Regression Analysis

The number of movie audiences according to the weather

2021-1 Regression Analysis

20176735 Surin Kim

01

Introduction

Topic and Variables

01

Introduction

Analysis of the number of movie audiences according to the weather

Because of the difficulty in getting data, the response variable is based on the country while the explanatory variable is based on Seoul.

variable name	detail
aud[1]	the number of movie audiences(in thousands)
avgtem, lowtem, hitem[2]	average/minimum/maximum temperature (℃)
rain / maxSn	precipitation (mm) / amount of snowfall (cm)
avgWS, maxWS	average/maximum wind speed (m/s)
avgHum	average relative humidity (%)
sumSun	sum of the duration of sunshine (hr)
sumSR	sum of solar radiation quantity (MJ/m2)
maxSun	maximum solar radiation quantity in an hour (MJ/m2)
tem15	1.5m soil temperature (℃)
avgCl	average amount of clouds (1/10)
dust	the concentration of fine dust (㎍/㎥)

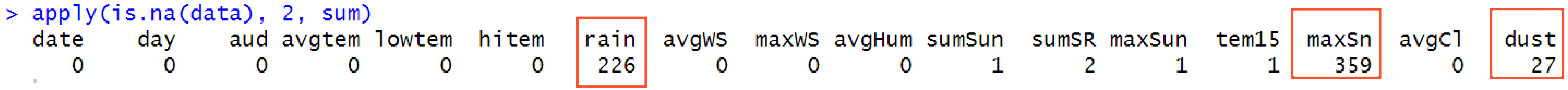
[1] source(movie audiences data): www.kobis.or.kr (KOBIS: KOREA Box-office Information System)

[2] source(all weather data): data.kma.go.kr (open MET data portal)

01

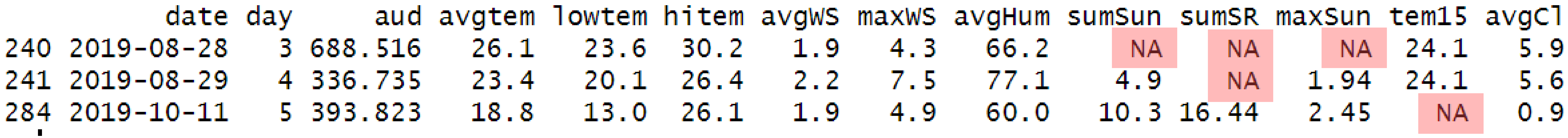
Variables : NA values

- Drop variables with many missing values: *rain, maxSn, dust*



```
> apply(is.na(data), 2, sum)
  date    day    aud avgtem lowtem  hitem   rain  avgWS  maxWS avgHum sumSun  sumSR maxSun  tem15  maxSn  avgCl   dust 
     0      0      0      0      0      0    226      0      0      0      1      2      1      1    359      0     27 
```

- Drop rows with other missing values: 240, 241, 284 (index)

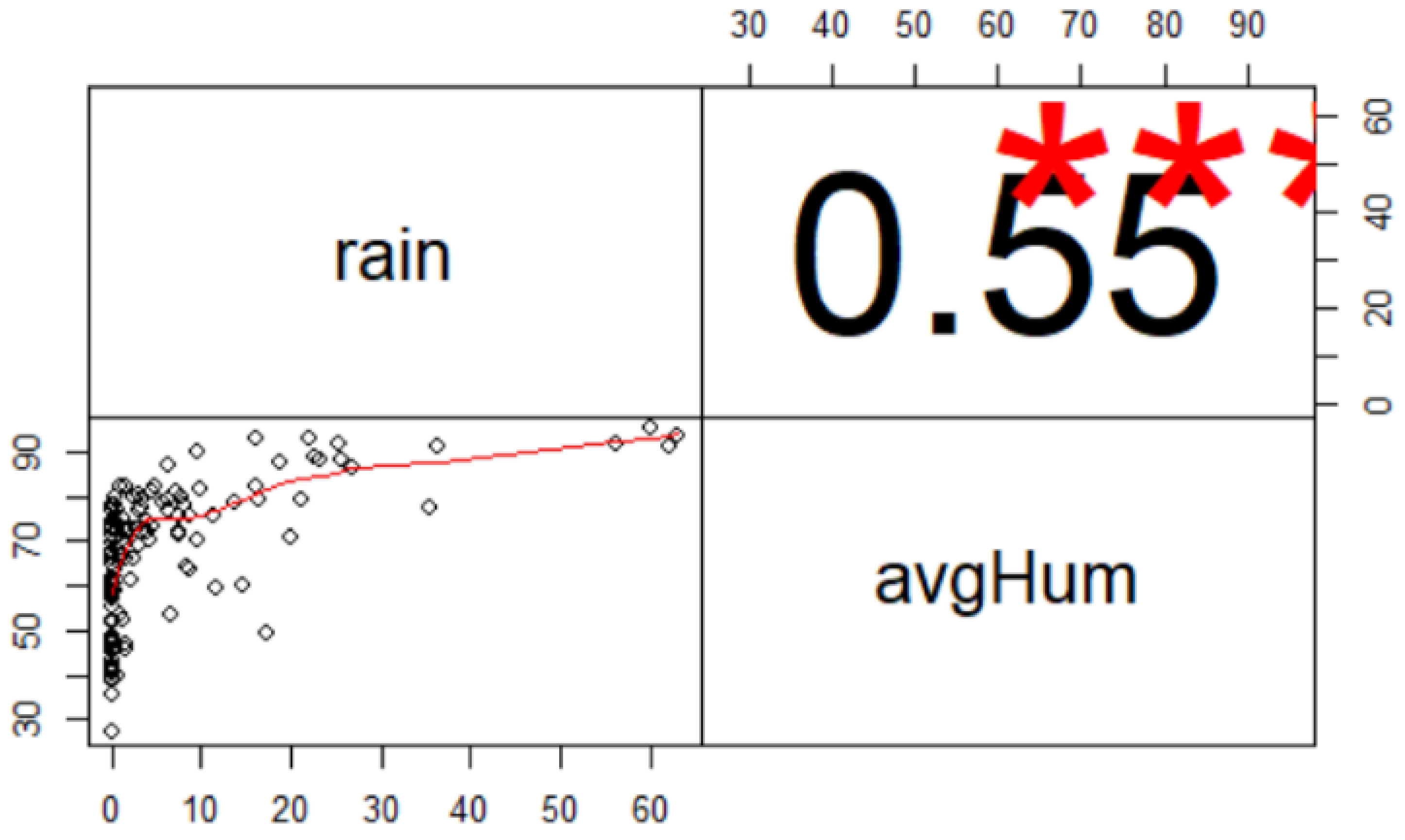


```
          date day     aud avgtem lowtem hitem avgWS maxWS avgHum sumSun sumSR maxSun tem15 avgCl
240 2019-08-28   3 688.516   26.1   23.6  30.2   1.9   4.3   66.2     NA    NA     NA  24.1   5.9
241 2019-08-29   4 336.735   23.4   20.1  26.4   2.2   7.5   77.1    4.9    NA   1.94  24.1   5.6
284 2019-10-11   5 393.823   18.8   13.0  26.1   1.9   4.9   60.0   10.3 16.44   2.45    NA   0.9
```

01

Variables : Correlation

- *rain* variable can be replaced by ***avgHum***

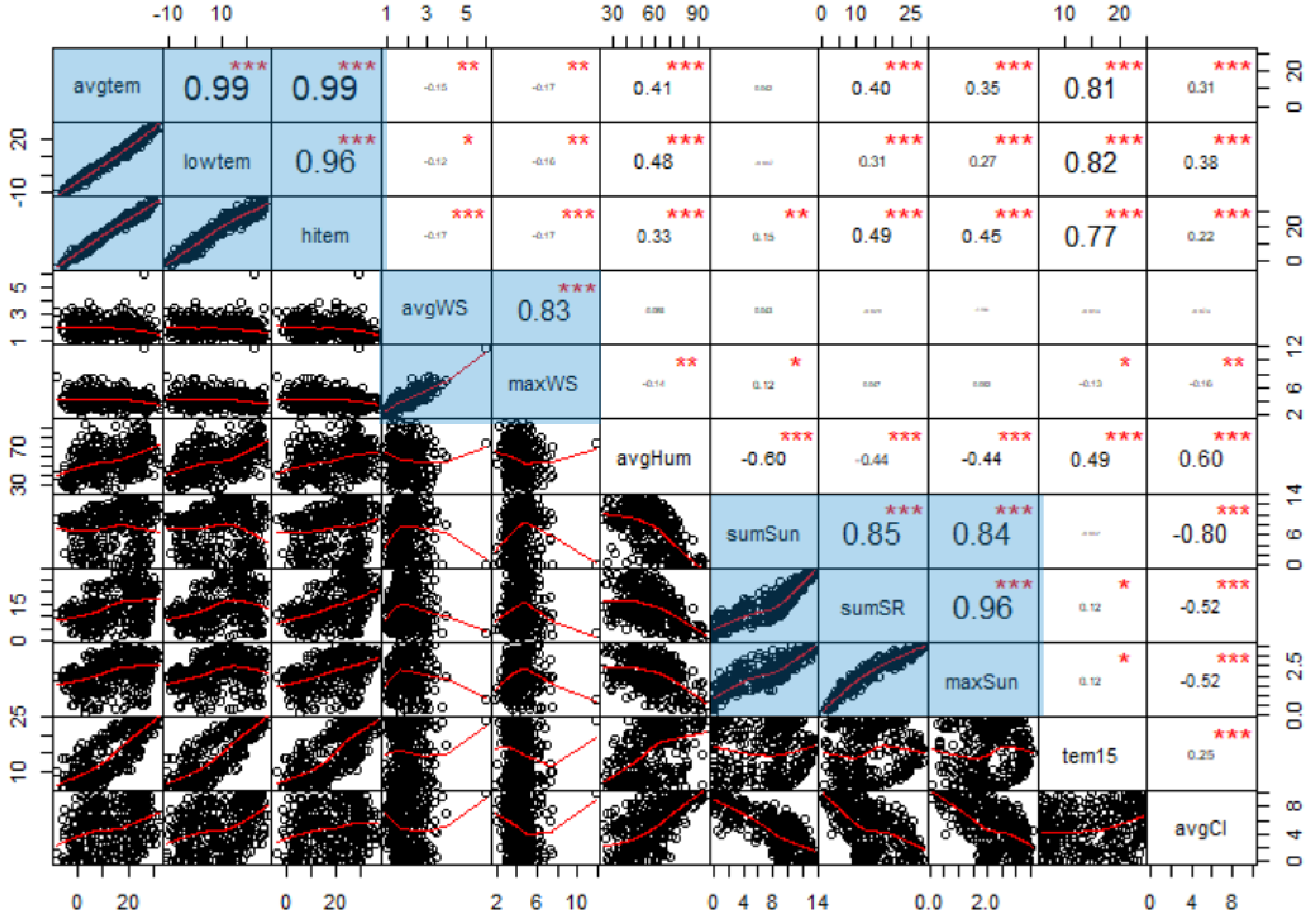


01

Variables : Correlation

- Check the correlation between explanatory variables

- avgtem lowtem hitem
- avgWS maxWS
- sumSun sunSR maxSun



variable name	detail
aud (Y)	the number of movie audiences (in thousands)
lowtem	minimum temperature (°C)
maxWS	maximum wind speed (m/s)
avgHum	average relative humidity (%)
maxSun	maximum solar radiation quantity in an hour (MJ/m2)
tem15	1.5m soil temperature (°C)
avgCl	average amount of clouds (1/10)

02

Find the best model

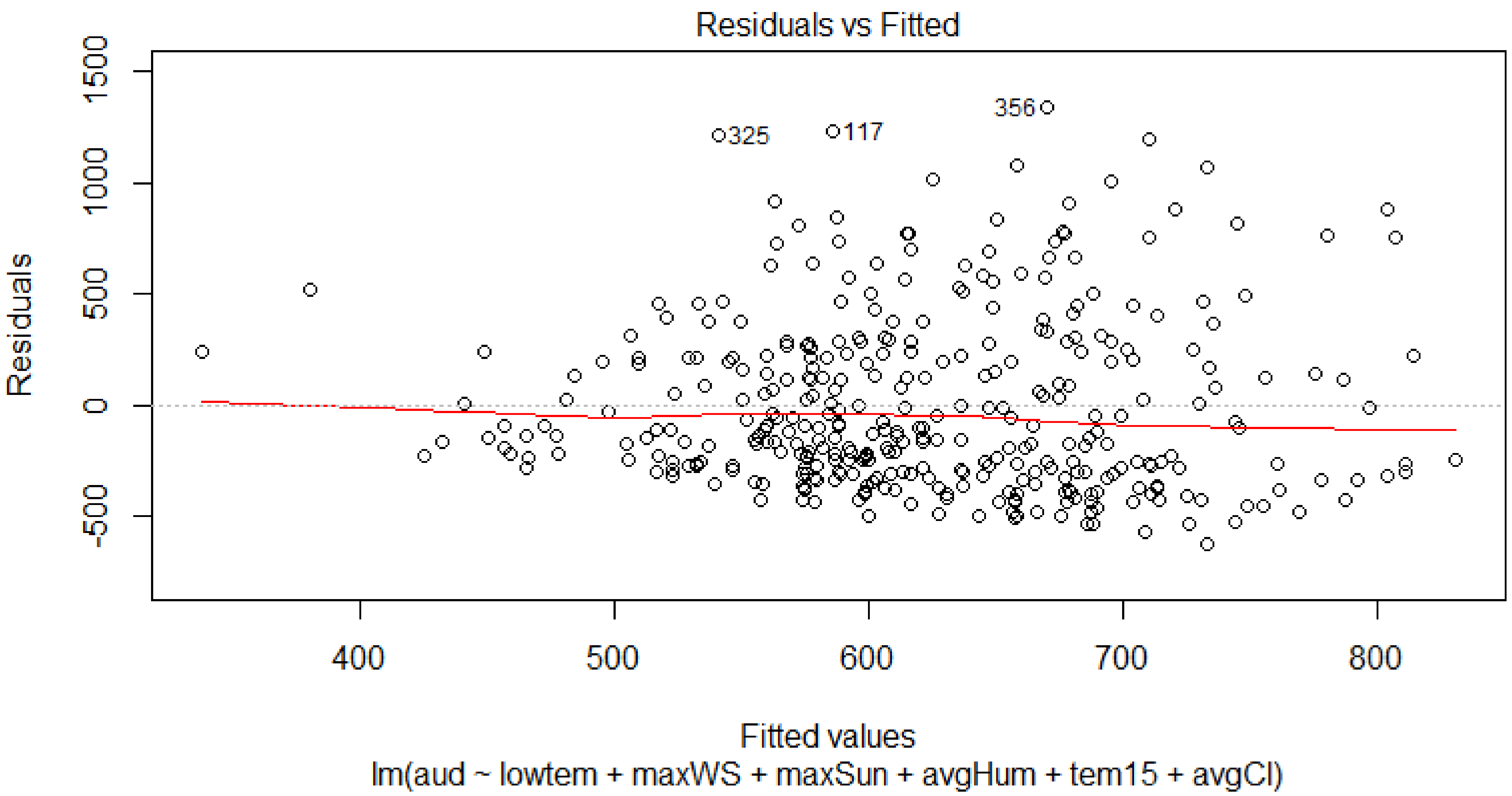
Model diagnostic

and developing the model

02

Model 1: Full Model

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \beta_5 \mathbf{X}_5 + \beta_6 \mathbf{X}_6 + \epsilon$$ (all 6 variables are included)



- It seems that the homogeneity of variance assumption is violated.

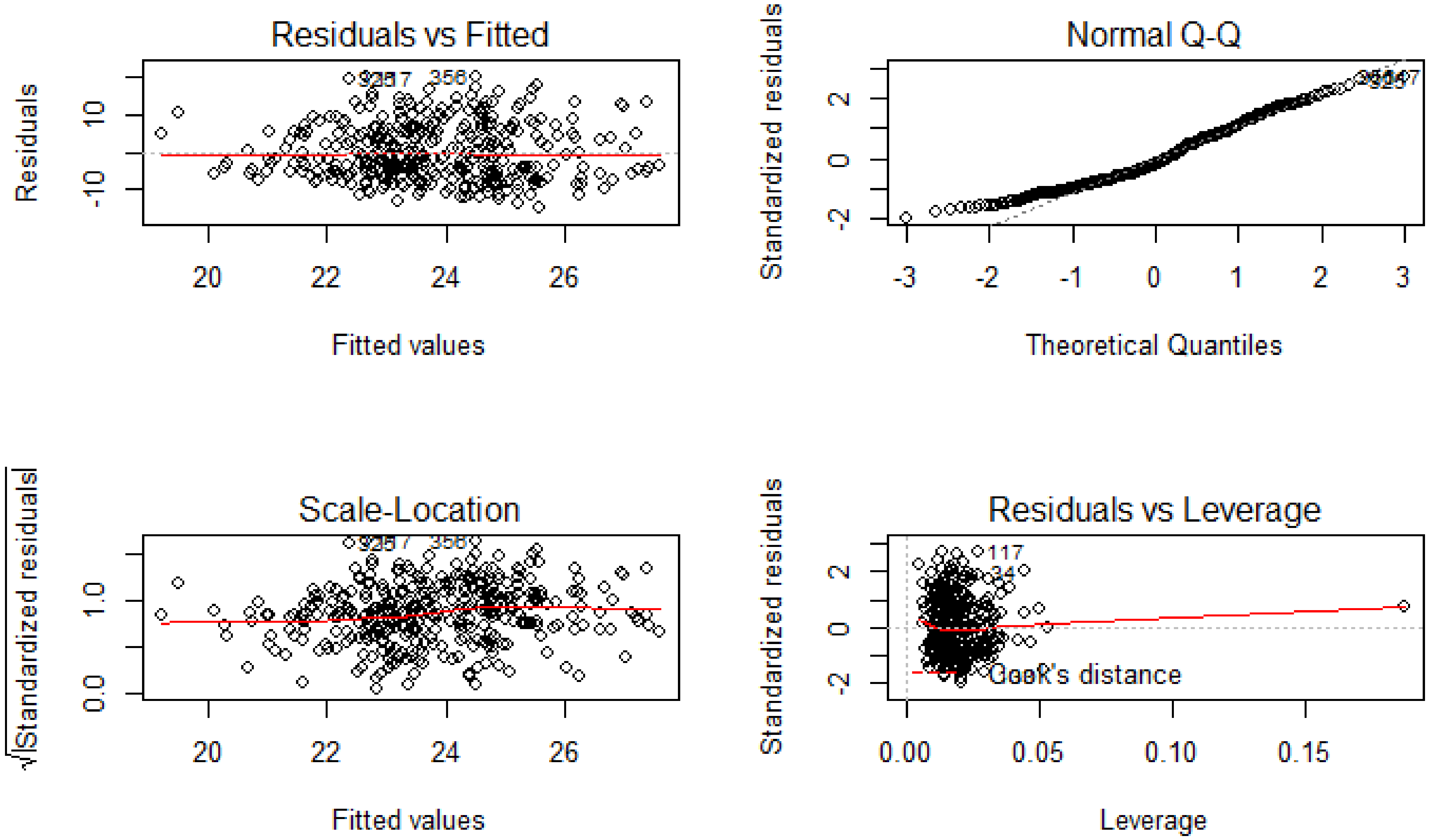
This is because the dependent variable is 'the number of movie audiences'

→ **Poisson distribution**

02

Model 2: Y → sqrt(Y)

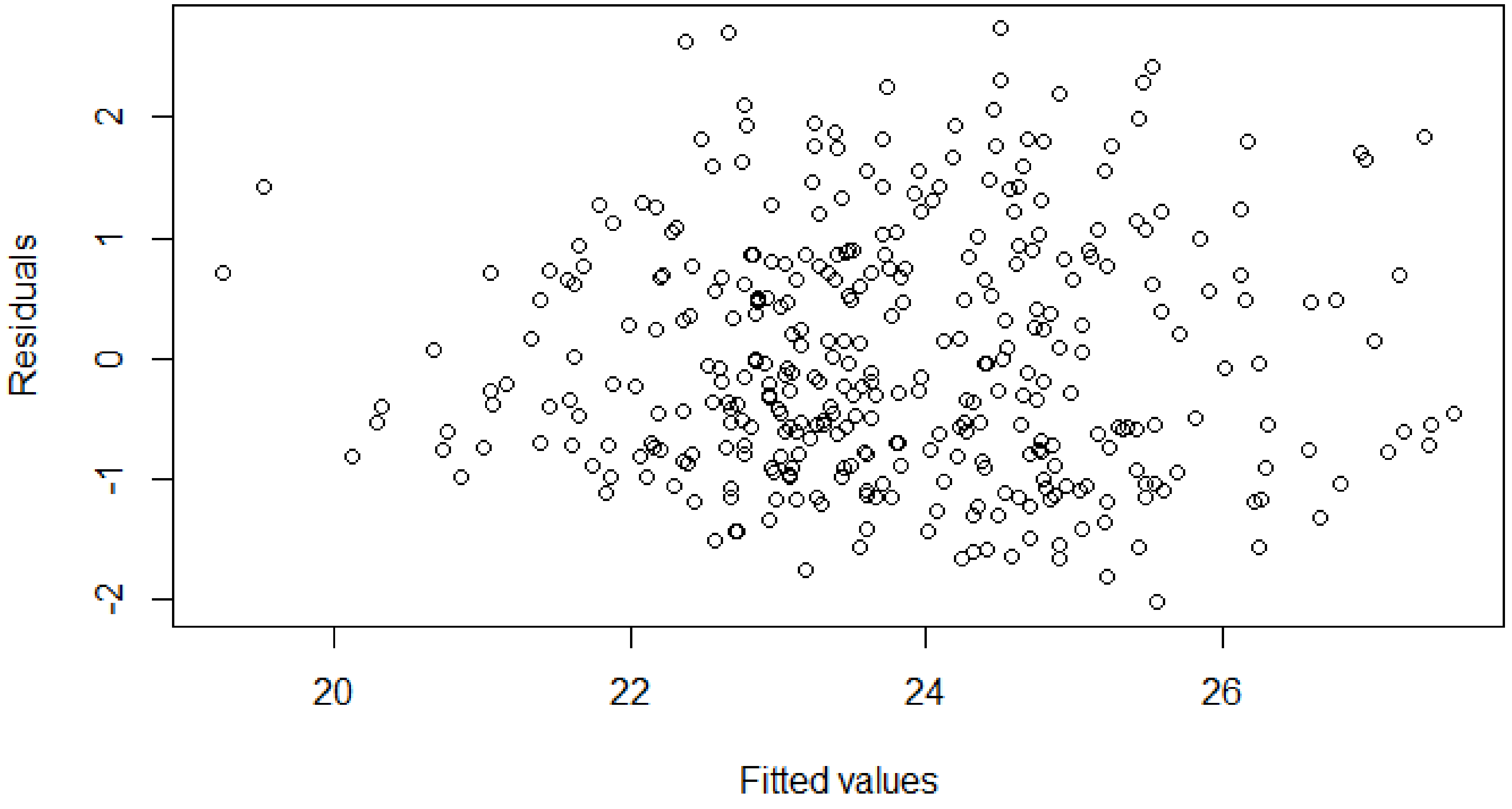
$$\sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$



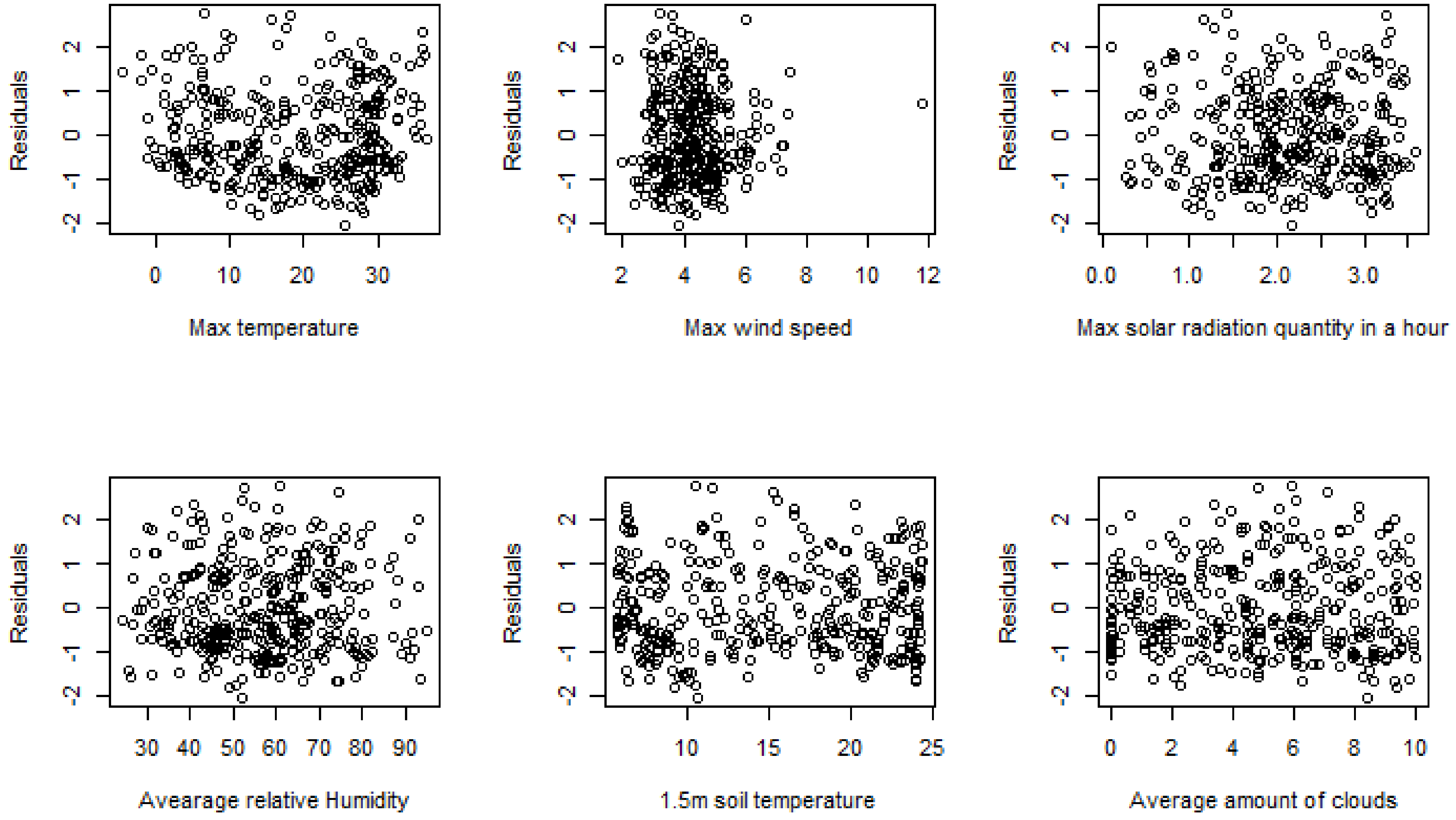
02

Model 2: Y → sqrt(Y)

Fitted values vs Residuals



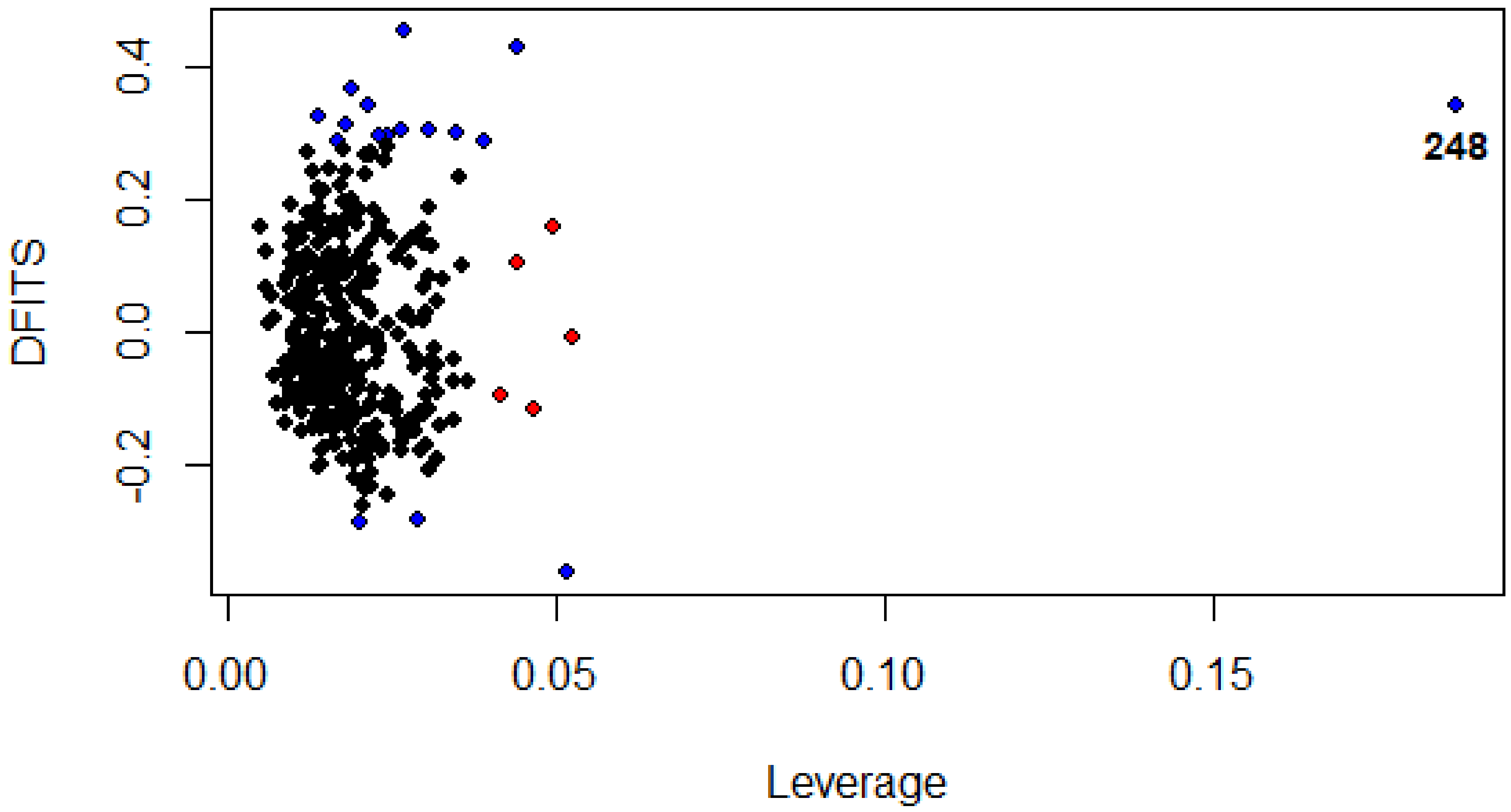
Explanatory variables vs Residuals



02

Model 2: Y → sqrt(Y)

Leverage vs DFITS



- Observation 248 is a high leverage point and influential point based on DFITS

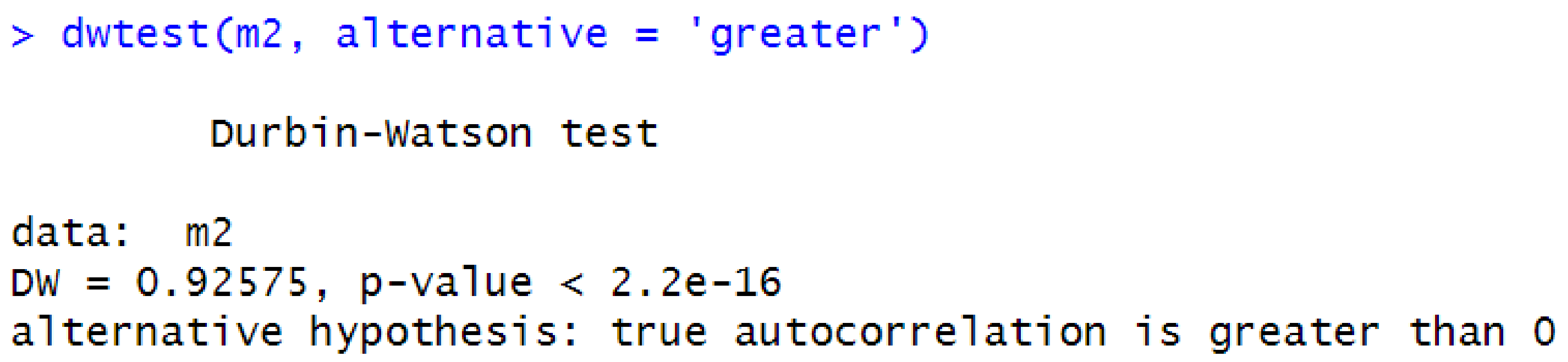
→ Drop 248th row

02

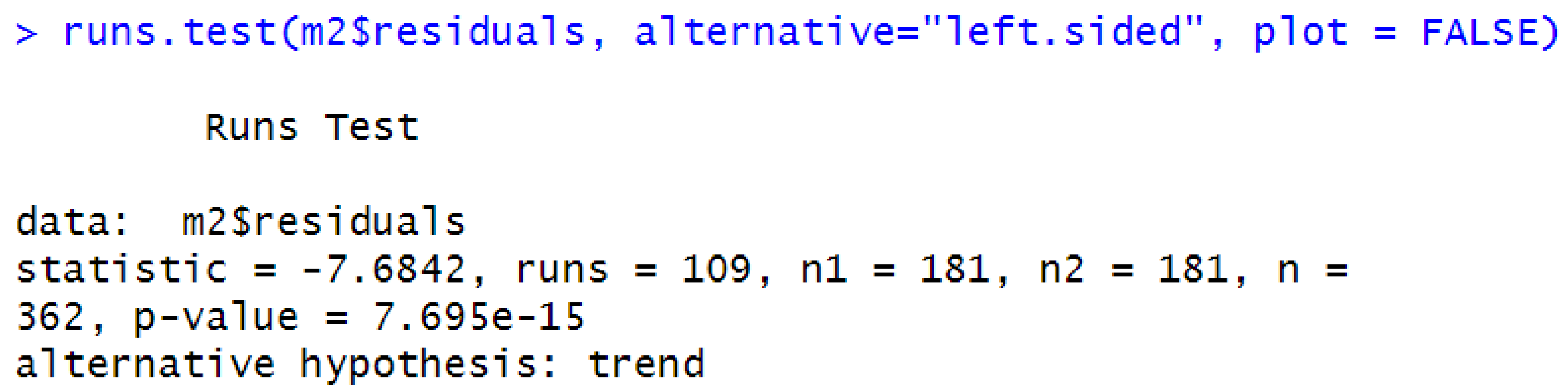
Model 2: Y → sqrt(Y)

Durbin-Watson test & Runs Test

- There is a strong positive correlation in adjacent errors.



```
> dwtest(m2, alternative = 'greater')

        Durbin-Watson test

data:  m2
DW = 0.92575, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

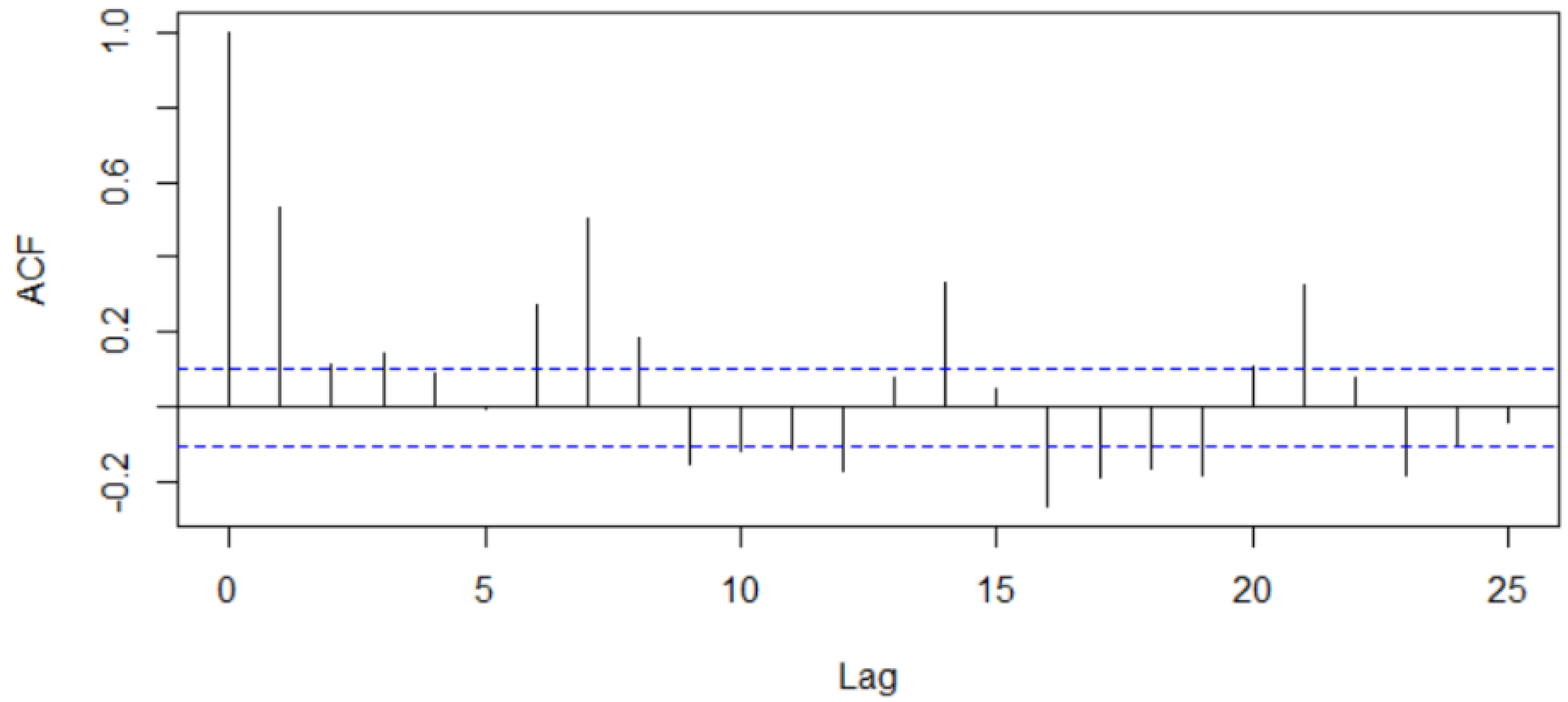


```
> runs.test(m2$residuals, alternative="left.sided", plot = FALSE)

        Runs Test

data:  m2$residuals
statistic = -7.6842, runs = 109, n1 = 181, n2 = 181, n =
362, p-value = 7.695e-15
alternative hypothesis: trend
```

02

Model 2: Y → sqrt(Y)

ACF vs lag



- lag 1, lag 7 autocorrelation is high

① AR(1) model

② 'day of the week' dummy variable

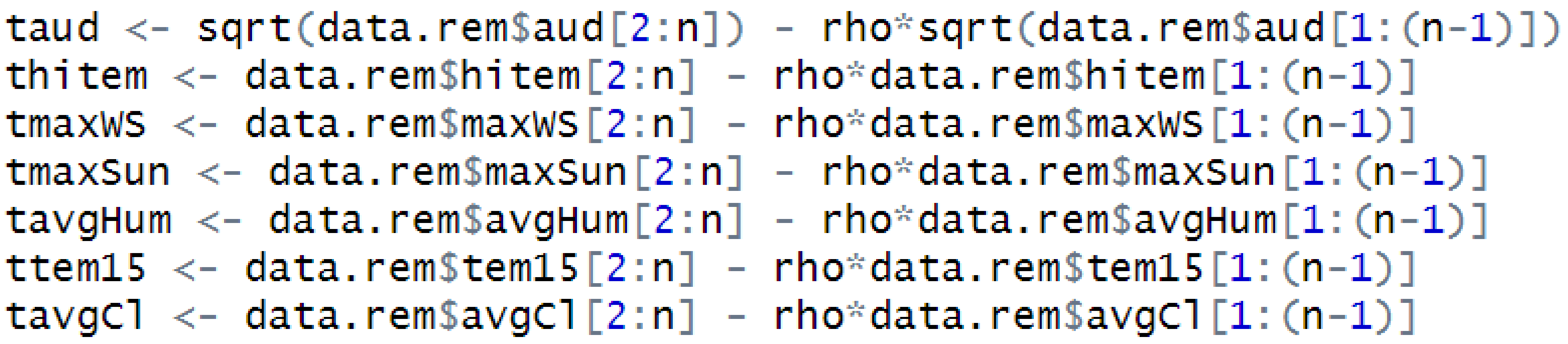
```
    0      1      2      3      4      5      6      7      8      9     10     11
1.000  0.531  0.116  0.144  0.089 -0.002  0.275  0.503  0.184 -0.153 -0.115 -0.112
```

02

Model 3-1: AR(1) Model

$$\mathrm{X}^*_{t,1} = \mathrm{X}_{t,1} - \hat{\rho}\mathrm{X}_{t-1,1}$$

transform all 6 variables in this way.

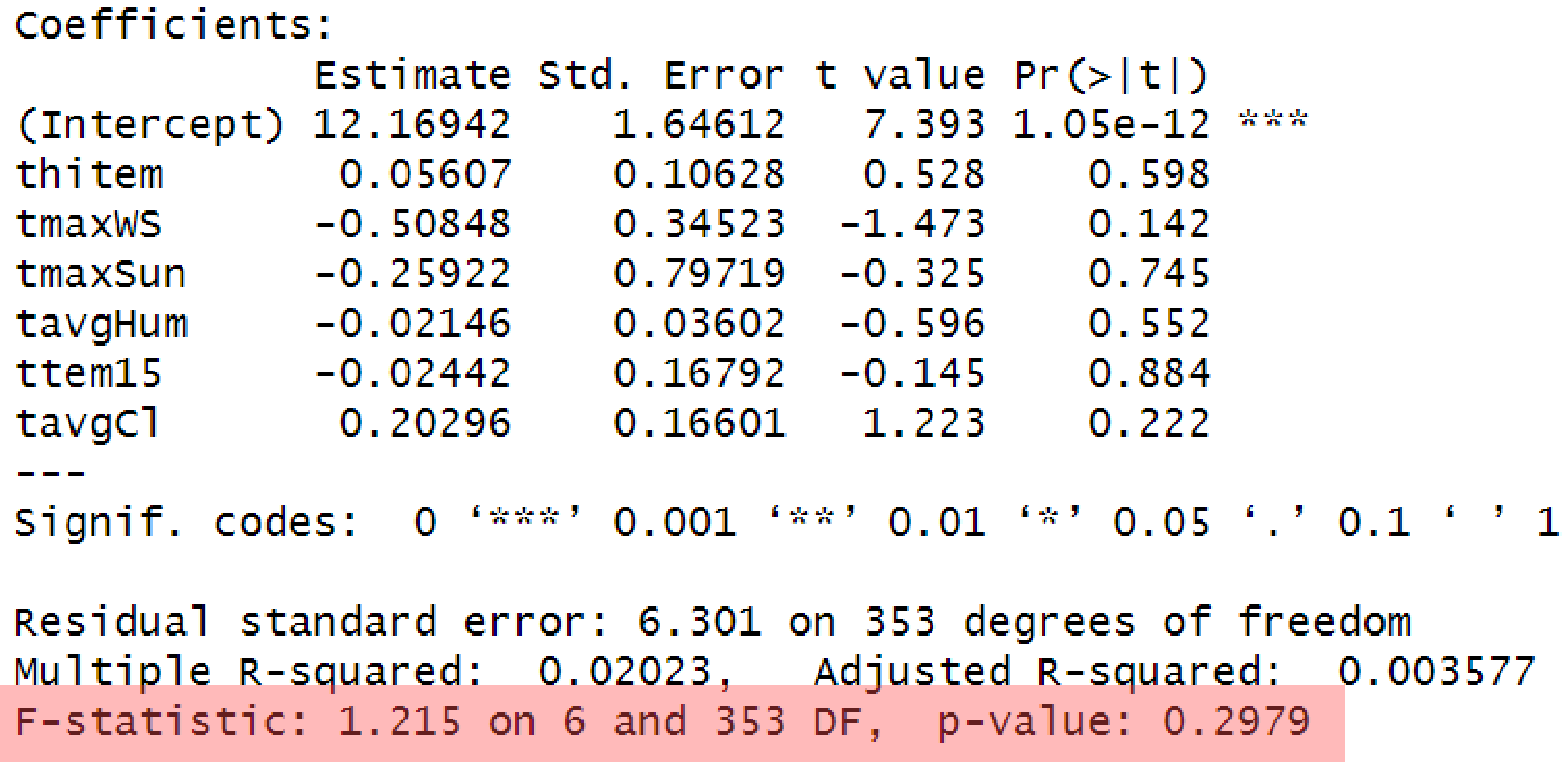


```
taud <- sqrt(data.rem$aud[2:n]) - rho*sqrt(data.rem$aud[1:(n-1)])
thitem <- data.rem$hitem[2:n] - rho*data.rem$hitem[1:(n-1)]
tmaxWS <- data.rem$maxWS[2:n] - rho*data.rem$maxWS[1:(n-1)]
tmaxSun <- data.rem$maxSun[2:n] - rho*data.rem$maxSun[1:(n-1)]
tavgHum <- data.rem$avgHum[2:n] - rho*data.rem$avgHum[1:(n-1)]
ttem15 <- data.rem$tem15[2:n] - rho*data.rem$tem15[1:(n-1)]
tavgCl <- data.rem$avgCl[2:n] - rho*data.rem$avgCl[1:(n-1)]
```

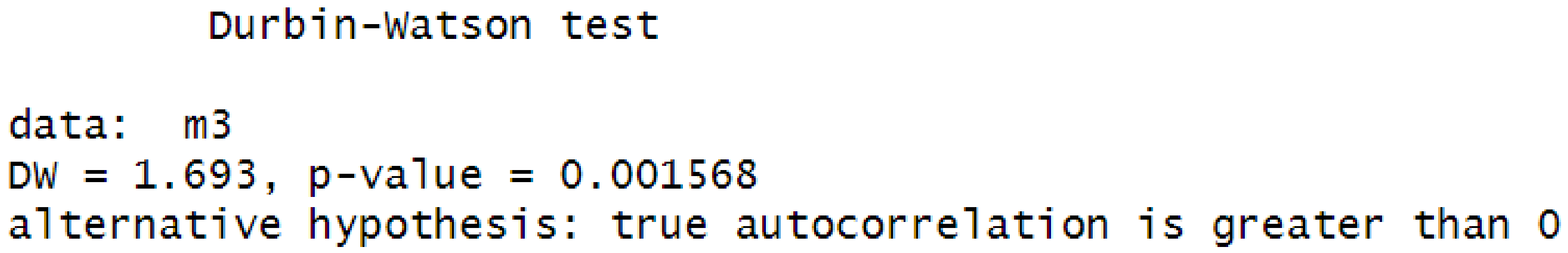
$$\sqrt{Y}^* = \beta^*_0 + \beta^*_1\mathrm{X}^*_1 + \beta^*_2\mathrm{X}^*_2 + \beta^*_3\mathrm{X}^*_3 + \beta^*_4\mathrm{X}^*_4 + \beta^*_5\mathrm{X}^*_5 + \beta^*_6\mathrm{X}^*_6 + \epsilon^*$$

02

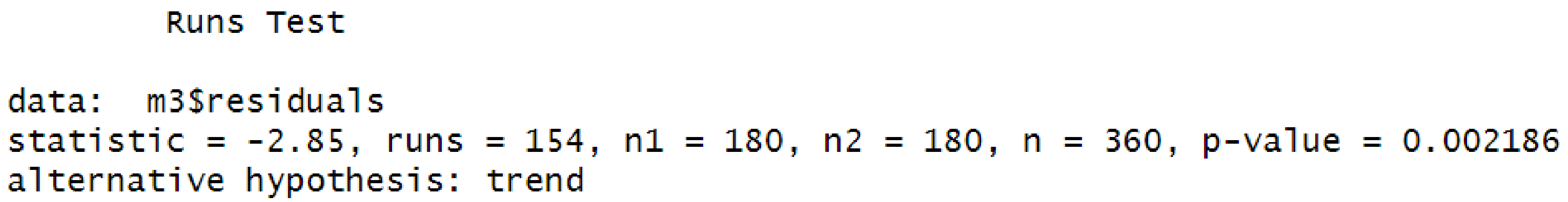
Model 3-1: AR(1) Model



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.16942    1.64612   7.393 1.05e-12 ***
thitem       0.05607    0.10628   0.528    0.598    
tmaxWS      -0.50848    0.34523  -1.473    0.142    
tmaxSun     -0.25922    0.79719  -0.325    0.745    
tavgHum     -0.02146    0.03602  -0.596    0.552    
ttem15      -0.02442    0.16792  -0.145    0.884    
tavgCl       0.20296    0.16601   1.223    0.222    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.301 on 353 degrees of freedom
Multiple R-squared:  0.02023,	Adjusted R-squared:  0.003577 
F-statistic: 1.215 on 6 and 353 DF,  p-value: 0.2979
```



```
	Durbin-Watson test

data:  m3
DW = 1.693, p-value = 0.001568
alternative hypothesis: true autocorrelation is greater than 0
```



```
	Runs Test

data:  m3$residuals
statistic = -2.85, runs = 154, n1 = 180, n2 = 180, n = 360, p-value = 0.002186
alternative hypothesis: trend
```

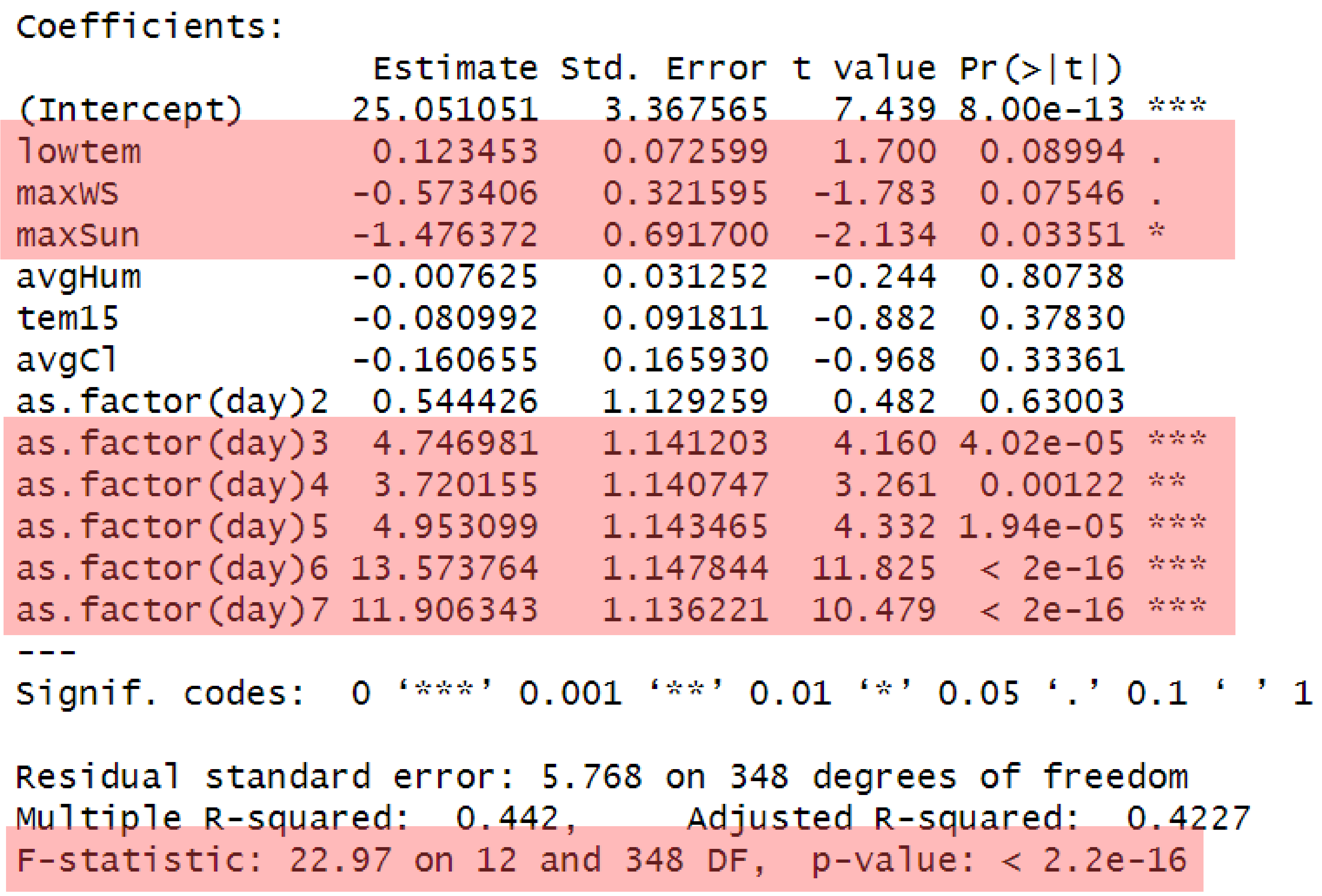
Still has a positive correlation in adjacent errors.

From F-test, this model is not significant at a 5% level of significance.

02

Model 3-2: Add dummy variable 'day of the week'

$$\sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \gamma_1 \text{day2} + \gamma_2 \text{day3} + \gamma_3 \text{day4} + \gamma_4 \text{day5} + \gamma_5 \text{day6} + \gamma_6 \text{day7} + \epsilon$$



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     25.051051   3.367565   7.439 8.00e-13 ***
lowtem           0.123453   0.072599   1.700  0.08994 .  
maxWS           -0.573406   0.321595  -1.783  0.07546 .  
maxSun          -1.476372   0.691700  -2.134  0.03351 *  
avgHum          -0.007625   0.031252  -0.244  0.80738    
tem15           -0.080992   0.091811  -0.882  0.37830    
avgCl           -0.160655   0.165930  -0.968  0.33361    
as.factor(day)2  0.544426   1.129259   0.482  0.63003    
as.factor(day)3  4.746981   1.141203   4.160 4.02e-05 ***
as.factor(day)4  3.720155   1.140747   3.261  0.00122 ** 
as.factor(day)5  4.953099   1.143465   4.332 1.94e-05 ***
as.factor(day)6 13.573764   1.147844  11.825  < 2e-16 ***
as.factor(day)7 11.906343   1.136221  10.479  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.768 on 348 degrees of freedom
Multiple R-squared:  0.442,	Adjusted R-squared:  0.4227 
F-statistic: 22.97 on 12 and 348 DF,  p-value: < 2.2e-16
```

- dummy variable 'day'

1: Monday

2: Tuesday

3: Wednesday

4: Thursday

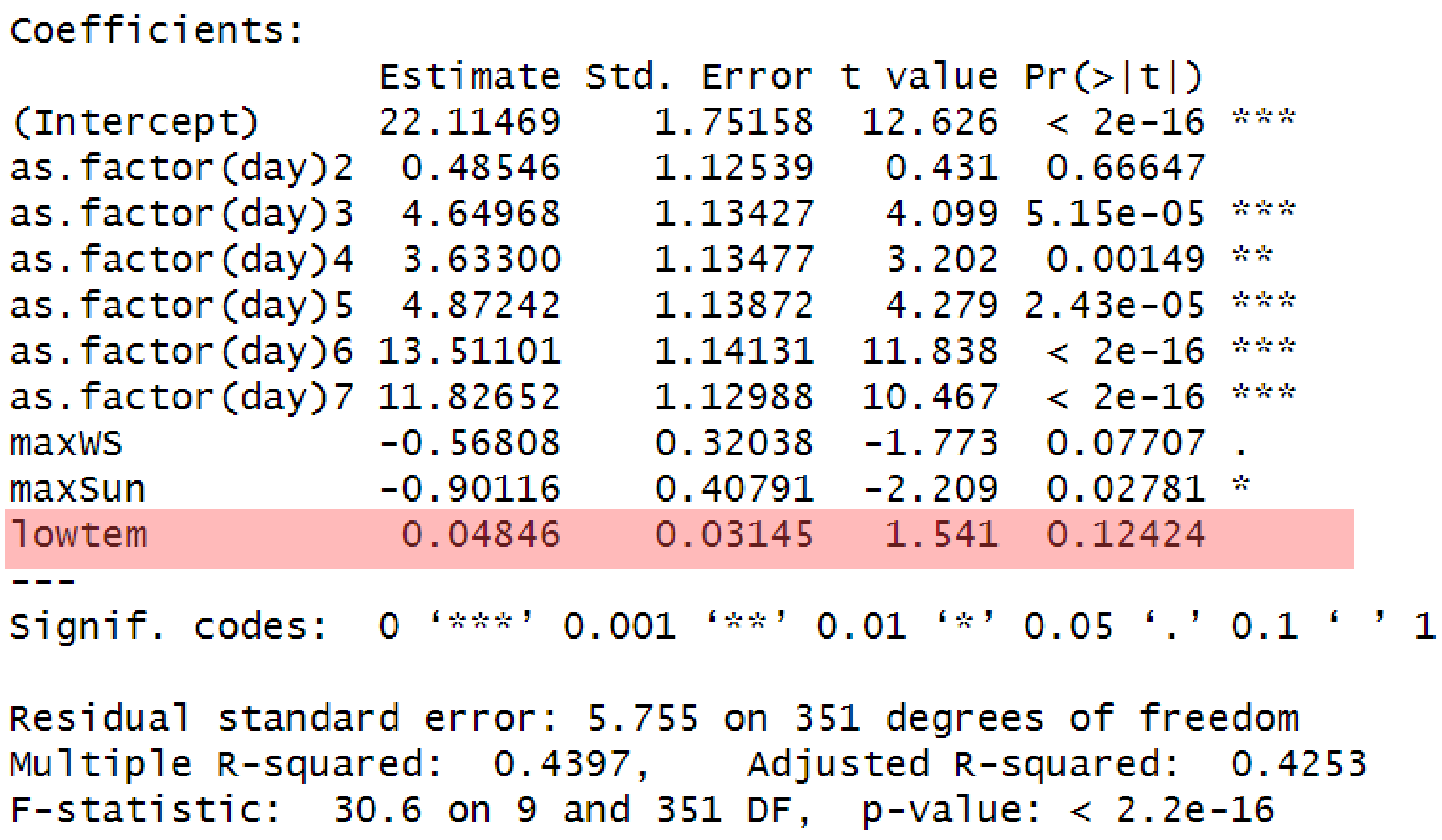
5: Friday

6: Saturday

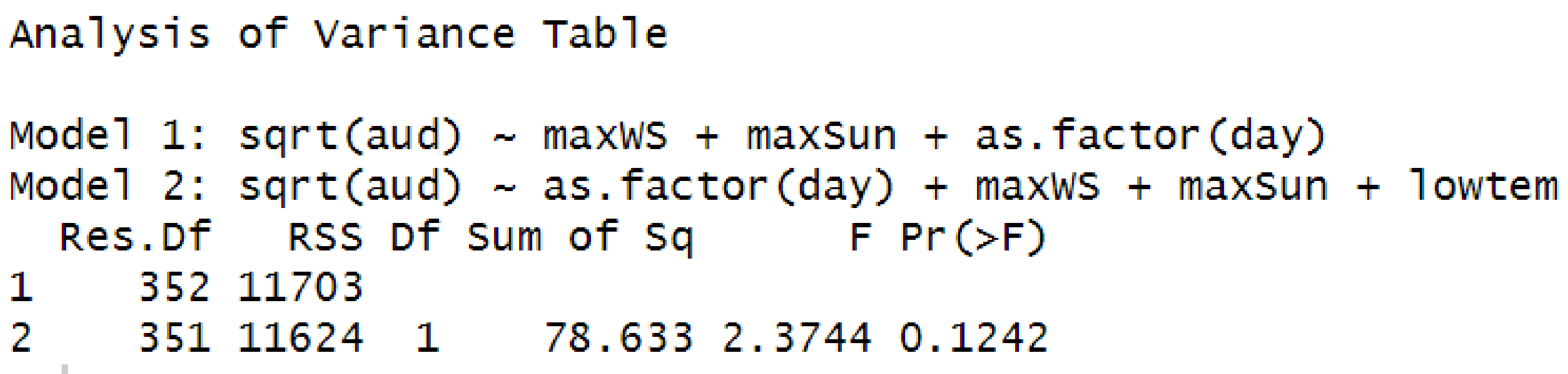
7: Sunday

02

Model 4: Model selection based on AIC



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      22.11469    1.75158  12.626  < 2e-16 ***
as.factor(day)2   0.48546    1.12539   0.431  0.66647    
as.factor(day)3   4.64968    1.13427   4.099 5.15e-05 ***
as.factor(day)4   3.63300    1.13477   3.202  0.00149 ** 
as.factor(day)5   4.87242    1.13872   4.279 2.43e-05 ***
as.factor(day)6  13.51101    1.14131  11.838  < 2e-16 ***
as.factor(day)7  11.82652    1.12988  10.467  < 2e-16 ***
maxWS            -0.56808    0.32038  -1.773  0.07707 .  
maxSun           -0.90116    0.40791  -2.209  0.02781 *  
lowtem            0.04846    0.03145   1.541  0.12424    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.755 on 351 degrees of freedom
Multiple R-squared:  0.4397,	Adjusted R-squared:  0.4253 
F-statistic:  30.6 on 9 and 351 DF,  p-value: < 2.2e-16
```



```
Analysis of Variance Table

Model 1: sqrt(aud) ~ maxWS + maxSun + as.factor(day)
Model 2: sqrt(aud) ~ as.factor(day) + maxWS + maxSun + lowtem
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    352 11703                           
2    351 11624  1    78.633 2.3744 0.1242
```

Finally, According to the result of the ANOVA test, drop 'lowtem' variable

03

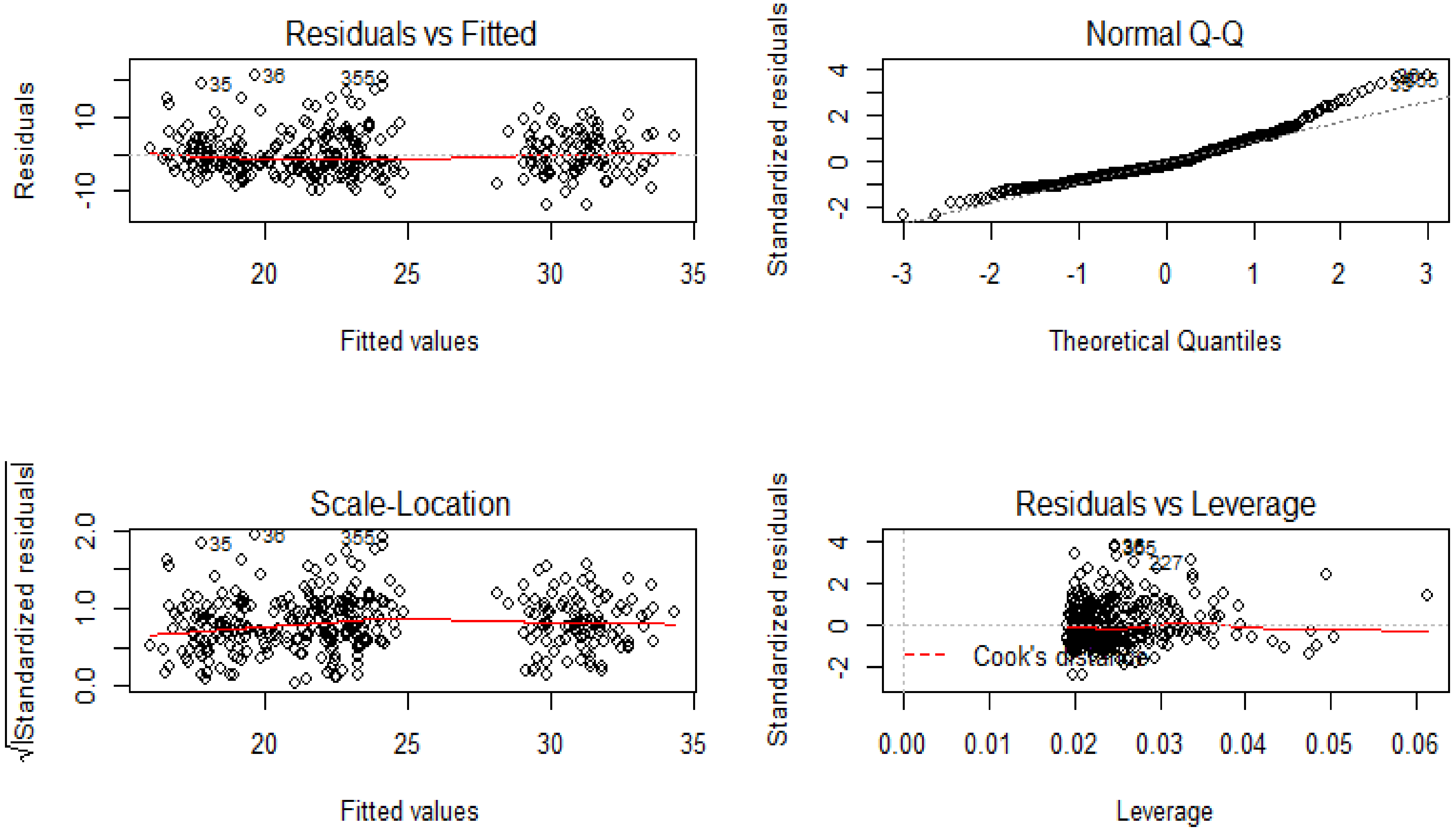
Final Model

Final Model Analysis

03

Final Model Analysis

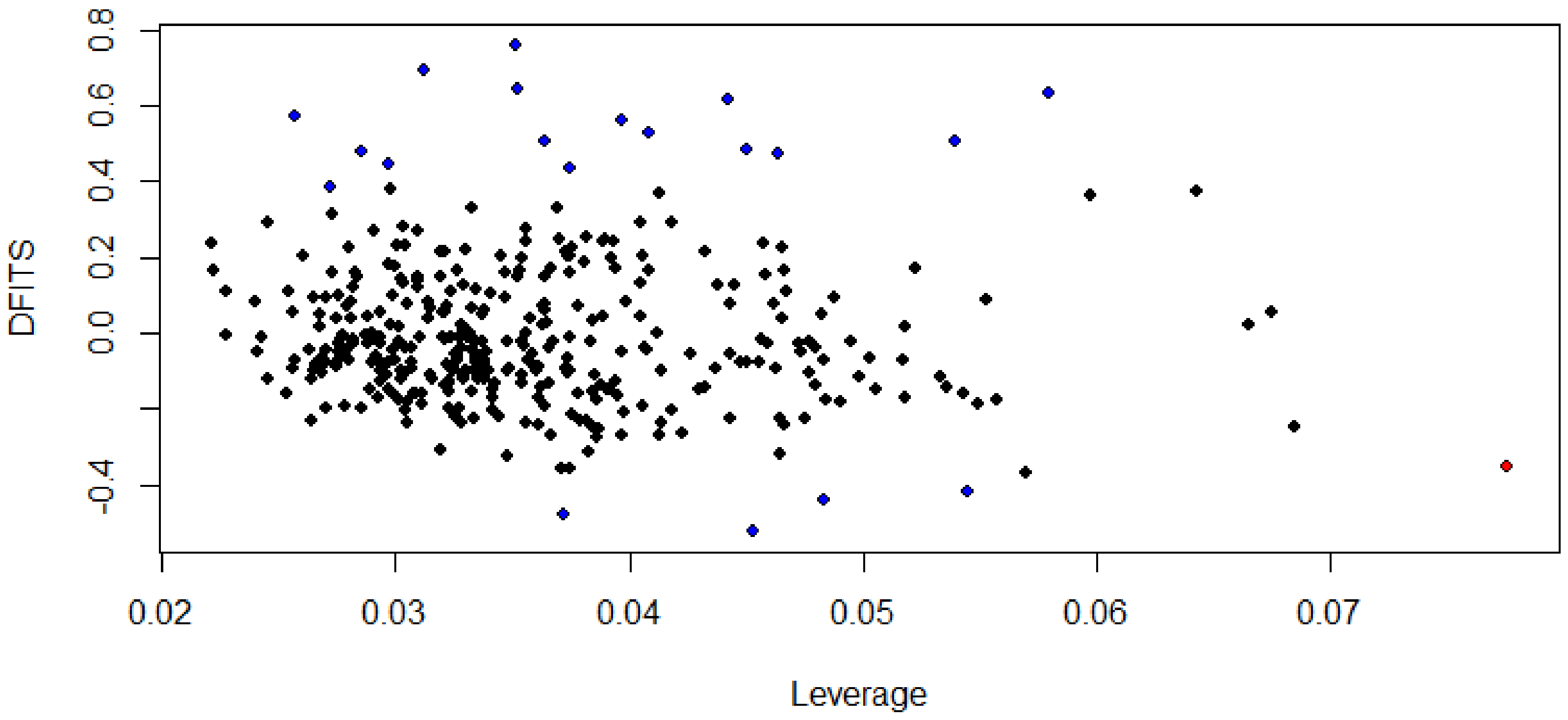
$$\sqrt{Y} = \beta_0 + \beta_1\mathrm{maxWS}_1 + \beta_2\mathrm{maxSun}_2 + \gamma_1\mathrm{day2} + \gamma_2\mathrm{day3} + \gamma_3\mathrm{day4} + \gamma_4\mathrm{day5} + \gamma_5\mathrm{day6} + \gamma_6\mathrm{day7} + \epsilon$$



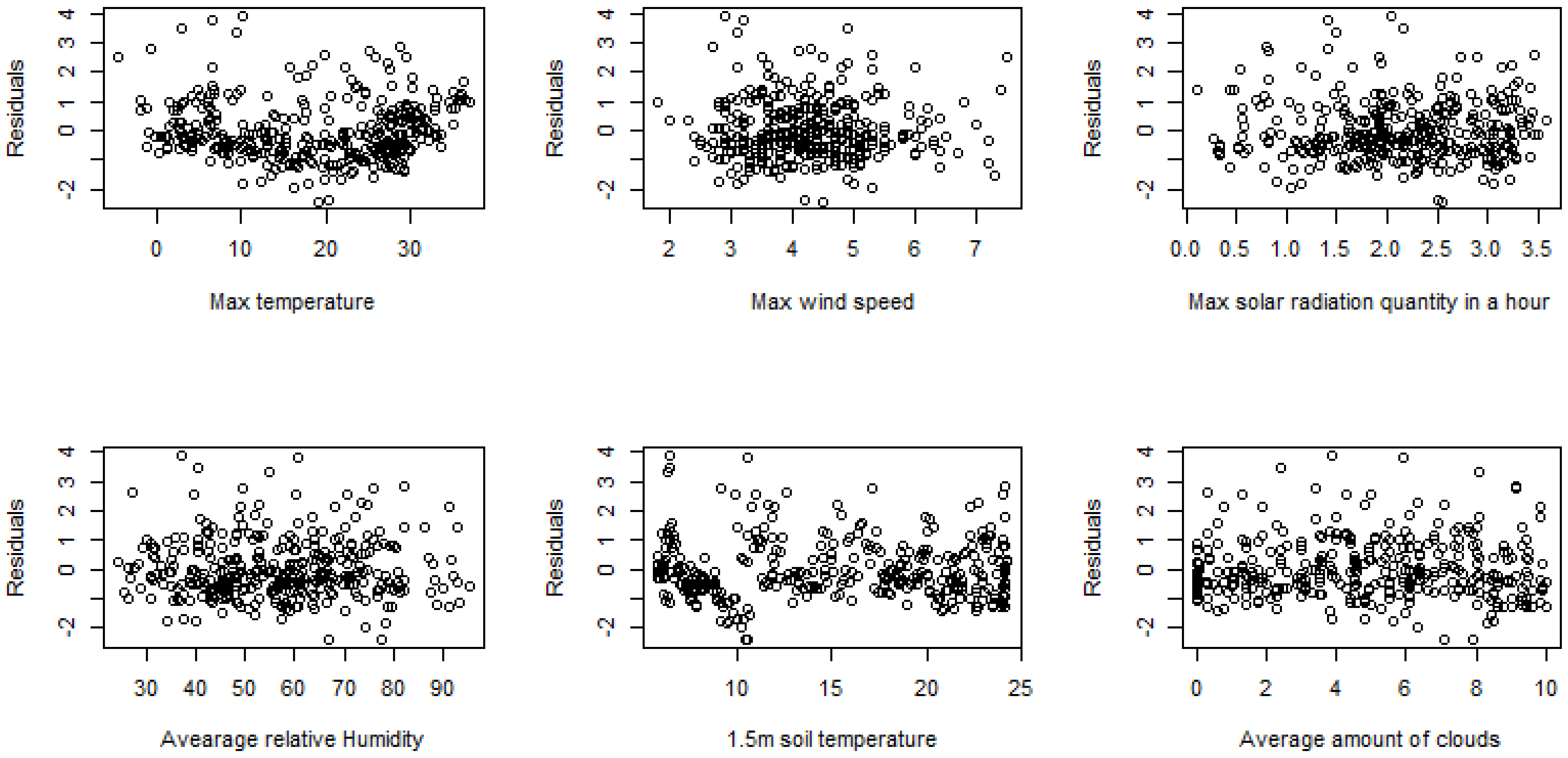
03

Final Model Analysis

Fitted values vs Residuals



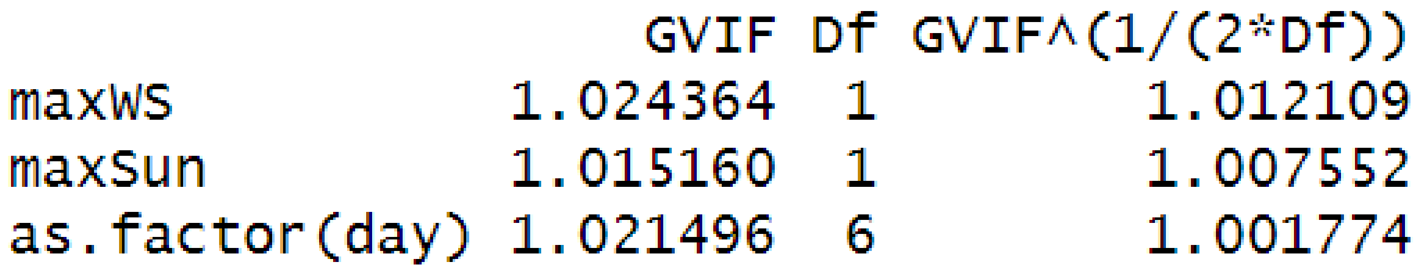
Explanatory variables vs Residuals



03

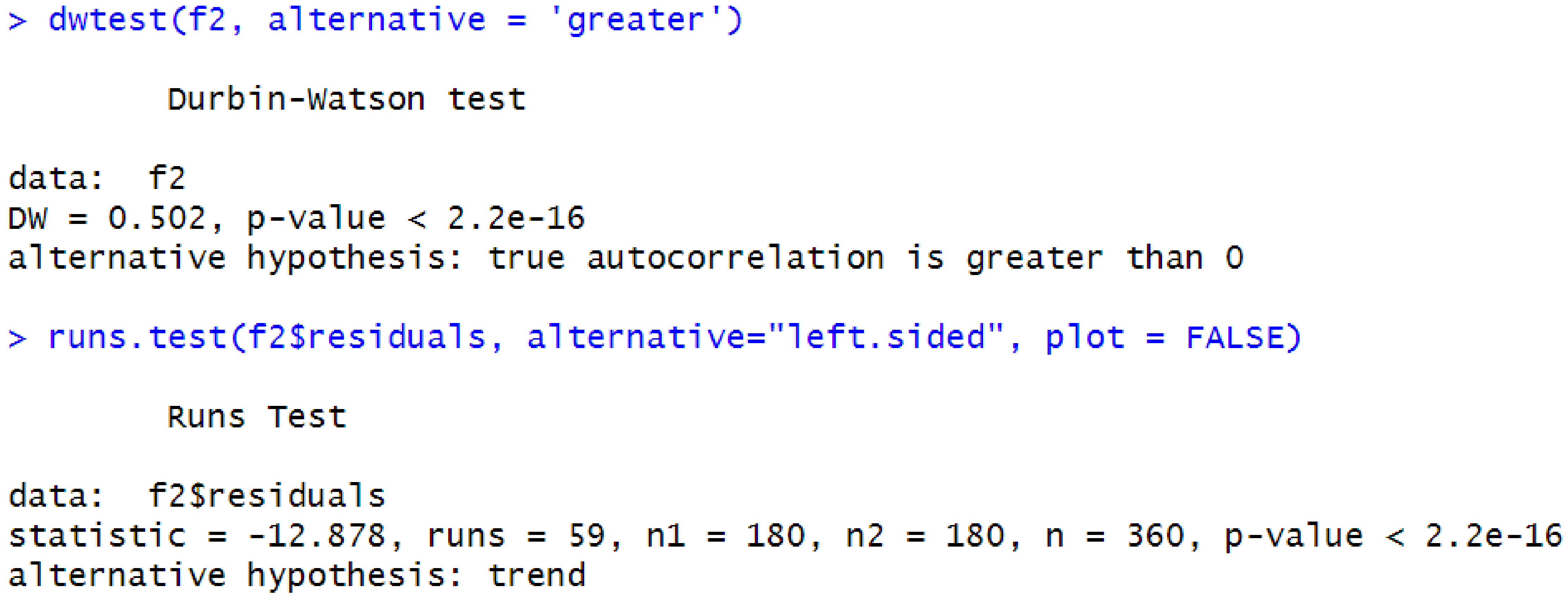
Final Model Analysis

Check Multicollinearity



```
                    GVIF Df GVIF^(1/(2*Df))
maxWS           1.024364  1        1.012109
maxSun          1.015160  1        1.007552
as.factor(day) 1.021496  6        1.001774
```

Check Autocorrelation



```
> dwtest(f2, alternative = 'greater')

	Durbin-Watson test

data:  f2
DW = 0.502, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

> runs.test(f2$residuals, alternative="left.sided", plot = FALSE)

	Runs Test

data:  f2$residuals
statistic = -12.878, runs = 59, n1 = 180, n2 = 180, n = 360, p-value < 2.2e-16
alternative hypothesis: trend
```

Still has a positive correlation in adjacent errors.

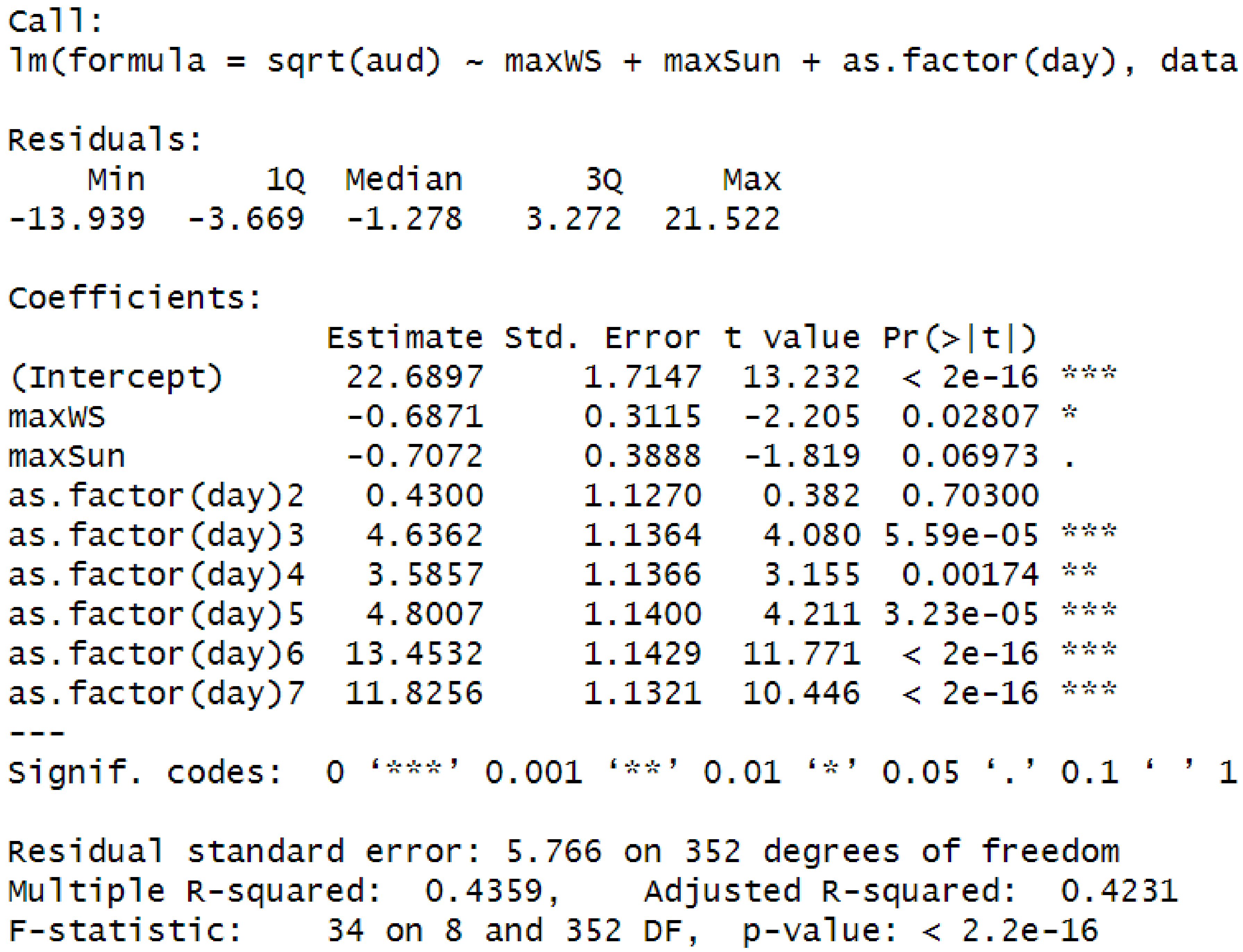
04

Conclusion

Conclusion and limitation of the model

04

Conclusion



```
Call:
lm(formula = sqrt(aud) ~ maxWS + maxSun + as.factor(day), data = data.rem)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.939  -3.669  -1.278   3.272  21.522 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      22.6897     1.7147  13.232  < 2e-16 ***
maxWS            -0.6871     0.3115  -2.205  0.02807 *  
maxSun           -0.7072     0.3888  -1.819  0.06973 .  
as.factor(day)2   0.4300     1.1270   0.382  0.70300    
as.factor(day)3   4.6362     1.1364   4.080 5.59e-05 ***
as.factor(day)4   3.5857     1.1366   3.155  0.00174 ** 
as.factor(day)5   4.8007     1.1400   4.211 3.23e-05 ***
as.factor(day)6  13.4532     1.1429  11.771  < 2e-16 ***
as.factor(day)7  11.8256     1.1321  10.446  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.766 on 352 degrees of freedom
Multiple R-squared:  0.4359,    Adjusted R-squared:  0.4231 
F-statistic:    34 on 8 and 352 DF,  p-value: < 2.2e-16
```

Among the weather variables,
the maxWS is significant at a 5% level of significance
and maxSun is significant at a 10% level of significance.

04

Conclusion

maxWS: Maximum wind speed
maxSun: maximum solar radiation quantity in an hour

$$\sqrt{\hat{Y}} = 22.69 - 0.69\text{maxWS} - 0.71\text{maxSun} + 0.43\text{day2} + 4.64\text{day3} + 3.59\text{day4} + 4.80\text{day5} + 13.45\text{day6} + 11.83\text{day7}$$

Interpretation of regression coefficients

- Except for Tuesday, the number of movie audiences on all days of the week is significantly different from that of Monday.

- Given that the other explanatory variables are the same, expected $\sqrt{\text{number of movie audiences}}$ is 0.69 decreased when maximum windspeed increased by 1(m/s)

- Given that the other explanatory variables are the same, expected $\sqrt{\text{number of movie audiences}}$ is 0.71 decreased when maximum solar radiation quantity in an hour increased by 1(MJ/m2)

Contrary to expectations, variables related to temperature or humidity were not significant.

04

Conclusion

maxWS: Maximum wind speed
maxSun: maximum solar radiation quantity in an hour

$$\sqrt{\hat{Y}} = 22.69 - 0.69\mathrm{maxWS} - 0.71\mathrm{maxSun} + 0.43\mathrm{day2} + 4.64\mathrm{day3} + 3.59\mathrm{day4} + 4.80\mathrm{day5} + 13.45\mathrm{day6} + 11.83\mathrm{day7}$$

Interpretation of regression coefficients

- Except for Tuesday, the number of movie audiences on all days of the week is significantly different from that of

Mond

- Give

decre

- Giv

people prefer indoor activities including watching movies on days with strong winds or sunlight

decreased when maximum solar radiation quantity in an hour increased by 1(MJ/m2)

Contrary to expectations, variables related to temperature or humidity were not significant.

04

Limitation

1. Adjusted R squared is 0.42

2. The response variable is based on the country while the explanatory variable is based on Seoul

3. Autocorrelation problem

Thank you

2021-1 Regression Analysis

20176735 Surin Kim