MATHS 7107 Data Taming Assignment Four Questions

Chang Dong

2023-03-21

Summary

The Melbourne Water Corporation (MWC) manages the water supply in Melbourne, and there is a need for MWC to build a model to help predict the evaporation at the reservoirs. In this paper, we mainly focus on this problem to build a model. there are 3 main parts to this paper. Firstly, we introduced the Method, which includes the data set introduction, Exploratory Data Analysis, Model selection, and diagnostics. Then in the result section, we interpreted the details of the final model. Finally, we give the prediction result of new data using the final model to predict a 95% probability of evaporation exceeding 10 mm or below 10 mm. And discussed the pros and cons of the model.

Method Section

Bivariate summaries

In this section, we mainly focused on exploratory data analysis, which including data cleaning, dataset investigation, and visulization to check relationships between Evaporation and other 5 features.

Data Cleaning: Firstly, the data set was loaded and then just 5 attributes were choosen, which are “Date”,“Maximum.Temperature..Deg.C.”, “Minimum.temperature..Deg.C.”,“X9am.relative.humidity….”,“Evaporation..mm.”. The dataset name is MWC_df. Next, we extract months and the day of week from date attribute. And the date attribute was not contained in this dataset. Then the 6 attributes were recoded as “month”, “wday”,“maxtemp”,“mintemp”,“humidity”,“Evaporation”. Next, we found month and wday are categorical data, while others are numerical. So these categorical data were retyped as factor. Finaly, we found and removed 8 missing values in Evaporation column.

Data Analysis: The relationship between Evaporation and other 5 attributes was visulized which can be seen in appendix below. 1) Evaporation ~ month. From fig1 we can find that month obviously influence the evaporation level. The evaporation is at the lowest average value in the Month of Jun, which is around 2mm. And as the month gets far away from June in the year, the average value of evaporation roughly gets bigger and bigger. The largest average value was achieved in Jan, which is around 8mm. 2) Evaporation ~ wday. From fig2 we can roughly see that the average evaporation value is similar as for different weekdays, which is around 4mm, and evaporation shows almost the same distribution on different weekdays. So there seems no obvious relationship between them. 3) Evaporation ~ Maximum temperature. From fig3 we can see that there seems a positive relationship between Maximum temperature and Evaporation. 4) Evaporation ~ Minimum

temperature.From fig4 we can see that there seems a positive relationship between Minimum temperature and Evaporation. 5) Evaporation ~ humidity. From fig5 we can see that there seems a negative relationship between relative humidity at 9am and Evaporation.

In summary, we can conclude that all the features except wday has some relationship with Evaporation.

Model selection

In this section, we build 3 models(lm1,lm2,lm3) to predict evaporation, using all of the predictors listed in the Bivariable summaries paragraph above. And all the latter models reduced one attribute compared with the former one. The reduction process followed by reducing the most insignificant attribute of that model, the significance was measured by a statistical value, p-value, and it shows less likely to influence the evaporation for the attribute with the big p-value. As for the numerical data maxtemp, mintemp and humidity, we use summary function to determine their significance, while for the rest of catergorical data, we use anova instead.

Round 1: we build lm1 using all the above five attributes and 1 interaction term of month and humidity, which shows the co-influence of them. From the summary and anova result, we can see the most insignificant attribute is maxtemp, whoes p-value is around 0.561, so we remove it in the first round then use the rest 5 attributes to build lm2.

Round 2: we build lm2 based on lm1, the only difference is we removed maxtemp. From the summary and anova result, we can see the most insignificant attribute is wday, whose p-value is around 0.102, so we remove it in the second round.

Round 3: we build lm3 based on lm2, the only difference is we removed wday. From the summary and ANOVA result, all the numerical attributes show considerable statistical significance whose p-value is less than 0.05, and the same as categorical attribues.

So the final model after 3 rounds of selection is lm3. We use month, mintemp, humidity, and the interaction of humidity and month to predict evaporation. As we mentioned, all the terms show considerable statistical significance whose p-value is less than 0.05. it has one difference that we conclude from the Bivariate summaries, which is maxtemp shows a positive influence on evaporation while the model tells us it has a very weak statistical significance, the possibility that we can reject them have a linear relationship is around 0.561. And the rest of the terms shows strong significance while wdays show week significance as we expected. The reason for the difference is maxtemp and mintemp show a similar influence on evaporation that we drew from Bivariate summaries, so mintemp can have a good explanation for it that maxtemp shows a weak significance, the correration(0.701) of mintemp and maxtemp reveal that they all belongs to temeprature, and temperature have influence to evaporation.

Result Section

Model interpretation

Let's interpret the model lm3 that we finally obtained.

$$Evaporation = month + mintemp + humidity + humidity{:}month$$

According the our Appendix round3 summary result, we can see that Jan is the reference month is January, which is 8.589 mm in the condition of mintemp, humidity = 0. and other estimate terms of mintemp and humidity indicate the coefficient between the term and evaporation, which is 0.369 and -0.010 respectively. The other month's estimate indicate that the interception increment compared with the reference month. And the intersection term of each month indicate the evaporation increment in that month with 1 unit increment of humidity.

Discussion Section

Prediction

In this part, we will use the lm3 model to predict the following new data. The details were shown in Table 1. Applying our model lm3, we get the predictions of our new data with best-fit value and prediction interval of 95% level including upper bound and lower bound. The output was arranged in Table 2.From our result, the day of "2020-01-13" get a highest prediction value of evaporation which is around 14.872 mm, on the contrast of the lowest value at the day of "2020-07-06" which is 2.265.

Now we have two events. Event 1 is there is 95% confidence that more than 10mm of evaporation will occur. Event 2 is there is 95% confidence that more than 10mm of evaporation will not occur. We need to use one-tail prediction interval to obtain our result. The one-tailed 95% prediction interval estimates the evaporation boundary with 95% prediction either above or below the bound in this case, not within a range. The result was show in Table 3. From Table 3 we can see, only the day of "2020-01-13" can satisfy event 1, because the lower bound of the one-tail prediction interval(0.05,1) is 13.114 mm (>10 mm), while the other not. Meanwhile, the day of "2020-02-29", "2020-07-06" and "2020-12-25" can all satisfy event 2, because the upper bound of one-tail prediction interval(0,0.95) are all less than 10 mm, while the day of "2020-01-13" not.

pros and cons

pros

1. Linear model has a good explainability, we can directly get how much of the influence of each factor.
2. Our data basically obey the linear model assumptions, so we can have a more credible trust in this model

cons

1. As we can see our prediction interval in some cases occurs at a negative value, this is inconsistence with our physical world.
2. There are other potential influences to evaporation, so we can not get a precise prediction of our evaporation.

Conclusion

In this paper, we build a linear model to predict evaporation at Cardinia Reservoir based on Melbourne's day-to-day weather observations. This paper outlines the factors that can have a significant influence on evaporation and demonstrates the ability to make predictions for each individual day. We also give prediction intervals to help forecast if the event of evaporation of more than 10 mm will happen or not on a specific day, Only the date "2020-01-13" shows more than 95% confidence to have more than 10 mm evaporation, and the rest of date shows 95% confidence to have less than 10 mm evaporation. Besides, we have noticed that it will have a relatively big evaporation value in around January with low humidity and high temperature. In conclusion, this model can well explain the impact of each factor on evaporation, and basically obeys the linear assumption, but there are still some drawbacks that need to think in our future work.

Appendix

Load all necessary pkgs

```
pacman::p_load(tidyverse,lubridate)
```

EDA(Exploratory Data Analysis)

Load, Select and Recode data

```
MWC <- read.csv("melbourne.csv")
MWC_df <- MWC[c(1,2,3,11,5)]
MWC_df["month"] = lapply(MWC_df[1], month)
MWC_df["wday"] =lapply(MWC_df[1], wday)
MWC_df <- MWC_df[c("month","wday","Maximum.Temperature..Deg.C.",
                   "Minimum.temperature..Deg.C.",
                   "X9am.relative.humidity....","Evaporation..mm.")]
colnames(MWC_df) <- c("month", "wday","maxtemp",
                      "mintemp","humidity","Evaporation")
MWC_df %>% str()
```

```
## 'data.frame':    365 obs. of  6 variables:
##  $ month      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ wday       : num  3 4 5 6 7 1 2 3 4 5 ...
##  $ maxtemp    : num  26.2 22.2 29.5 42.6 21.2 22.1 23.1 24.1 20.5 21.4 ...
##  $ mintemp    : num  15.5 18.4 15.9 18 17.4 14.6 17.1 16.7 16.1 13.5 ...
##  $ humidity   : int  74 64 75 31 63 55 55 72 62 53 ...
##  $ Evaporation: num  7 7 6.6 7.8 15.4 6.4 9 7.2 7.4 8.2 ...
```

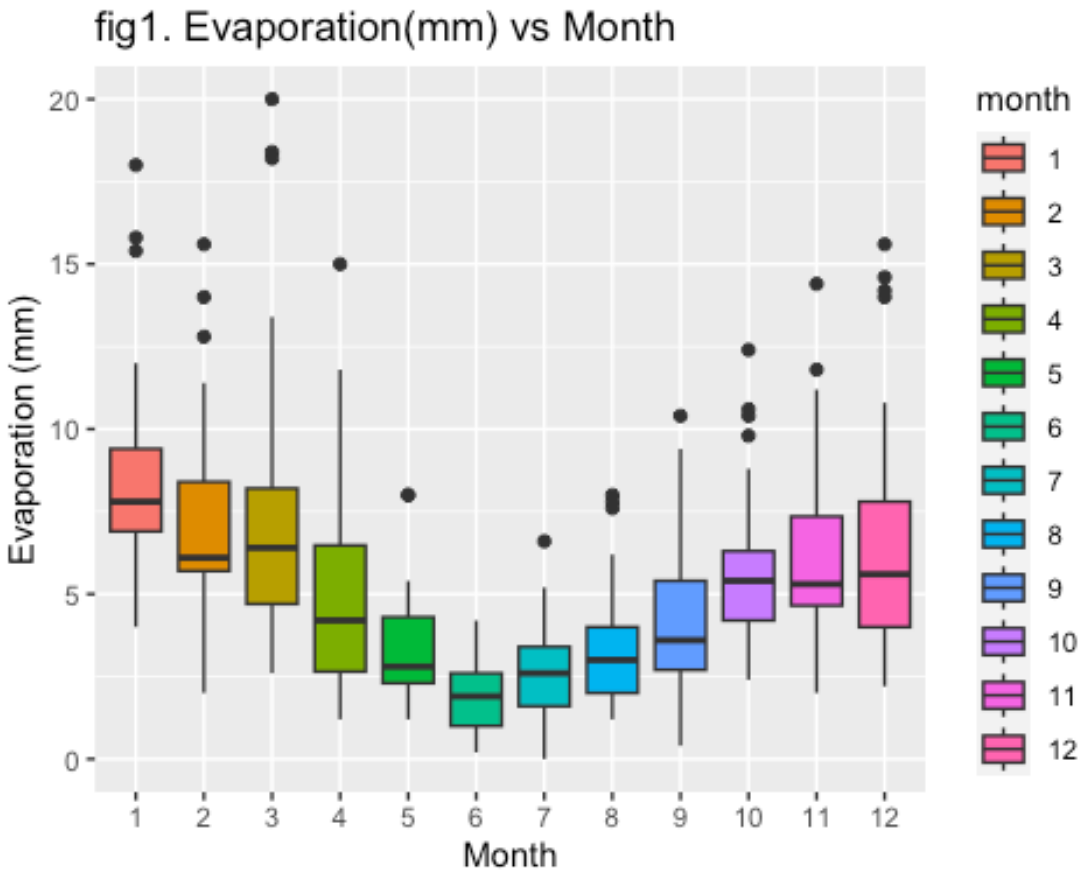
```
MWC_df$month <- as.factor(MWC_df$month)
MWC_df$wday <- as.factor(MWC_df$wday)
MWC_df %>% head()

##   month wday maxtemp mintemp humidity Evaporation
## 1     1    3    26.2    15.5       74         7.0
## 2     1    4    22.2    18.4       64         7.0
## 3     1    5    29.5    15.9       75         6.6
## 4     1    6    42.6    18.0       31         7.8
## 5     1    7    21.2    17.4       63        15.4
## 6     1    1    22.1    14.6       55         6.4
```

Remove missing value

```
MWC_df %>% is.na() %>% colSums()

##       month        wday     maxtemp     mintemp    humidity Evaporation
##           0           0           0           0           0           8

MWC_df <- MWC_df %>% na.omit()
MWC_df %>% is.na() %>% colSums()

##       month        wday     maxtemp     mintemp    humidity Evaporation
##           0           0           0           0           0           0
```

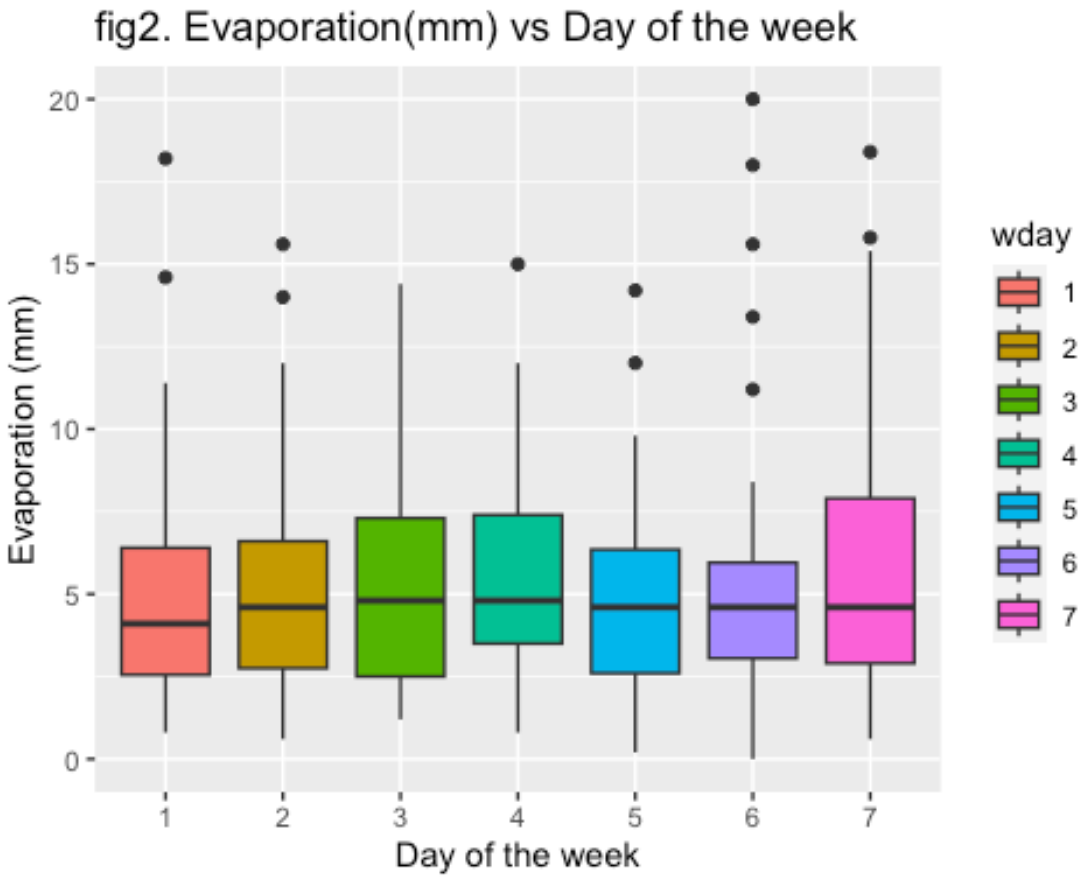
1. Evaporation(mm) vs Month

```
ggplot(data = MWC_df, aes(x = month, y = Evaporation, fill = month)) +
  geom_boxplot() +
  ggtitle("fig1. Evaporation(mm) vs Month") +
  ylab("Evaporation (mm)") +
  xlab("Month")
```



2. Evaporation(mm) vs Day of the week

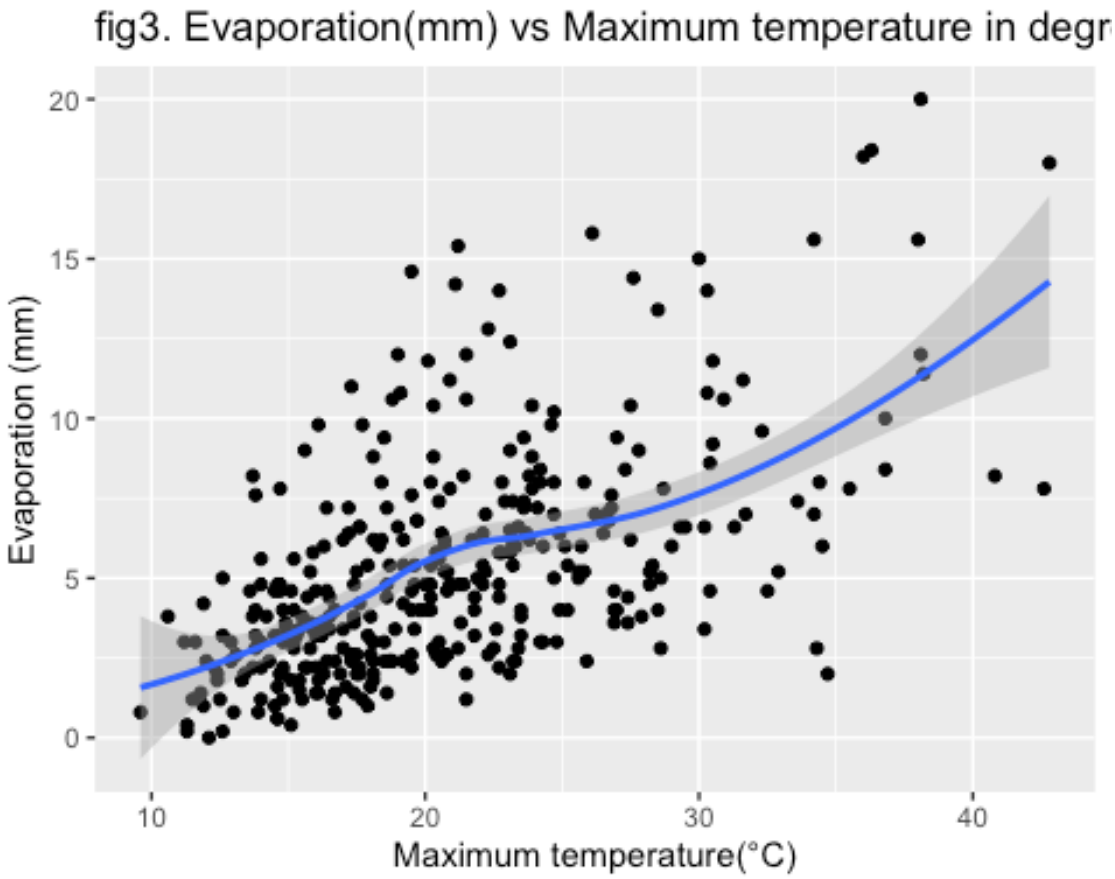
```
ggplot(data = MWC_df, aes(x = wday, y = Evaporation, fill = wday)) +
  geom_boxplot() +
  ggtitle("fig2. Evaporation(mm) vs Day of the week") +
  ylab("Evaporation (mm)") +
  xlab("Day of the week")
```



3. Evaporation(mm) vs Maximum temperature in degrees Celsius

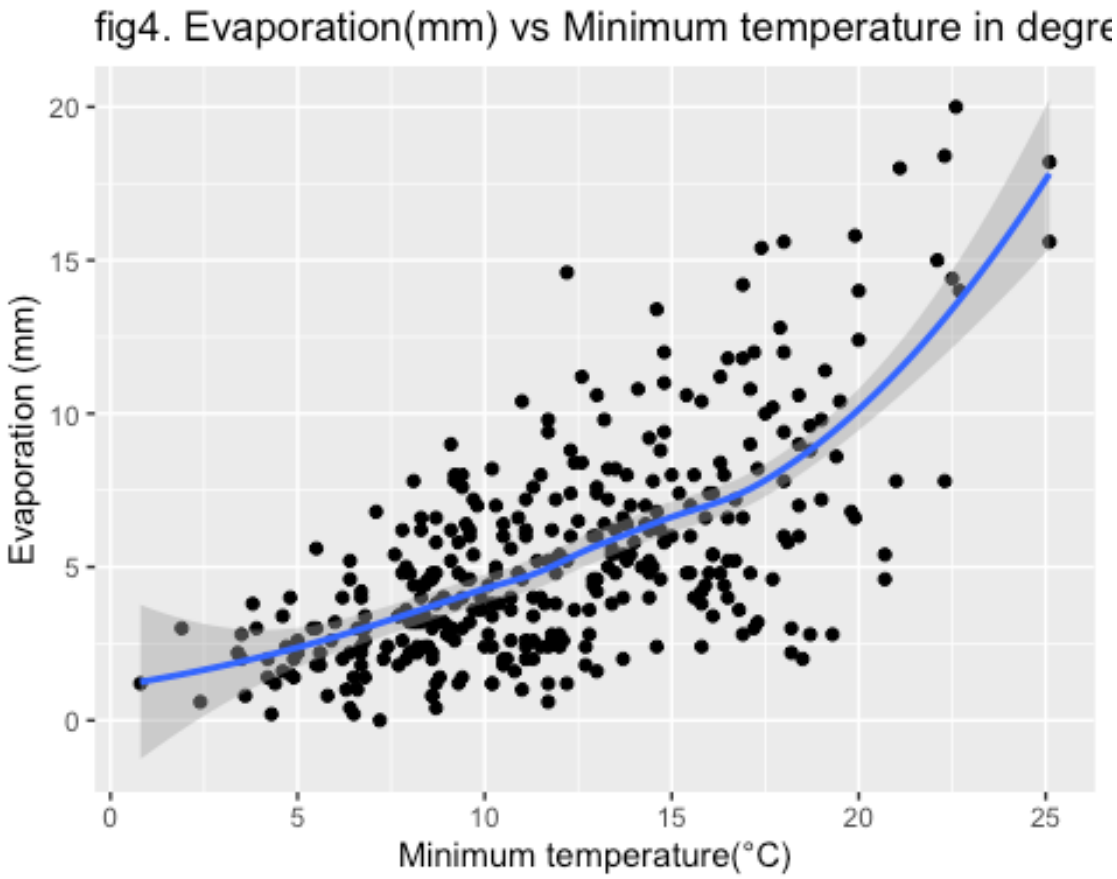
```
ggplot(data = MWC_df, aes(x = maxtemp, y = Evaporation)) +
  geom_point() + geom_smooth()+
  ggtitle("fig3. Evaporation(mm) vs Maximum temperature in degrees Celsius")+
  ylab("Evaporation (mm)")+
  xlab("Maximum temperature(°C)")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



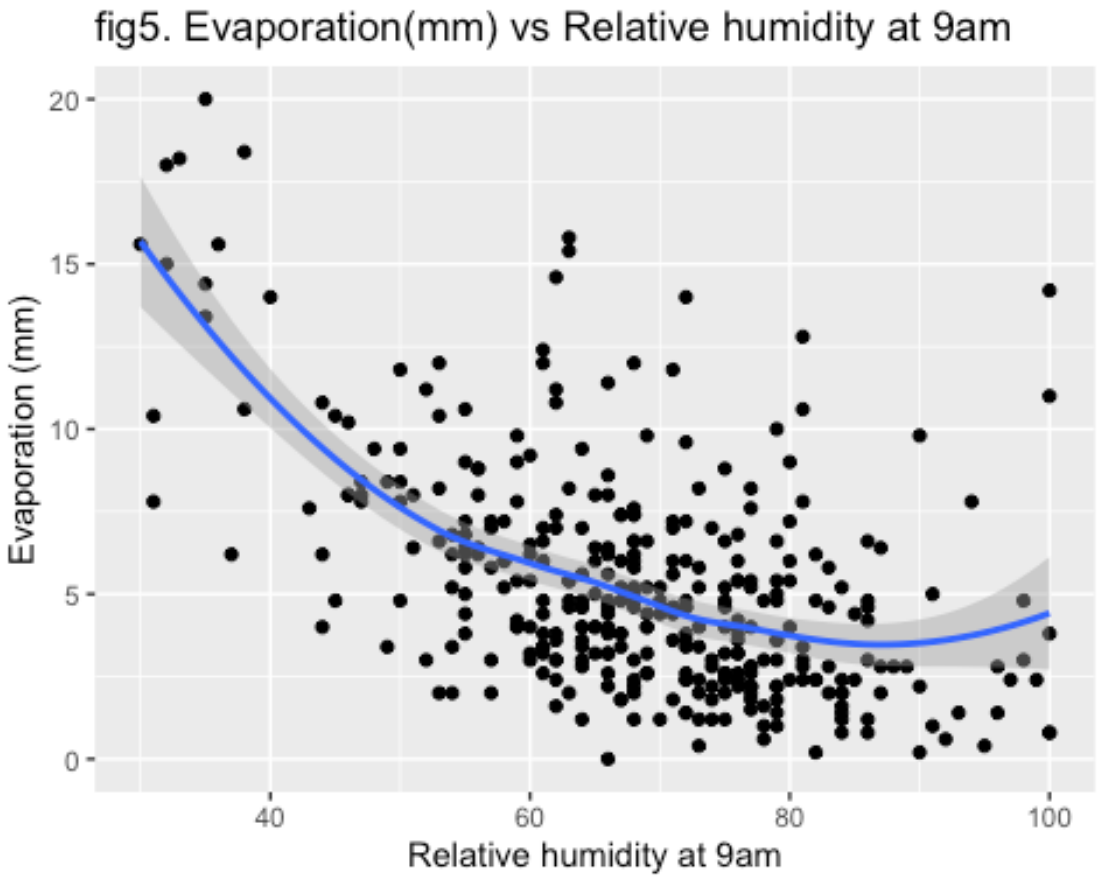
4. Evaporation(mm) vs Mnimum temperature in degrees Celsius

```
ggplot(data = MWC_df, aes(x = mintemp, y = Evaporation)) +
  geom_point() + geom_smooth()+
  ggtitle("fig4. Evaporation(mm) vs Minimum temperature in degrees Celsius")+
  ylab("Evaporation (mm)")+
  xlab("Minimum temperature(°C)")

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



5. Evaporation(mm) vs Relative humidity at 9am.

```
ggplot(data = MWC_df, aes(x = humidity, y = Evaporation)) +
  geom_point() + geom_smooth()+
  ggtitle("fig5. Evaporation(mm) vs Relative humidity at 9am")+
  ylab("Evaporation (mm)")+
  xlab("Relative humidity at 9am")

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Model Selection

round1

```
lm1 <- lm(Evaporation ~ month + wday+ mintemp + maxtemp +
            humidity + humidity:month , data = MWC_df)
summary(lm1)

##
## Call:
## lm(formula = Evaporation ~ month + wday + mintemp + maxtemp +
##     humidity + humidity:month, data = MWC_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5166 -1.1713 -0.0523  1.0677 11.0447
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.313165   2.375833   3.499 0.000532 ***
## month2            1.122982   3.341422   0.336 0.737028
## month3            5.340251   2.630467   2.030 0.043155 *
## month4            1.729320   3.102811   0.557 0.577679
## month5           -4.255253   3.347211  -1.271 0.204537
```

```
## month6           -7.914716   3.972809  -1.992 0.047183 *
## month7           -4.930279   3.580302  -1.377 0.169442
## month8           -6.310577   3.222937  -1.958 0.051083 .
## month9           -0.544108   3.157664  -0.172 0.863298
## month10          -6.307800   3.112895  -2.026 0.043546 *
## month11          -1.080420   2.787061  -0.388 0.698525
## month12           0.667154   2.793904   0.239 0.811420
## wday2            -0.272388   0.432537  -0.630 0.529304
## wday3            -0.083051   0.436596  -0.190 0.849252
## wday4            -0.078214   0.436180  -0.179 0.857801
## wday5            -0.536148   0.435847  -1.230 0.219539
## wday6            -0.408977   0.443221  -0.923 0.356828
## wday7             0.499638   0.432760   1.155 0.249127
## mintemp           0.357912   0.044596   8.026 1.86e-14 ***
## maxtemp           0.017765   0.030507   0.582 0.560738
## humidity         -0.098209   0.032565  -3.016 0.002765 **
## month2:humidity  -0.026262   0.050976  -0.515 0.606776
## month3:humidity  -0.080822   0.039559  -2.043 0.041850 *
## month4:humidity  -0.043164   0.047080  -0.917 0.359914
## month5:humidity   0.034968   0.047799   0.732 0.464966
## month6:humidity   0.078436   0.052691   1.489 0.137560
## month7:humidity   0.049674   0.051370   0.967 0.334276
## month8:humidity   0.079397   0.047371   1.676 0.094686 .
## month9:humidity  -0.006753   0.049154  -0.137 0.890813
## month10:humidity  0.092502   0.047400   1.952 0.051853 .
## month11:humidity  0.015097   0.041694   0.362 0.717527
## month12:humidity -0.018916   0.041366  -0.457 0.647783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.175 on 325 degrees of freedom
## Multiple R-squared:  0.6458, Adjusted R-squared:  0.612
## F-statistic: 19.12 on 31 and 325 DF,  p-value: < 2.2e-16

anova(lm1)

## Analysis of Variance Table
##
## Response: Evaporation
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## month           11 1478.85  134.44  28.4288 < 2.2e-16 ***
## wday             6   50.51    8.42   1.7801 0.1025018
## mintemp          1  588.63  588.63 124.4719 < 2.2e-16 ***
## maxtemp          1   74.85   74.85  15.8275  8.56e-05 ***
## humidity         1  448.57  448.57  94.8548 < 2.2e-16 ***
## month:humidity  11  160.95   14.63   3.0941 0.0005645 ***
## Residuals      325 1536.94    4.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

round2

```
lm2 <- lm(Evaporation ~ month + wday+ mintemp + humidity +
            humidity:month , data = MWC_df)
summary(lm2)

##
## Call:
## lm(formula = Evaporation ~ month + wday + mintemp + humidity +
##     humidity:month, data = MWC_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.619 -1.194 -0.085  1.098 11.063
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.754075   2.249680   3.891 0.000121 ***
## month2             1.070387   3.336814   0.321 0.748582
## month3             5.349365   2.627753   2.036 0.042587 *
## month4             1.736304   3.099641   0.560 0.575753
## month5            -4.410731   3.333162  -1.323 0.186667
## month6            -8.038559   3.963090  -2.028 0.043337 *
## month7            -5.201366   3.546311  -1.467 0.143422
## month8            -6.473398   3.207531  -2.018 0.044390 *
## month9            -0.610357   3.152414  -0.194 0.846597
## month10           -6.286771   3.109529  -2.022 0.044016 *
## month11           -1.139353   2.782399  -0.409 0.682452
## month12            0.781062   2.784222   0.281 0.779248
## wday2             -0.277981   0.431992  -0.643 0.520361
## wday3             -0.096705   0.435524  -0.222 0.824420
## wday4             -0.101325   0.433930  -0.234 0.815516
## wday5             -0.537121   0.435402  -1.234 0.218233
## wday6             -0.397814   0.442357  -0.899 0.369154
## wday7              0.485861   0.431674   1.126 0.261194
## mintemp            0.366245   0.042195   8.680  < 2e-16 ***
## humidity          -0.099383   0.032470  -3.061 0.002391 **
## month2:humidity   -0.025880   0.050920  -0.508 0.611617
## month3:humidity   -0.081594   0.039496  -2.066 0.039631 *
## month4:humidity   -0.044248   0.046996  -0.942 0.347135
## month5:humidity    0.035445   0.047744   0.742 0.458383
## month6:humidity    0.078315   0.052637   1.488 0.137765
## month7:humidity    0.051360   0.051236   1.002 0.316883
## month8:humidity    0.079649   0.047321   1.683 0.093302 .
## month9:humidity   -0.007641   0.049081  -0.156 0.876378
## month10:humidity   0.091053   0.047287   1.926 0.055028 .
## month11:humidity   0.014932   0.041651   0.358 0.720201
## month12:humidity  -0.021128   0.041150  -0.513 0.607989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.172 on 326 degrees of freedom
## Multiple R-squared:  0.6454, Adjusted R-squared:  0.6128
## F-statistic: 19.78 on 30 and 326 DF,  p-value: < 2.2e-16

anova(lm2)

## Analysis of Variance Table
##
## Response: Evaporation
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## month           11 1478.85  134.44  28.4865 < 2.2e-16 ***
## wday             6   50.51    8.42   1.7837 0.1017458
## mintemp          1  588.63  588.63 124.7247 < 2.2e-16 ***
## humidity         1  519.30  519.30 110.0348 < 2.2e-16 ***
## month:humidity  11  163.47   14.86   3.1488 0.0004588 ***
## Residuals      326 1538.54    4.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

round3

```
lm3 <- lm(Evaporation ~ month + mintemp + humidity  +
            humidity:month , data = MWC_df)
summary(lm3)

##
## Call:
## lm(formula = Evaporation ~ month + mintemp + humidity + humidity:month,
##     data = MWC_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0316 -1.1560 -0.1263  1.0184 10.6597
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.589140   2.202471   3.900 0.000117 ***
## month2            0.822148   3.297575   0.249 0.803268
## month3            5.263051   2.610525   2.016 0.044596 *
## month4            1.971572   3.040391   0.648 0.517136
## month5           -4.377344   3.261415  -1.342 0.180461
## month6           -8.376118   3.924447  -2.134 0.033547 *
## month7           -5.360039   3.479608  -1.540 0.124412
## month8           -7.102852   3.189591  -2.227 0.026625 *
## month9           -1.243475   3.090815  -0.402 0.687712
## month10          -6.158396   3.068813  -2.007 0.045585 *
## month11          -1.036904   2.737218  -0.379 0.705066
## month12           0.926791   2.748164   0.337 0.736149
## mintemp           0.368846   0.041819   8.820  < 2e-16 ***
## humidity         -0.099750   0.031724  -3.144 0.001815 **
## month2:humidity  -0.021806   0.050276  -0.434 0.664760
## month3:humidity  -0.079813   0.039166  -2.038 0.042360 *
```

```
## month4:humidity  -0.047469   0.046050  -1.031 0.303377
## month5:humidity   0.035145   0.046597   0.754 0.451246
## month6:humidity   0.083313   0.052006   1.602 0.110113
## month7:humidity   0.054069   0.050199   1.077 0.282219
## month8:humidity   0.089054   0.047045   1.893 0.059234 .
## month9:humidity   0.003411   0.048049   0.071 0.943452
## month10:humidity  0.089443   0.046676   1.916 0.056194 .
## month11:humidity  0.013451   0.040881   0.329 0.742336
## month12:humidity -0.022341   0.040556  -0.551 0.582087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.175 on 332 degrees of freedom
## Multiple R-squared:  0.638,  Adjusted R-squared:  0.6119
## F-statistic: 24.38 on 24 and 332 DF,  p-value: < 2.2e-16
```

```
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: Evaporation
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## month           11 1478.85  134.44  28.4160 < 2.2e-16 ***
## mintemp          1  608.93  608.93 128.7068 < 2.2e-16 ***
## humidity         1  510.03  510.03 107.8030 < 2.2e-16 ***
## month:humidity  11  170.74   15.52   3.2808 0.0002758 ***
## Residuals      332 1570.74    4.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explain the difference between model result and the visual analysis

```
cor(x = MWC_df$maxtemp,MWC_df$mintemp)
```

```
## [1] 0.6972069
```

Model diagnostics

In this part, we will test all of the assumptions of my linear model lm3(the final model). 1) Linearity;2) Homoscedasticity; 3) Normality; 4) Independence

1) The Assumption 1 Linearity was shown in fig6. The x-axis is Fitted values which is the evaporation prediction using our model, and the y-axis, residuals refer to the error between our prediction and ground true value. From the graph, we can see that the residuals are roughly evenly distributed near 0, corresponding with the red trend line that is almost straight and near 0 with a slightly upwards bent at the end). And only a small number of points are located far away from the red line. So we can justify that the assumption 1 is basically true.

2) The Assumption 2 Homoscedasticity was shown in fig7. The x-axis is Fitted values which is the evaporation prediction using our model, and the y-axis, the root

squared of standardized residuals, the transformation of error that we only care about the standardized absolute distance between y_pred and y_ture, that we can check whether these errors are same in different prediction values. From fig7 we can see that the error level has a small upwards trend as predictions rise up(the red line). But as our experience, that trend is acceptable for a linear model. This assumption is roughly established.

3) The Assumption 3 Normality was shown in fig8. The x-axis is the theoretical quantile, and the y-axis is the standardized residuals which is what we observed. If these transformed errors follow the same quantile as the theoretical quantile(normally distributed), which is these points lie in the line y=x , we can justify our observations obey the assumption of normality. From the graph we can see, within 1.5 quantiles, the assumption is perfectly established, while apart from this range, these points are spread around and even far away. But it is also acceptable, most of the values follow the theoretical distribution. So we can justify that assumption 3 is roughly true.

4) The Assumption 4 Independence. Typically in a time series model, this assumption can not be established, because day 1 usually has an influence on a continuous day. the error do not independent means our observations are not independent, so this assumption is not justified.

fig6. Check Assumption 1: Linearity

```
plot(lm3, which =1,
     sub = "fig6. Check Assumption 1: Linearity")
```

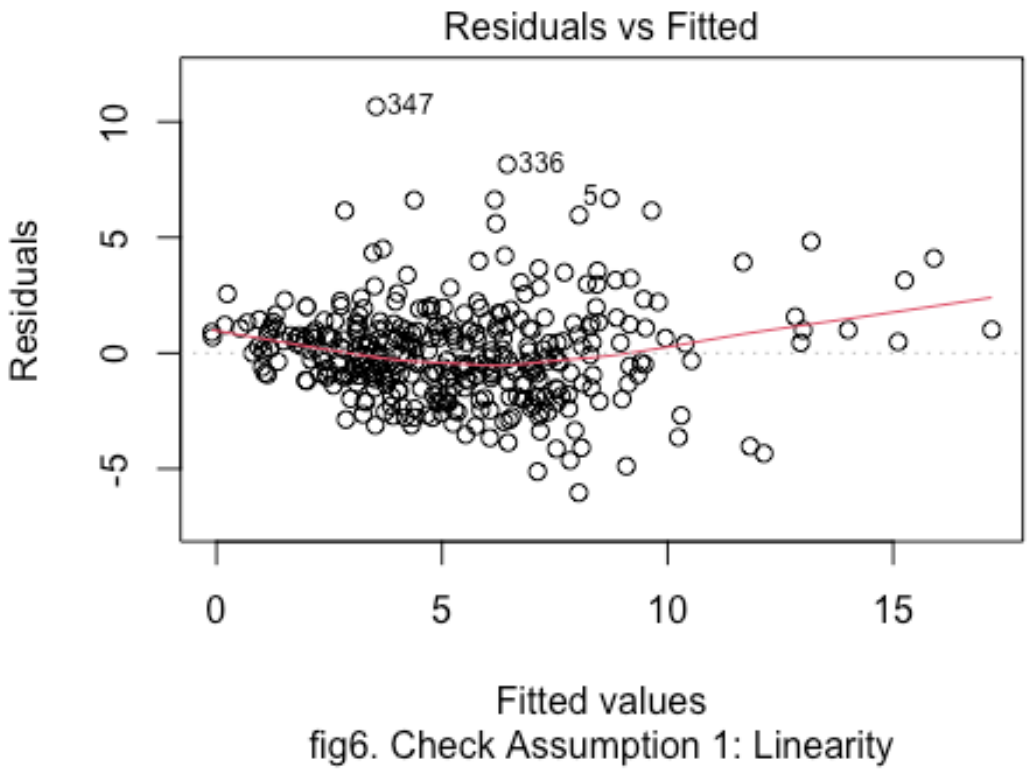


fig6. Check Assumption 1: Linearity

fig7. Check Assumption 2: Homoscedasticity

```
plot(lm3, which =3,
     sub = "fig7. Check Assumption 2: Homoscedasticity")
```

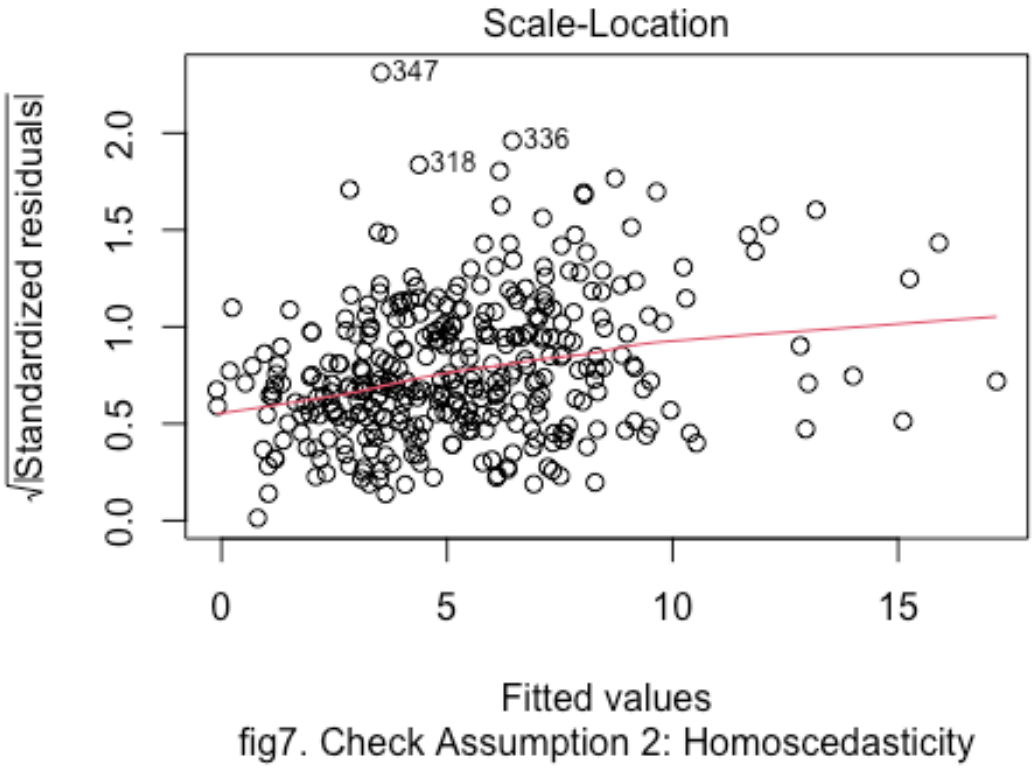


fig7. Check Assumption 2: Homoscedasticity

fig8. Check Assumption 2: Normality

```
plot(lm3, which =2,
     sub = "fig8. Check Assumption 2: Normality")
```

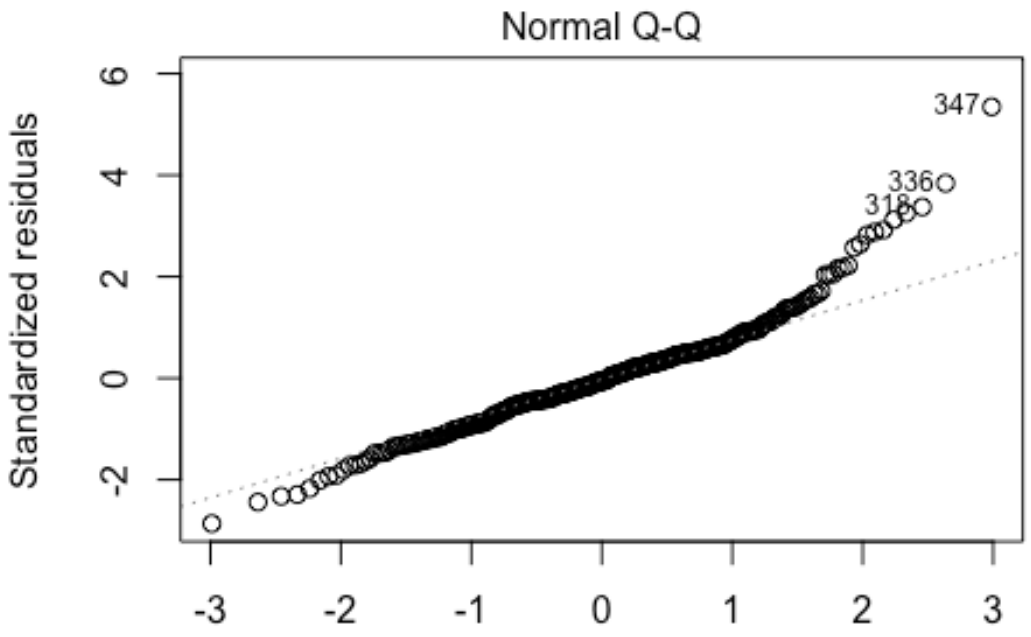


fig8. Check Assumption 2: Normality

Prediction Result

Table 1: The newdata infomration

```
date <- c(ymd("2020-02-29"), ymd("2020-12-25"),
          ymd("2020-01-13"), ymd("2020-07-06"))

newdata <- tibble(
  date = date,
  month = as.factor(month(date)),
  wday = as.factor(wday(date)),
  maxtemp = c(23.2, 31.9, 44.3, 10.6),
  mintemp = c(13.8, 16.4, 26.5, 6.8),
  humidity = c(74, 57, 35, 76)
)

newdata %>% arrange(date)
```

```
## # A tibble: 4 × 6
##   date       month wday  maxtemp mintemp humidity
##   <date>     <fct> <fct>   <dbl>   <dbl>    <dbl>
## 1 2020-01-13 1     2        44.3    26.5       35
## 2 2020-02-29 2     7        23.2    13.8       74
```

```
## 3 2020-07-06 7     2        10.6     6.8       76
## 4 2020-12-25 12    6        31.9    16.4       57
```

Table 2: Prediction of newdata using lm3

```
Evaporation_pred <- round(predict(lm3, newdata[,2:6],
                                  interval = "prediction",
                                  level = 0.95),3)

result_df <- data.frame(date = date,
                        Evaporation_pred = Evaporation_pred)
result_df %>% arrange(date)
```

```
##         date Evaporation_pred.fit Evaporation_pred.lwr
Evaporation_pred.upr
## 3 2020-01-13               14.872               10.105
19.640
## 1 2020-02-29                5.506                1.089
9.923
## 4 2020-07-06                2.265               -2.111
6.642
## 2 2020-12-25                8.606                4.209
13.003
```

Table 3: One-tailed prediction interval

```
predictions <- predict(lm3, newdata, se.fit = TRUE)

alpha_1 <- 0.05
z_1 <- qnorm(1 - alpha_1)
uppr_0.95 <- round(predictions$fit +
                     z_1 * predictions$se.fit,3)


alpha_2 <- 0.95
z_2 <- qnorm(1 - alpha_2)
lwr_0.05 <- round(predictions$fit +
                    z_2 * predictions$se.fit,3)

predictions_df <- data.frame(
  index <- date,
  uppr_0.95 = uppr_0.95,
  lwr_0.05 = lwr_0.05
)

predictions_df %>% arrange(date)
```

```
##   index....date uppr_0.95 lwr_0.05
## 3    2020-01-13    16.631   13.114
## 1    2020-02-29     6.423    4.589
## 4    2020-07-06     3.037    1.494
## 2    2020-12-25     9.453    7.758
```