SIT 742

ASSIGNMENT 2

Adekola Oluwatade (215383256)

Deakin University

Background

As a result of the ED demands calculated across all nine variables in the first assignment and selecting Y (target variable). Total attendance, which is the sum of all patients' attendance across all nine hospitals will be used the main target variable (Y) to calculate ED demands in Perth Area. Further insights will be generated as more variables are included to improve the model and make models generated more meaningful and better than the previous ones.

Task 1

1. Just as the WA ED app has been put in place to be developed due to low budget and other factors affecting the government's ability to provide huge budget for the healthcare system, the predictive model that would be built will further help the emergency department prepare well before hand. The primary aim is to reduce the amount of pressure in the ED by providing not just better information to the public about the situation at the ED at any given time but to also provide insights to the ED health authorities about the particular times of the year when there is peak attendance of patients in the ED.
   So, what the model does is that it shows trends across all the months in the analysis where there is relationship between attendance and weather (i.e. temperature and precipitation). After building the predictive model using regression, future years can be used to estimate periods when there is huge demand of resources in the ED based on attendance.
   More so, this will greatly solve the overcrowding problem in the ED. The WA ED app together with the predictive model will function under one unit. Where the WA App provides detailed information to potential patients about the time to wait before they will be attended to, the predictive model will provide insights to the ED authorities on the relationship between weather and time as it affects attendance of patients. This will enable correct estimate on times of the year when there would be huge need for more staff, more resources, more common disease to provide medications for etc.
   The potential users are the ED authorities and the government who are looking to minimize cost, reduce overcrowding and provide effective resources to the people.

2. Here, we are going to predict an estimate of the number of attendance at any given time in the ED. The response variables are the variables that measures weather (e.g. Temperature and Precipitation). These variables (both predictors and target) are as a result of daily collection of both weather and ED data. As a result of this, there will be regular updates to the model as more and more data tend to shift the data points and this in turn will change the model.

3. The model which is based on historical data will highly be influenced by the pattern of relationship between the predictors and target over the period the data was collected. So, this model will be trained based on these data and used for future predictions.
   The model due to its numeric attribute and the type of model used in building it, will maintain the same data type throughout.

4. In this statistical method, the best method to use will be Regression analysis. This is because all the data types are numerical where y is the target variable and x1…xn

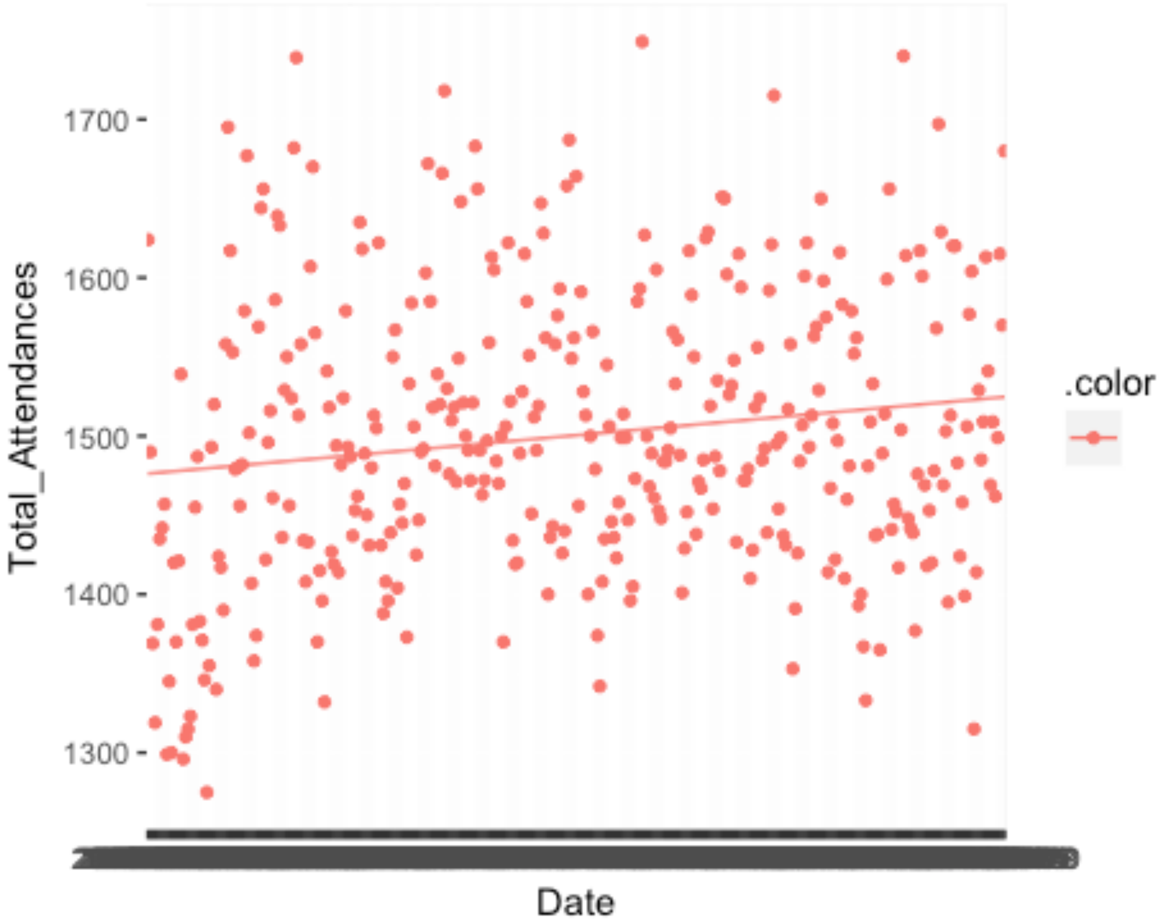
represents the predictors. Therefore, a change in x will influence y. So we can say that a change in weather condition will influence the number of attendance in the emergency department.

Task 2

Task 2.1

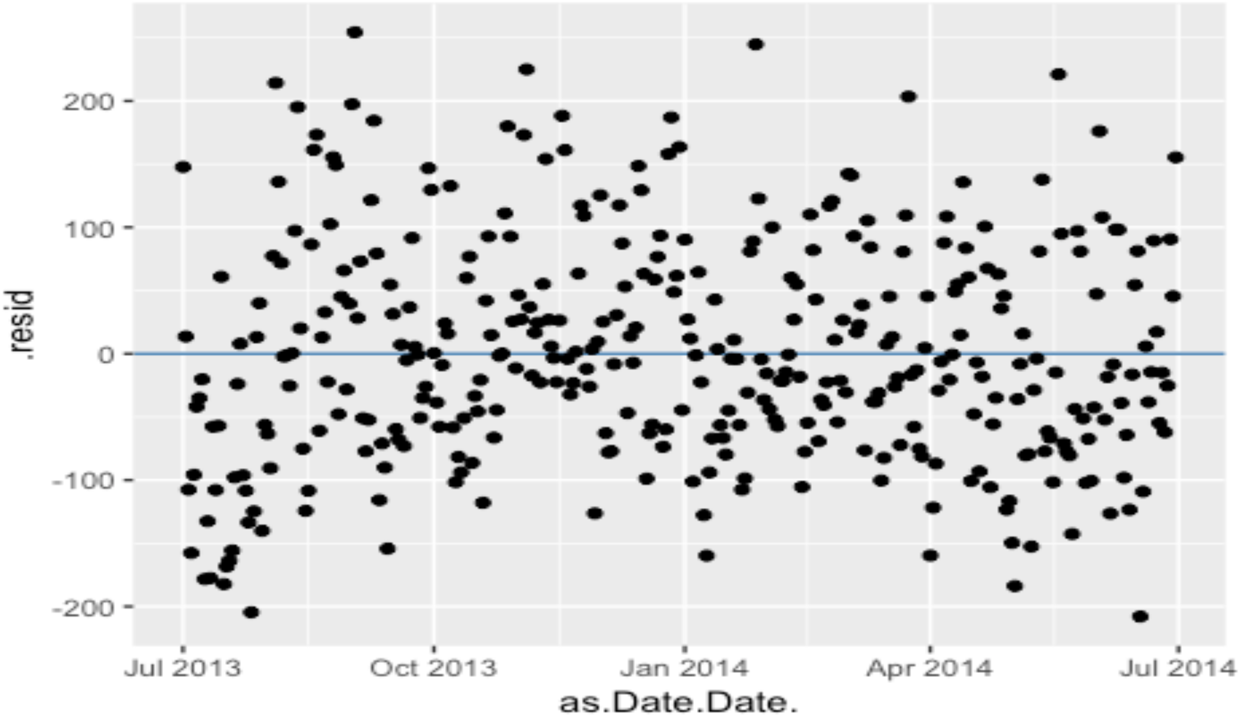
1. The plot below shows the Linear model for Y (Total Attendance as per assignment 1) and date (Predictor variable).

Linear Model



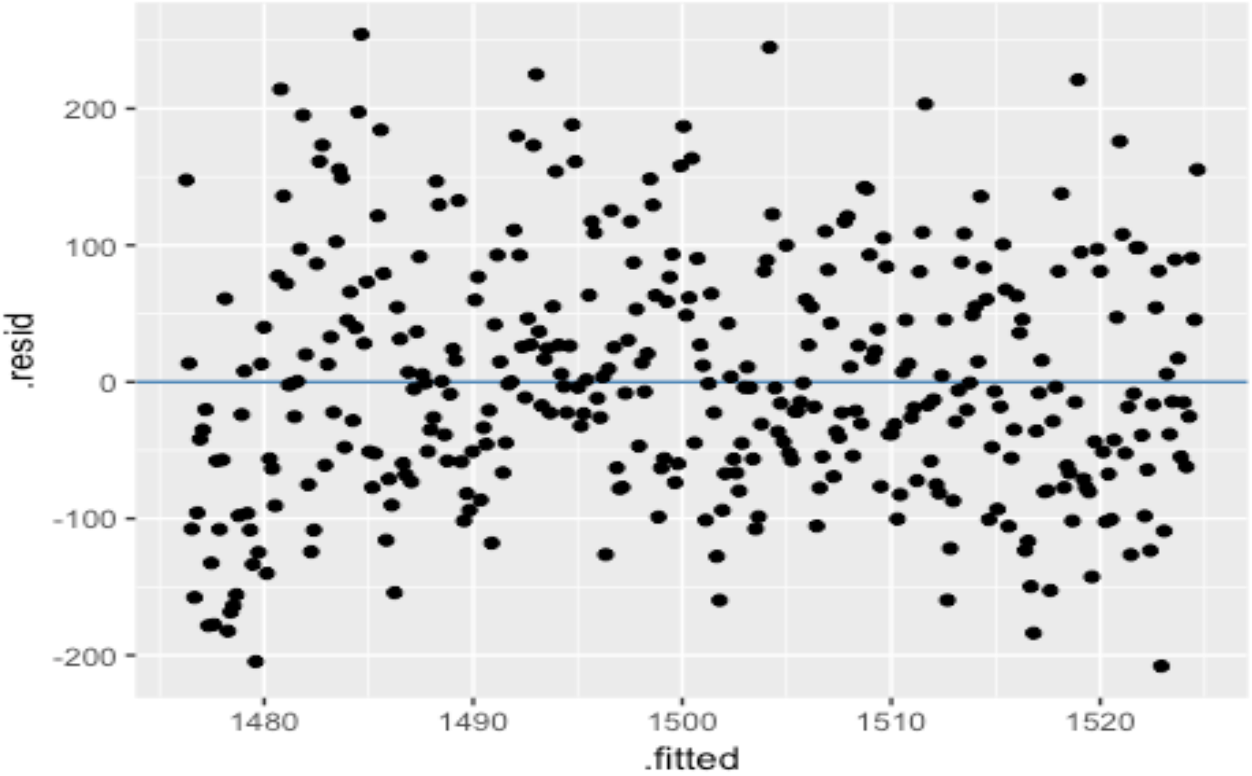
Residual vs Predictor

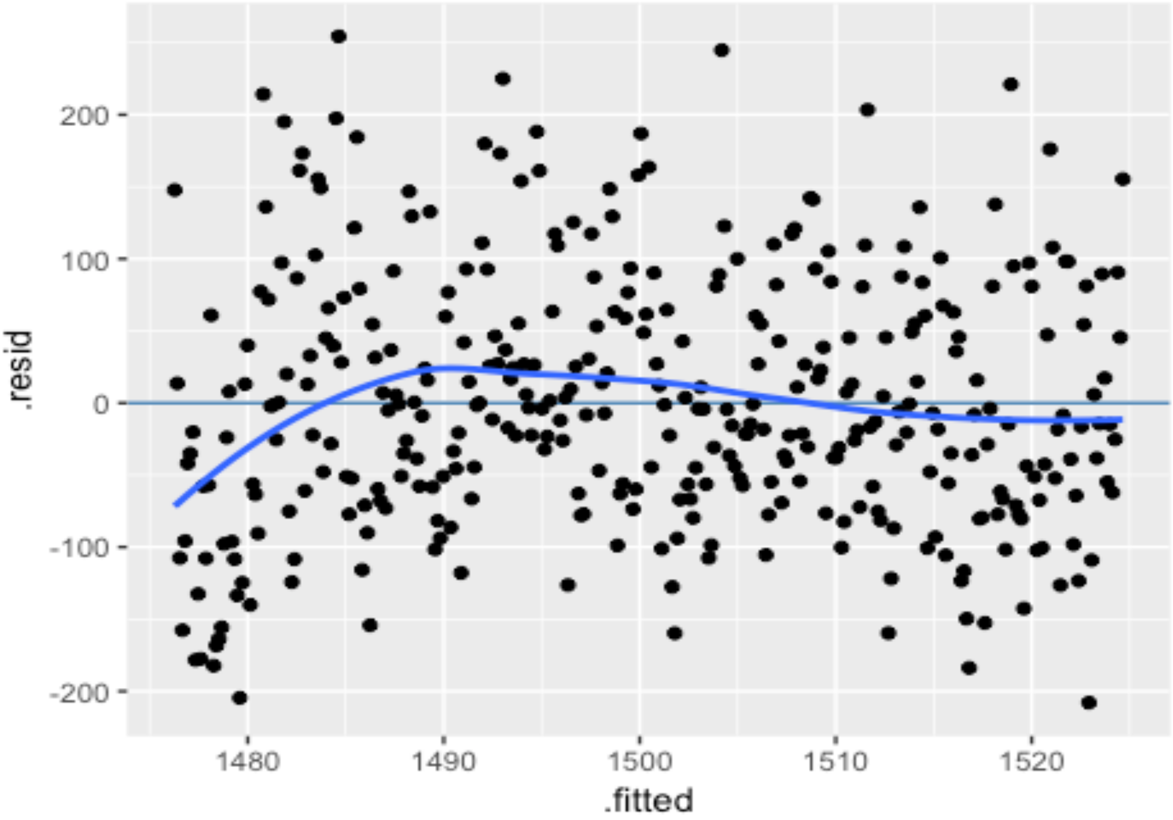
The below figure shows the plot of the fitted value against the Predictor (Date)



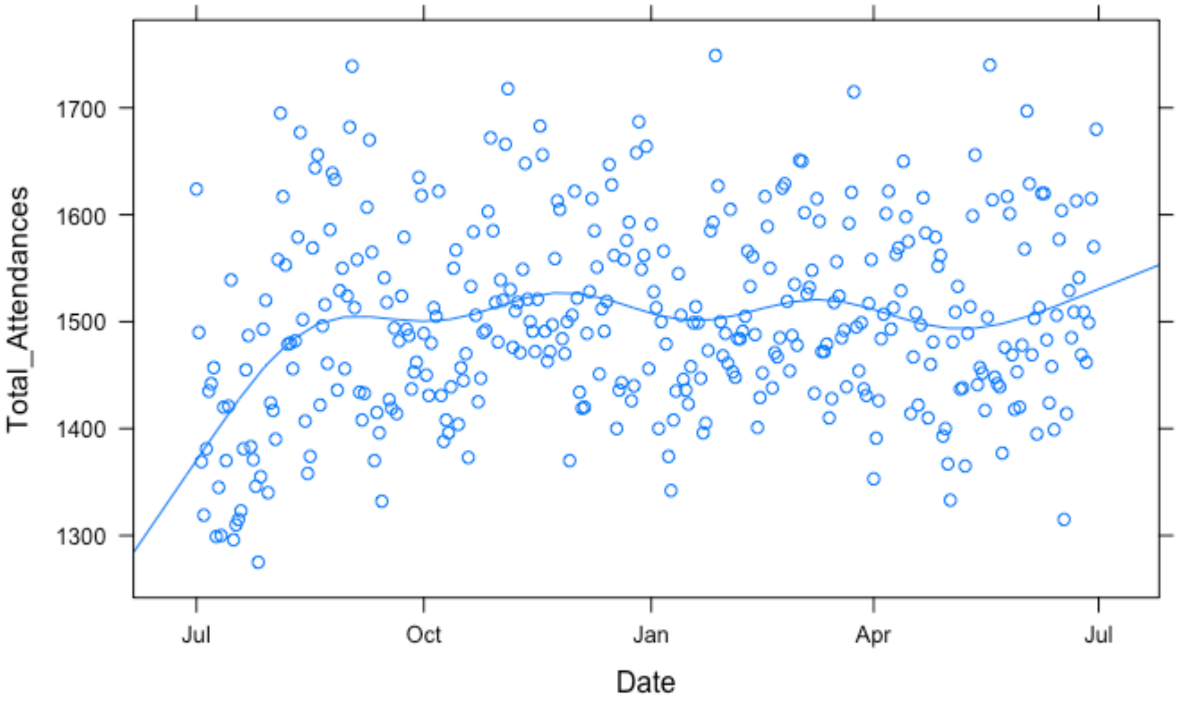
Fitted vs Residual

The below figure shows the plot for residual against the fitted values.

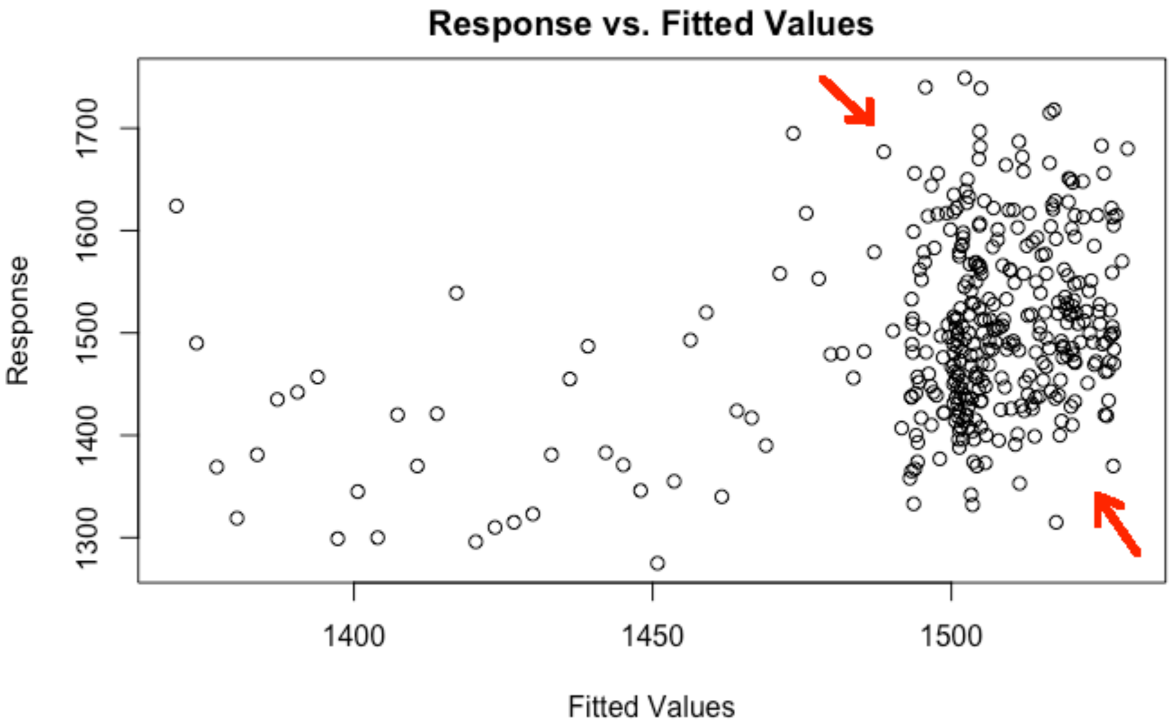




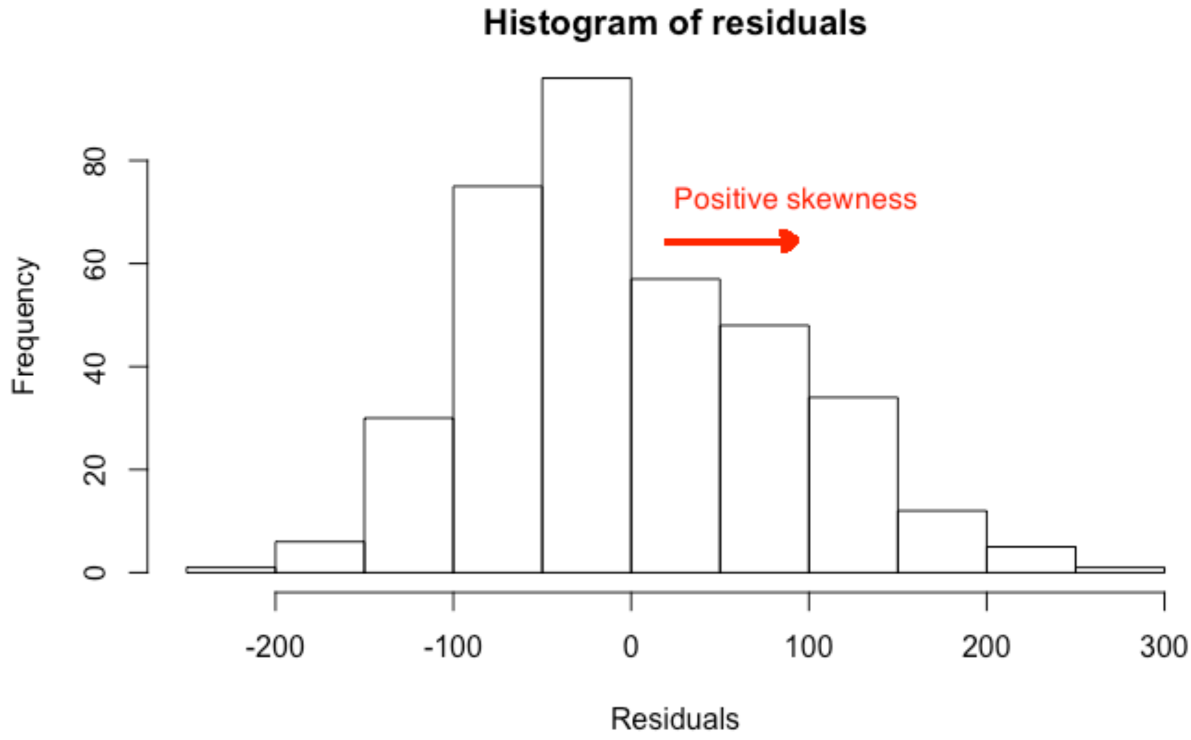
2. The plot below shows the Fitted generalised additive model (GAM) for the target (Total Attendance) and predictor (date)



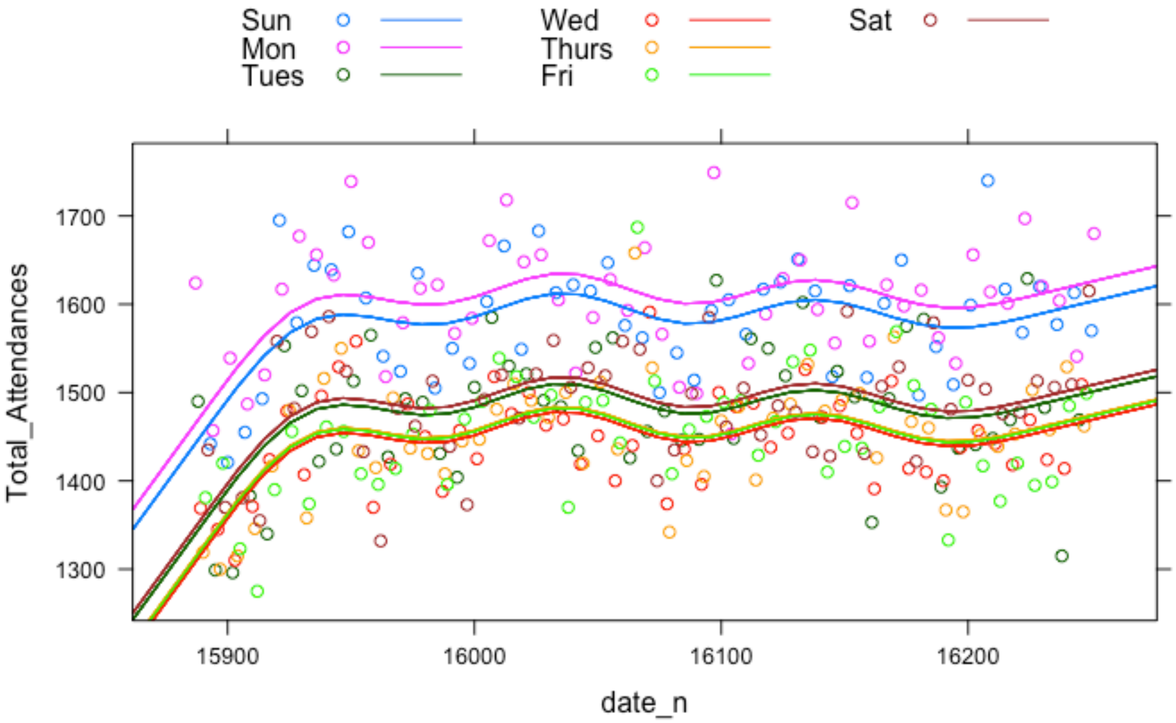
After assessment of the model fit was done by plotting the residuals against the response variable i.e Date, it was discovered that most of the data point are not normally distributed as most data points are highly clustered which indicates insufficient model fit.

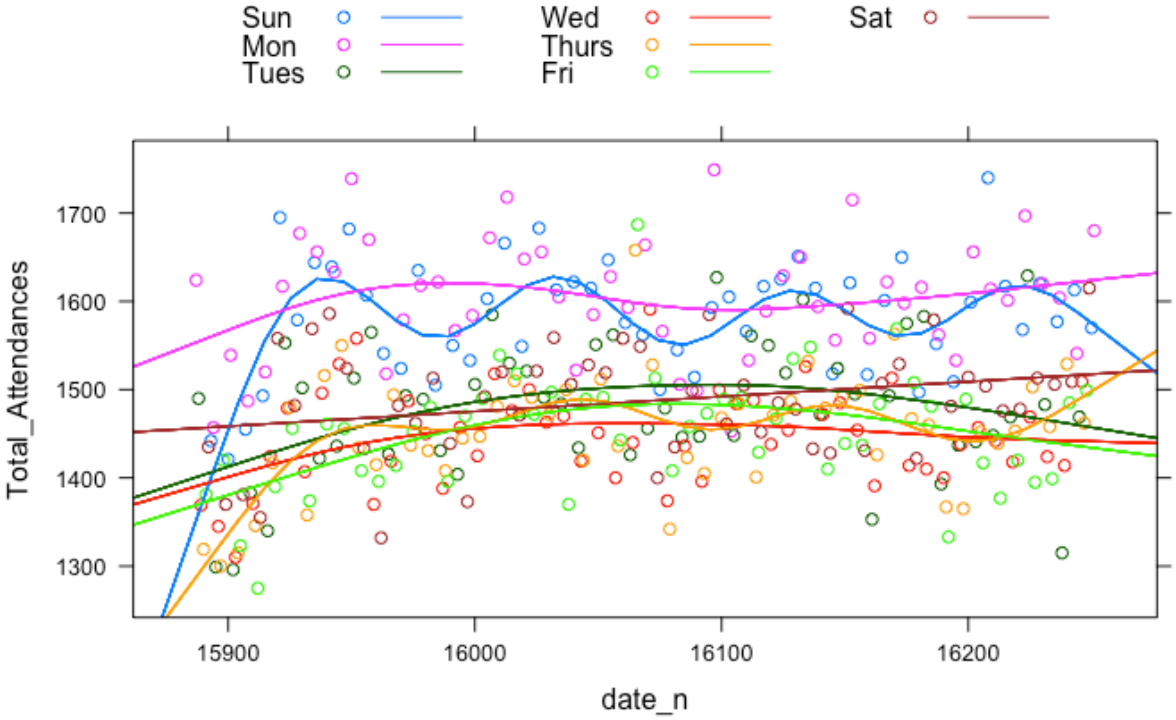


The below shows a histogram of the residuals. It’s seen that plot is positively skewed. A plot which is not normalized

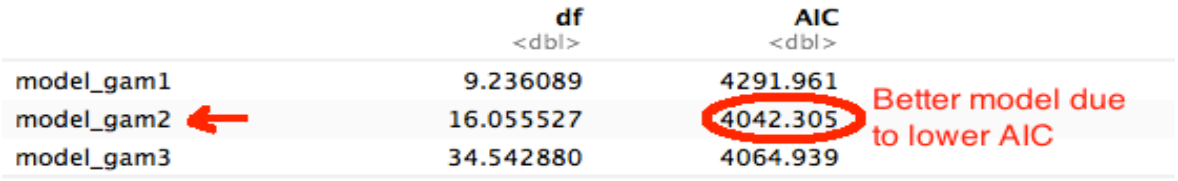


3. Incorporated weekly seasonality is shown below





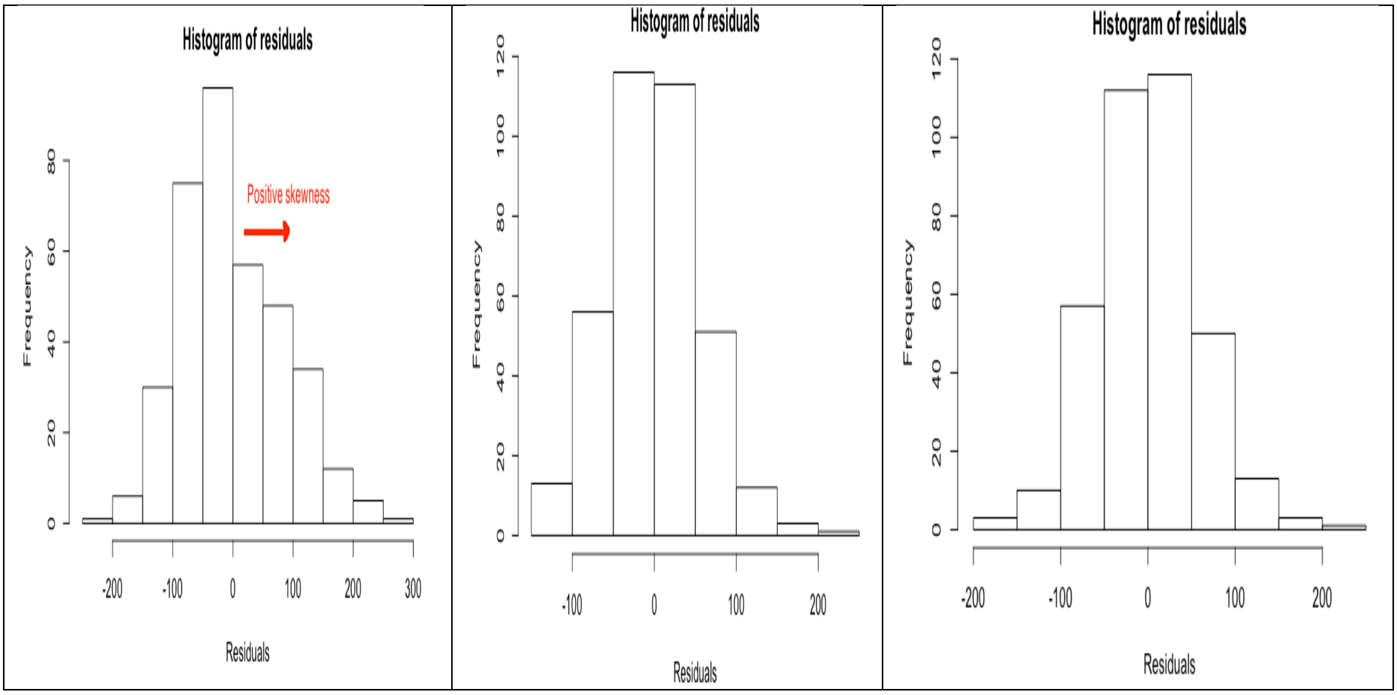
Model Comparison using AIC coefficient estimate



	df <dbl>	AIC <dbl>
model_gam1	9.236089	4291.961
model_gam2	16.055527	4042.305
model_gam3	34.542880	4064.939

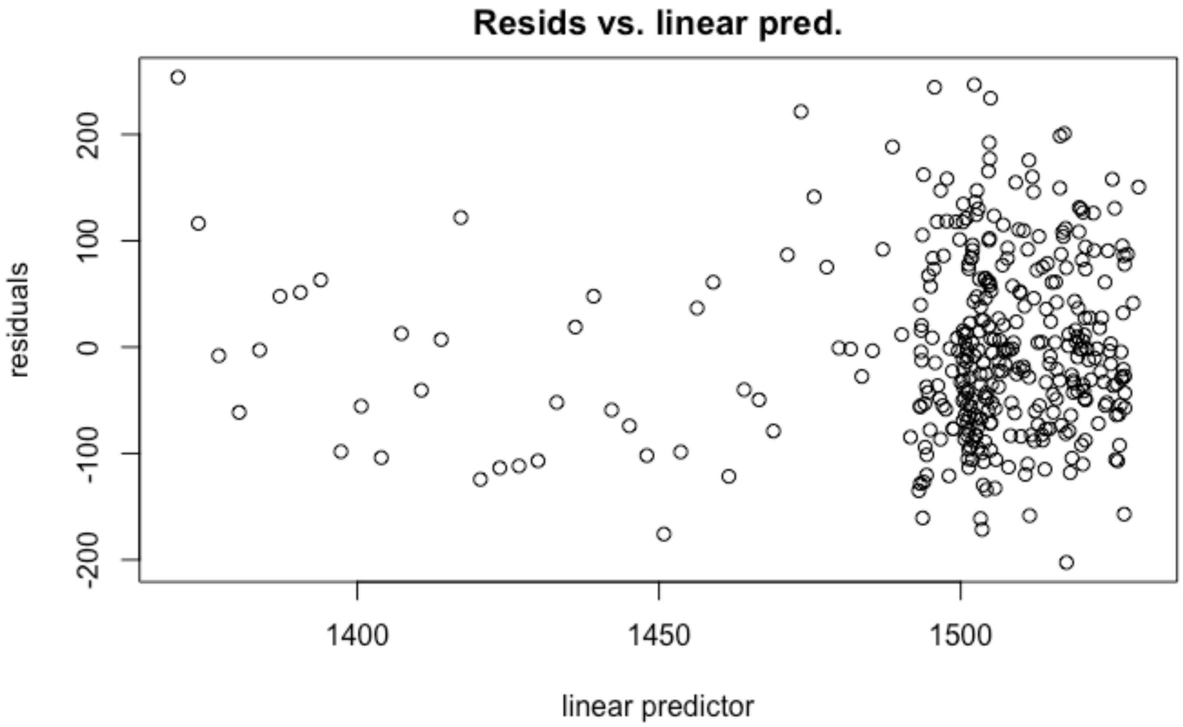
Better model due to lower AIC

Model comparison using Histogram plots

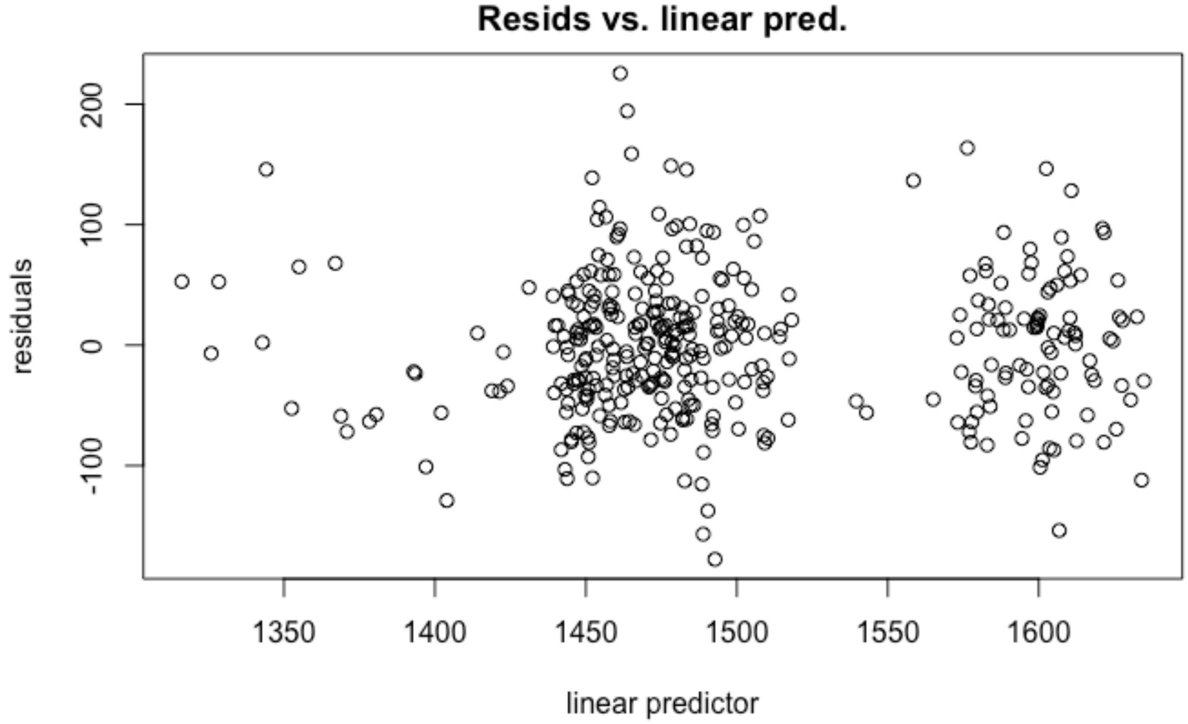


4. Analysing the residuals and checking for correlation. As seen below, it can be understood that the linear predictor tends towards higher values.

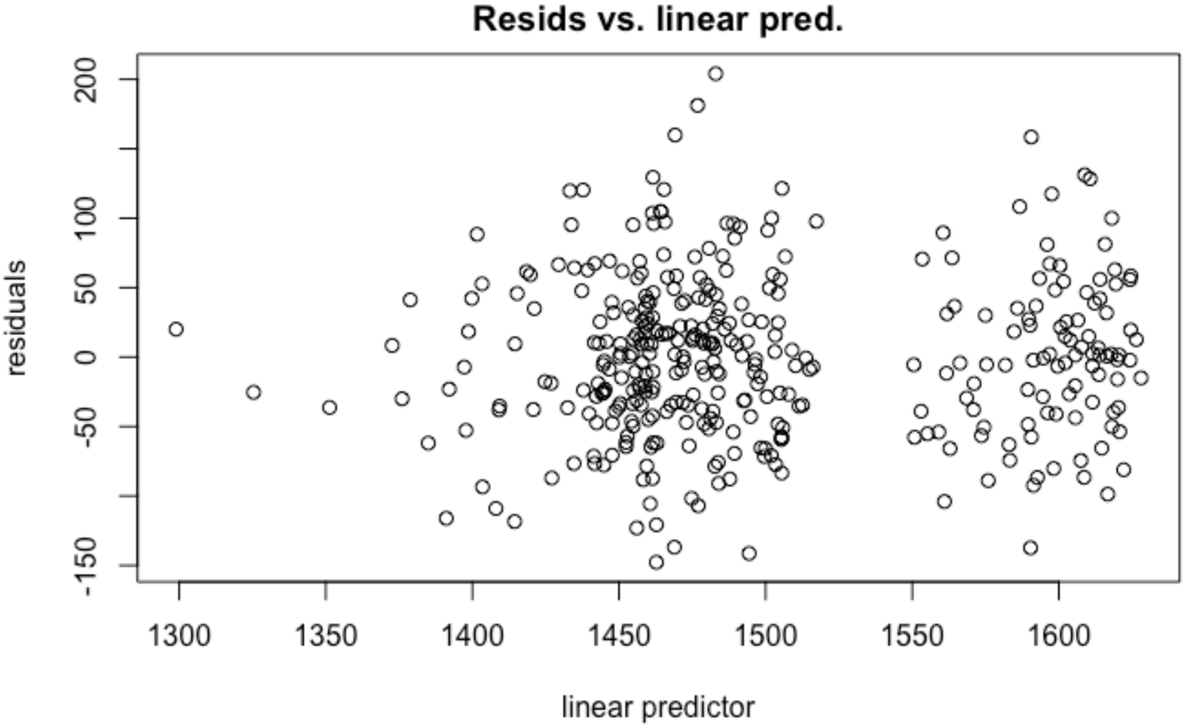
Model 1



Model 2

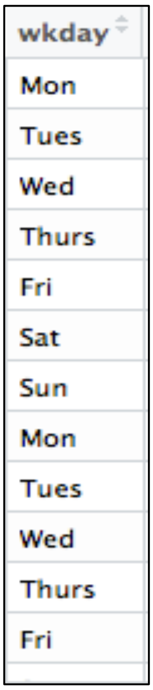


Model 3



5. The day of the week variable created in the “weekly” data frame to incorporate weekly seasonality is a **categorical variable** as all values are in non-numerical format. Incorporating weekly seasonality didn’t affect the model fit as the target variable will

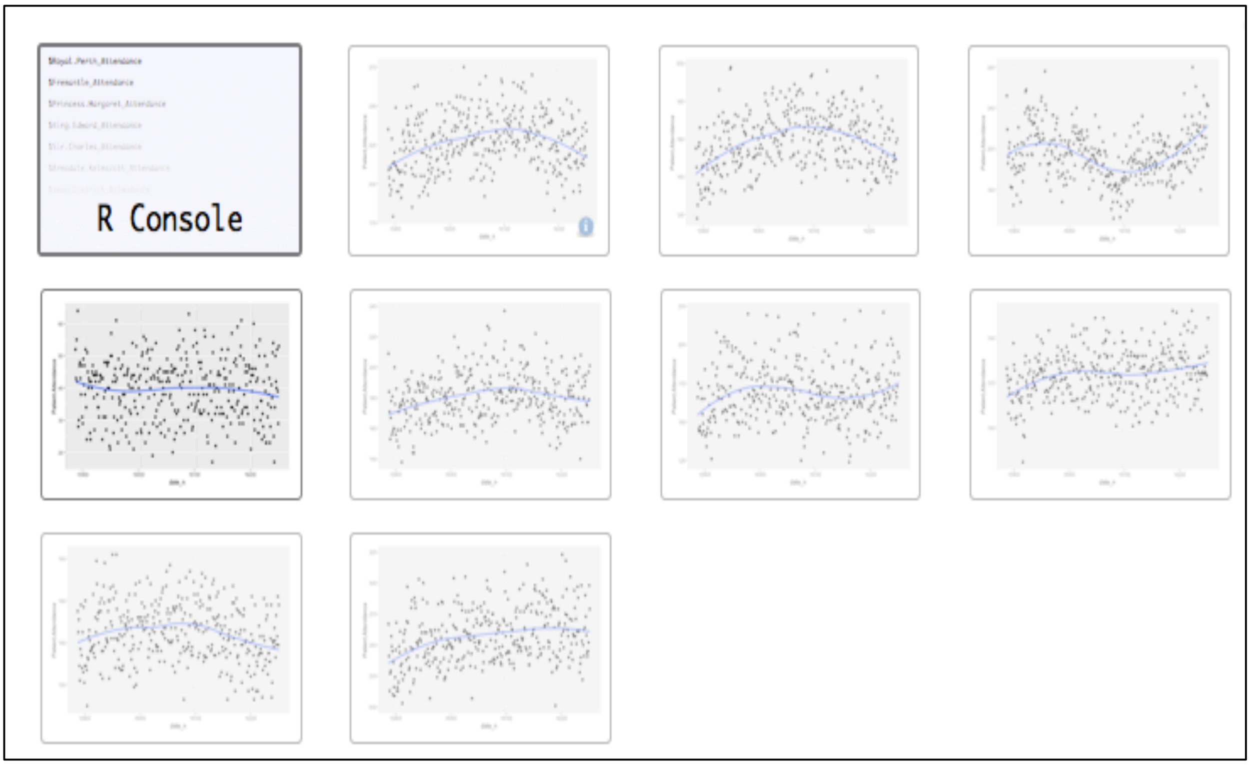
always have the same effect on the continuous variable (date) no matter the format it takes. Below shows the representation of the day of the week in the data frame used in the analysis.



wkday
Mon
Tues
Wed
Thurs
Fri
Sat
Sun
Mon
Tues
Wed
Thurs
Fri

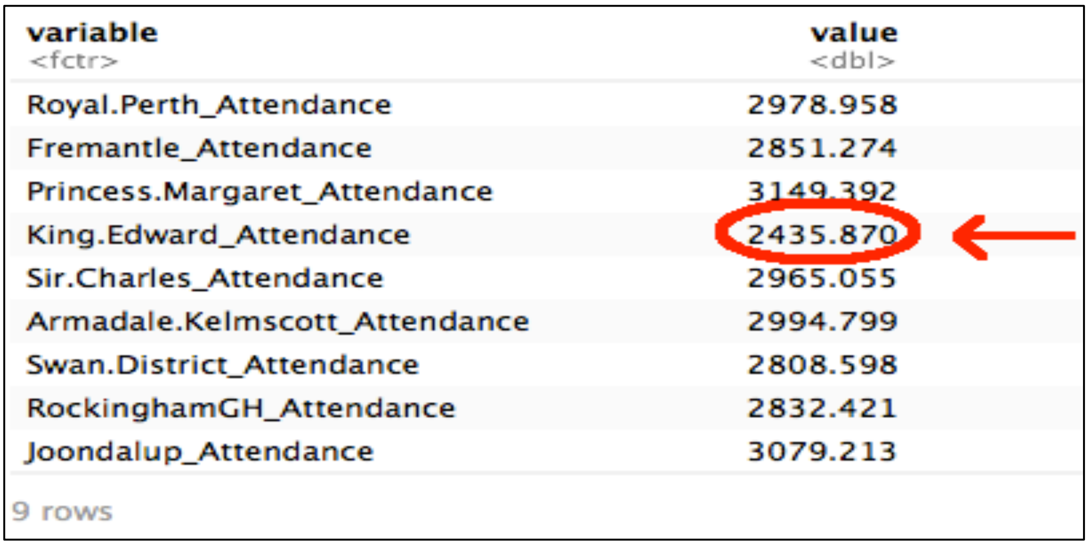
Task 2.2

1. After running the map function on all nine hospitals, the following plots were the final outcome.



```
$Royal.Perth_Attendance
$Fremantle_Attendance
$Princess.Margaret_Attendance
$King.Edward_Attendance
$Sir.Charles_Attendance
$Armadale.Kelmscott_Attendance
$Swan.District_Attendance
```

Also, using map function, the table below shows an AIC comparison across all nine hospitals with King Edward Hospital attendance coming out with the lowest AIC coefficient. This signals a best predictive model.

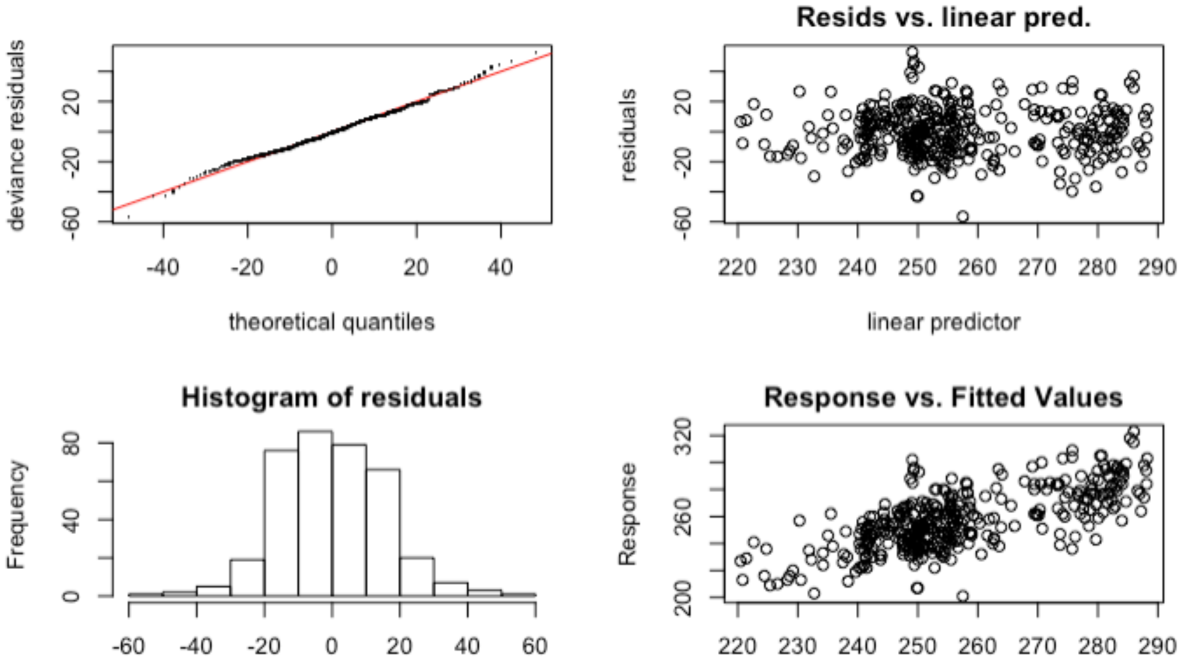


variable <fctr>	value <dbl>
Royal.Perth_Attendance	2978.958
Fremantle_Attendance	2851.274
Princess.Margaret_Attendance	3149.392
King.Edward_Attendance	2435.870
Sir.Charles_Attendance	2965.055
Armadale.Kelmscott_Attendance	2994.799
Swan.District_Attendance	2808.598
RockinghamGH_Attendance	2832.421
Joondalup_Attendance	3079.213

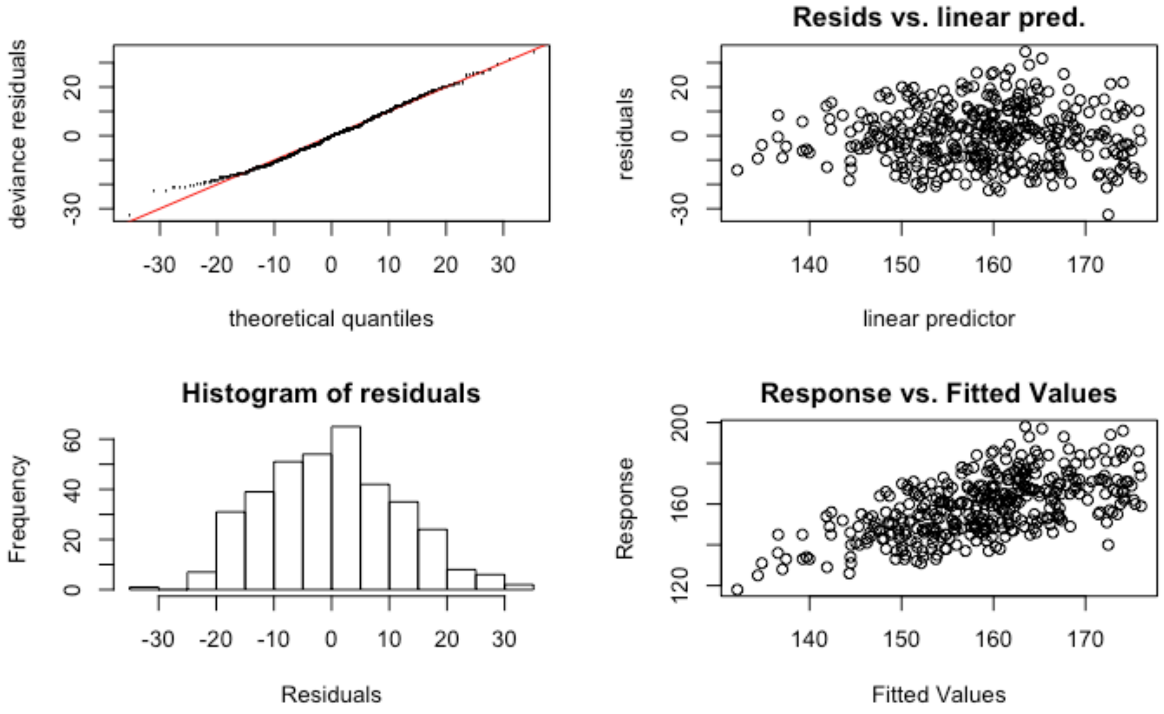
9 rows

2. Plotting the trends and residual, we have the following output for all nine hospitals

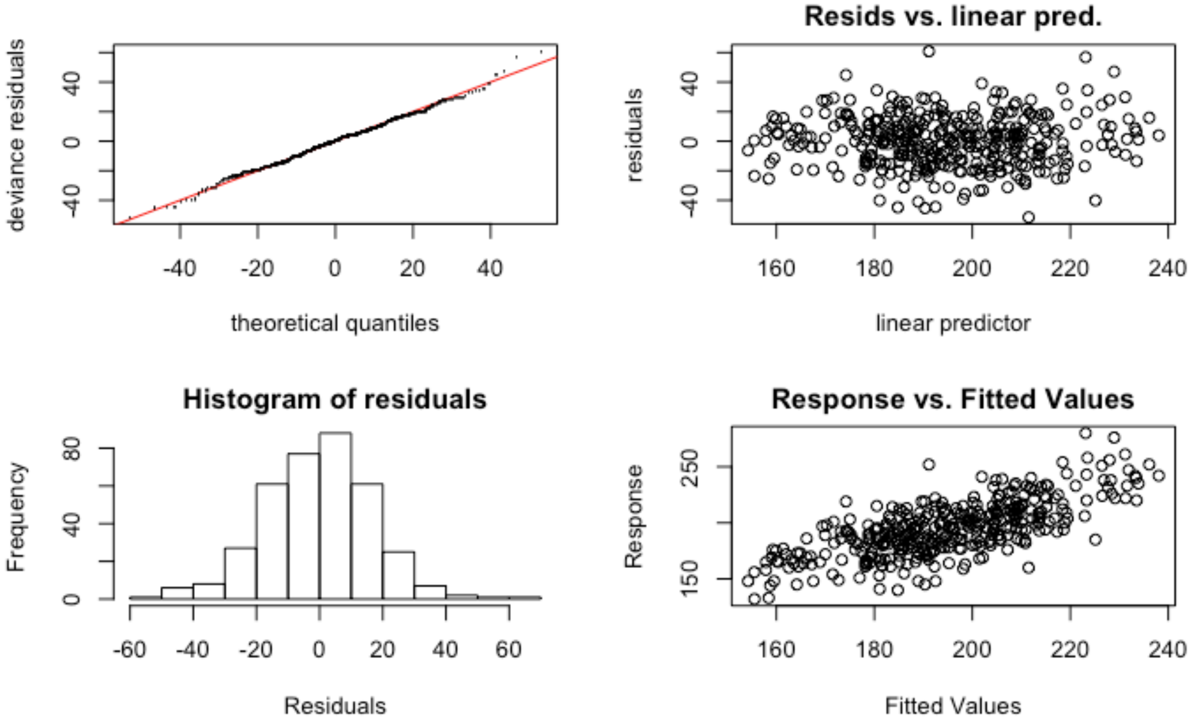
Royal Perth Hospital



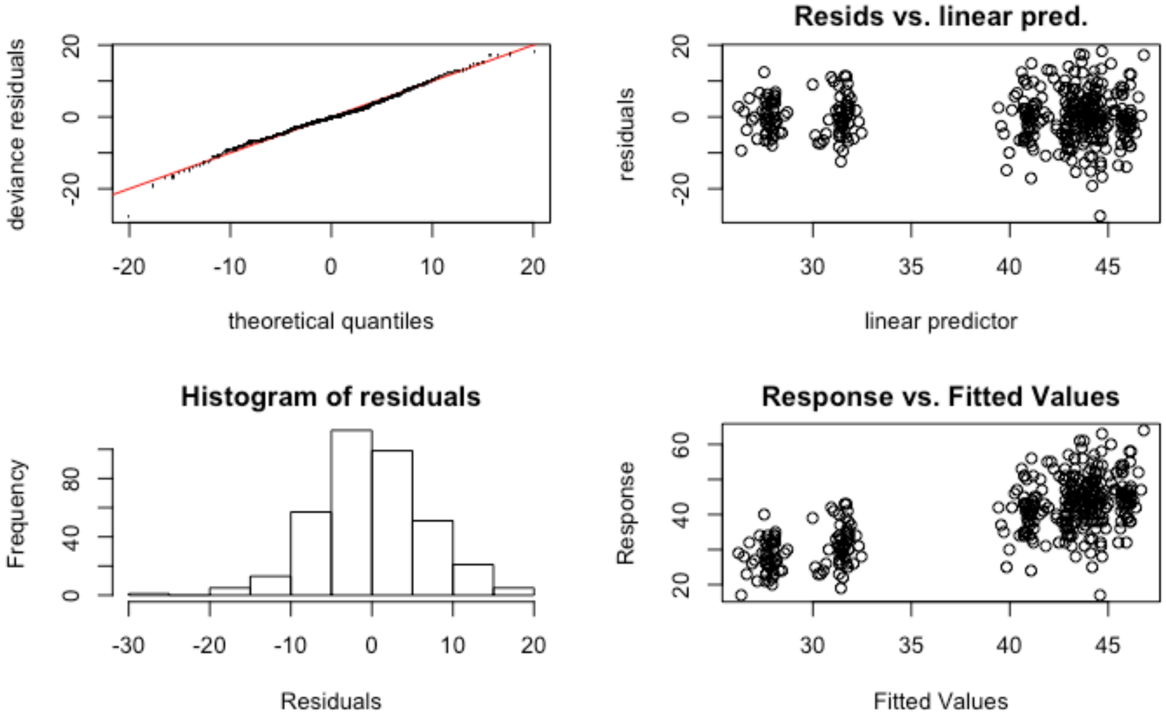
Fremantle Hospital



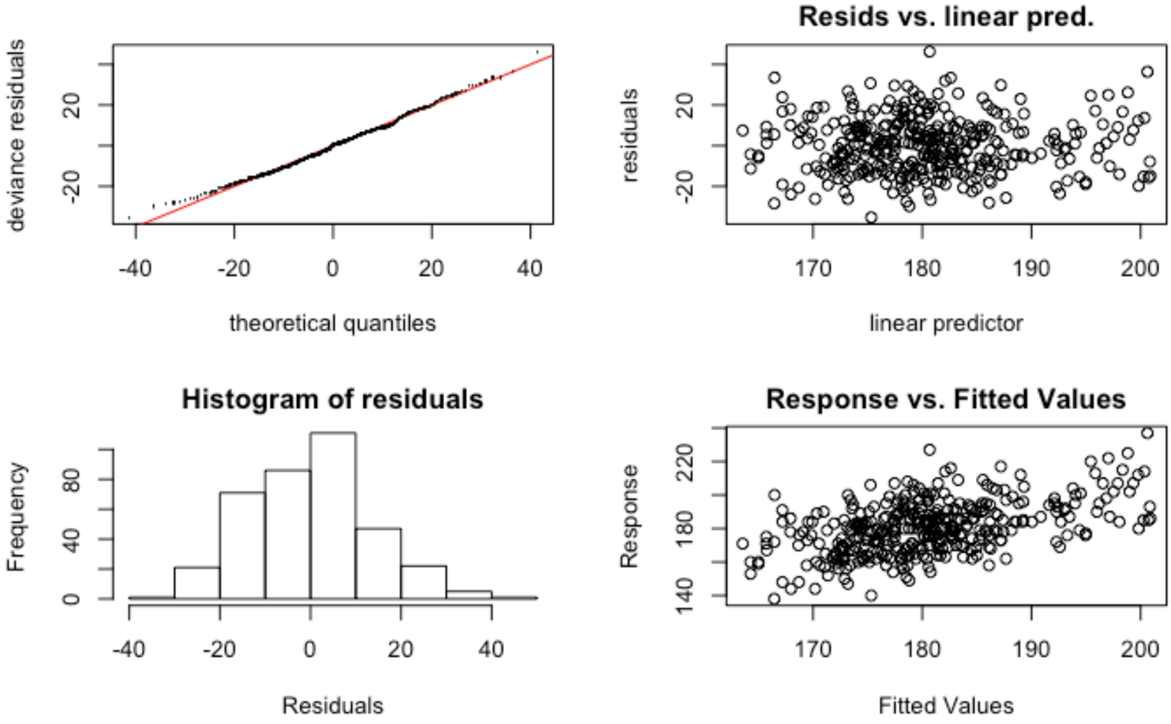
Princess Margaret Hospital



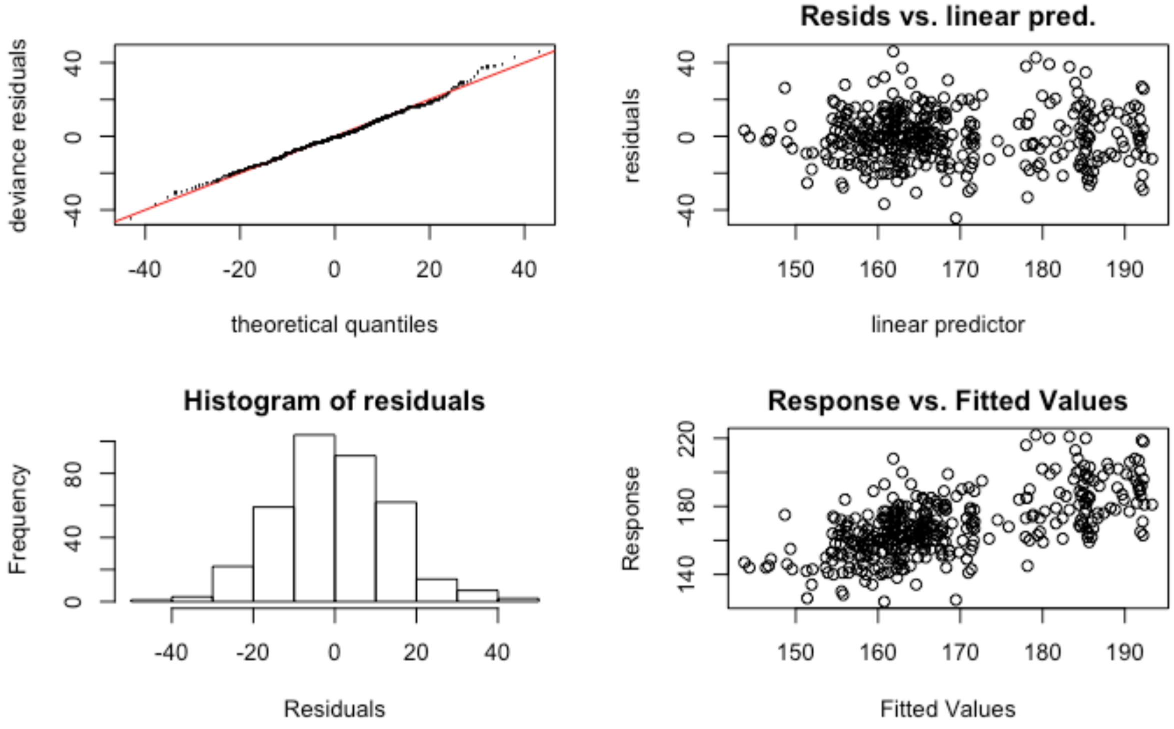
King Edward Hospital



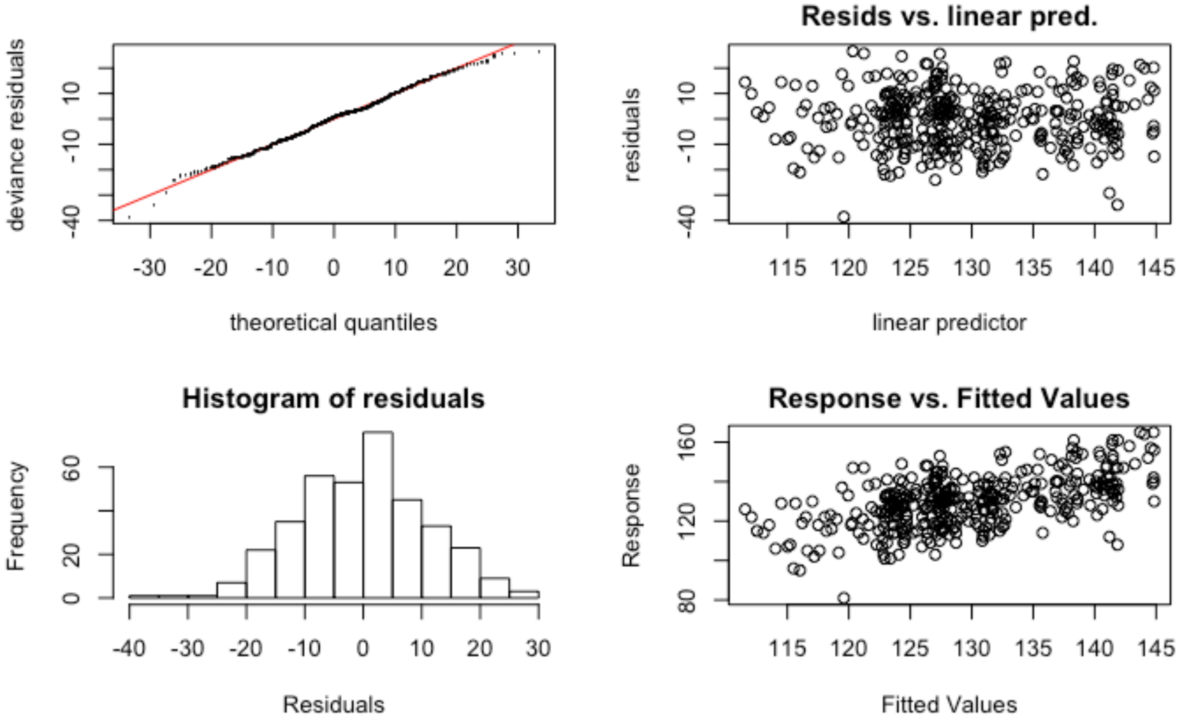
Sir Charles Hospital



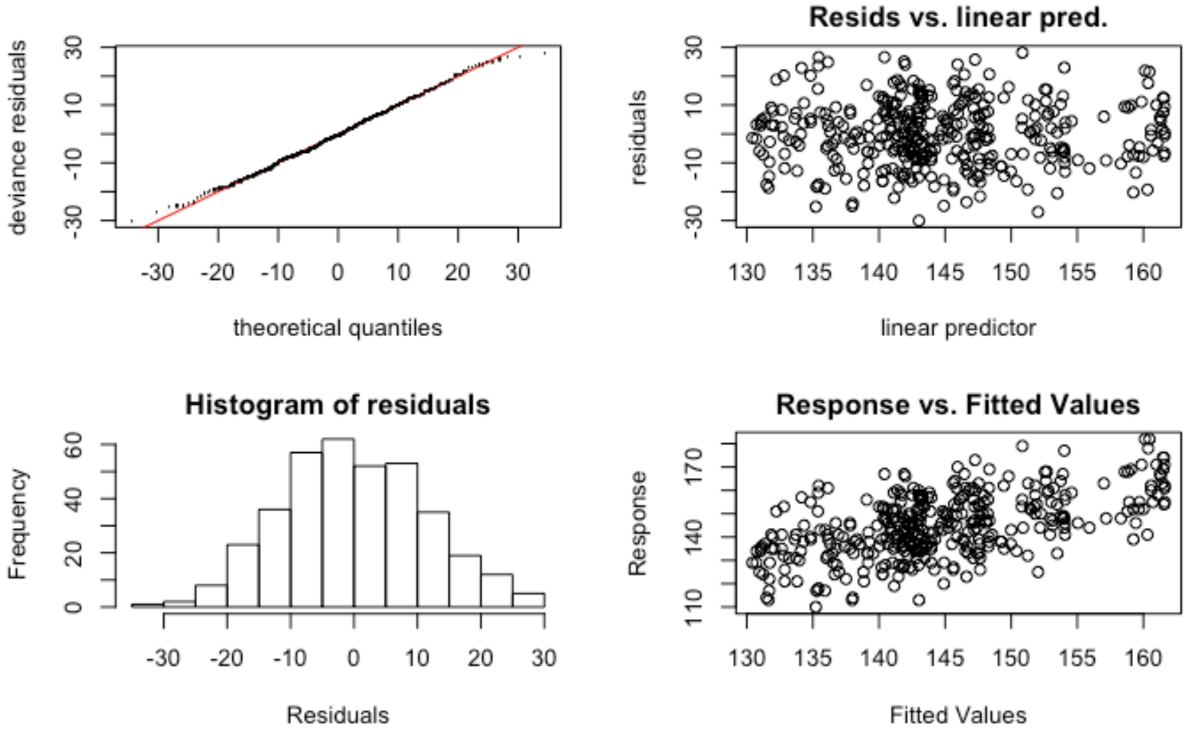
Armadale Kelmscott Hospital



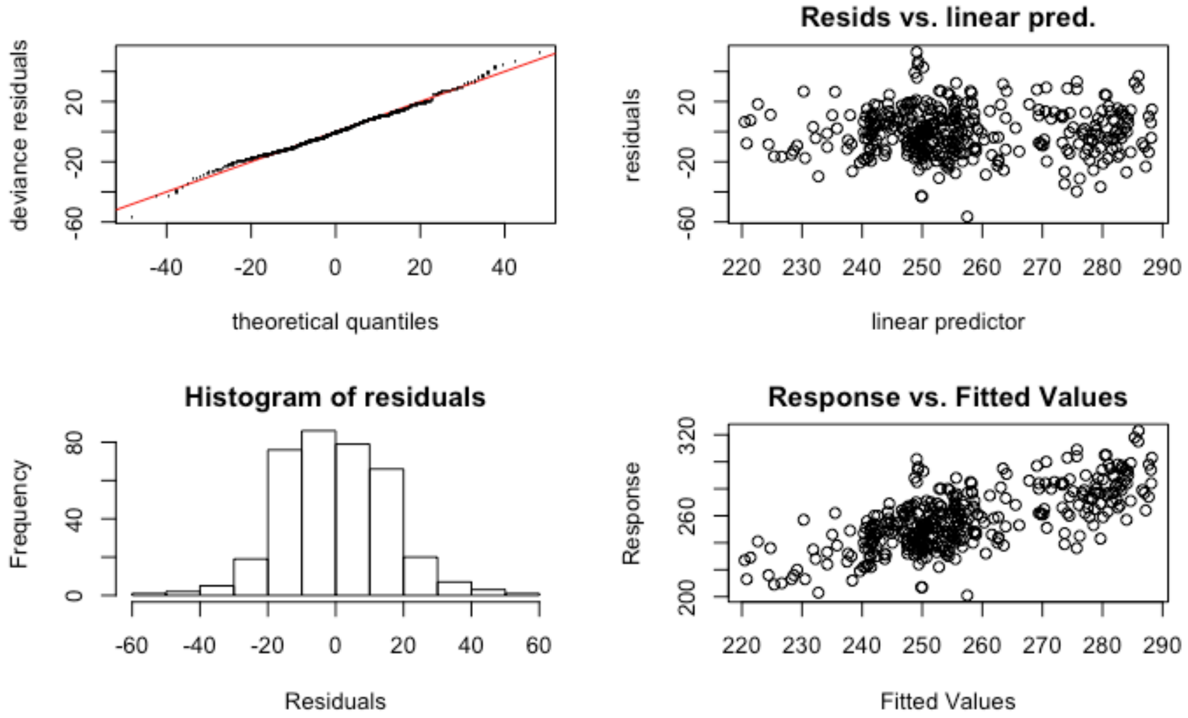
Swan District Hospital



RockinghamGH Hospital



Joondalup Hospital



Task 3

1. The EHF was used to measure heatwave intensity, incorporating two ingredients. The first ingredient often called **significance index** was a measure of how hot a three-day period (TDP) is with respect to an annual temperature threshold at each particular location. If the daily mean temperature (DMT) averaged over the TDP is higher than the climatological 95th percentile for DMT (hereafter T_{95}), then the TDP and each day within in it are deemed to be in heatwave conditions. On average, around 18 days per year will have a DMT exceeding T_{95}, but it is necessary to have three high DMTs in succession in order to form a heatwave according to this characterisation. Also, another ingredient often called **acclimatisation index** is done by measuring a TDP of the previous 30-day period (Nairn, 2014).

   The two ingredients in the EHF calculation, as described above, **are called excess heat indices (EHIs)** and calculated as follows:

$$\text{EHIsig} = (Ti+Ti+1+Ti+2)/3 - T95$$

   and:

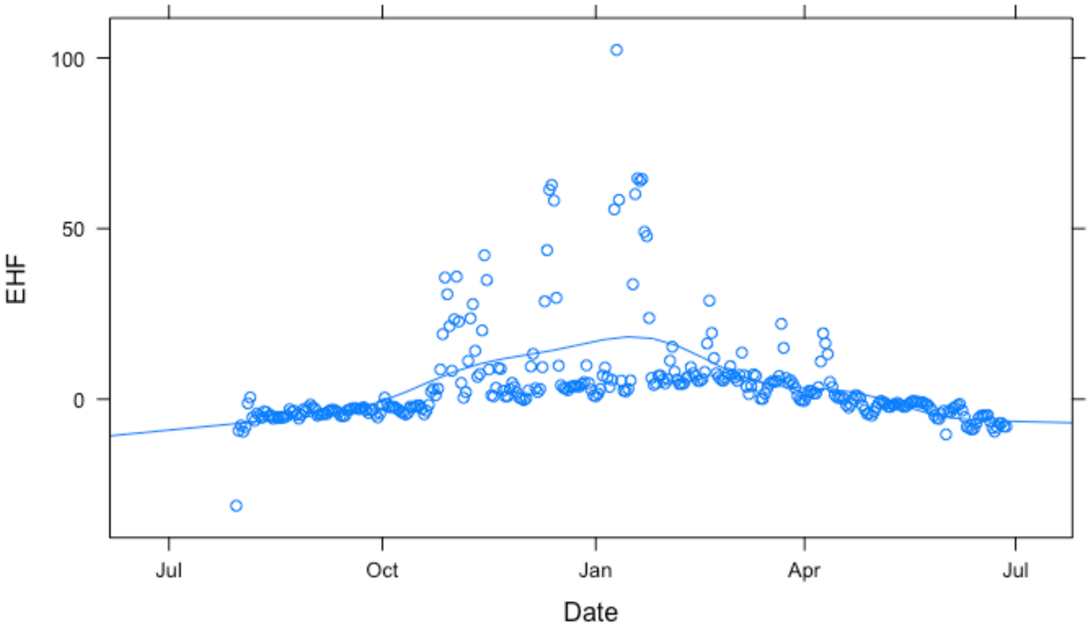
$$\text{EHIaccl} = (Ti+Ti+1+Ti+2)/3 - (Ti-1+ \ldots +Ti-30)/30$$

   (Nairn, 2014)

   Finally, the **Excess Heat Facto**r (EHF) is calculated using the below formula:

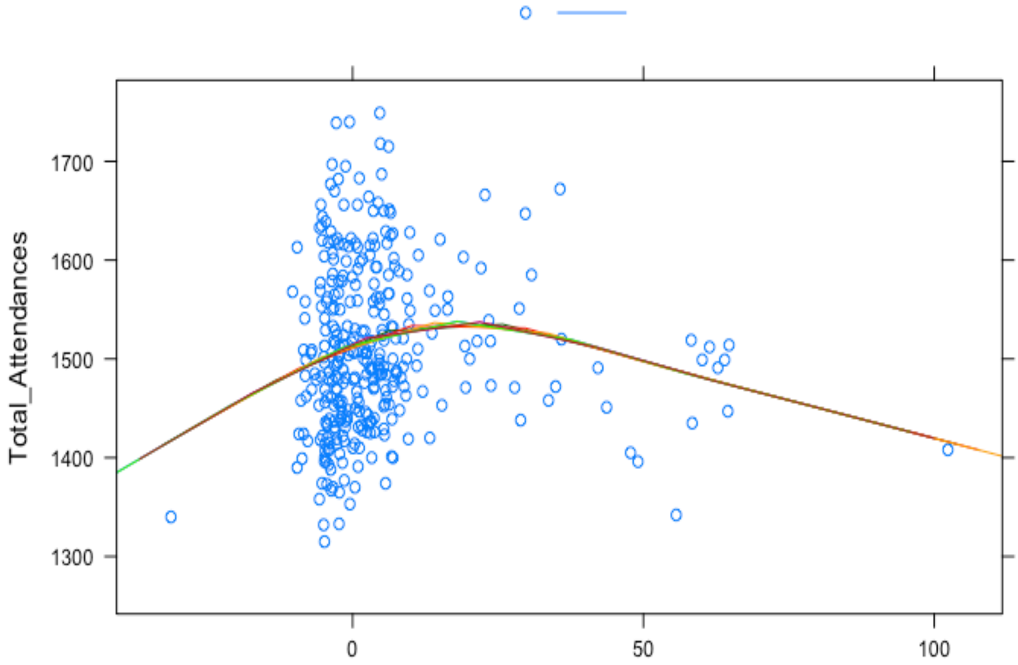
$$\mathbf{EHF = EHIsig \times max\ (1,\ EHIaccl)}$$

2. Firstly, let's observe the effect of EHF across the given time period. We'll discover that during the relevant time period for heat wave (summer), we realise a massive increase in excess heat factor as seen in the plot below.



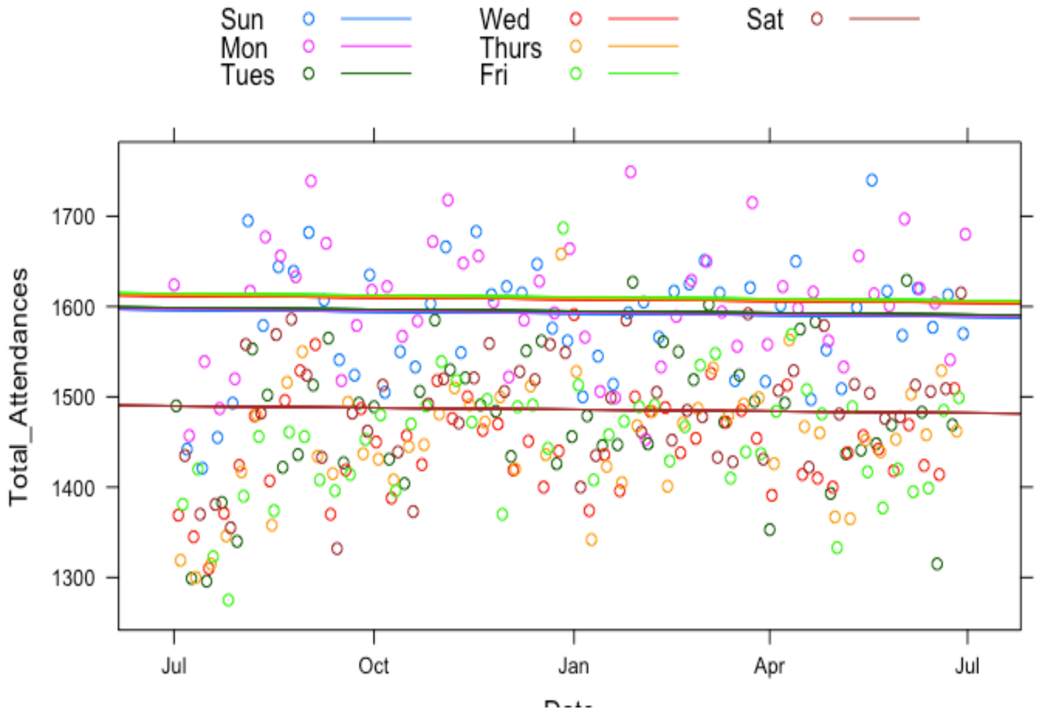
As shown in the plot above, we'll see a drastic rise in the EHF between October and late January/February. This makes sense as this is the Australian summer period.

Plotting Total Attendance against EHF we also discover some relationships as seen in the plot below:



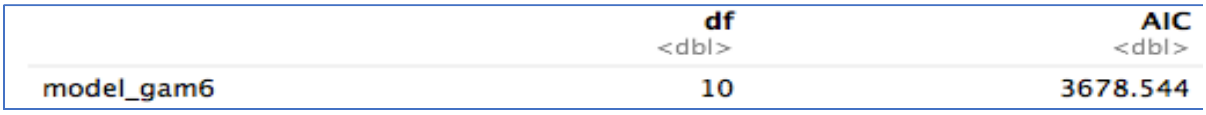
Task 3.2

Using EHF as an additional predictor to augment the previous models, we have the following plot:



Checking whether the extra predictor improves the model by using the AIC coefficient values as a means of comparison. Evidently, it's seen that the AIC coefficient has drastically reduced with the inclusion of EHF as an additional predictor as opposed to the AIC value without EHF. This is seen in the table below:

After EHF inclusion



	df	AIC
	<dbl>	<dbl>
model_gam6	10	3678.544

Task 4

1. The limitations of historical data asper the data used in this analysis is that no detailed information was given about how the data were transformed during collection. After deeper analysis were done on the data collection, it was realised that the temperature values were multiplied by 10. Also, there could be some bias during the collection of the data or inaccurate input. Overall, some insights were generated about the relationships between heatwaves, date and total attendance of patients across all hospitals.

2. As most of the variables are numeric, Regression models are the best when it comes to handling numeric variables. Regression models try to find the best model fit between a target non-categorical variable and a predictor(s) non-categorical variable. So, choosing a different predictive method will adversely affect model accuracy and in turn affect the final outcome and insight generated.

3. So far, the regression model has done a good job through the use of GAM (generalised additive model) in helping us gain more insight. As more variables were included in the model, improved. First was the inclusion of weekdays and finally the inclusion of EHF (Excess Heat Factor).

References

1. Nairn, J.R. and Fawcett, R.J., 2014. The excess heat factor: a metric for heatwave intensity and its use in classifying heatwave severity. *International journal of environmental research and public health*, *12*(1), pp.227-253.
2. Peter Hannam, K. (2017). *'Going to be a big day': Heatwave to strain hospitals, power supplies in NSW*. [online] The Sydney Morning Herald. Available at: http://www.smh.com.au/environment/going-to-be-a-big-day-heatwave-to-strain-hospitals-power-supplies-in-nsw-20170209-gu99le.html