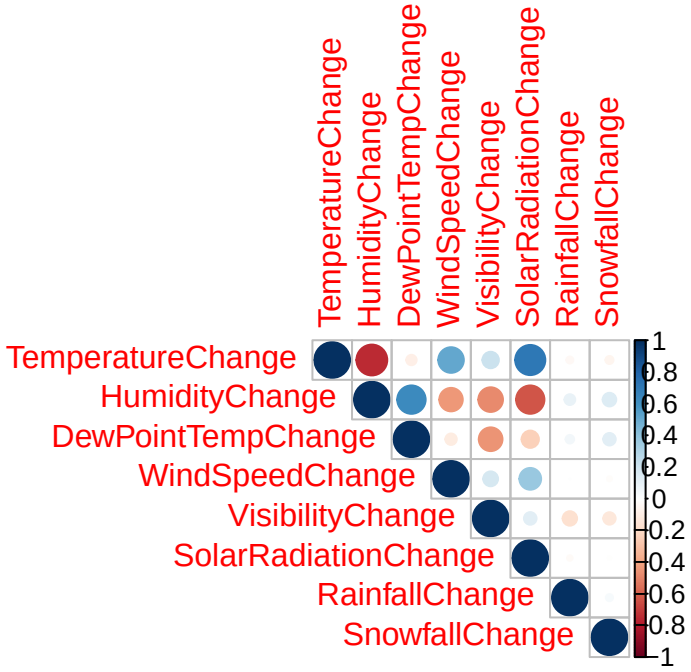
1.3 Regression Methods & Model Determination/Selection Procedures

We decided to run a first differences model to model the fluctuations in short-run temperature. where $\Delta\text{Temperature}_t = \text{Temperature}_t - \text{Temperature}_{t-k}$. We found that the optimal choice of k was 5 (see A.0 for optimal selection process on k).

Thus, we wish to estimate,

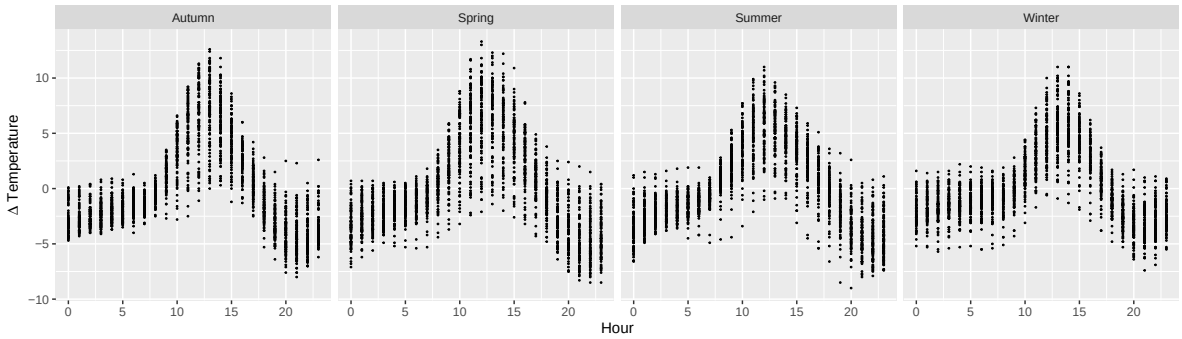
(1) $\Delta\text{Temperature}_{st} = \beta_1\Delta\text{Humidity}_{st} + \beta_2\Delta\text{DewPointTemp}_{st} + \beta_3\Delta\text{WindSpeed}_{st} + \beta_4\Delta\text{SolarRadiation}_{st} + \beta_5\Delta\text{Rainfall}_{st} + \beta_6\Delta\text{Snowfall}_{st} + \Delta\beta_7 I(\text{Hour}_t = 1) + ... + \Delta\beta_{29} I(\text{Hour}_t = 23) + \Delta\eta_{st}$

Where $\Delta\text{Temperature}_{st} = \text{Temperature}_{st} - \text{Temperature}_{t-5}$



With our basic multilinear model (1) the correlation matrix indicates that there may be some potential multicolinearity between $\text{DewPointTempChange}_{st}$ and $\text{HumidityChange}_{st}$. For the most part, the indicators look good with changes in $\text{SolarRadiation}_{st}$ potentially affecting flucutations in temperature the most.

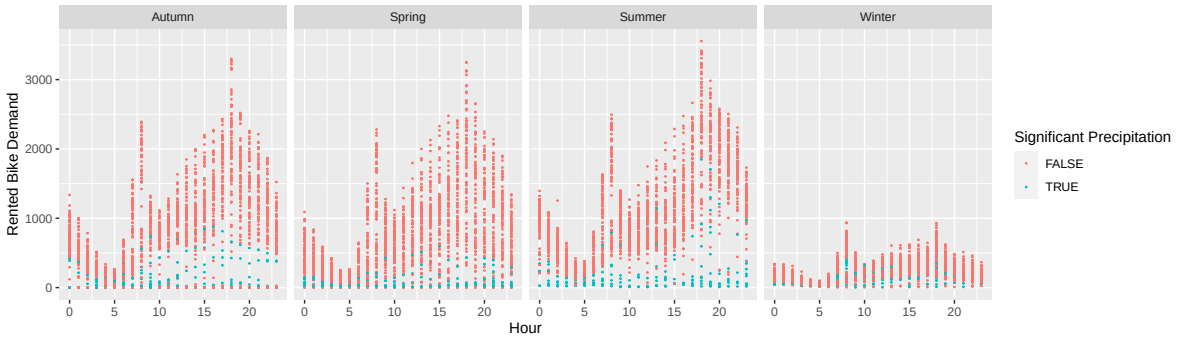
Change in Temperature Across Hours of the Day by Season



According to our model, changes in temperature not only depend on temperature but also on season. While there are some general linear trends across some ranges of hours, the non-linearity of the hours in general with respect to temperature change led us to include indicators for each hour as specified in Model (1) using Hour 0 (12:00 a.m.) as our base year.

We used elastic net regression to select interaction terms to see if the influential points (and thus, Normality) could be remedied. We used elastic net regression over lasso regression due to the potential colinearity between $\text{HumidityChange}_{st}$ and other weather effects. (See A.2.0 for model output)

Similar to our model for temperature, we have unobserved seasonal effects in ϵ. Instead of using a kth-difference approach, we will use a fixed effects models to include the seasonal effects for $s-1$ (3) seasons to obtain unbiased estimates. We will use Spring as our base season. However, exploratory data analysis has shown that current bike demand is strongly dependent on the bike demand from previous hours. Hence, in order to include seasonal dummy variables, we also need to include a set of lagged effects (δ_p) conditional on Season on $\text{RentedBikeCount}_{st} \quad \forall s \in \text{Seasons}, s \neq \text{Spring}$ to adjust for the time dependency. (See A.7 to see how we chose our set of lagged effects).



Preliminary EDA shows that precipitation (rain and snow) will have a significant negative effect on rented bike demand.

Hence, we will first estimate the hourly demand for rented bikes in Seoul with the following model (see A.8.0 for full model description),

(3) $\text{RentedBikeCount}_{st} = \beta_0 + \beta_1 \text{Humidity}_{st} + \beta_2 \text{Temperature}_{st} + ... + \beta_{33} I(\text{Season}_s = \text{Winter}) + \beta_{34} I(\text{Season}_s = \text{Summer}) + \beta_{35} I(\text{Season}_s = \text{Autumn}) + \delta_1 \text{RentedBikeCount}_{st-1} + \delta_{23} \text{RentedBikeCount}_{st-23} + \sum_{p=1}^{9} \delta_p + \eta_{st}$

$\eta_{st} \overset{iid}{\sim} N(0, \sigma^2)$

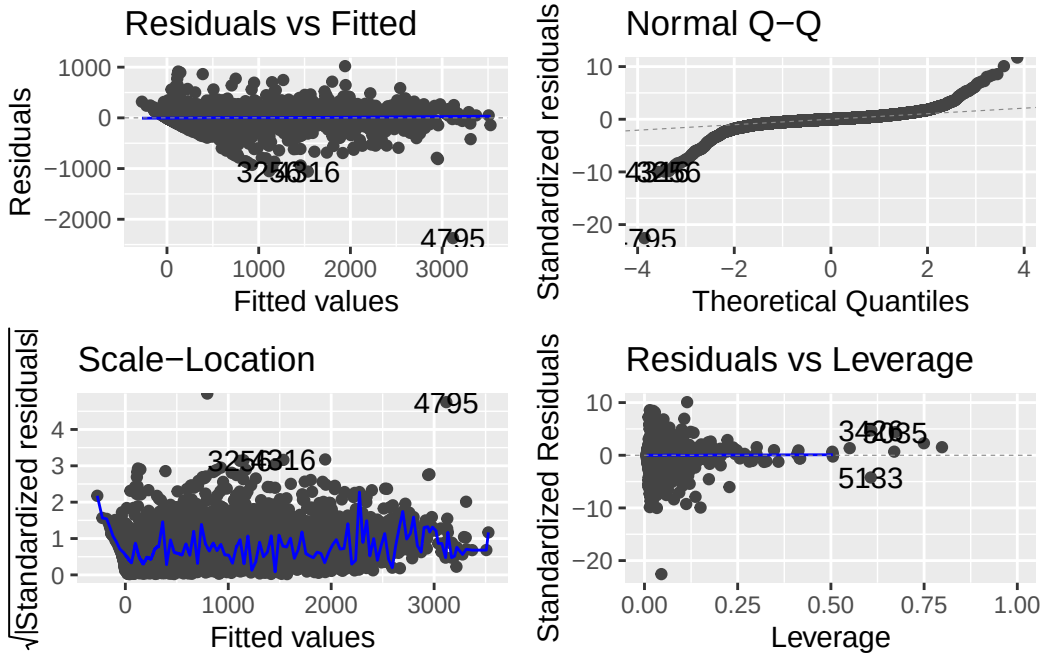
Diagnostics for fitting this model to this data indicate that linearity is met A.8. Additionally, due to the large sample size, the leverage plot also indicates that there were not any influential points. However, since homoskedasticity was blatantly violated, especially for fitted values < 1000, we modified the model by including interaction terms.

We ran an elastic net regression model to control for colinearity. (See model A.9 for model output and diagnostics). Since $\text{RentedBikeCount}_{st}$ is likely poisson-distributed, we considered log-transforming the response, but diagnostics show that model assumptions did not improve A.10.0.

However, we noticed non-linear trends in Humidity_{st} against $\text{RentedBikeCount}_{st}$ and introduced categorical transformations A.10.1. Using an inverse response plot, we also found that we could optimally transform Visibility_{st} using a *log* transformation A.10.2. Lastly, we transformed WindSpeed_{st} since we noticed that extreme values were causing some non-linearity between the response. Hence, we introduced a categorical variable called ExtremeWind_{st} to capture this effect A.10.3.

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 8 rows containing missing values (`geom_line()`).
```



4. Analyses, Results and Interpretation:

Thus, using our elastic regression results, we will modify Model (1) to include the following interaction terms. We made sure to include the lower-order regression coefficients for the coefficients that elastic net regression left out when elastic net included interaction terms involving a lower-order term. We are interested in estimating the most assailant factors (A.5). See A.6 for the full model.

(2) $\Delta\text{Temperature}_{st} = ...+\beta_5\Delta\text{SolarRadiation}_{st}+...+\beta_{32}\Delta\text{Humidity}_{st}\text{SolarRadiation}_{st}+...+\beta_{39}\Delta\text{SolarRadiation}_{st}\text{Rainfall}_{st}+ ... + \sum_{n=1}^{23}\beta_{131+n}\Delta\text{SolarRadiation}_{st}\Delta I(\text{Hour}_t = n) + ... + \Delta\eta_{st}$

$\Delta\eta_{st} \overset{iid}{\sim} N(0,\sigma^2)$

Using an elastic net regression method to determine once again the appropriate coefficients to include A.11, we have determined our final model for measuring the demand for rented bikes in Seoul,

(4) $\text{RentedBikeCount}_{st} = \beta_0 + ... + \beta_{11}\text{Rainfall}_{st} + ... + \beta_{17}\text{SolarRadiation}_{st} + \beta_{19}I(\text{Season}_s = \text{Winter}) + \beta_{20}I(\text{Season}_s = \text{Summer}) + \beta_{21}I(\text{Season}_s = \text{Autumn}) + ... + \beta_{240}\text{SolarRadiation}_{st}I(\text{Hour}_t = 8) + \beta_{241}\text{SolarRadiation}_{st}I(\text{Hour}_t = 19) + \beta_{242}\text{Rainfall}_{st}I(\text{Hour}_t = 9) + \beta_{243}\text{Rainfall}_{st}I(\text{Hour}_t = 13) + \beta_{244}\text{Rainfall}_{st}I(\text{Hour}_t = 17) + \beta_{245}\text{Rainfall}_{st}I(\text{Hour}_t = 19) + \beta_{246}\text{Rainfall}_{st}I(\text{Hour}_t = 20) + ... + \eta_{st}$

$\eta_{st} \overset{iid}{\sim} N(0,\sigma^2)$

(The full model is referenced in A.13, only assailant terms are included for simplicity.

Diagonostic Checks

For both models, model (2) (Temperature Change) and model (4) (Bike Sharing Demand),

Based on the Residuals vs. Fitted Values plot below, the linearity assumption appears to be met. There is a straight blue curve across most of the fitted values with majority of the residuals randomly distributed around the horizontal line at y = 0.

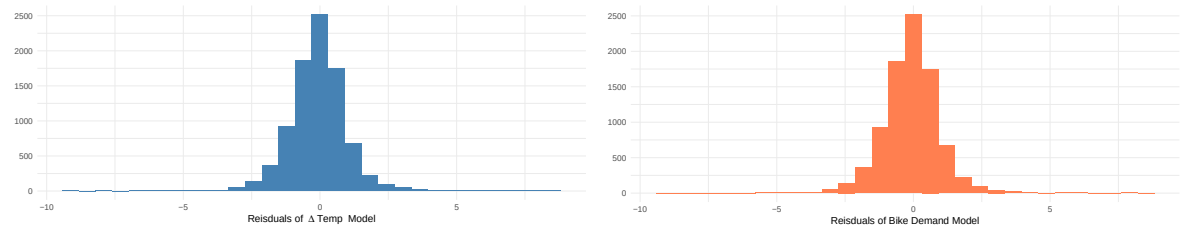
From the Normal Q-Q plot below, the plot shows that the points deviate from the reference line, especially in the tails. This suggests that the data may have heavier tails than a normal distribution, indicating the presence of outliers or a non-normal distribution of residuals.

However, even with selected interaction terms included, Normality did not drastically improve (see A.2.1 for QQ-Plot).

We considered transforming the model (1) (A.3) to better meet normality and somewhat homoskedasticity. However, due to the nature of our X variable matrix, since each independent variable is a change in a variable, we thus had to shift the entire vector of that corresponding variable in order to derive optimal transformations. Hence, interpreting any transformations we would make would prove difficult and impractical.

However, with large sample size ($n = 8740$), normality could be appealed to. The residuals do appear to be normal, though extreme values cause deviations from normality in the tails as seen in the Normal Q-Q plot. We considered alternatives to meeting the assumptions such as linear robust regression to reduce the weights of outliers. One alternative is further discussed in A.4, where we considered estimating a model for each season.

Histogram of Residuals from Elastic Net Regression

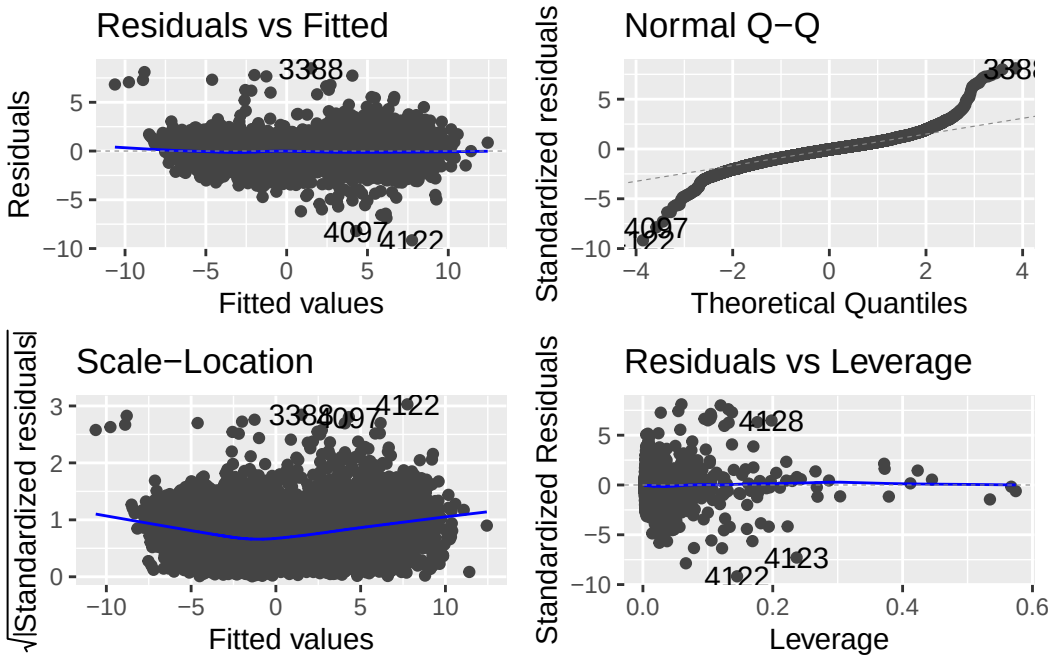


Since the histogram of the residuals above appears to be normal, aside from the extreme values in the tails in the Normal Q-Q plot, it is important to note that with large datasets, such as ours, the central limit theorem suggests that the sampling distribution of the regression coefficients will tend to be normally distributed, even if the residuals are not. This implies that the violation of the normality assumption may not substantially affect the validity of our inferential statistics, like confidence intervals and hypothesis tests..

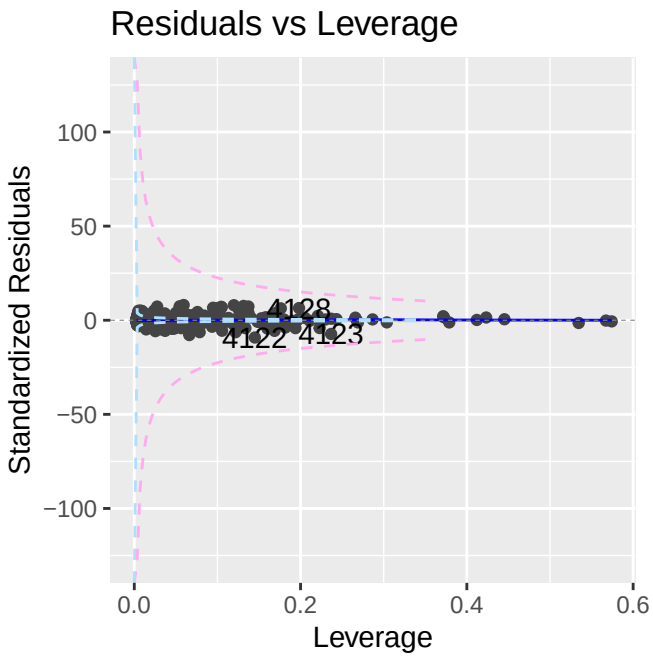
Additionally, the homoscedasticity assumption appears to be met. In the Scale-Location plot (A.2.1), there is a straight and horizontal blue curve for most of the fitted values. The points are evenly spread out above and below the curve. Even though there is a slight curvature in the blue line shown in the plot due to effects of some outliers, overall, I think that the homoscedasticity assumption is met.

Lastly, from the Residuals vs. Leverage plot (A.2.1), none of the standardized residuals falls outside the 0.5 Cook's distance thresholds. Therefore, we can conclude that there is no unduly influential points.

```
autoplot(seoul.elastic)
```



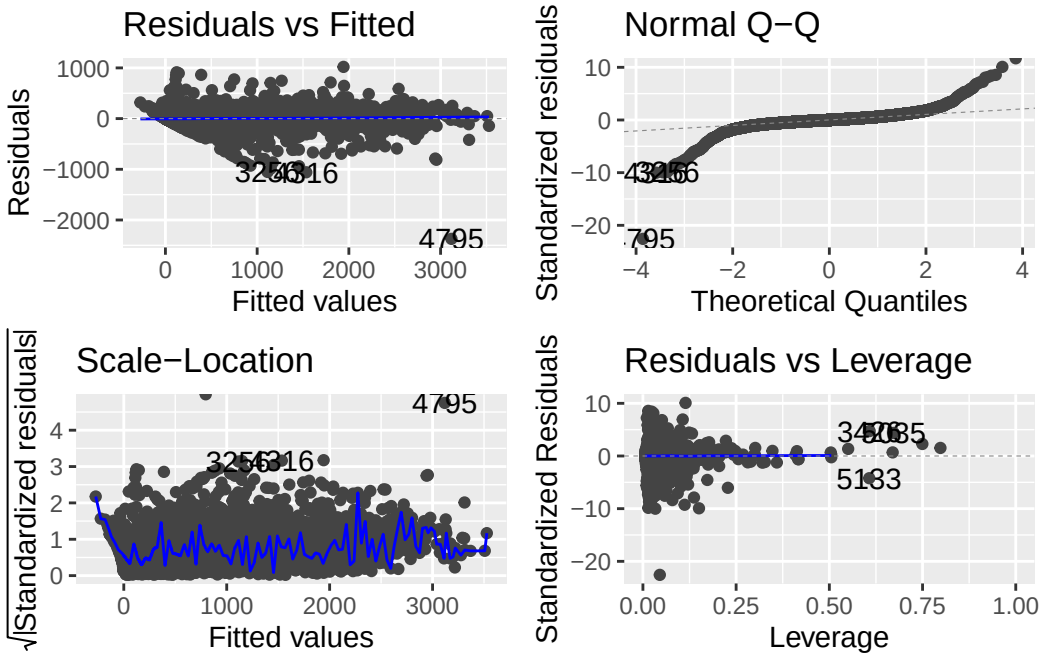

```
show_leverage(seoul.elastic)
```



```
autoplot(seoul.bike.elastic)
```

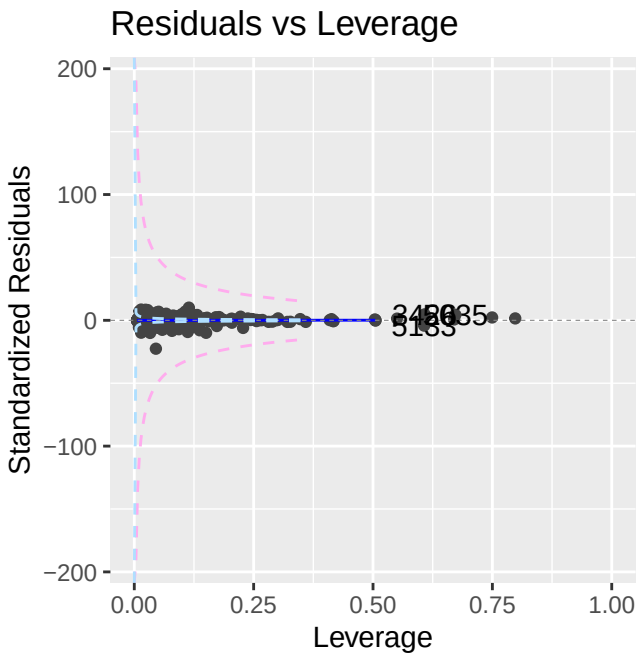
```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 8 rows containing missing values (`geom_line()`).
```




```
show_leverage(seoul.bike.elastic)
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
Removed 8 rows containing missing values (`geom_line()`).
```



Inferential Statistics

95% Confidence Interval of β_5 for Solar Radiation Change

```
conf_ints <- confint(seoul.elastic, level = 0.95)
conf_ints['SolarRadiationChange', ]
```

```
    2.5 %    97.5 %
0.8726519 1.0023804
```

Holding all else constant, we are 95% confident that the true estimated average temperature change in Seoul will be between 0.8726519 and 1.0023804 Celsius with every additional one MJ/m^2 increase in solar radiation change. The 95% confidence interval of β_5 for solar radiation change does not include zero. It suggests that the effect of solar radiation change on the average temperature change is statistically significant. In other words, there is a significant association between these variables.

95% Confidence Interval of β_241 for Solar Radiation at Hour 19

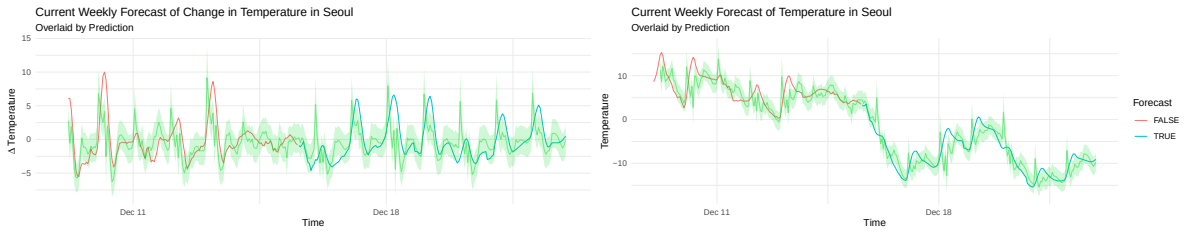
```
conf_ints <- confint(seoul.bike.elastic, level = 0.95)

solar_radiation_ci <- conf_ints['SolarRadiationXHour19', ]
print(solar_radiation_ci)
```

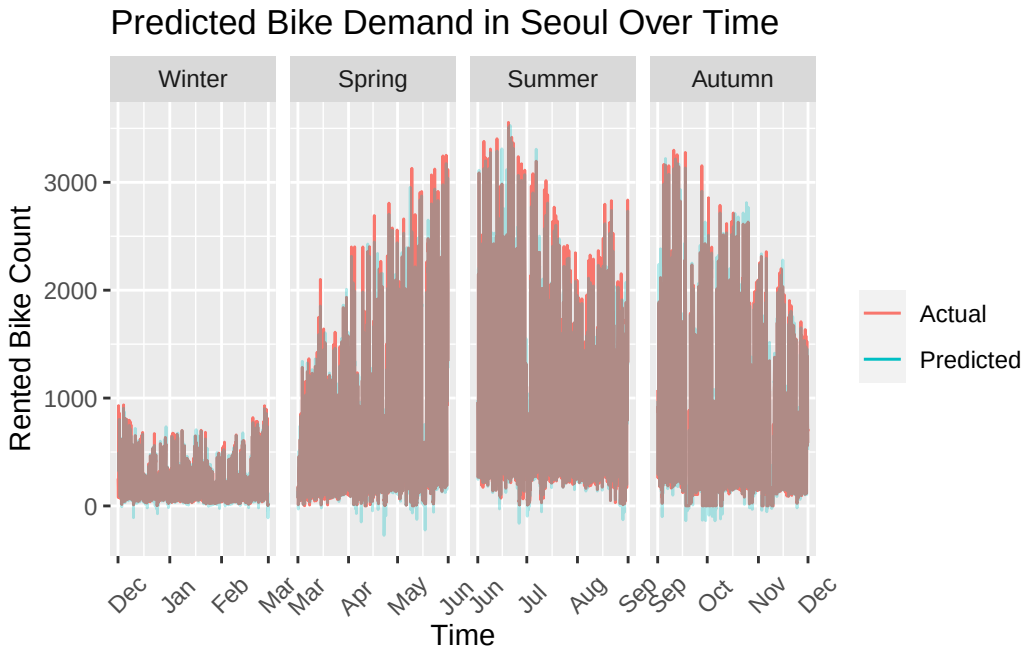
```
   2.5 %   97.5 %
527.1191 728.6566
```

Holding all else constant, at hour 19, we are 95% confident that the true estimated bike sharing demand in Seoul will be between and bikes with every additional one MJ/m^2 increase in solar radiation. The 95% confidence interval of $\beta_2 41$ for solar radiation at hour 19 does not include zero. It suggests that the effect of solar radiation at hour 19 on the average bike sharing demand is statistically significant. In other words, there is a significant association between these variables.

95% Prediction Interval (see A.15 for API code/data retrieval)




```
         fit       lwr      upr
1 -0.3562851 -2.484287 1.771717
```



5. Conclusions

For temperature changes, our multilinear model suggests a significant correlation with solar radiation, particularly during evening hours. The confidence interval for the effect of solar radiation on temperature change does not include zero, indicating a substantial impact. Interestingly, while solar radiation generally increased

temperature, its interaction with other factors like rainfall showed a negative effect on bike demand, except during peak hours where the effect was positive. This nuanced relationship highlights the complexity of weather's impact on urban activities.

Regarding bike-sharing demand, our analysis revealed strong seasonality and time dependencies. Notably, humidity and solar radiation played significant roles, with solar radiation's impact on demand varying by hour and season. The confidence interval for the interaction between solar radiation and bike-sharing demand at 7 PM underlines this point, demonstrating a significant effect during this time.

Given that we controlled for time-series independence using regression techniques such as kth-differences and fixed effects, our findings are believed to be reliable within the context of Seoul's climate and urban dynamics. While the results are specific to Seoul, they may also provide valuable insights for similar metropolitan areas with comparable climates and bike-sharing systems. However, caution should be exercised when generalizing beyond these conditions.

In conclusion, the study's results offers a deeper understanding of the intricate interplay between weather conditions, time factors, and urban mobility in Seoul. This information is crucial for city planners and policymakers aiming to promote sustainable transportation and urban living. Further research could explore these relationships in different contexts or extend the analysis to other forms of urban mobility.

6. Contributions