

**SIT 741**  
**Statistical Data Analysis**  
**Assignment-2**

**Shantanu Gupta**  
**218200234**  
**Deakin University**

## Q1.

1.

I have used all the websites that are mentioned in the application scenario.

- Global Historical Climatology Network (GHCN).
- National Centers for Environmental Information (NOAA)
- Data.gov.au/dataset/emergency-department-admissions-and-attendances

We need to access all three websites for successfully completing this assignment.

3.

- 1000 rows in the dataset with 8 variables
- The time period covers the 1<sup>st</sup> of July, 2013 to 7<sup>th</sup> March, 2014. Approximately 9 months of data we have.

## Q2.

1.

A final model is a model that you can use to make predictions on new data. The goal of the final model that performs the best, where “best” is defined by:

- **Data:** the available historical data.
- **Procedure:** the data preparation steps, and the chosen algorithm configurations.

Yes, we were able to solve the overcrowding problem in Western Australia hospital. We have an impact on major hospital metrics like improving customer satisfaction, reducing the number of patients leaving before being seen to doctors, and gives the hospital a real strategic market advantage over others. The challenge is how other hospitals adapt this innovation.

Potential users are:

- **Hospital & Health Authorities:** To better balance demand and load
- **General Public:** How much time they need to spend in waiting
- **State Government:** To improve efficiency and bring innovations in the health area. It also helps the government to reduce down its expenditure over health.

2.

It is better if we get a linear relationship between Response and predictor variables. Because we have so many statistical tools for the linear model, need fewer data to train and it also less computationally expensive. I wanted to predict attendances from date.

**Response Variable:** Date

**Predictor Variable:** Attendance- the number of patients recorded as arriving at a public emergency department.

Yes, all the variables in my model are collected on a daily basis. Yes, it readily available for prediction from the hospital website or from the government website.

3.

My model is based on insight on the past data. Also, my model has the ability to “Predict” what might happen in future.

4.

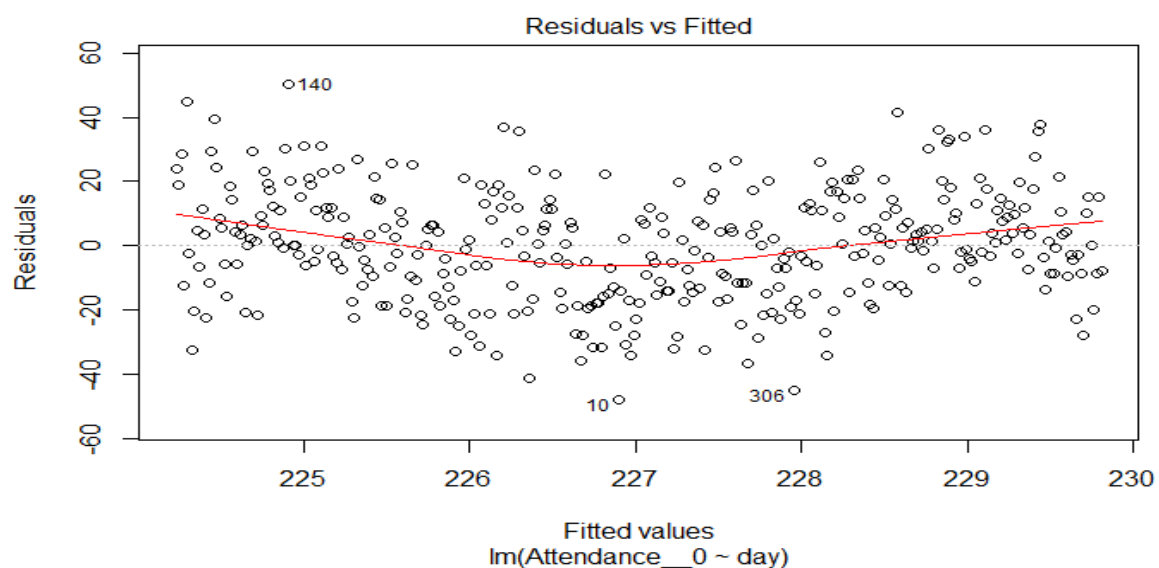
- Linear Model (LM)
- Generalized Linear Model (GLM) especially Poisson and Negative Binomial Distribution
- General Additive Model (GAM)

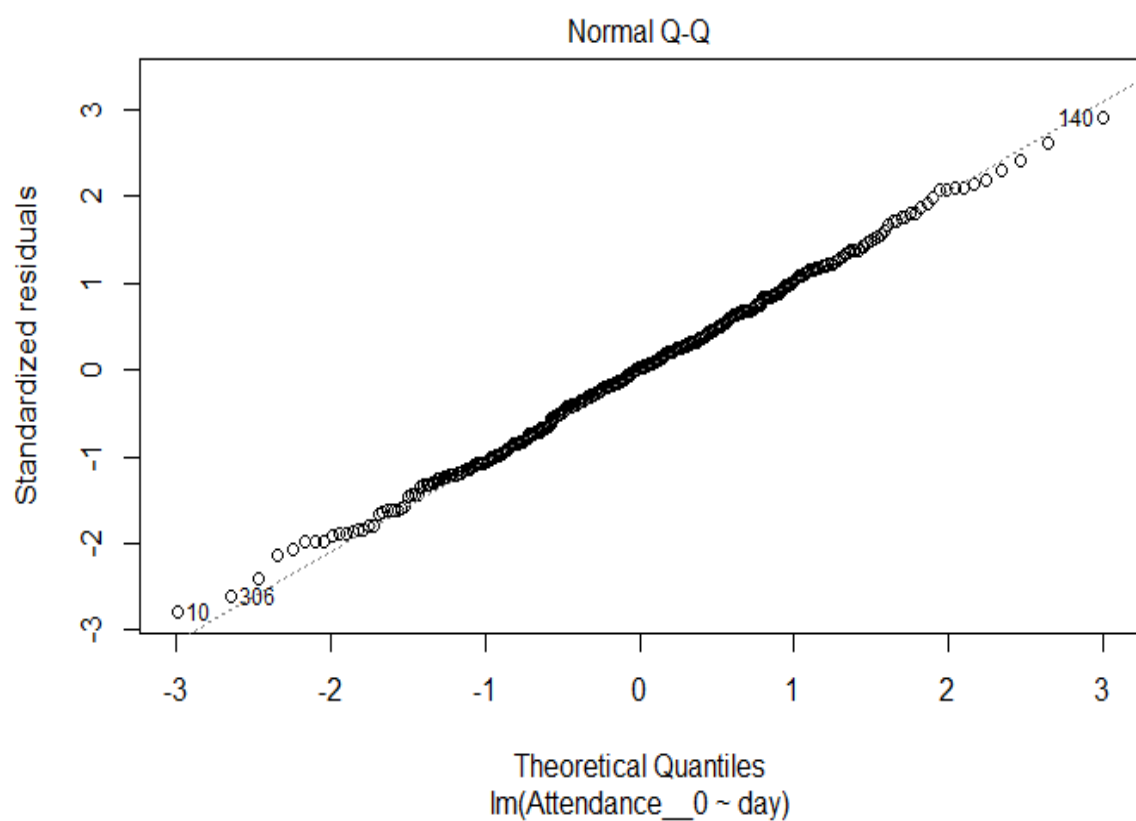
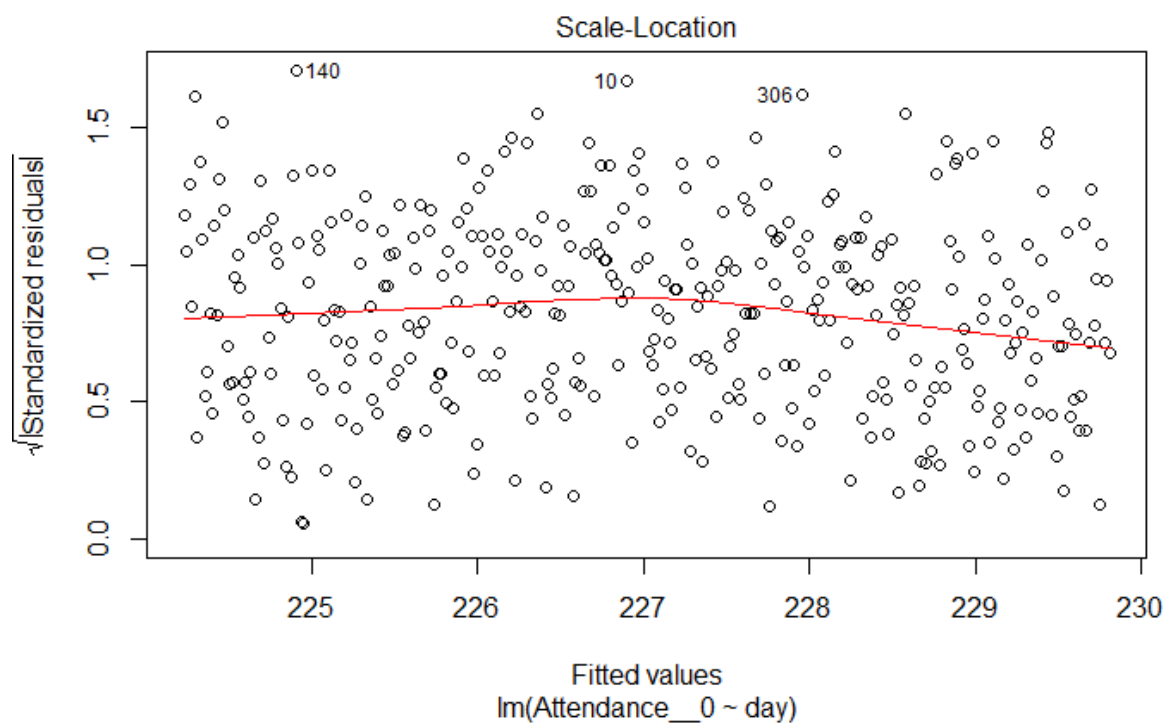
### Q3.1

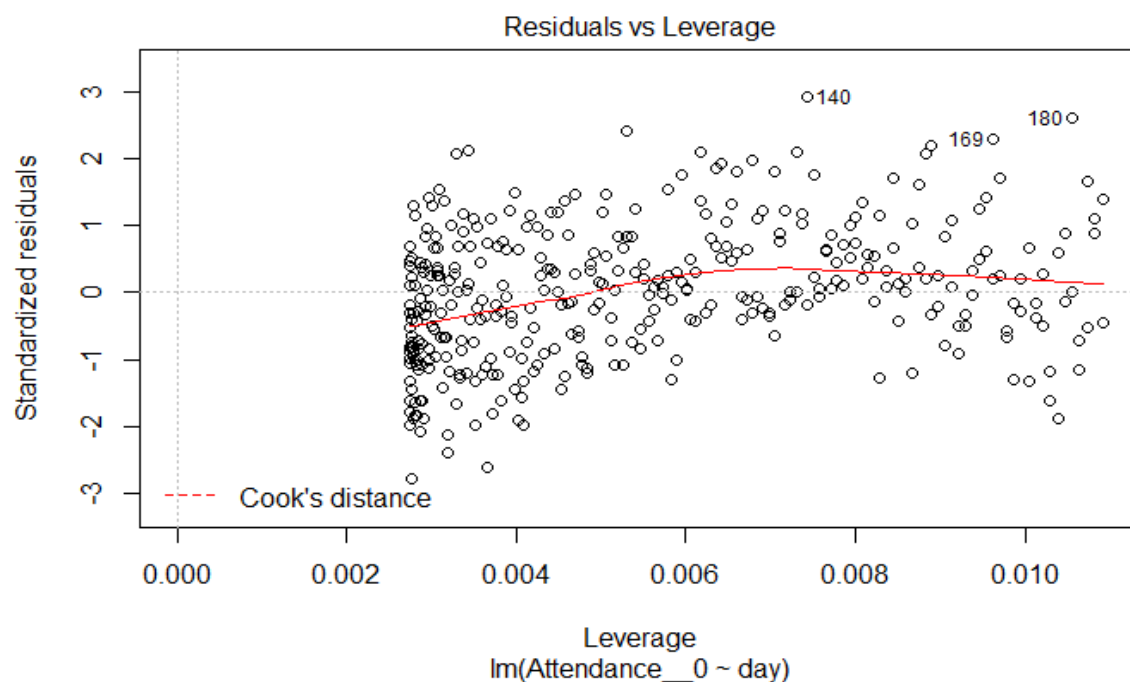
1.

Royal Perth Hospital

2.







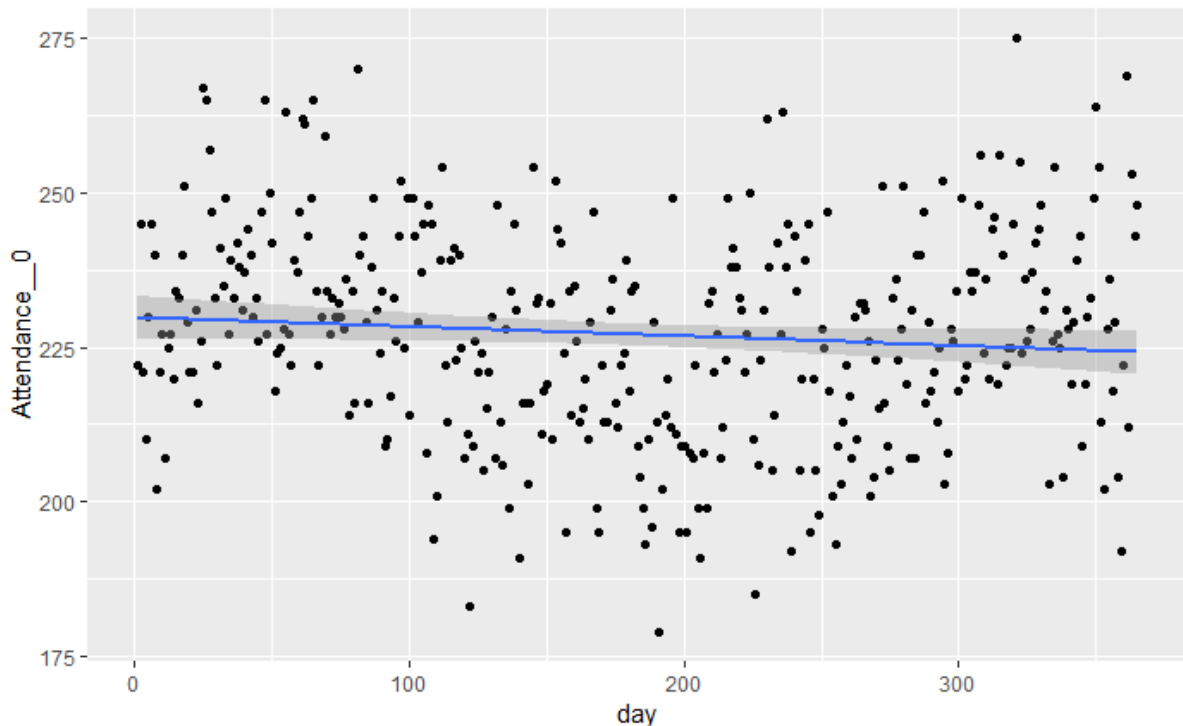
In our data, we may have some influential points that might skew our model. The most common are the outliers. We can see 140, 169 and 180 observations have high cook distance. We can see there are 3 points that are considered as an outlier. If we have 1 or 2 points, then it is better if we can remove that point or correct it. If it's 3, it's better if we don't delete it, as it may have a valid point. We can try a non-linear model like GAM. In our case, we should not delete this point as it may consider important for our application. But as we see cook distance there are more points that are very close to the range of the values of 140, 169, and 180.

180	140	169	209	210	178	231	265	182	170	249	192	239	306
0.03633	0.03193	0.02584	0.02154	0.01916	0.01867	0.01610	0.01551	0.01527	0.01457	0.01379	0.01359	0.01341	0.01255
154	245	141	134	211	246	127	10	184	168	253	45	195	172
0.01243	0.01200	0.01164	0.01156	0.01142	0.01119	0.01075	0.01071	0.01062	0.00985	0.00927	0.00927	0.00886	0.00853
293	55	202	149	74	177	324	188	49	113	58	152	183	157
0.00811	0.00789	0.00748	0.00746	0.00740	0.00733	0.00729	0.00716	0.00710	0.00701	0.00696	0.00671	0.00656	0.00634
25	65	120	126	99	217	233	5	163	132	341	91	139	68
0.00619	0.00609	0.00596	0.00584	0.00559	0.00551	0.00537	0.00536	0.00534	0.00519	0.00518	0.00513	0.00508	0.00508
294	148	114	353	20	17	87	212	320	296	7	281	248	186
0.00507	0.00506	0.00495	0.00491	0.00482	0.00479	0.00470	0.00461	0.00460	0.00454	0.00446	0.00435	0.00432	0.00431
106	73	190	131	230	147	164	352	27	24	329	88	76	4
0.00429	0.00421	0.00420	0.00420	0.00403	0.00402	0.00382	0.00378	0.00375	0.00371	0.00366	0.00366	0.00364	0.00364
244	271	94	283	162	327	285	311	275	103	337	101	115	43
0.00362	0.00359	0.00349	0.00326	0.00324	0.00323	0.00321	0.00312	0.00310	0.00306	0.00302	0.00301	0.00300	0.00298

There are some of the points in all the above-generated plots that there are some highly influential points. Observation 10, 140, 169, 180, 306 all are unusual points.

I don't think we need to use linear regression models for our data. As the regression doesn't fit the data well. We can't apply machine learning algorithms like a neural net, or decision tree etc. because they are

computationally expensive and needs a lot of data. It is better to use General Additive Model (GAM) as this provides a middle ground without losing interpretability and flexibility. I also think there is some kind of non-linearity is present in our data. I suppose the relationship will be complex and also tracks the non-linear relationship in our data.



I have also used Poisson and negative binomial Generalized Linear model as our data represents counts data (i.e. Number of patients arrive in a single day in one hospital). I tried this because there are impossible predicted values for our system or there may be heterogeneity in the residuals as we see in the residual analysis of linear plot. When we compare the AIC I find that Poisson regression performs worst in our case while the linear model and negative binomial gives similar AIC value. I think it is better to use negative binomial distribution as this distribution is more flexible as we have an extra parameter  $\theta$  which relaxes the assumptions of equality between the mean and the variance. Even negative binomial is a more general distribution family.

```
Call:
glm(formula = Attendance__0 ~ day, family = "poisson", data = daily_attendance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3024  -0.8305   0.0221   0.7627   3.2262

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.437e+00  6.941e-03 783.382  <2e-16 ***
day         -6.743e-05  3.297e-05  -2.045  0.0408 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 479.63  on 364  degrees of freedom
Residual deviance: 475.45  on 363  degrees of freedom
AIC: 3129.6

Number of Fisher Scoring iterations: 3
```

```
tidy(poissonOut)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	5.437349e+00	6.940869e-03	783.381594	0.00000000
day	-6.743052e-05	3.297064e-05	-2.045169	0.04083821

2 rows

```
coef(poissonOut)
```

```
(Intercept)      day
5.437349e+00 -6.743052e-05
```

```
confint(poissonOut)
```

```
waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) 5.4237258903 5.450934e+00
day         -0.0001320527 -2.809601e-06
```

```

{r}
neg_binom_glm<-glm.nb(Attendance_0~day ,
  data = daily_attendance)
neg_binom_glm

```

```

Call: glm.nb(formula = Attendance_0 ~ day, data = daily_attendance,
  init.theta = 754.9895458, link = log)

```

```

Coefficients:
(Intercept)      day
  5.437e+00   -6.713e-05

```

```

Degrees of Freedom: 364 Total (i.e. Null); 363 Residual
Null Deviance:      369
Residual Deviance: 365.8      AIC: 3118

```

```

{r}
summary(neg_binom_glm)

```

```

Call:
glm.nb(formula = Attendance_0 ~ day, data = daily_attendance,
  init.theta = 754.9895458, link = log)

```

```

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.92085  -0.73076   0.01881   0.66731   2.80847

```

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.437e+00  7.922e-03  686.384  <2e-16 ***
day          -6.713e-05  3.760e-05  -1.785   0.0742 .
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

(Dispersion parameter for Negative Binomial(754.9895) family taken to be 1)

```

```

    Null deviance: 368.99  on 364  degrees of freedom
Residual deviance: 365.79  on 363  degrees of freedom
AIC: 3117.9

```

```

Number of Fisher Scoring iterations: 1

```

```

      Theta: 755
    Std. Err.: 242

```

```

2 x log-likelihood: -3111.876

```

### 3.

The first part of the summary describes the model we fit. As we can see the algorithm chooses Gaussian or normal distribution of errors. The link function is an identity that shows the model doesn't transform prediction. Both model and smooth terms are significant to our model. 15.4% of the variation are explained by this model. Remaining 84.6% variation is explained by other parameters. As we see the distribution follows to a



normal distribution. It well fitted to data in the middle but it is not following the extreme values.

```
3. General Additive Model
##{r}
ct1 <- gam(Attendance_0 ~ s(day), data=daily_attendance, method="REML")
ct1
```

```
Family: gaussian
Link function: identity

Formula:
Attendance_0 ~ s(day)

Estimated degrees of freedom:
5.56 total = 6.56

REML score: 1533.887
```

```
##{r}
summary(ct1)
```

```
Family: gaussian
Link function: identity

Formula:
Attendance_0 ~ s(day)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 227.0192    0.8381   270.9   <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(day)  5.558  6.705  8.898  1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.141  Deviance explained = 15.4%
-REML = 1533.9  Scale est. = 256.37    n = 365
```

```
##{r}
coef(ct1) #Coefficients
```

```
(Intercept)      s(day).1      s(day).2      s(day).3      s(day).4      s(day).5      s(day).6      s(day).7      s(day).8
227.019178      3.822030      49.902857      -1.538648      23.965933      6.319981      20.648998      -6.908263      52.342633
s(day).9
8.707612
```

```
##{r}
augment(ct1)
```

Attendance_0	day	.fitted	.se.fit	.resid	.hat	.sigma	.cooks
205	232	220.5984	1.971688	-15.5983761	0.01516407	NA	2.262500e-03
214	233	220.6937	1.973450	-6.6937387	0.01519118	NA	4.174140e-04
242	234	220.7887	1.975249	21.2112506	0.01521890	NA	4.199312e-03
227	235	220.8834	1.977044	6.1166163	0.01524656	NA	3.498494e-04
263	236	220.9776	1.978791	42.0223741	0.01527352	NA	1.654287e-02
238	237	221.0715	1.980450	16.9285306	0.01529914	NA	2.689296e-03
245	238	221.1649	1.981980	23.8350833	0.01532279	NA	5.339804e-03
192	239	221.2580	1.983346	-29.2579792	0.01534392	NA	8.057452e-03
243	240	221.3507	1.984515	21.6493232	0.01536202	NA	4.416980e-03
234	241	221.4430	1.985459	12.5569615	0.01537663	NA	1.487413e-03

51-60 of 365 rows

```
##{r}
tidy(ct1)
```

term	edf	ref.df	statistic	p.value
s(day)	5.558395	6.704721	8.898027	1.00366e-09

1 row

```
##{r}
glance(ct1)
```

df	logLik	AIC	BIC	deviance	df.residual
6.558395	-1526.859	3070.559	3103.396	91892.28	358.4416

1 row

To examine the quality of the model. Model is full converged around 6 iterations. An effective number of basis function is the 9. Small value of p indicates that residuals are not randomly distributed. This means there is not enough basis function. The problem is that they have some significant function but not enough basis functions. We can change by using larger

values of the basis number. We need to add one parameter in the smoothing parameter (i.e.  $s(\cdot, k=)$ ) as this will make a much better model.

```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-2.014897e-07,3.584328e-09]
(score 1545.33 & scale 0.005018186).
Hessian positive definite, eigenvalue range [1.071215,181.8331].
Model rank = 10 / 10

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'.

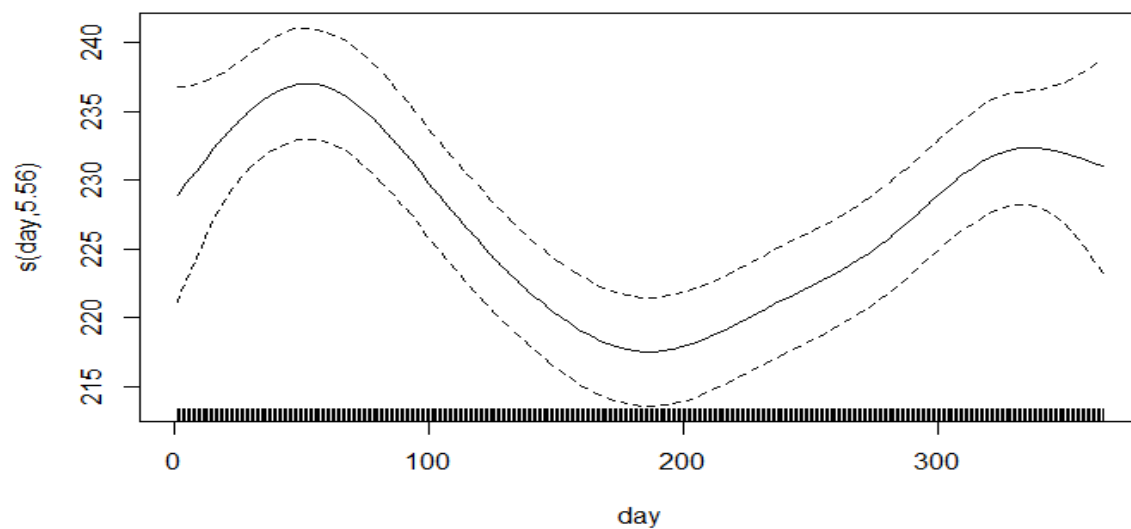
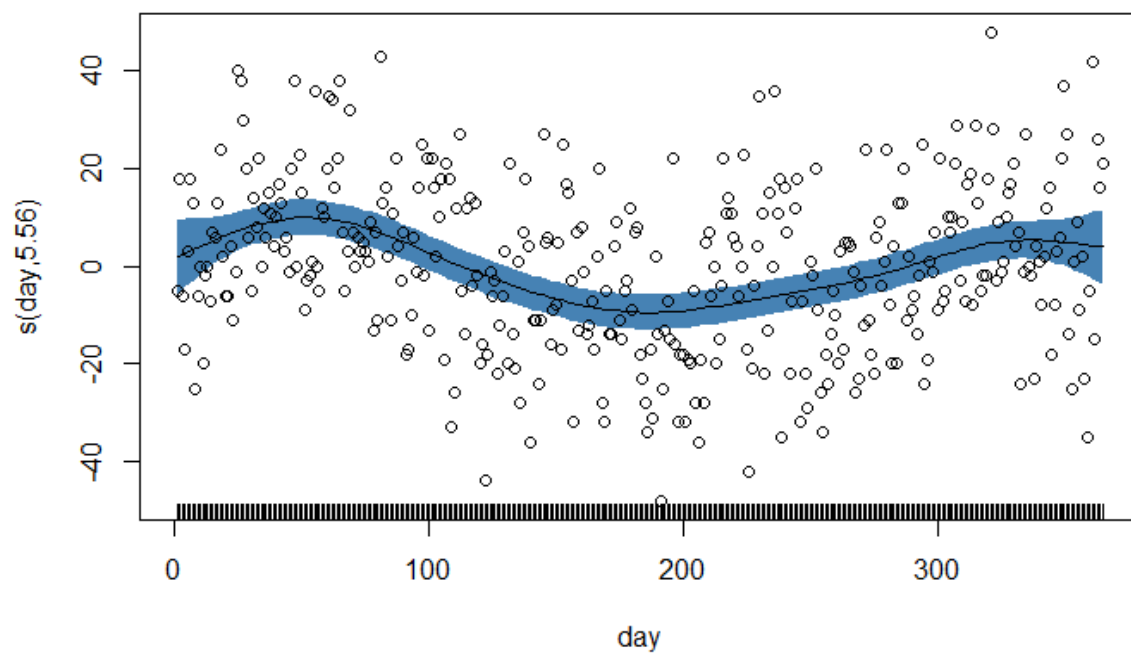
      k'   edf k-index p-value
s(day) 9.00 5.49    0.8 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In regression, we need to remove multi-collinearity in our model. Same applies to GAM data. In this, we need to remove multi-collinearity and concurvity. Concurvity means one may be a smooth curve of another. Here we didn't need to worry as we are dealing with only one variable. In order to confirm my result, as all the values are less than 0.8, so, we can conclude that there is no concurvity in our model.

```
## {r}
concurvity(ct1,full=TRUE) #Checking Concurvity
```

	para	s(day)
worst	3.040639e-25	3.061451e-25
observed	3.040639e-25	9.414018e-28
estimate	3.040639e-25	6.425409e-28

The below plot is called a partial effect plot. It shows us the prediction of the number of patients arrive in a single given day, assuming other variables are at their average value. This plot shows that the average number of people arrive in Royal Perth hospital on any day is about more than 220, considering all else being equal.



4.

```
4. Akaike Information Criteria(AIC) and Bayesian Information Criteria(BIC)
## {r}
AIC(ct1)
AIC(lmfit_royal)
BIC(ct1)
BIC(lmfit_royal)
```

```
[1] 3071.924
[1] 3117.647
[1] 3104.457
[1] 3129.346
```

The lower value of AIC suggests "better" is the model, but it is a relative measure of model fit. This is an error, so less is better. It is used

for model selection, i.e. it lets you compared different models estimated on the same dataset. By seeing that value I would say Generalized Additive Models (GAM) is a better model for this data as compared to linear models.

```
```{r}
coef(poissonOut)
```
```

---

|              |               |
|--------------|---------------|
| (Intercept)  | day           |
| 5.437349e+00 | -6.743052e-05 |

---

```
Call:
lm(formula = Attendance__0 ~ day, data = daily_attendance)
```

```
Coefficients:
(Intercept)          day
  229.82052      -0.01531
```

#####Negative Binomial and General Linear Model

```
```{r}
neg_binom_glm<-glm.nb(Attendance__0~day ,
                      data = daily_attendance)
neg_binom_glm
```
```

```
Call: glm.nb(formula = Attendance__0 ~ day, data = daily_attendance,
             init.theta = 754.9895458, link = log)
```

```
Coefficients:
(Intercept)          day
  5.437e+00      -6.713e-05
```

```
Degrees of Freedom: 364 Total (i.e. Null);  363 Residual
Null Deviance:      369
Residual Deviance: 365.8      AIC: 3118
```

```
```{r}
coef(ct1) #Coefficients
```
```

---

|             |          |           |           |
|-------------|----------|-----------|-----------|
| (Intercept) | s(day).1 | s(day).2  | s(day).3  |
| 227.019178  | 3.822030 | 49.902857 | -1.538648 |
| s(day).4    | s(day).5 | s(day).6  | s(day).7  |
| 23.965933   | 6.319981 | 20.648998 | -6.908263 |
| s(day).8    | s(day).9 |           |           |
| 52.342633   | 8.707612 |           |           |

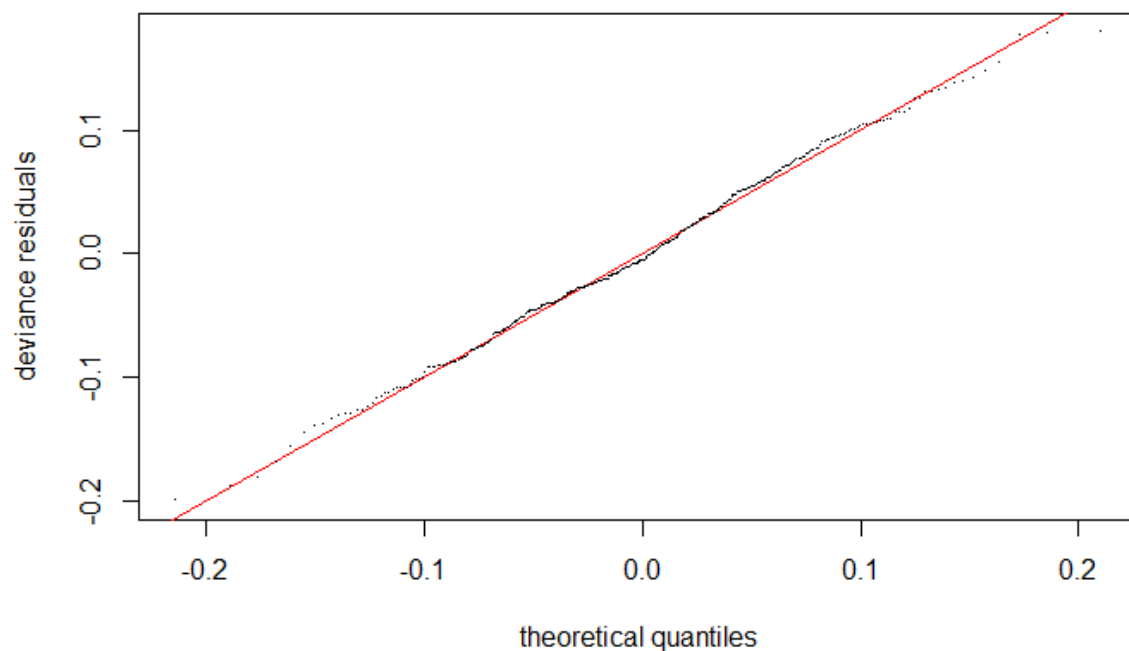
---

I think the best coefficients are from negative binomial and GAM. In GAM we have 9 basis functions that are why 9 coefficients. Best coefficients are from GAM as it is more accurate and capture more subtle data points.

5.

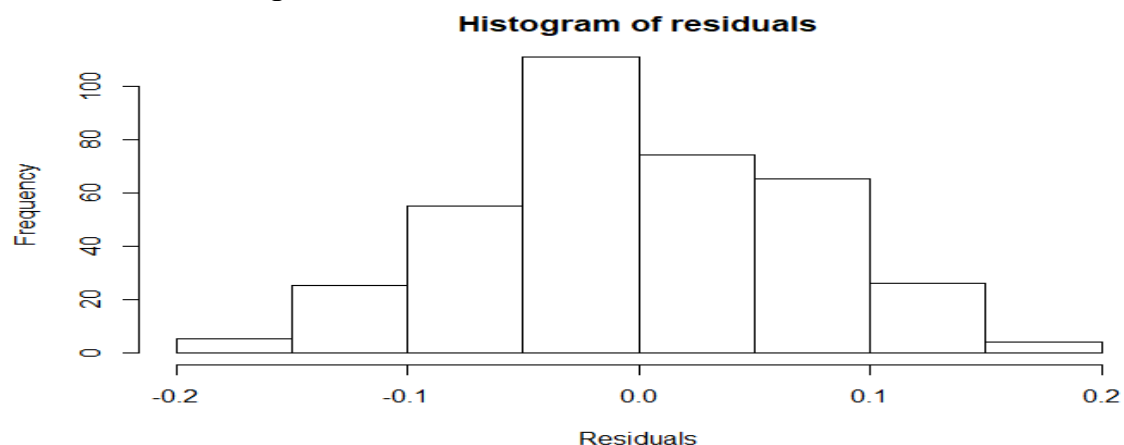
### Q-Q Plot

This graph compares the model residuals to a normal distribution. A well fit models residuals will be close to a straight line. Most of the values are fitting towards the straight line except extreme ends where some points deviate from a straight line.



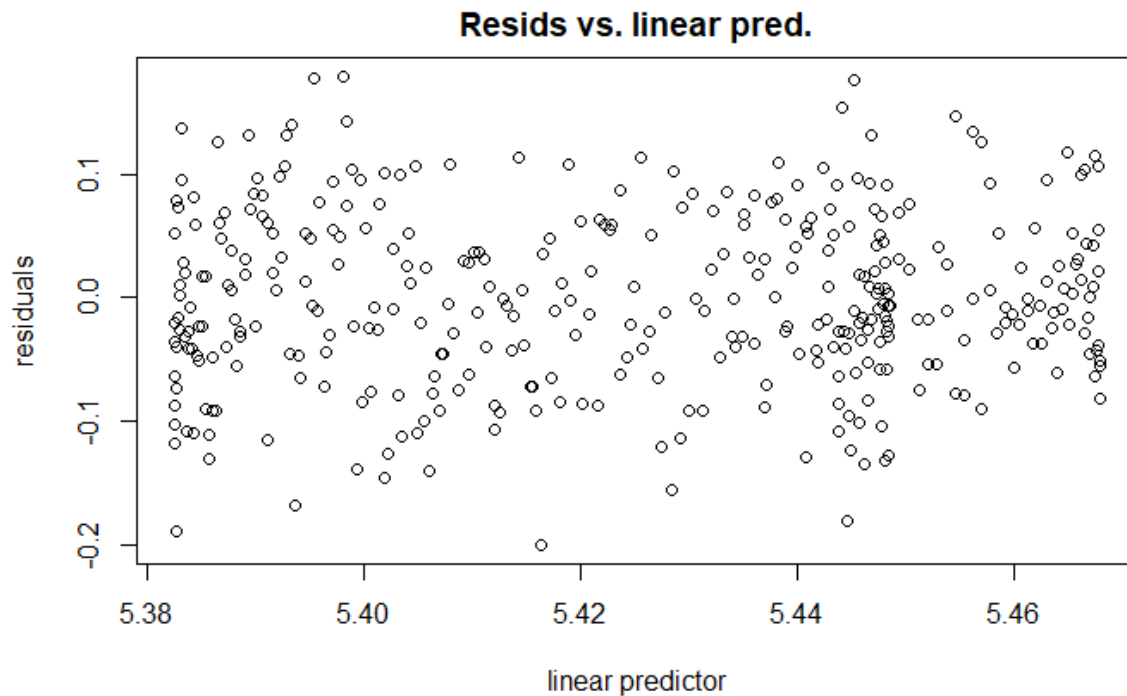
### Histogram of Residuals

The ideal curve should be a symmetrical bell shape. As we can see the curve is following the normal curve.



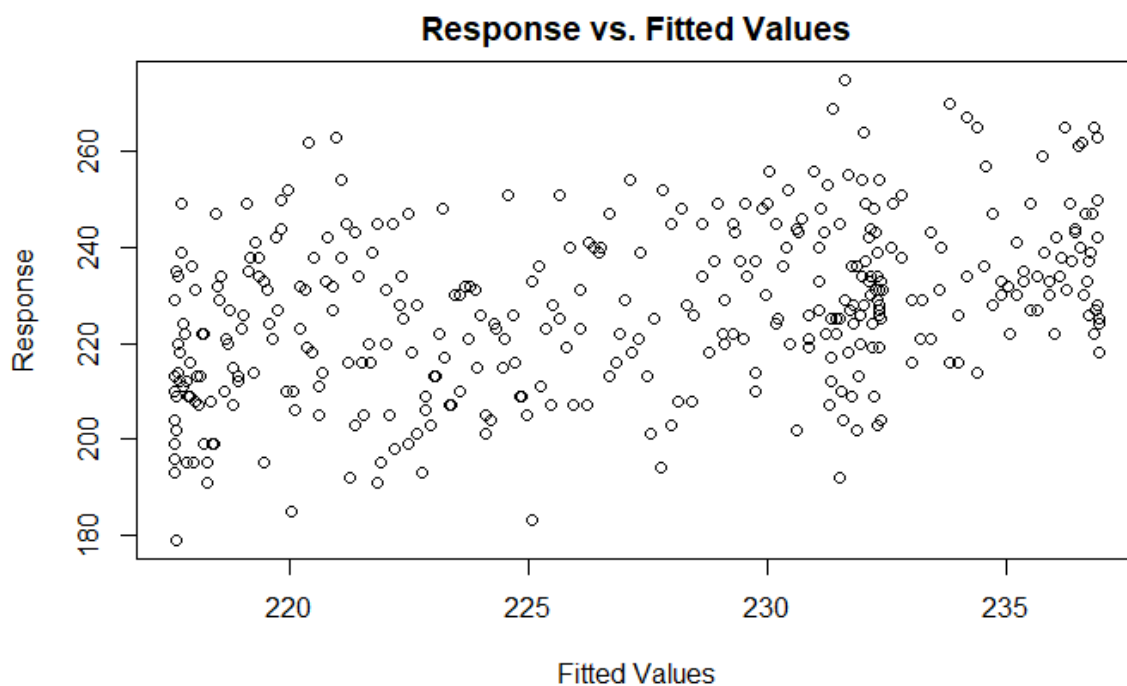
## Residual values

These should be evenly distributed around zero. We can see there is an evenly distribution.



## Response vs. fitted values

A perfect model forms a straight line. We don't expect a perfect model, but we do expect a pattern that clusters around the 1-to-1 line. As we can see it is not a perfectly straight line. Some of the values are away from zero.



I have analysed GAM residuals only. I have already explained the linear model residuals in Task 3.1 part 2.

6.

My day-of-the-week variable is numeric. Initially, it is a character. It's problematic as we can't do computation over it. Second, we can't do regression on character data. However, I breakdown dates into other variables, like years, months and days. When you do this, then you can treat them individually as scales. So, I changed the date in character to numerical. It can't be ordinal or categorical because we can't say "The year "2000 is four times as much as the year 500". It doesn't make sense in time-series.

### (Optional)Q3.2

2.

#### General Points:

- GAM best fit on King Edward Memorial Hospital for women followed by Princess Margaret Hospital for children
- In King Edward hospital for women, approximately 37% of the variation in attendance can be explained by the regression model (i.e., variables included in the regression model). Approximately 63% of the variation in productivity would be explained by other factors not included in the model. The value of deviance residuals is not high, thus this is not a strong predictive model. 48% of Standard error we can commit when using this regression model. The average error we commit by using this model. Similar interpretation for other hospitals.

- **Residual Analysis-** General Residual analysis for all hospitals

**QQ-Plot-Normal**, extreme values deviate

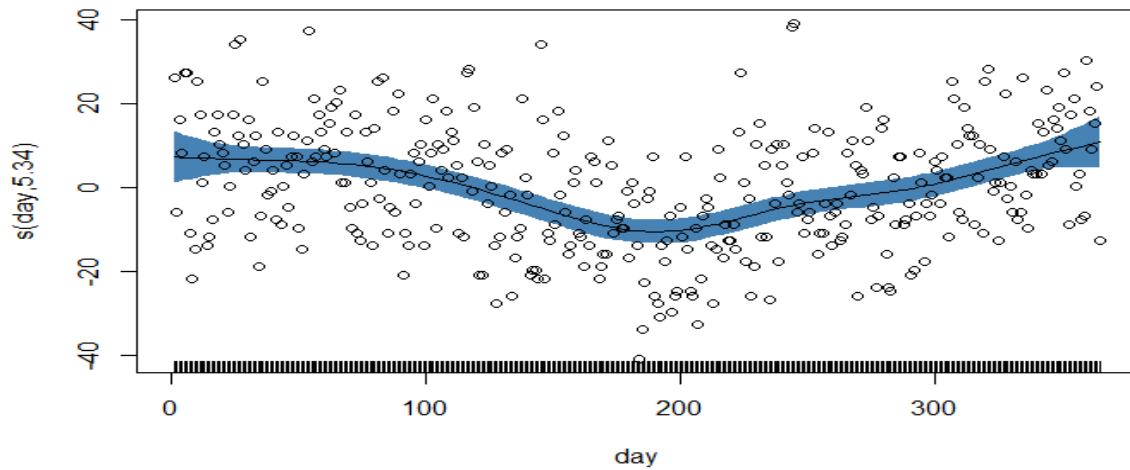
**Histogram of Residuals-Normal**

**Residual values-** Evenly distributed around zero

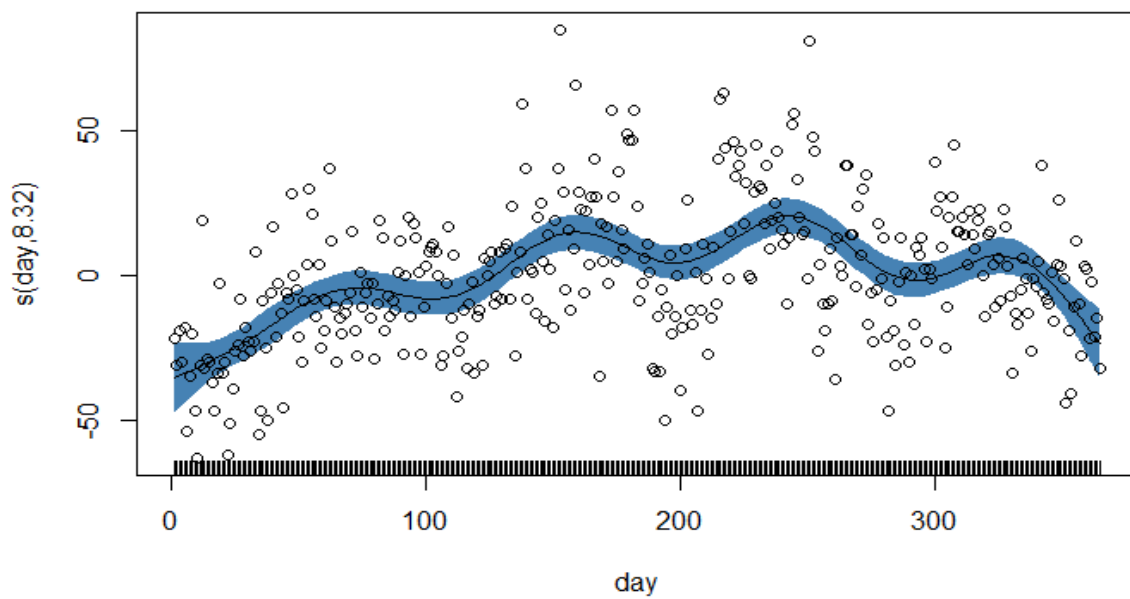
**Response vs. fitted values-** Not a perfectly straight line. But some hospital shows a 1 to 1 relationship.

- When I compare AIC values of different hospitals the best hospital in terms of AIC is Princess Margaret hospital for children, closely followed by swan district hospital and Rockingham general hospital. AIC always less that means the best model for fit.
- Fully converged all models within 6-10 iterations.

- All other hospitals have less predictive powers as deviance explained is not very high.
- All hospitals are statistically significant.

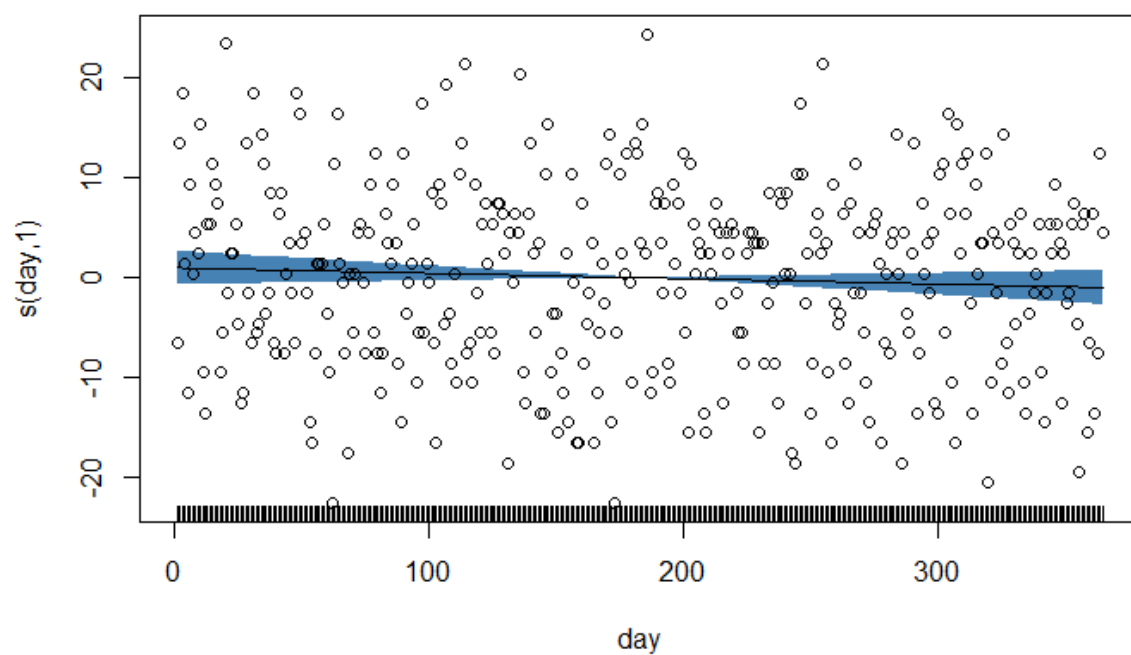


**Fig: Fremantle Hospital**

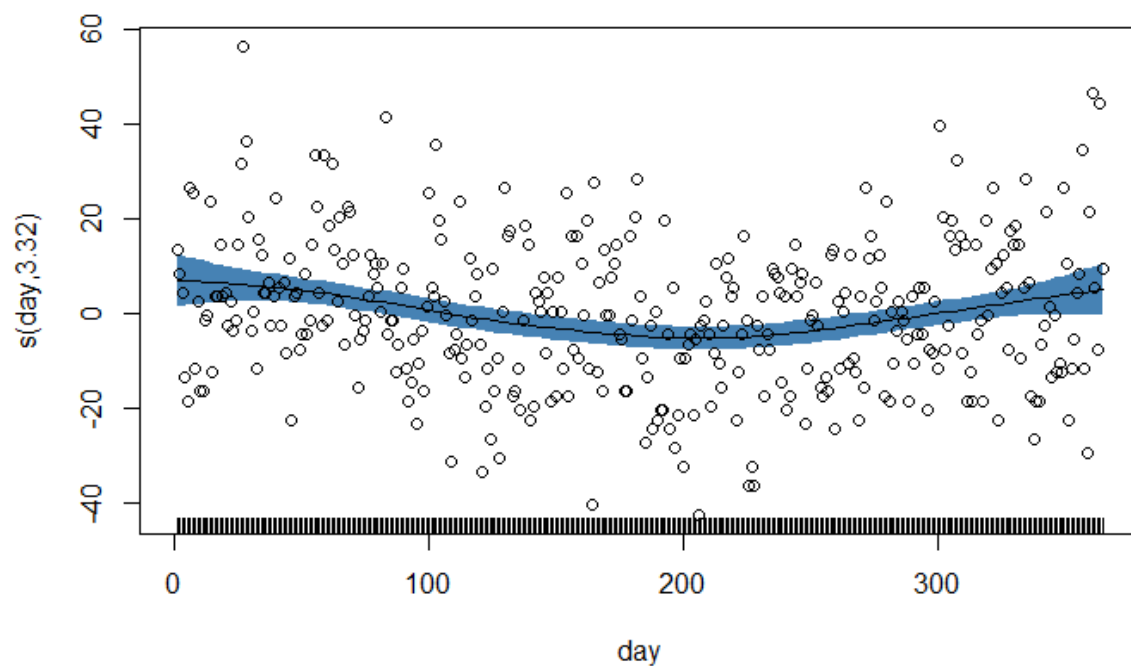


**Fig: Princess Margaret Hospital for Children**

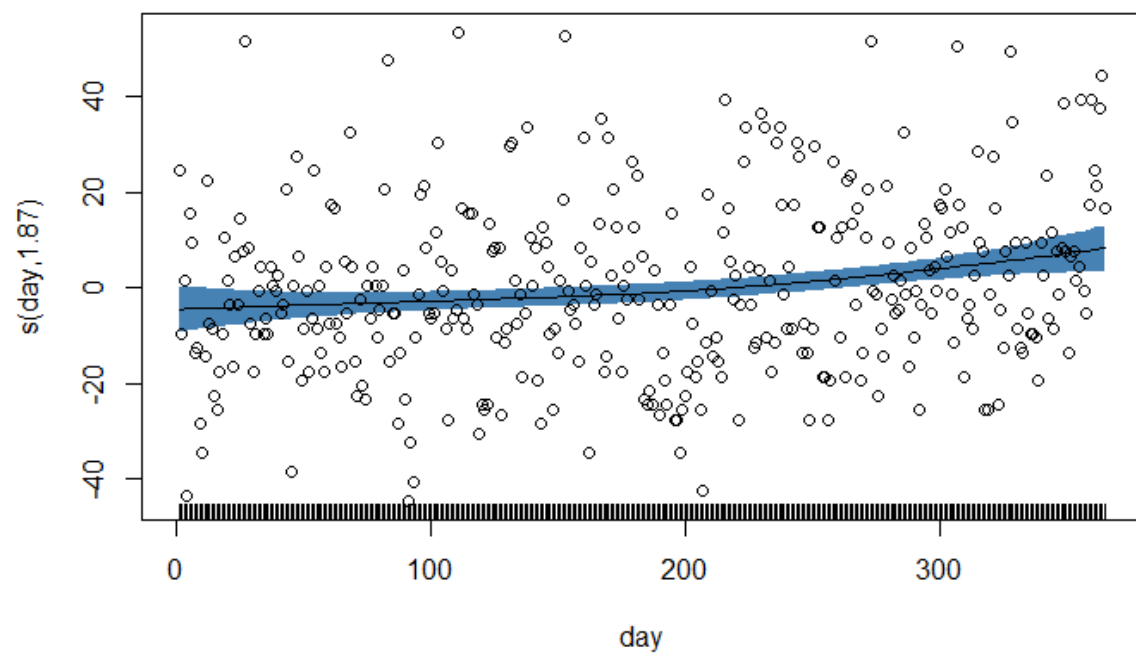




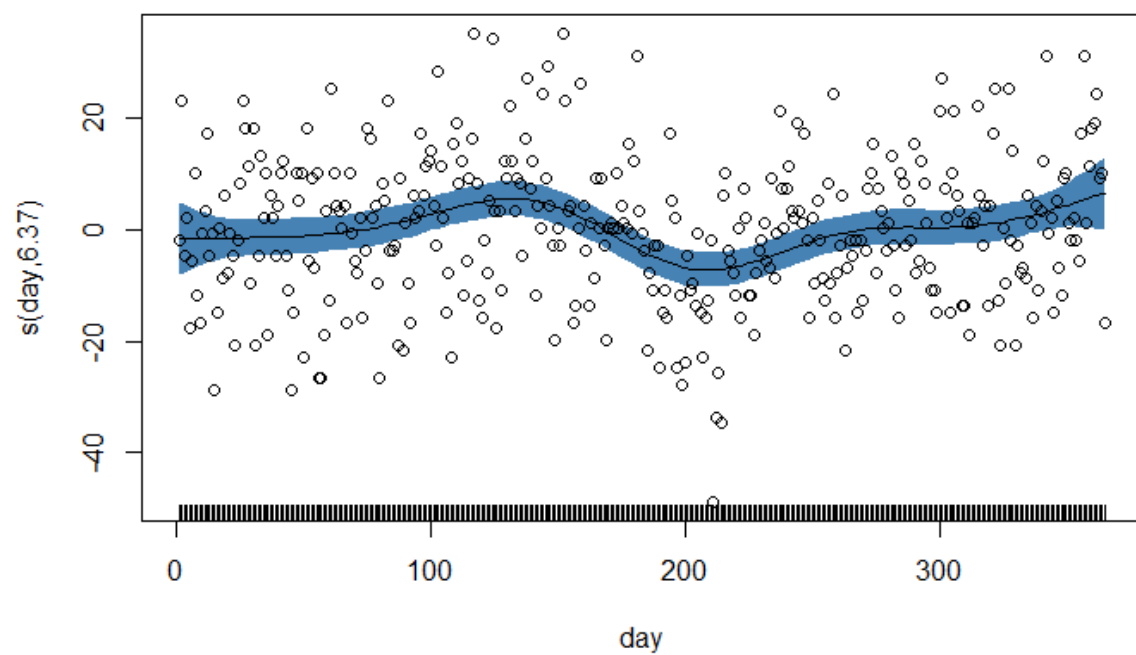
**Fig: King Edward Memorial Hospital for women**



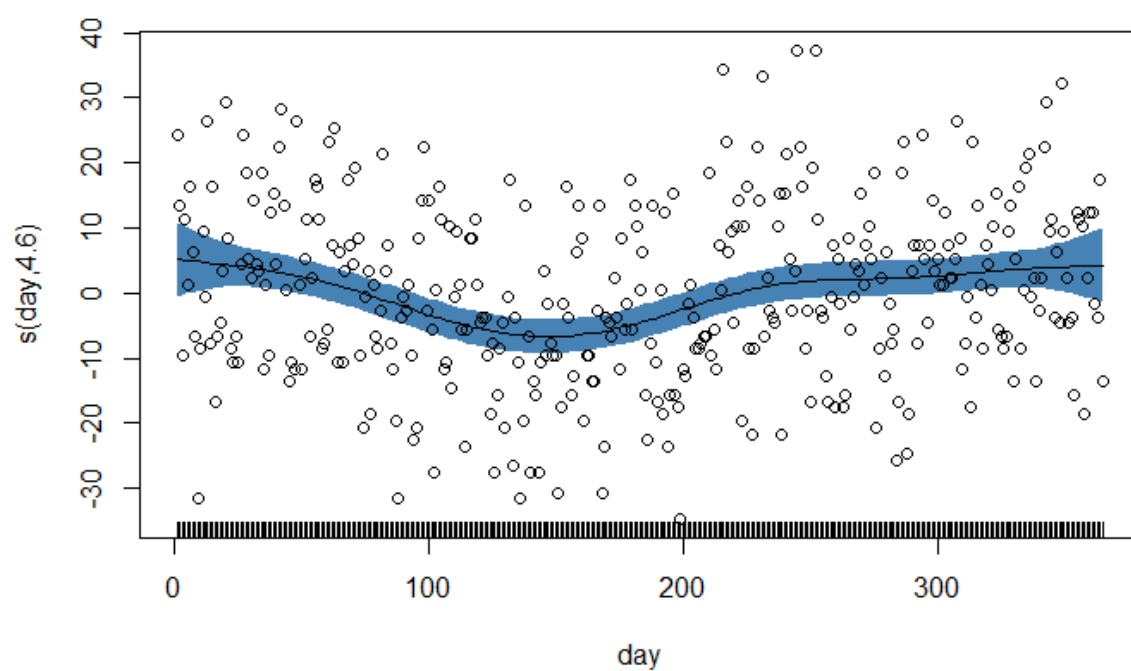
**Fig: Sir Charles Gairdner Hospital**



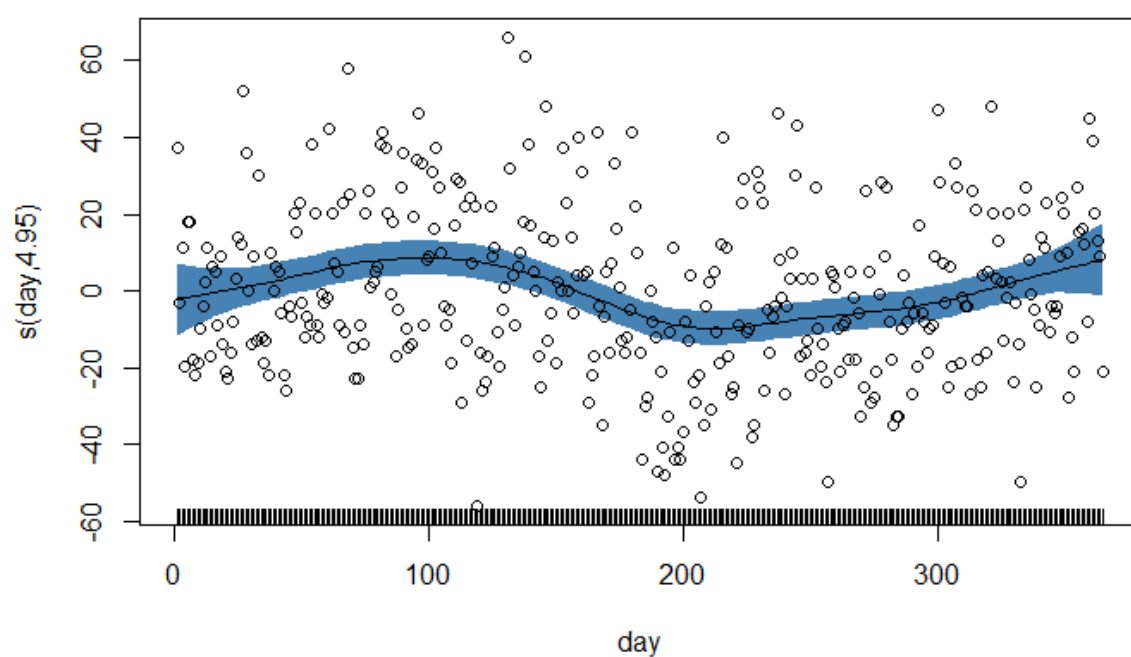
**Fig: Armadale/Kelmscott District Memorial Hospital**



**Fig: Swan District Hospital**



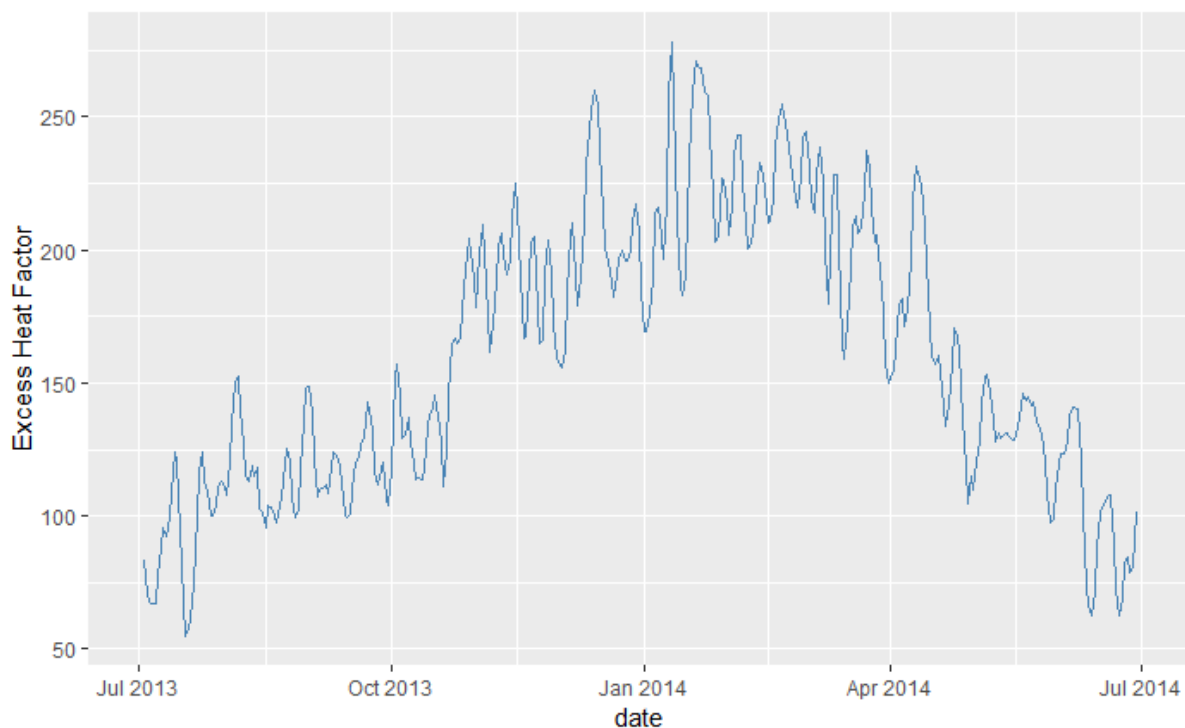
**Fig: Rockingham General Hospital**



**Fig: Joondalup Health Campus**

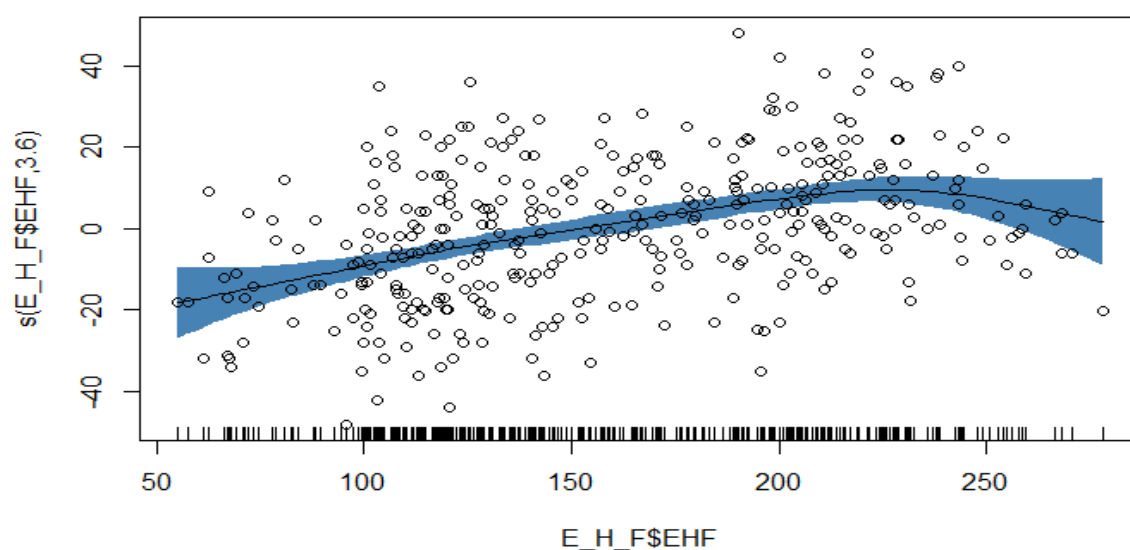
#### Q4.1

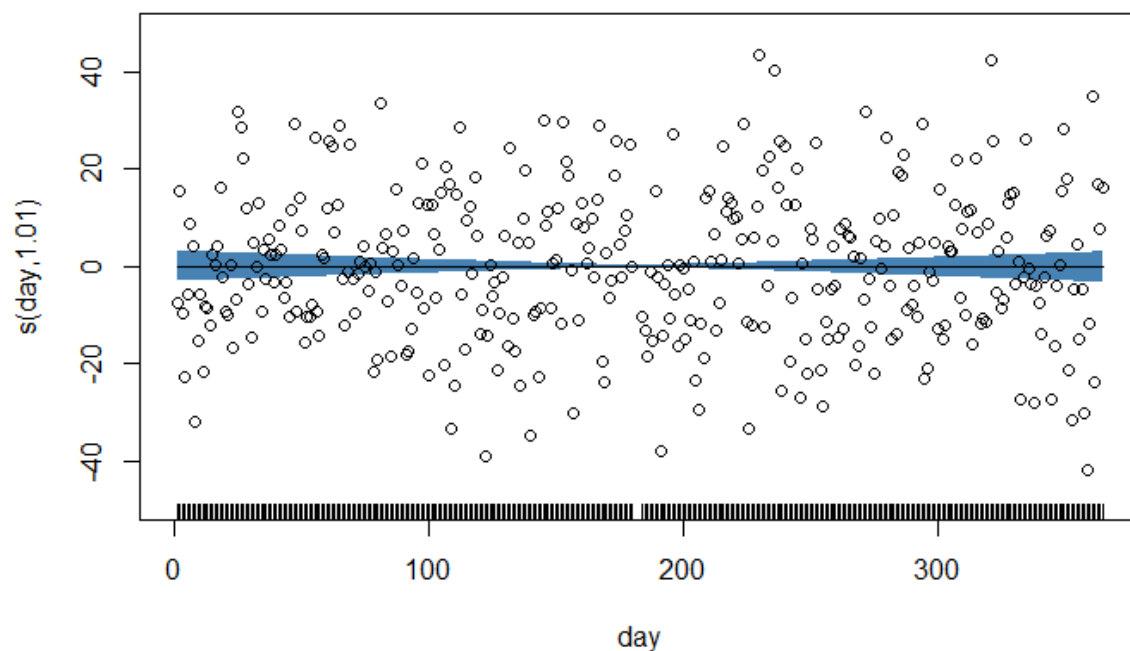
2.



In this section, we explore a significant heatwave which occurred across Western Australia from July 2013 to July 2014. In January/February 2014 values of EHF are very high. In consequence, the EHF exceeded the severity threshold in the Perth region by some factor.

#### Q4.2





It improves the model somewhat but it is not the best model. As we can see smooth terms of day and EHF are not significant. We make them significant by changing the number of the basis function (i.e. by changing the higher value of  $k$ ) to make it more significant.

```
Family: gaussian
Link function: identity
```

```
Formula:
Attendance__0 ~ s(day) + s(E_H_F$EHF)
```

```
Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  227.0276    0.8235   275.7   <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(day)         1.012  1.023  0.006    0.95
s(E_H_F$EHF)   3.596  4.476 16.871 1.02e-13 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) = 0.181   Deviance explained = 19.1%
-REML = 1508.7   Scale est. = 245.49    n = 362
```

```
Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-0.0005881069,0.004036958]
(score 1508.671 & scale 245.4854).
Hessian positive definite, eigenvalue range [0.0006124679,179.5054].
Model rank = 19 / 19
```

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'.

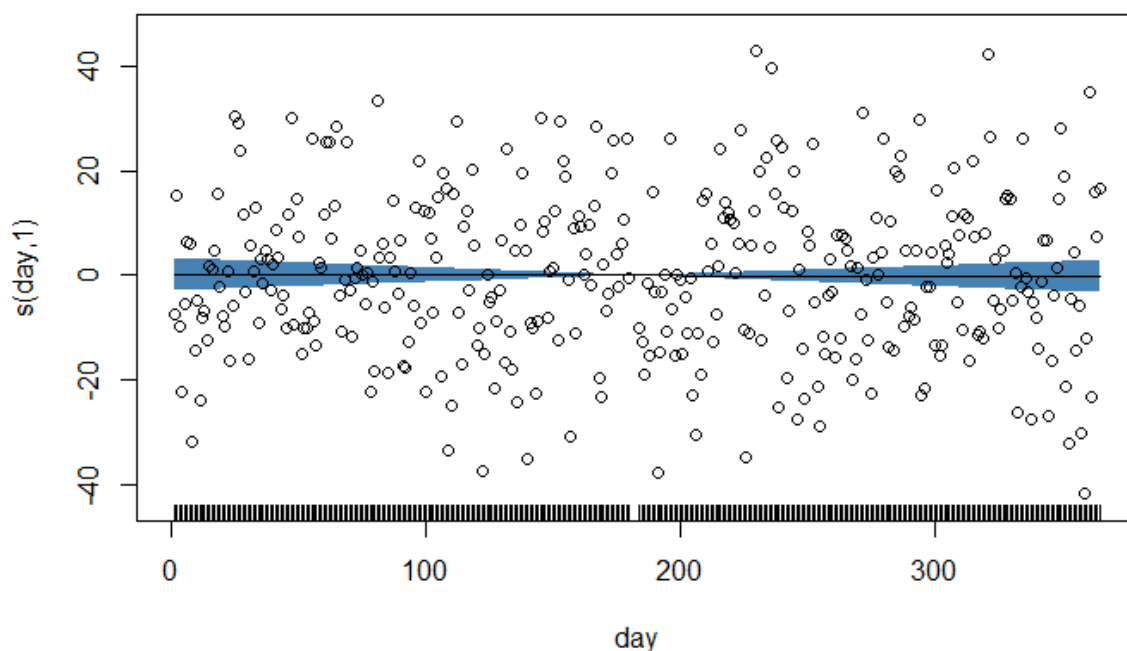
|               | k'   | edf  | k-index | p-value    |
|---------------|------|------|---------|------------|
| s(day)        | 9.00 | 1.01 | 0.83    | <2e-16 *** |
| s(E_H_F\$EHF) | 9.00 | 3.60 | 1.05    | 0.83       |

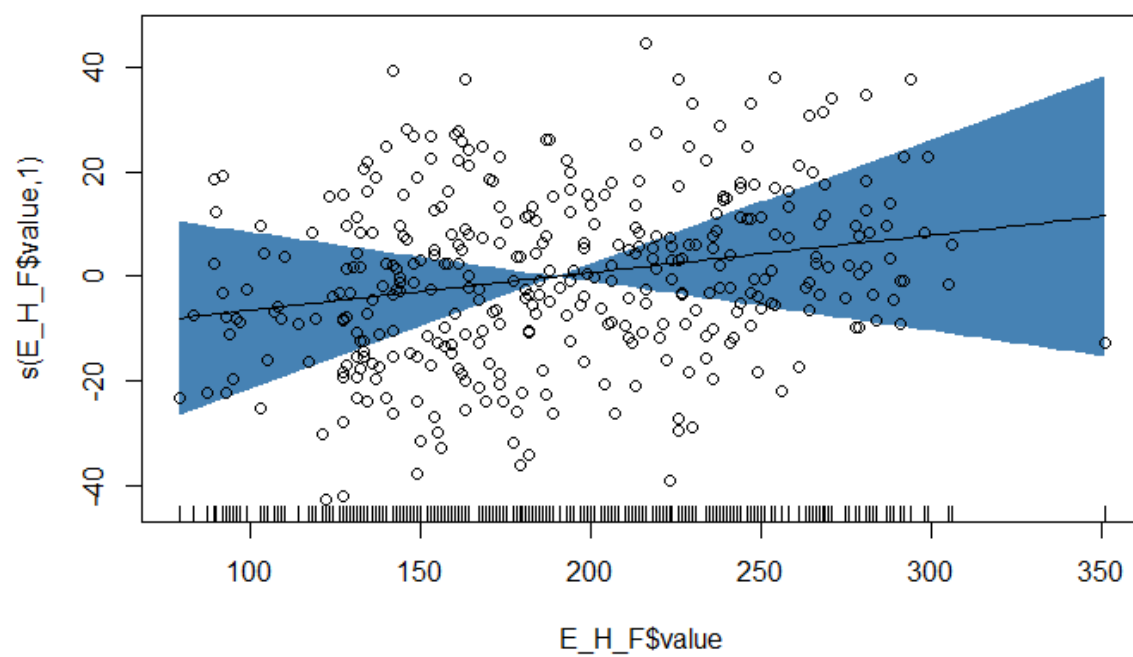
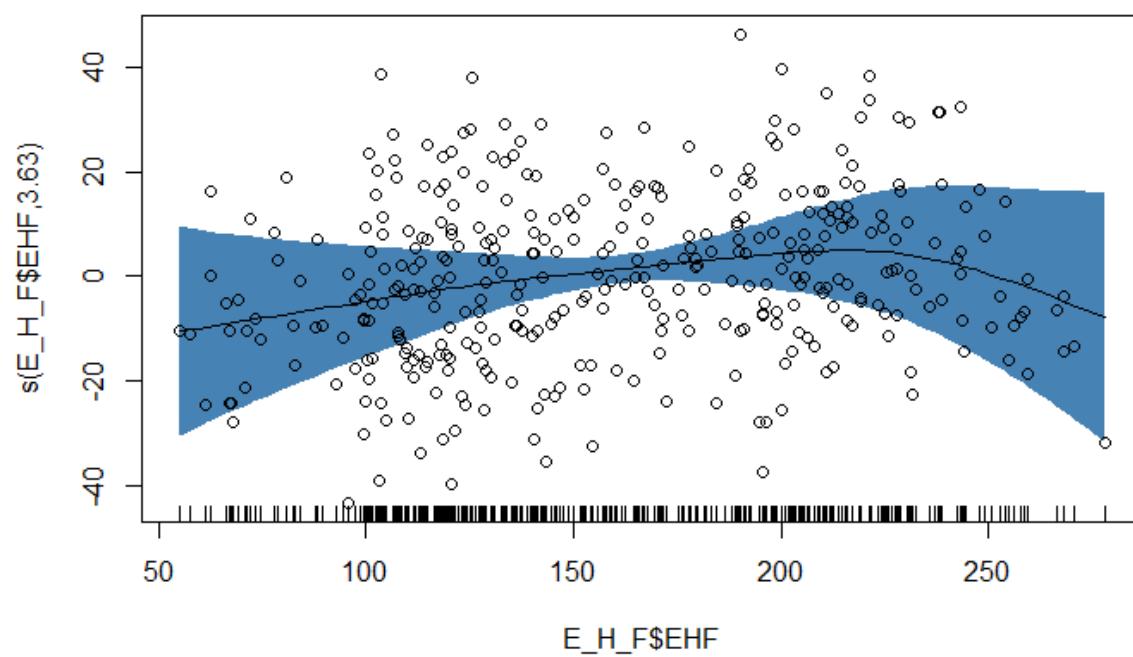
---  
 signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### (Optional)Q4.3

**Extra Features:** Relative Humidity, Wind Velocity, Extreme temperatures (heat waves)

I have researched online to find similar data from this period but I am unable to find the data with these variables values. So, in order to implement extra weather feature what I have put is that the average temperature of the day in the smoothing variable. As we can see from the plot's and model output's results quite improved a lot. Even the value of AIC and BIC has improved a lot.





Family: gaussian  
Link function: identity

Formula:

Attendance\_\_0 ~ s(day) + s(E\_H\_F\$EHF) + s(E\_H\_F\$value)

Parametric coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 227.0276 | 0.8237     | 275.6   | <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(day)          | 1.002 | 1.004  | 0.018 | 0.8943   |
| s(E_H_F\$EHF)   | 3.629 | 4.518  | 2.180 | 0.0471 * |
| s(E_H_F\$value) | 1.001 | 1.002  | 0.744 | 0.3888   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.181    Deviance explained = 19.3%  
-REML = 1505.9    Scale est. = 245.59    n = 362

Method: REML    Optimizer: outer newton

full convergence after 10 iterations.

Gradient range [-0.0002814096,0.001052474]

(score 1505.877 & scale 245.5851).

Hessian positive definite, eigenvalue range [0.0002235063,179.0086].

Model rank = 28 / 28

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(day)          | 9.00 | 1.00 | 0.82    | <2e-16 *** |
| s(E_H_F\$EHF)   | 9.00 | 3.63 | 1.05    | 0.78       |
| s(E_H_F\$value) | 9.00 | 1.00 | 1.09    | 0.94       |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Q5

### 1.

- Historical data may not give a true picture of an underlying trend. An obvious event like hurricanes fani, Katrina and Sandy will distort a normal business trend line.
- Long-term projections need more data to support them.
- The fact is that more outdated data is and more inaccurate in the calculations.

### 2.



Regression analysis is an important tool for looking at relationships between sets of variables. It is a mathematical method that helps in modelling and analysing data. In other words, it helps to determine which specific independent variable influence dependent variables.

Regression analysis does the following things:

- Model multiple independent variables
- Include continuous and categorical variables
- Use polynomial terms to model curvature
- Assess interaction terms to determine whether the effect of one independent variable depends on the value of another variable

We can easily make predictions by splitting the data into three sets. Training, Testing and Validation set. So from this, we can check the accuracy of the model. So when in the future if new data comes we will be able to predict future attendances in the assigned hospital.

### **3.**

Yes, all of my questions have been answered. In this project, I conduct the study for the impacts of temperature on emergency department visits in Western Australian Hospitals. It helps the government to improve its public health services management like to design an app to give the alerts, for pre-hospital health services and clinical resource allocation.