Assignment Submission Coversheet

Faculty of Science, Engineering and Built Environment

Student ID: 220129254

Student Name: Alina Kriutchenko

Campus: ☒ Burwood ☐ Waterfront ☐ Waurn Ponds ☐ Warrnambool ☐ Cloud

Assignment Title: Assignment 2

Due Date: Friday 25th September 2020, 17:00 pm

Course Code/Name: Master of Data Science

Unit Code/Name: SIT741

Unit Chair/Campus Coordinator: Sergiy Shelyag

Task 1: Source weather data

Which data source do you plan to use? Justify your decision.

I am planning to use The NOAA Climate Data since it is more convenient to process. Since our hospital is Swan, I tried to select the data by the "SWAN VIEW, AS".

It turns out that this weather station does not have the records from the whole year. It seems that the "SWAN VIEW, AS" area (where the "Swan District Hospitals" is located) has only 246 observations. The closest station "MIDLAND, AS" also has only 327 observations.

The weather station of Perth Airport is located close to the "Swan District Hospital" and contain all 365 observations. Therefore, I plan to use the "PERTH AIRPORT, AS" as a source.

How many rows are in the data and what time period does the data cover?

There are 365 rows in the data. The data cover the period from the 2013-07-01 to the 2014-06-30

Task 2: Model planning

How will the final model be used:
To further analyze the 2013-2014 emergency department (ED) demands at Perth and its connection with weather events?

How will it be relevant to the overcrowding problems at our EDs
It may be useful for better planning and organization of medical staff shifts based on the weather forecast.

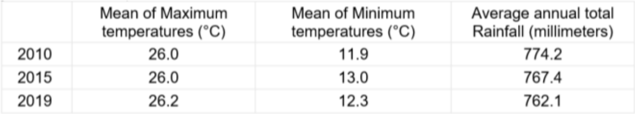
Who are the potential users of your model?
The hospitals authority can use this model for managing the emergency staff among hospitals based on weather.

What relationship do you plan to model or what do you want to predict?
I want to predict the relationships between weather events and hospital ED workload.

What is the response and the predictor variable?
The response variable is the admissions. The predictor variables are date, weekly seasoning, mean average temperature and precipitation data.

Will the variables in your model be routinely collected and made available soon enough for prediction?
The variables that will be collected and used for the prediction are the temperature and the precipitation data.

As you are likely to build your model on historical data, will the data in the future have similar characteristics?
If we look at the data from the Perth Airport Weather station in 2010, 2015, 2020, then we can see that there is no noticeable changes throughout the las decade. Therefore, if we consider to make predictions about the nearest future then the historical data is useable. The data source:
http://www.bom.gov.au/climate/current/annual/wa/perth.shtml



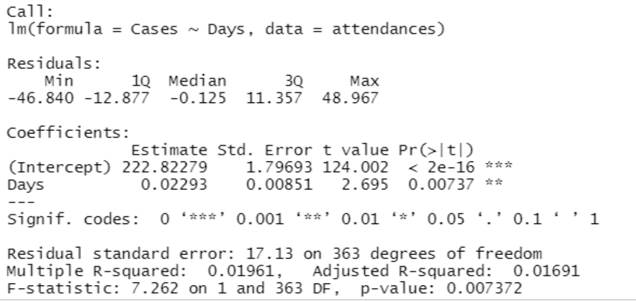
	Mean of Maximum temperatures (°C)	Mean of Minimum temperatures (°C)	Average annual total Rainfall (millimeters)
2010	26.0	11.9	774.2
2015	26.0	13.0	767.4
2019	26.2	12.3	762.1

What statistical method(s) will be applied to generate the model? Why?
Possibly, the regression analysis. It is useful for prediction of continuous dependent variable from a number of independent variables.

Task 3: Model the ED demands

I picked the “Swan District Hospital”. X: "DATE" or "Days" is the predictor or independent variable. Y: The ED demand variable(s) or dependent variable. In the first case it is attendances

Linear model 1 (x: Date)



```
Call:
lm(formula = Cases ~ Days, data = attendances)

Residuals:
    Min      1Q  Median      3Q     Max
-46.840 -12.877  -0.125  11.357  48.967

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 222.82279    1.79693 124.002  < 2e-16 ***
Days          0.02293    0.00851   2.695  0.00737 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.13 on 363 degrees of freedom
Multiple R-squared:  0.01961,	Adjusted R-squared:  0.01691
F-statistic: 7.262 on 1 and 363 DF,  p-value: 0.007372
```

Residuals: The median is not far from zero and the sum of absolute values of min and max is close to being equal, which is good. The sum of residuals equals to 0.482

According to the **Multiple R-squared**, the model can explain approximately 1.96% of variation in Cases.

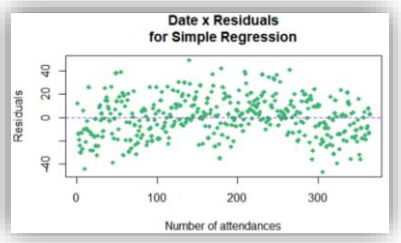
Adjusted R-squared: 0.01691

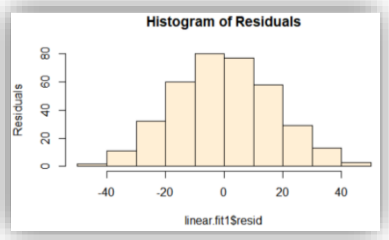
The p-value is 0.007374506, which is less than 0.05. Therefore, the model is statistically significant. We can reject the Null hypothesis.

However, the Days coefficient is small and close to zero and has a really week relationships with Y.

Residual standard error: 17.13 on 363 degrees of freedom, shows how far the observed Cases(Y values) are far from the predicted or fitted cases.

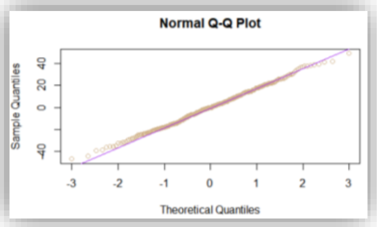
The intercept is 222.82279. **It means the estimated mean Y value is 222.82279 when x equal to zero.** The slope for the date (days) or the effect on the number of cases is 0.02293. The increase in cases in one day is 0.02293 or 2%. Based on the t-value and p-value the slope for the days is zero.



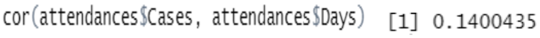


The plot shows that the there is no or week relationships between the residuals and the Days.

The histogram has a bell-curve shape which means that it probably normally distributed.

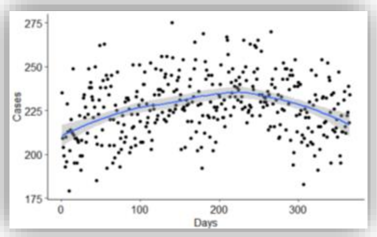


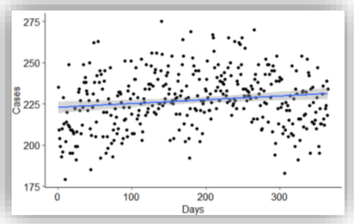
The Q-Q plot has point very close to the line and it means that residuals probably normally distributed.



```
cor(attendances$Cases, attendances$Days)  [1] 0.1400435
```

It means that there is too much of variation between the response variable Y (Cases) and predictor X (Days or Date)





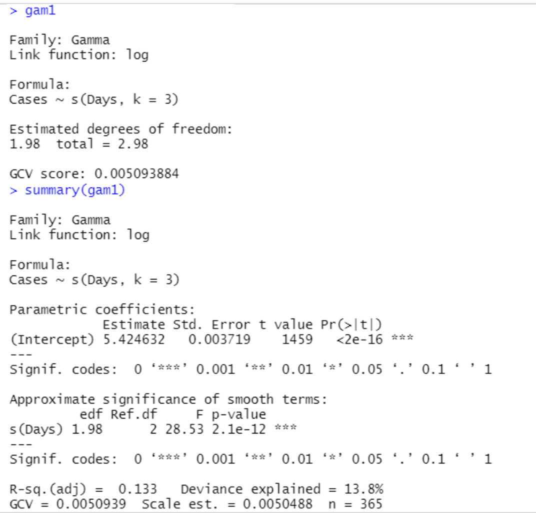
The graph indicates that there is a linear relationship between Cases and Dates variables, where it at first increased and then decreased. However, the correlation coefficient between these two variables is 0.1400435, which indicates a week relationships.

The plot above shows how the regression line is almost horizontal. Therefore, the linear function is not sufficient for modelling the trend of Y.

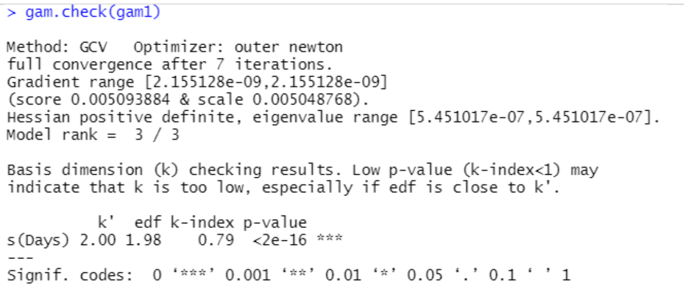
Task 3.3

As we are not interested in the trend itself, relax the linearity assumption by fitting a generalised additive model (GAM). Assess the model fit. Do you see patterns in the residuals indicating insufficient model fit?

Gam 1 (x: Date)



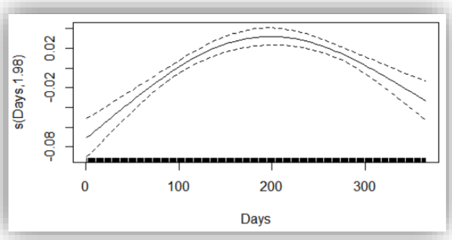
```
> gam1

Family: Gamma 
Link function: log 

Formula:
Cases ~ s(Days, k = 3)

Estimated degrees of freedom:
1.98  total = 2.98 

GCV score: 0.005093884     
> summary(gam1)

Family: Gamma 
Link function: log 

Formula:
Cases ~ s(Days, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.424632   0.003719    1459   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value    
s(Days) 1.98      2 28.53 2.1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.133   Deviance explained = 13.8%
GCV = 0.0050939  Scale est. = 0.0050488  n = 365
```



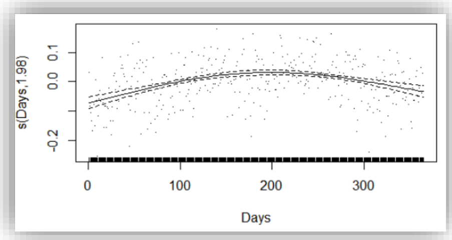
```
> gam.check(gam1)

Method: GCV   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [2.155128e-09,2.155128e-09]
(score 0.005093884 & scale 0.005048768).
Hessian positive definite, eigenvalue range [5.451017e-07,5.451017e-07].
Model rank =  3 / 3 

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

           k'  edf k-index p-value    
s(Days) 2.00 1.98    0.79  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The edf equals to 1.98, therefore the values are close to the line. Therefore, the significance of the smooth term “Days” is 1.98. The adjusted R-square is 0.133.
The precent of deviance explained by the model is only 13.8%.

Plot

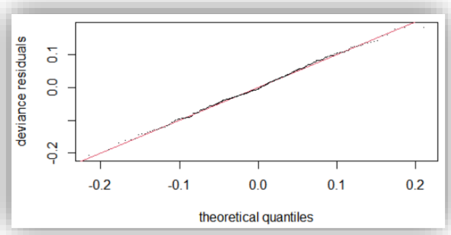


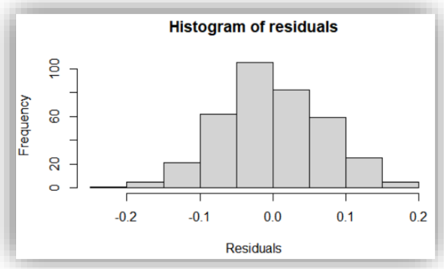
Residuals

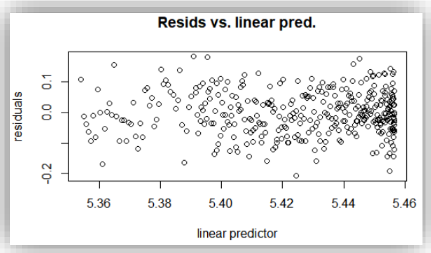


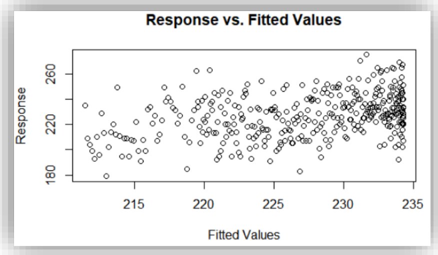
The model fits the data just slightly better.

The histogram of residuals indicate significant model fit.



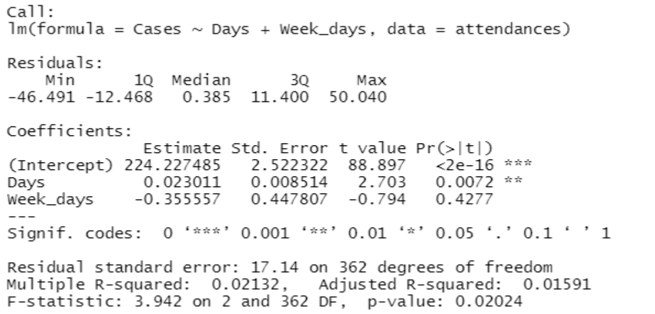






Task 3.4

Linear model 2 (x: Date and Weekly seasonality)

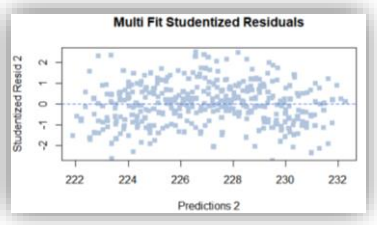


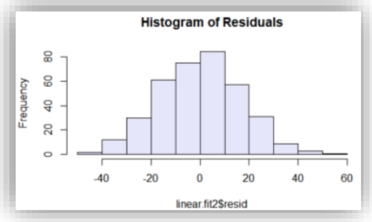
```
Call:
lm(formula = Cases ~ Days + Week_days, data = attendances)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.491 -12.468   0.385  11.400  50.040 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 224.227485   2.522322  88.897   <2e-16 ***
Days          0.023011   0.008514   2.703   0.0072 ** 
Week_days    -0.355557   0.447807  -0.794   0.4277    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.14 on 362 degrees of freedom
Multiple R-squared:  0.02132,   Adjusted R-squared:  0.01591 
F-statistic: 3.942 on 2 and 362 DF,  p-value: 0.02024
```

Residuals: The median is not far from zero and the sum of absolute values of min and max is close to being equal, which is good. The sum of residuals equals to 2.86. This is worse than for the previous model linear model.
According to the **Multiple R-squared**, the model can explain approximately 2.13% of variation in Cases. This is better than for the previous model linear model.

Adjusted R-squared: 0.01591 – it slightly less than the previous linear model. Therefore, the model did not improve with the new term.
The p-value is 0.02024, which is less than 0.05. The model has some statistical significance.
We can reject the Null hypothesis.
The Days coefficient is small and close to zero and has a really week relationships with Y.
The Week_days coefficient is small and close to zero and has a relatively week relationships with Y.

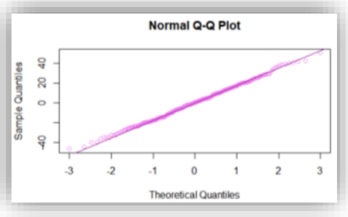
Residual standard error: 17.14 on 363 degrees of freedom, shows how far the observed Cases(Y values) are far from the predicted or fitted cases. This is almost the as in the previous liner model.
The intercept is 224.227485. It means the estimated mean Y value is 224.227485 when all x equal to zero. The slope for the date (days) or the effect on the number of cases is 0.02293. The slope for the week or the effect on the number of cases is -0.355557.



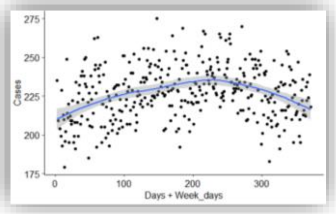


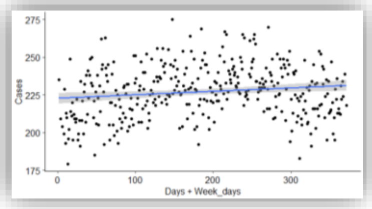
The plot shows that there is no or week relationships between the residuals and Days and Week_days

The histogram has a bell-curve shape which means that it probably normally distributed. Possibly the histogram of residuals is slightly better fit for the first model.



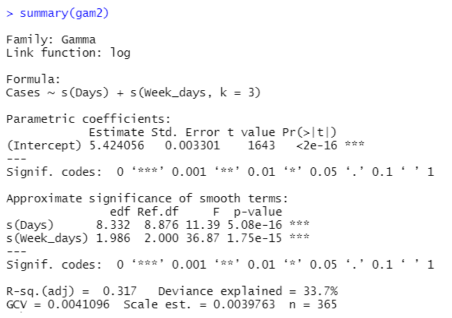
The Q-Q plot has point very close to the line and it means that residuals probably normally distributed for the second model as well.



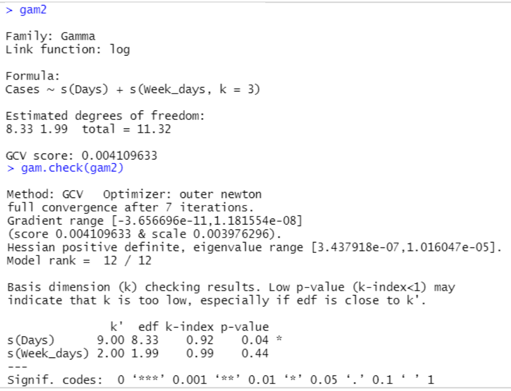


The graph indicates that there is a linear relationship between Cases and Dates variables, where it at first increased and then decreased.

GAM 2 (x: Date and Weekly seasonality)



```
> summary(gam2)

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Week_days, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.424056   0.003301    1643   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df     F  p-value
s(Days)      8.332  8.876 11.39 5.08e-16 ***
s(Week_days) 1.986  2.000 36.87 1.75e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.317   Deviance explained = 33.7%
GCV = 0.0041096  Scale est. = 0.0039763  n = 365
```



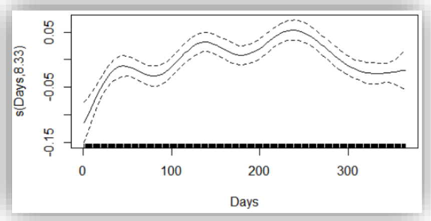
```
> gam2

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Week_days, k = 3)

Estimated degrees of freedom:
8.33 1.99  total = 11.32

GCV score: 0.004109633
> gam.check(gam2)

Method: GCV   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-3.656696e-11,1.181554e-08]
(score 0.004109633 & scale 0.003976296).
Hessian positive definite, eigenvalue range [3.437918e-07,1.016047e-05].
Model rank =  12 / 12

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

               k'  edf k-index p-value
s(Days)      9.00 8.33    0.92    0.04 *
s(Week_days) 2.00 1.99    0.99    0.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

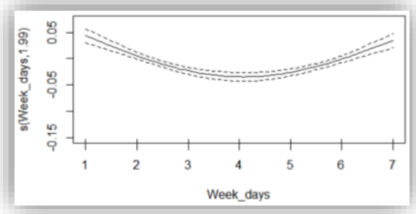
The edf is higher for the Days rather than Week_days. Therefore, the significance of the smooth term “Days” is higher. The adjusted R-square is 0.317, which is higher than previous 0.133 in the GAM 2. Therefore, the term “Week_days” slightly improved the model.

The precent of deviance explained by the model is noticeably higher: 33.7%.

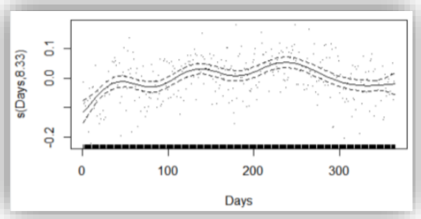
summary(gam2)

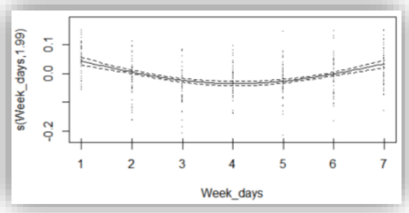
plot(gam2)



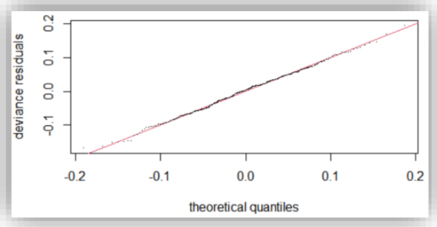


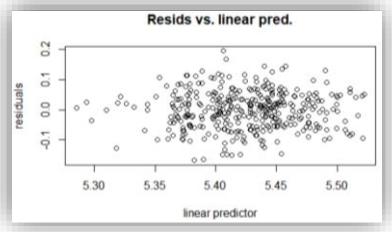
plot(gam2, residuals = TRUE)

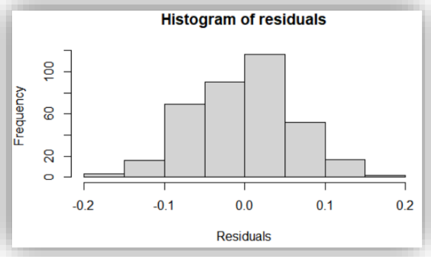


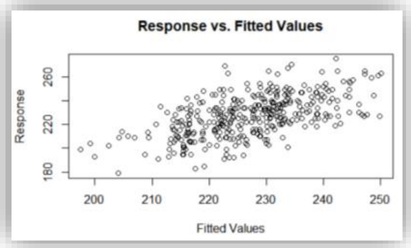


gam.check(gam2)



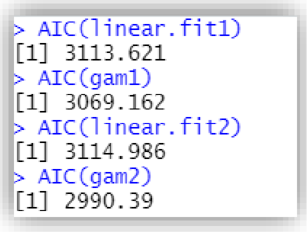






Compare the models using the Akaike information criterion (AIC).

Based on the AIC, one model is not significantly better than another is since the Akaike information criterion is not more than two time higher.



```
> AIC(linear.fit1)
[1] 3113.621
> AIC(gam1)
[1] 3069.162
> AIC(linear.fit2)
[1] 3114.986
> AIC(gam2)
[1] 2990.39
```

Task 3.6

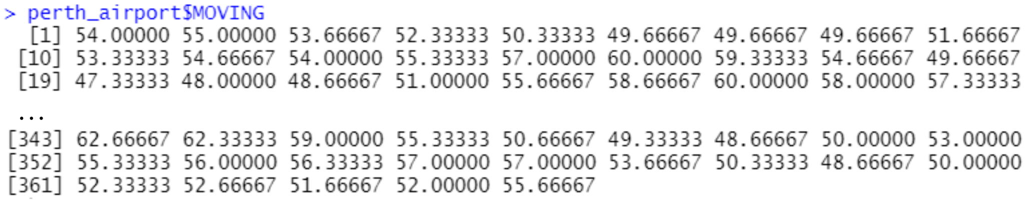
Is your day-of-the-week variable numeric, ordinal, or categorical? Does the decision affect the model fit?

In this case, the day-of-the-week is numerical type. The categorical or categorical variables like “Monday”, “Tuesday” etc need to be converted before being used in the linear regression.

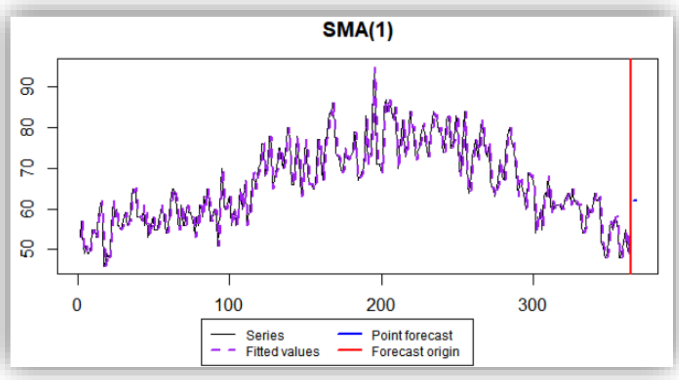
Task 4 Heatwaves and ED demands

Task 4.1: Measuring heatwave

Use the NOAA data to calculate the daily EHF values for the Perth area during the relevant time period. Plot the daily EHF values.

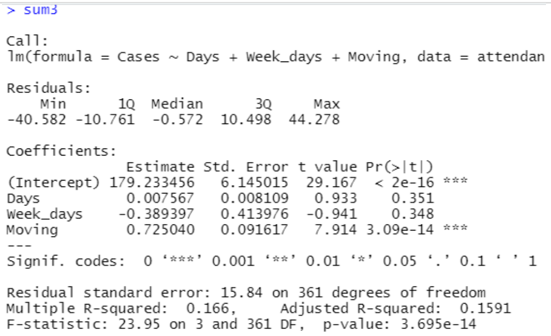


```
> perth_airport$MOVING
  [1] 54.00000 55.00000 53.66667 52.33333 50.33333 49.66667 49.66667 49.66667 51.66667
 [10] 53.33333 54.66667 54.00000 55.33333 57.00000 60.00000 59.33333 54.66667 49.66667
 [19] 47.33333 48.00000 48.66667 51.00000 55.66667 58.66667 60.00000 58.00000 57.33333
 …
[343] 62.66667 62.33333 59.00000 55.33333 50.66667 49.33333 48.66667 50.00000 53.00000
[352] 55.33333 56.00000 56.33333 57.00000 57.00000 53.66667 50.33333 48.66667 50.00000
[361] 52.33333 52.66667 51.66667 52.00000 55.66667
```



Task 4.2: Models with EHF

Linear model 3 (x: Date, Weekly seasonality and Moving (ETF))



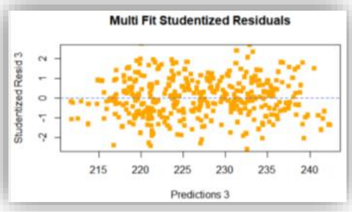
```
> sum3

Call:
lm(formula = Cases ~ Days + Week_days + Moving, data = attendan

Residuals:
    Min      1Q  Median      3Q     Max 
-40.582 -10.761  -0.572  10.498  44.278 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 179.233456   6.145015  29.167  < 2e-16 ***
Days          0.007567   0.008109   0.933    0.351    
Week_days    -0.389397   0.413976  -0.941    0.348    
Moving        0.725040   0.091617   7.914 3.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.84 on 361 degrees of freedom
Multiple R-squared:  0.166,	Adjusted R-squared:  0.1591 
F-statistic: 23.95 on 3 and 361 DF,  p-value: 3.695e-14
```

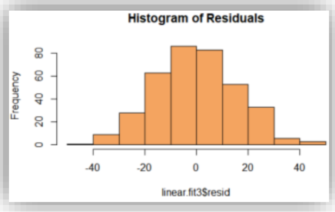
Residuals: The median is not far from zero and the sum of absolute values of min and max is close to being equal, which is good. The sum of residuals equals to 2.86 and this is similar to linear model 2.

According to the **Multiple R-squared**, the model can explain approximately 16.6% of variation in Cases. This is significantly improved number in comparison to linear models 1 and 2.

Adjusted R-squared: 0. 1591. It is significantly higher than the previous linear model. Therefore, the model significantly improved with the ETF.
The p-value is 3.695e-14, which is less than 0.05 or close to zero. The model has statistical significance. We can reject the Null hypothesis.
The Days coefficient is small and close to zero and has a really week relationships with Y.
The Week_days coefficient is small and close to zero and has a relatively week relationships with Y.
The Moving (ETF) coefficient has from medium to high relationship level with Y.

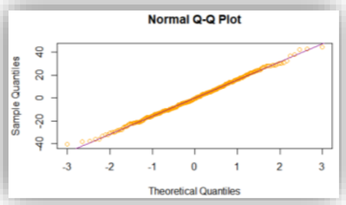
Residual standard error: 15.84 on 361 degrees of freedom. Therefore, the observed Cases(Y values) are close to the predicted or fitted cases than linear model 1 and 2.
The intercept is 179.233456. It means the estimated mean Y value is 179.233456 when all x equal to zero.



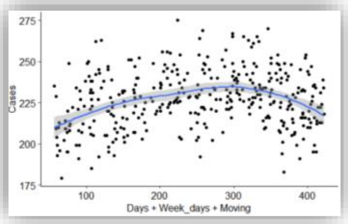


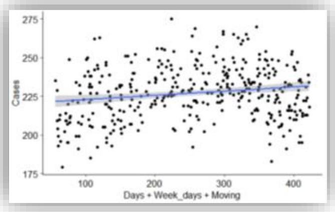
The plot shows that there are week relationships between the residuals and Days, Week_days and Moving (ETF)

The histogram has a bell-curve shape which means that it probably normally distributed.



The Q-Q plot has points very close to the line (even close than in the model 1 and 2). It means that residuals probably normally distributed.

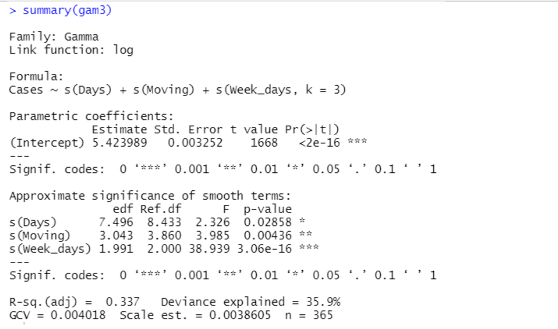




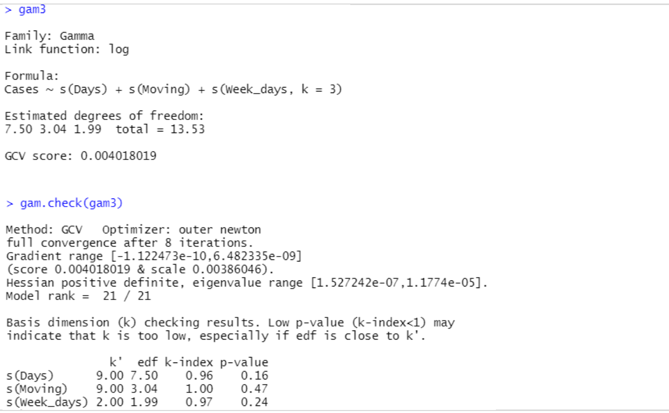
The graph indicates that there is a week linear relationship between Cases and Dates variables, where it at first increased and then decreased.

The plot above shows how the regression line has a very small angle, but it higher than in the model 1.

GAM 3 (x: Date, Weekly seasonality and Moving)



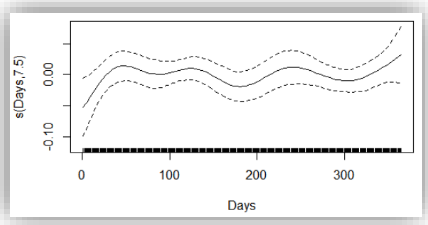
```
> summary(gam3)

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Moving) + s(Week_days, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.423989   0.003252    1668   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df      F  p-value
s(Days)      7.496  8.433  2.326  0.02858 *
s(Moving)    3.043  3.860  3.985  0.00436 **
s(Week_days) 1.991  2.000 38.939 3.06e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.337   Deviance explained = 35.9%
GCV = 0.004018  Scale est. = 0.0038605  n = 365
```

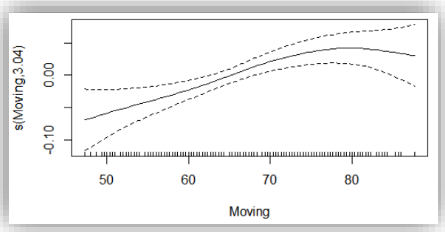


```
> gam3

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Moving) + s(Week_days, k = 3)

Estimated degrees of freedom:
7.50 3.04 1.99  total = 13.53

GCV score: 0.004018019


> gam.check(gam3)

Method: GCV   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-1.122473e-10,6.482335e-09]
(score 0.004018019 & scale 0.00386046).
Hessian positive definite, eigenvalue range [1.527242e-07,1.1774e-05].
Model rank =  21 / 21

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

               k'  edf k-index p-value
s(Days)      9.00 7.50    0.96    0.16
s(Moving)    9.00 3.04    1.00    0.47
s(Week_days) 2.00 1.99    0.97    0.24
```

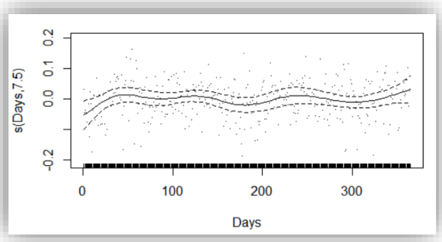
The edf is higher for the “Days” rather than Week_days. Therefore, the term “Moving” has a higher edf and higher significance. The GAM 3 explain 35.9% of the deviance, which is higher, number than in the previous gam model without the EHF.

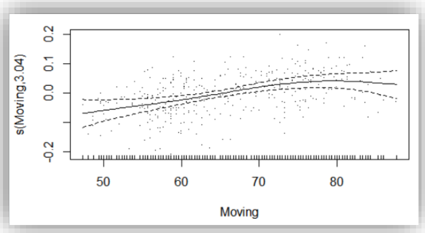
plot(gam3)



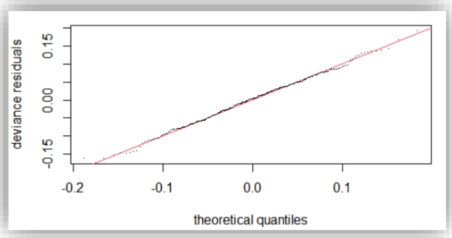


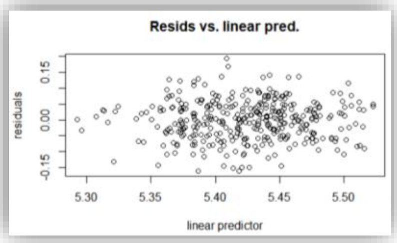
plot(gam3, residuals = TRUE)

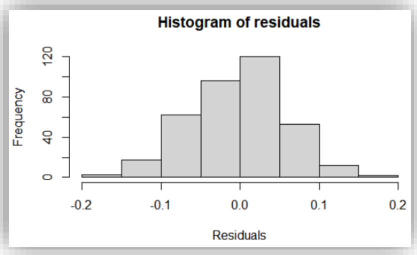


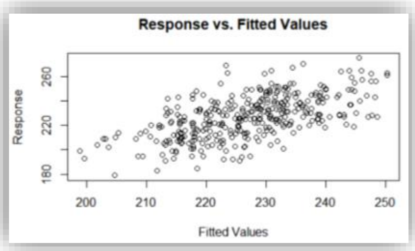


gam.check(gam3)





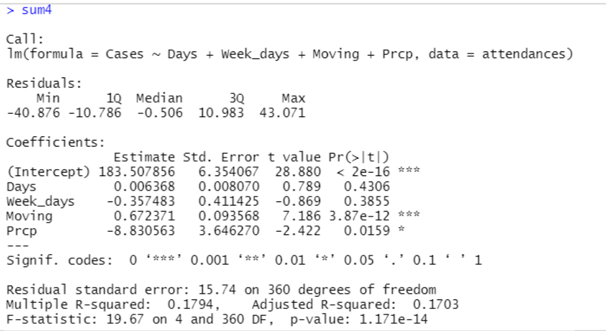




Task 4.3: Extra weather features

The precipitation alongside with extreme temperatures increase the possibility of health problems.

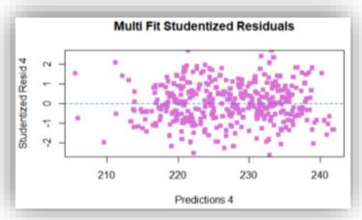
Linear model 4 (x: Date, Weekly seasonality, Moving and Prcp)

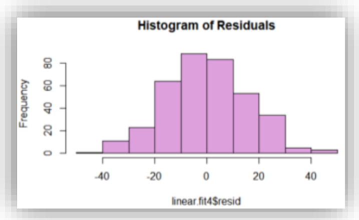


```
> sum4

Call:
lm(formula = Cases ~ Days + Week_days + Moving + Prcp, data = attendances)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.876 -10.786  -0.506  10.983  43.071 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 183.507856   6.354067  28.880  < 2e-16 ***
Days          0.006368   0.008070   0.789   0.4306    
Week_days    -0.357483   0.411425  -0.869   0.3855    
Moving        0.672371   0.093568   7.186 3.87e-12 ***
Prcp         -8.830563   3.646270  -2.422   0.0159 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.74 on 360 degrees of freedom
Multiple R-squared:  0.1794,	Adjusted R-squared:  0.1703 
F-statistic: 19.67 on 4 and 360 DF,  p-value: 1.171e-14
```

Residuals: The median is not far from zero and the sum of absolute values of min and max is close to being equal, which is good. The sum of residuals equals to 2.86 and this is similar to linear model 2.
According to the **Multiple R-squared**, the model can explain approximately 17.94% of variation in Cases. This slightly higher than linear model 3 and significantly higher than linear models 1 and 2.

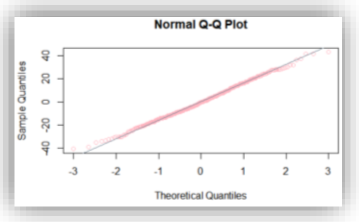
Adjusted R-squared: 0.1703. It is noticeably higher than the previous linear model 4. Therefore, the model improved with the Prcp.
The p-value is 1.171e-14, which is less than 0.05 and even close to zero than linear model 3. The model has statistical significance. We can reject the Null hypothesis.
The Days coefficient is small and close to zero and has a really week relationships with Y.
The Week_days coefficient is small and close to zero and has a relatively week relationships with Y.
The Moving (ETF) coefficient has medium relationship level with Y.
The Prcp has strong relationships with Y.
Residual standard error: 15.74 on 360 degrees of freedom. Therefore, the observed Cases(Y values) are also close to the predicted or fitted cases as the linear model 3. The intercept is 183.507856. It means the estimated mean Y value is 183.507856 when all x equal to zero.



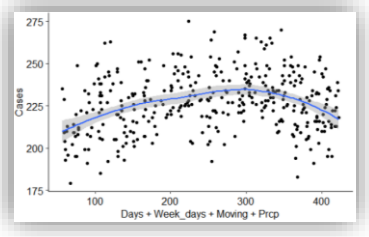


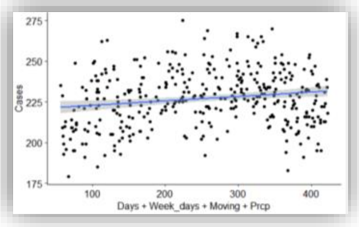
The plot shows that there are week relationships between the residuals and Days, Week_days, Moving (ETF) and Prcp.

The histogram has a bell-curve shape which means that it probably normally distributed as well.



The Q-Q plot has points very close to the line (even close than in the model 1 and 2). It means that residuals probably normally distributed.

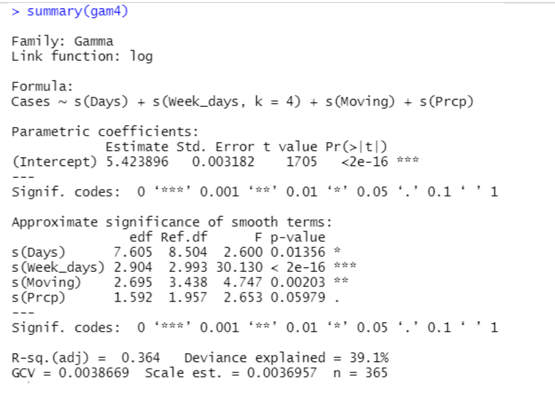




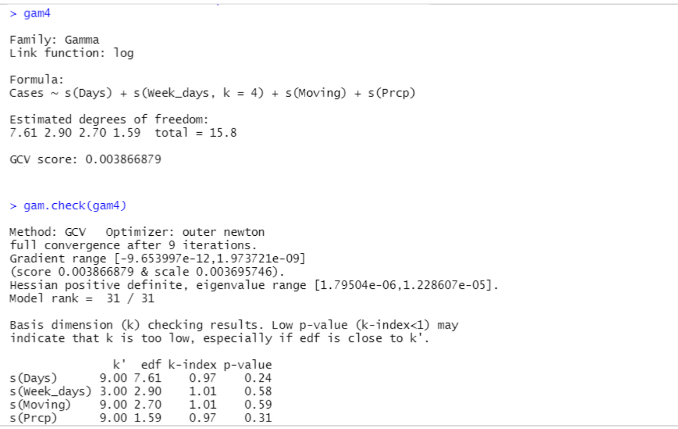
The graph indicates that there is a week linear relationship between Cases and Dates variables, where it at first increased and then decreased.

The plot above shows how the regression line has a very small angle, but it close to the model 3.

GAM 4 (x: Date, Weekly seasonality, Moving and Prcp)



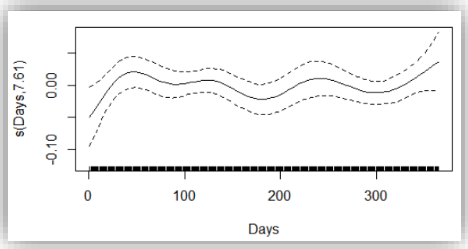
```
> summary(gam4)

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Week_days, k = 4) + s(Moving) + s(Prcp)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.423896   0.003182    1705   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df      F p-value
s(Days)      7.605  8.504  2.600 0.01356 *
s(Week_days) 2.904  2.993 30.130 < 2e-16 ***
s(Moving)    2.695  3.438  4.747 0.00203 **
s(Prcp)      1.592  1.957  2.653 0.05979 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.364   Deviance explained = 39.1%
GCV = 0.0038669  Scale est. = 0.0036957  n = 365
```

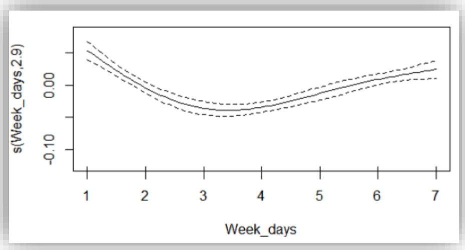


```
> gam4

Family: Gamma
Link function: log

Formula:
Cases ~ s(Days) + s(Week_days, k = 4) + s(Moving) + s(Prcp)

Estimated degrees of freedom:
7.61 2.90 2.70 1.59  total = 15.8

GCV score: 0.003866879


> gam.check(gam4)

Method: GCV   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-9.653997e-12,1.973721e-09]
(score 0.003866879 & scale 0.003695746).
Hessian positive definite, eigenvalue range [1.79504e-06,1.228607e-05].
Model rank =  31 / 31

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                k'  edf k-index p-value
s(Days)       9.00 7.61    0.97    0.24
s(Week_days) 3.00 2.90    1.01    0.58
s(Moving)    9.00 2.70    1.01    0.59
s(Prcp)      9.00 1.59    0.97    0.31
```

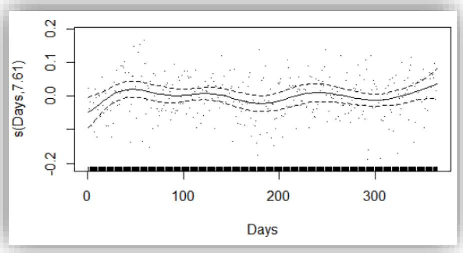
The edf is higher for the “Prcp” is 1.592, which is not so significant. However, the adjusted R-square is the highest and equals to 0.364, which is higher than in the previous GAMs.
Also this model has the higher percent of deviance explained: 39.1%.

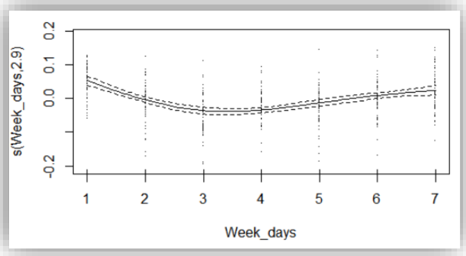
plot(gam4)



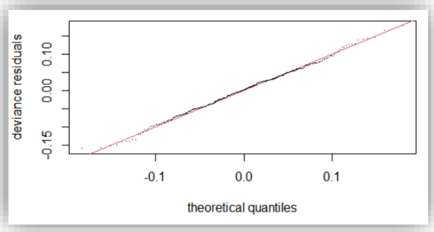


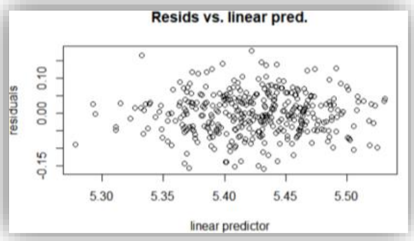
plot(gam4, residuals = TRUE)

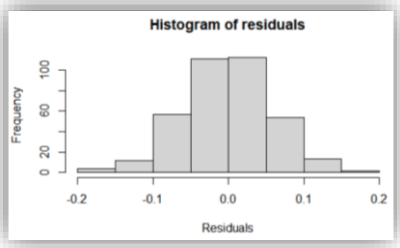


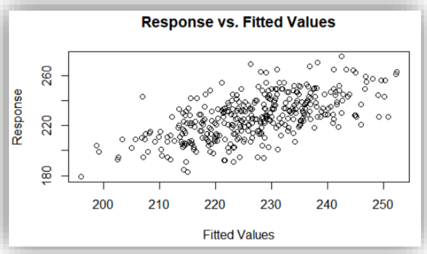


gam.check(gam4)



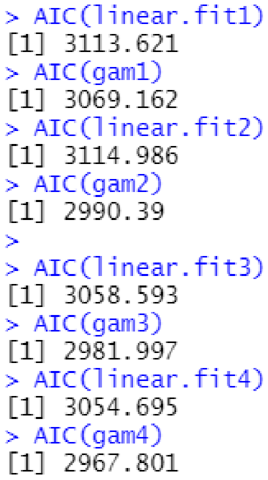






Compare models using the Akaike information criterion (AIC).

Overall, one model is not significantly better than another is since the Akaike information criterion is not more than two time higher.



```
> AIC(linear.fit1)
[1] 3113.621
> AIC(gam1)
[1] 3069.162
> AIC(linear.fit2)
[1] 3114.986
> AIC(gam2)
[1] 2990.39
>
> AIC(linear.fit3)
[1] 3058.593
> AIC(gam3)
[1] 2981.997
> AIC(linear.fit4)
[1] 3054.695
> AIC(gam4)
[1] 2967.801
```

In conclusion, the term “Week_days” had very little effect on the model improvement, while the EHF or “Moving” has a significant effect on the model improvement.

Task 5: Reflection

We used some historical data to fit regression models. What are the limitations of such data, if any?

The trend may change overtime and the older data can give inaccurate prediction results than newer data.

Regression models can be used for 1) understanding a process, or 2) making predictions. In this assignment, do we have reasons to choose one objective over the other? How would the decision affect our models?

Since we are not interested in the trend itself in this question, then we can say that in this assignment we use the regression models to understand the process and evaluation the models itself.

The prediction should not be make when the model statistically insignificant or week. In this case, if we choose to make the prediction then I can be invalid.

Overall, have your analyses answered the questions that you set out to answer?

In this assignment, I set out to find out the relationships between the days of the year and the number of cases by linear regression and by generalized additive model (GAM). By investigation of two models with the different set of predictors X, it concluded that the relationship is really week.

In addition, I set out to investigate the connection between heatwaves and the ED demands. As a result of this analysis, I find out that heatwaves has a noticeable effect on the ED demands.