**Swan district hospital**

Task 1

1. I chose the NOAA data source for my assignment as there was an R package called ‘rnoaa’ for which I could pull the data source directly from and into R. The government BOM website could only be accessed by website and only had 4 options of variables. Rnoaa had multitudes more variables for me to use. The readme text was easier to read so i can extract the variables i wanted with ease.I could also skip the csv import into R phase as well. THe weather station from NOAA is near my hospital as well in perth
2. I requested a token and downloaded precipitation and daily temperature
3. There were 365 rows in both datasets
4. The time period is from “2013-07-01” to “2014-06-30”

Task 2

1. The final model will try and use variables(IV) to get the best fit for ED demands(DV) as it was suggested in our last assignment that attendance has an effect on ED demands. We want to use the model to find meaningful relationships between weather variables and ED demands. Overcrowding is an increase in attendance at EDs, It is relevant because we can explore what affects attendance  
    If the model shows meaningful relationship between environmental variables and ED demands the model can assist researchers in ED demands as our assignment is not complex or in depth enough to be used by hospitals
2. I want to model the relationship between multiple variables with attendance which indicates ED demand and overcrowding.   
   The response variable is attendance which indicates ED demand

The predictor variables in the assignment are date, weekday, EHF, and extra research variables at the end.  
The variables in the model are collected daily and store at NOAA and can be accessed using API for prediction

Most of the variables such as date and weather are cyclical in nature and will share similar characteristics to data in the future. Though climate change may change that.

1. Two statistical methods used in this assignment will be fitting a linear model and generalised additive model. Fitting a basic linear model is simple to find linear relationships between variables. Often the most simple method is the best. A generalised additive model is used when the relationship is non-linear and it has smooth functions that smooth out different sections on the model.

Task 3

1. I picked the Swan district hospital

| Linear model | lm(formula = Attendance ~ Date, data = one\_hospital) |
| --- | --- |
|  | |
|  |  |
|  |  |
| Residual standard error: 12.91 on 363 degrees of freedom  Multiple R-squared: 0.05954, Adjusted R-squared: 0.05695  F-statistic: 22.98 on 1 and 363 DF, p-value: 2.392e-06 | AIC score 2907.146 |
| No pattern to residuals compared with fitted values  QQ plot suggest slightly Fat tails with data peaked in middle  Residual chart suggest high amount of uncorrelated outliers  Low adjusted R squared value of 0.05 with a AIC score of 2907  Poor model needs more variables  Insufficient model fit | |



| Gam 1 | Attendance = s(date) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.0832 Deviance explained = 9.43%  UBRE = 0.26609 Scale est. = 1 n = 365 | AIC score 2907.331 |
| Normal distribution of residuals  Slightly fat tails middle data peak  Residuals clump on right side of fitted and predicted values  R-sq.(adj) = 0.0832 and aic score of 2907 slightly better than linear model but not by much  Insufficient model fit | |

| weekdays | gam(Attendance = s(date) + weekdays) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.299 Deviance explained = 31.6%  UBRE = -0.0023093 Scale est. = 1 n = 365 | AIC score 2809.367 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.0832 and aic score of 2809 better than both previous models  Relatively decent model fit | |

| MODEL | AIC SCORE |
| --- | --- |
| LM | 2907.146 |
| GAM(date) | 2907.331 |
| GAM(date+weekday) | 2809.367 |
| Best fitting model using gam model with date and weekday as predictors | |

|  | histogram | Residuals vs fitted | QQ plot |
| --- | --- | --- | --- |
| LM |  |  |  |
| GAM(date |  |  |  |
| GAM2(date+weekday) |  |  |  |
| LM and Gam2 residuals seem to be spread. Gam(date) residuals clump up around certain predictor values. All Q-Q plots seem to indicate most residuals sit around 0 to 1. However there high number residual spread indicating poor fitting for many values. | | | |

1. Day of the week is a categorical data type and is used as a parametric term as it cannot be used as a smooth function. No penalties are applied to these parametric terms compared to smooth functions. This causes the fitted lines to be less smooth compared to if the variable was included as a smooth function.

Task 4.1

EHIsig = (t1 +t2+t3)/3 - T95

EHIaccl = (t1 +t2+t3)/3-(t1 +t2+t-30)/30

EHF = EHIsig \* max(1,EHIaccl)

Task 4.2

|  |
| --- |
| EHF values seem to increase during the start of the year which coincides with summer, EHF decreases during the colder seasons |

|  |  |
| --- | --- |
|  |  |
|  | |
| Approximate significance of smooth terms:  edf Ref.df Chi.sq p-value  s(EHF) 1.000 1.000 0.044 0.833  s(as.numeric(Date)) 4.292 5.317 37.800 2.93e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  R-sq.(adj) = 0.298 Deviance explained = 31.6%  UBRE = 0.0028829 Scale est. = 1 n = 365 | AIC SCORE 2811.262 |
| K may have been too restrictive looking at the plot  Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.0832 and aic score of 2811 worse than previous models  Relatively worse model fit  the extra predictor did not improve the model fit  EHF is not a good predictor of attendance which indicates ED demand  That doesn't not mean weather feature don’t have relationship with ED demand  Heatwaves which EHF measures | |
|

Task. 4.3

Extra weather features that may be more predictive of ED demands:

3 day rolling temperature average - temperature average may cause more accidents increasing attendance count affecting ED demand

Rainfall - high temperatures may cause more accidents increasing attendance count affecting ED demand

Temperature range - large ranges in temperature may cause more accidents increasing attendance count affecting ED demand

Maximum temperature- high temperatures may cause more accidents increasing attendance count affecting ED demand

Minimum Temperature- low temperatures may cause more accidents increasing attendance count affecting ED demand

| Rainfall | Gam (date weekday EHF precipitation ) |
| --- | --- |
|  |  |
|  |  |
|  |  |
| R-sq.(adj) = 0.297 Deviance explained = 31.7%  UBRE = 0.0068745 Scale est. = 1 n = 365 | 2812.719 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.297, aic score of 2812.719 worse than previous models  Relatively not decent model fit | |

| max temperature | Gam (date weekday EHF max temperature) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.311 Deviance explained = 33.7%  UBRE = -0.0038354 Scale est. = 1 n = 365 | 2808.81 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.311, aic score of 2808.81 better than previous models  Relatively decent model fit | |

| Minimum temperature | Gam (date weekday EHF min temperature) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.299 Deviance explained = 31.9%  UBRE = 0.0043793 Scale est. = 1 n = 365 | 2811.808 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.299 and aic score of 2811 worse than previous models  Relatively not decent model fit | |

| Temperature range | Gam (date weekday EHF range temperature) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.297 Deviance explained = 31.7%  UBRE = 0.0074978 Scale est. = 1 n = 365 | 2812.947 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.297 and aic score of 2812.947 worse than previous models  Relatively not decent model fit | |

| 3 day average temperature | Gam (date weekday EHF 3 day temperature) |
| --- | --- |
|  |  |
|  |  |
| R-sq.(adj) = 0.311 Deviance explained = 33.6%  UBRE = -0.0023412 Scale est. = 1 n = 365 | 2809.355 |
| Residuals show Normal distribution of residuals with high peak in middle  QQ plot shows normally distributed data  Residuals spread out, no correlations or pattern between residuals  R-sq.(adj) = 0.311 and aic score of 2809 better than both previous models  Relatively decent model fit | |

Task 5

There are many limitations of using historical data for my analysis.

In this assignment many limitations include missing data. I was unable to retrieve snowfall data from NOAA for 2013-2014. There are also missing years from 1971 and 1972 from the dataset. From 1971 to 2000 many daily average temperatures are missing from the dataset.

Using historical weather data for future forecasts assumes everything will be constant. Even though weather features are cyclical in nature, with climate change on the rise, past weather trends may not apply in the future. There are also biases in collecting the weather data. Even though I trust the NOAA dataset there may be different operators or sensory devices which could alter readings.

The objective assignment is mainly used to understand the process of ED demands. This assignment does not aim to predict what the dependent variable will be, but to find independent variables that have some sort of relationship such as EHF, precipitation and date to ED demand. We are exploring different variables and seeing if they affect the ED demand. The main objective of this assignment is also to both define a family of the model and fit model as accurately as possible. No section is dedicated to predicting ED demands.

Making predictions would require more data research into how the independent variables affect ED demand. We would have to split our data into 3 sections, 60% into training, 20% into query and 20% for testing. We explored the variables multiple times and as soon as you use the observation twice our objective is no longer for confirmation but exploration. If we were to predict we would call functions such as gam.predict

My analyses have so far increased my understanding of different variables that affect attendance which indicates ED demand. In my limited analysis, weekdays have an effect on attendance. I also found that EHF is a poor indicator of attendance demand. However, strangely Max temperature and 3 day average has a relationship with attendance but Min temperature did not. Suggesting high temperatures affected attendance. Rainfall did not have a strong relationship with attendance and decreased model fit. Overall my analyses have pointed me towards the right direction to exploring the relationship between dates, weather features and ED demands. However further in depth exploration analysis as well as confirmation studies is required to understand ED demands.