Generalized Addictive Models

Alberto Leiro
Thao Nhan

July 2022

1 Introduction

When performing any kind of predictive analysis, it is important to consider the type of model we use in order to best display or interpret our regression results. Linear regression is an extremely useful tool to perform such kinds of predictive analysis. However, in real world problems and data sets, there comes a point where performing linear regression is not sufficient to model our data. There are times when a linear regression model does not properly provide the right fit for the data points. This is mainly due to the data itself being nonlinear. So how can we properly perform a regression model on a nonlinear relationship? One option is to use a generalized additive model(GAM).

The purpose of this report is to explain how GAMs work in data science by giving its definition and methodology. This report also shows two examples on how to perform a GAM analysis using R programming code. We will be using packages in R specifically tailored towards performing GAM analysis. R code and graphs of the GAMs will be provided in the examples along with their interpretations and processes. Finally, we will be stating the limitations of using GAMs compared to other types of regression models.

The final goal is to convince more people to use GAM when possible since the technique posts three advantage. Generalized Additive Models are easy to interpret. It contains flexible predictor functions that can uncover hidden patterns in data; and also has regularization of predictor functions that helps avoid overfitting [Lar15].

1.1 What are GAMs?

GAMs were invented by Trevor Hastie and Robert Tibshirani in 1986.[HT86] GAMs are used when instead of a linear response variable depending on a linear function, instead it depends on smooth functions called splines, which are polynomial functions that cover a small range[Sha21]. There are advantages to using a GAM rather than other regression models.For example, GAMs can use multiple different link functions and could cover incredible amounts of predictor variables[Mol22]. GAMs are also very easily analyzed through programming languages such as R. The main aspect of GAMs is seeing how different covariates or predictor variables can be inferred with a more fluid function that would allow for nonlinear associations[Woo17]. This is important since it allows us to deduce predictor variables that are significant when compared to the response variable or even how predictor variables relate to each other, if needed.

1.2 Applications of GAMs in previous studies

One of the many advantages of GAMs is how it can be applied to real world phenomena. These types of models have been found to be able to cover large data sets with multiple predictor variables[WGS15]. Not only that, GAMs have been widely applied to studies in the health, environmental, data science, and statistical fields. A case study done in Australia used GAMs to make water quality assessments[RHGT13]. In another field, GAMs were used to study and analyze time series of air pollution and health[Dom02]. Additionally, in 2017 scholar

Xi Gong, Aaron Kaulfus, Udaysankar Nair, and Daniel A. Jaffe used Generalize Additive Models to analyze the Quantifying O_3 impacts in urban area due to wildfires. GAMs were able to give helpful information about the daily average O_3 in this case study. The results of the GAMs method was directly applicable to the EPA guidance on excluding data due to an uncontrollable source [GKNJ17].

Generalized Additive Models have shown its practicality and flexibility in the previous studies. In the next section, Methods, we will discuss in detail how Generalized Additive Models (GAM) works and its advantages by comparing it to its brother, the Generalized Linear Models (GLMs).

2 Methods

2.1 How do GAMs work?

2.1.1 GLMs

GAMs are an extension of Generalized Linear Models(GLMs), which in turn are extensions of the linear regression model [Sha21]. GLMs are shown as[Sha21]:

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{1}$$

The GLM consists of the following parts: g is the link function, $x_1, x_2, ..., x_n$ are the predictor variables, $\beta_1, \beta_2, ..., \beta_n$ are the weight of each function, and $E(Y)$ is the probability distribution from the exponential family that defines it[Sha21]. GLMs are categorized as a weighted sum of linear combinations of variables. Each variable is paired with the weight represented by β, which in turn is added together[Mol22].

2.1.2 GAMs

GAMs are very similar to the GLM but with one major difference. Instead of the variable being paired with a weight β, it is paired with a spline(smooth nonparametric functions) s, which could be linear if needed[Sha21]. GAMs do not assume the relationship to be linear and instead of trying to create a linear model or find a linear relationship between variables, GAMs use a non-linear combination of variables. We can write the GAM structure as

$$g(E(Y)) = \beta_0 + s_0 + s_1 x_1 + s_2 x_2 + ... + s_n x_n \tag{2}$$

where Y is the dependent variable (i.e., what we are trying to predict), $E(Y)$ denotes the expected value, and $g(Y)$ denotes the link function that links the expected value to the predictor variables x_1,. . . ,x_p. The term x_1,. . . ,x_p denote smooth, non-parametric functions [Lar15]. Keep in mind when using non-parametric functions, the shape of the predictor functions are fully determined by the data. This means one does not know what the patterns look like until some tests are run. This is the opposite with parametric functions where the predictor functions are defined by a typical small set of parameters.

The spline function equation s can be determined by:

$$s(x) = \sum_{k=1}^{k} \beta_k b_k(x) \tag{3}$$

Where k is the weight and $b_k(x)$ is a function [Sha21]. What must be noted is that we can have as many k weights and functions for each variable as required, allowing us to have more flexibility and smooth regression when compared to regular linear regression [Mol22]. This is one of the biggest strengths in using GAMs.

2.2 R and R packages

The examples shown in this report will be analyzed using R studio and utilizing the R programming language. R is free software useful for performing statistical analysis, computing and graphing [R C21]. R Studio is also a free software that creates an integrated environment to perform R programming functions in an organized manner [RSt20]. Two R packages were used in order to perform the analysis in our examples. Data sets used as examples that can properly showcase and utilize GAMs were retrieved from the *gamair* package [Woo06]. The GAM functions and graphs created will be derived from the *mgcv* package [Woo17].

```
library(gamair)
library(mgcv)
```

3 Illustrative examples with R

3.1 Example 1: Atmospheric Carbon Dioxide at the South Pole

```
data(co2s)
co2 = co2s$co2
time = co2s$c.month
month = co2s$month
```

The first example will be using the data set *co2s* which contains data on the amount of Carbon Dioxide concentrations per million on the South Pole over months[Woo06]. The following are the variables that are being used:

- **co2**: atmospheric CO2 concentration in parts per million
- **time**: cumulative number of months since January 1957
- **month**: month of the year

We will be using *co2* as our dependent variable and the others as independent predictor variables. In order to showcase the benefits of using GAMs instead of a regular linear regression model, I will be showing both and discussing the differences.

3.1.1 Showing Linear Analysis

```
#Regular Linear regression model
m1 = lm(co2 ~ time + month)
summary(m1)
plot(m1)
```

The following are the results from performing a regular linear regression analysis along with the plots:

```
## Call:
## lm(formula = co2 ~ time + month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9125 -1.3509 -0.2237  0.9640  5.1877
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.071e+02  2.552e-01 1203.44  < 2e-16 ***
## time         1.074e-01  6.268e-04  171.36  < 2e-16 ***
## month        8.594e-02  2.498e-02    3.44  0.00064 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.746 on 424 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9857
## F-statistic: 1.469e+04 on 2 and 424 DF,  p-value: < 2.2e-16
```

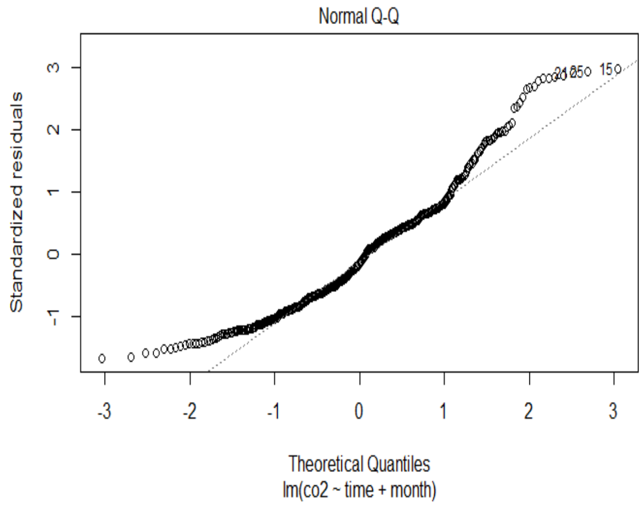

Figure 1: QQ plot-Co2

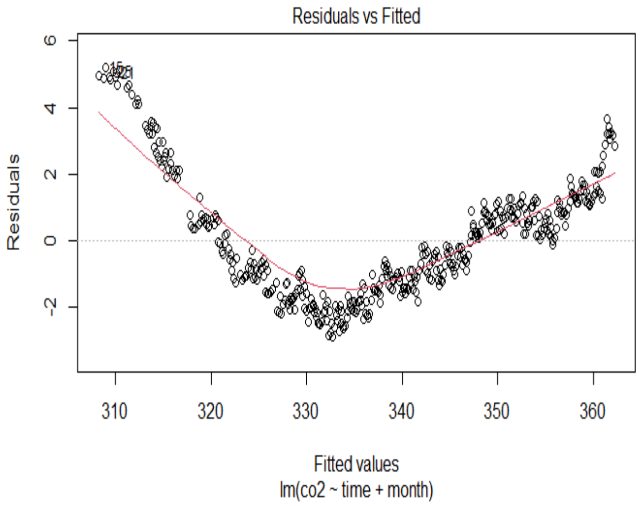

Figure 2: Residuals vs Fitted

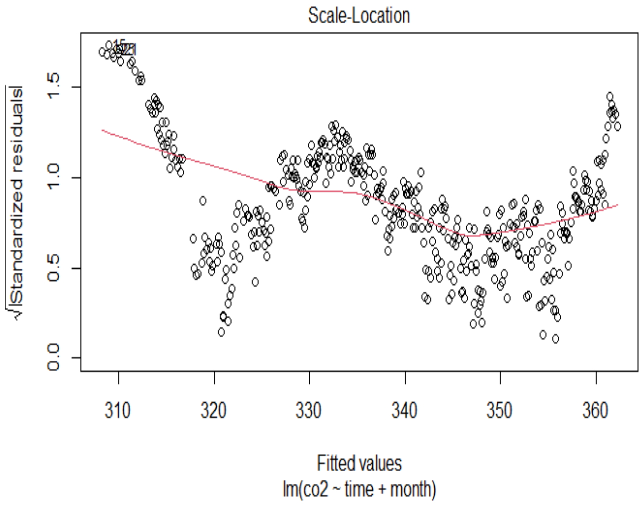

Figure 3: Scale-Location

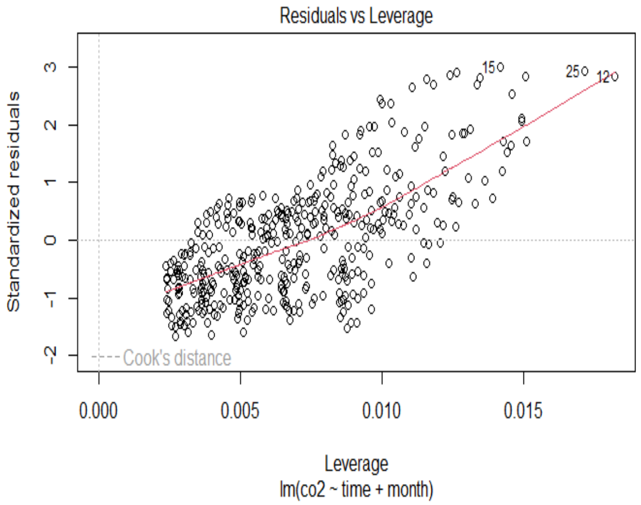

Figure 4: Residuals vs Leverage

The linear regression model shows that both the overall time and the months of the year are significant predictors of CO2 levels at the south pole. This makes sense since the predictor variables can be defined as linear functions.

3.1.2 GAM Analysis

Now the same analysis will be performed using GAMs:

```
#GAMs model
m2 = gam(co2 ~ s(time) + s(month), data = co2s)
summary(m2)
plot(m2)
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## co2 ~ s(time) + s(month)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 338.24515    0.01308   25869   <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(time)  8.980  9.000 137550.3  <2e-16 ***
## s(month) 5.921  7.084    120.5  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 0.075826  Scale est. = 0.073002  n = 427
```

The results are interesting since not only does it show similar results to the regular linear regression, stating that the two independent variables are also significant, but it shows a higher r-squared result of 1 with 100% of the deviance explained!

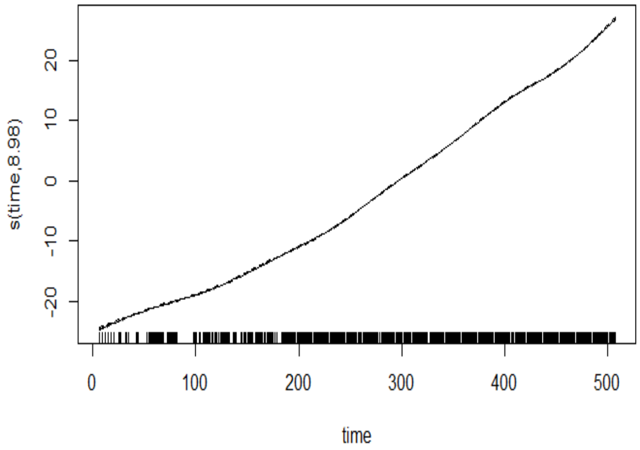

Figure 5: time vs CO2

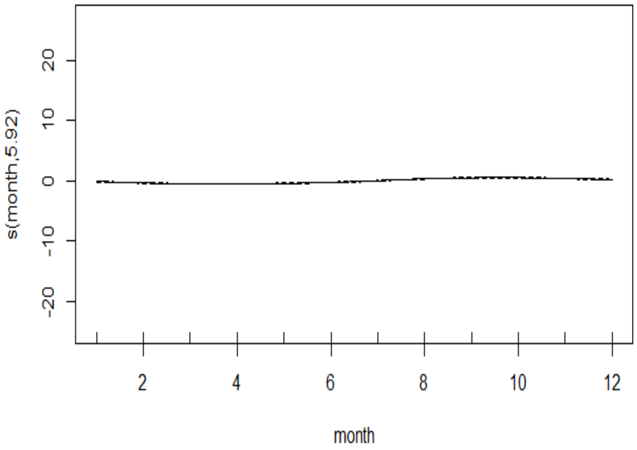

Figure 6: month vs CO2

When looking specifically at the GAM plots, the graphs show that the amount of CO2 levels increased over time since 1957. When it comes to the months of the year, there seems to be a small decrease in the Spring months while fall to winter months show a slight increase in CO2 levels.

3.2 Example 2: Chicago Air Pollution

```
data(chicago)
deaths = chicago$death
pm10 = chicago$pm10median
pm25 = chicago$pm25median
o3 = chicago$o3median
so2 = chicago$so2median
time = chicago$time
temp = chicago$tmpd
```

The second example that we will be using is the data set *chicago* which contains data on the daily air pollution and death rates for *chicago*[Woo06]. The following are the variables that will be used:

- *deaths*: total amount of deaths per day
- *pm10*: median particles in 2.5-10 mg per cubic meter
- *pm25*: median particles less than 2.5 mg per cubic meter (considered more dangerous)
- *o3*: ozone in parts per billion
- *so2*: Median Sulfur dioxide measurement
- *time*: time in days
- *temp*: temperature in Fahrenheit

We will be using *deaths* as the dependent variable and the others as predictor variables. Much like the first example, I am going to show a Linear regression Analysis in order to showcase the difference between it and the GAM.

3.2.1 Showing Linear Regression Analysis

```r
#Regular Linear regression model
m1 = lm(deaths ~ pm10 + pm25 + o3 + so2 + time + temp)
summary(m1)
plot(m1)
```

The following are the results from performing a regular linear regression analysis along with the plots:

```
## Call:
## lm(formula = deaths ~ pm10 + pm25 + o3 + so2 + time + temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.726  -7.844  -0.203   7.319  48.149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.365785   3.564862  37.692  < 2e-16 ***
## pm10          0.204558   0.034279   5.968  3.8e-09 ***
## pm25         -0.040480   0.063989  -0.633    0.527
## o3            0.059210   0.054956   1.077    0.282
## so2           0.065087   0.191244   0.340    0.734
## time         -0.000158   0.001360  -0.116    0.908
## temp         -0.455797   0.031930 -14.275  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.91 on 712 degrees of freedom
##   (4395 observations deleted due to missingness)
## Multiple R-squared:  0.2928, Adjusted R-squared:  0.2869
## F-statistic: 49.13 on 6 and 712 DF,  p-value: < 2.2e-16
```

As you can see, the linear regression does not show much in terms of results. The r-squared value is low and shows that the results are not entirely reliable. It also shows a majority of our predictor variables as insignificant. The only significant values are temperature and pm10.

Next, we will plot the residual results from the Linear Regression Model. The following are the residual plots from the models:

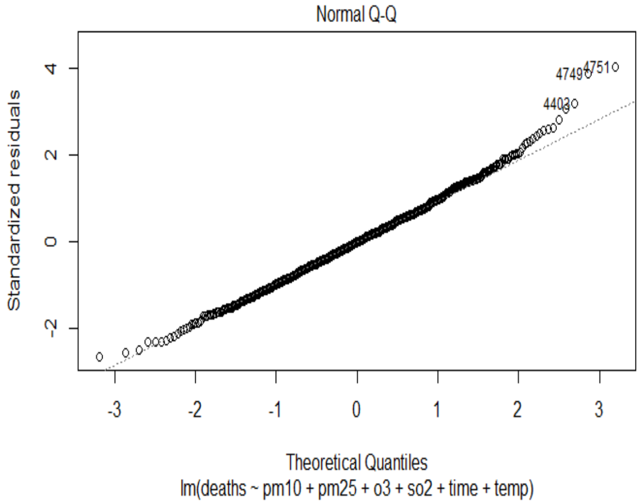

Figure 7: QQ plot-Chicago

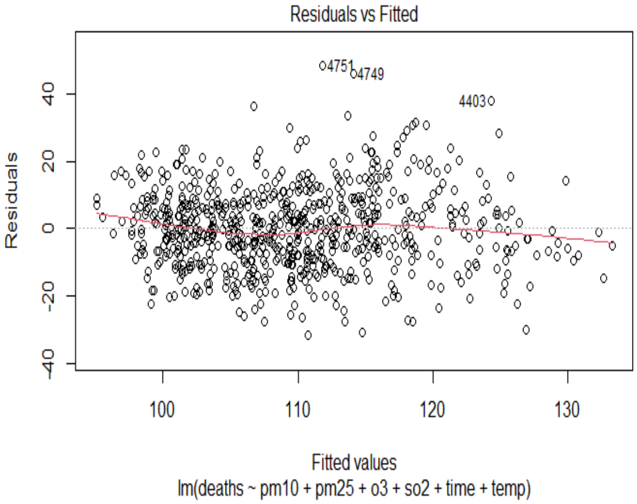

Figure 8: Residuals vs Fitted

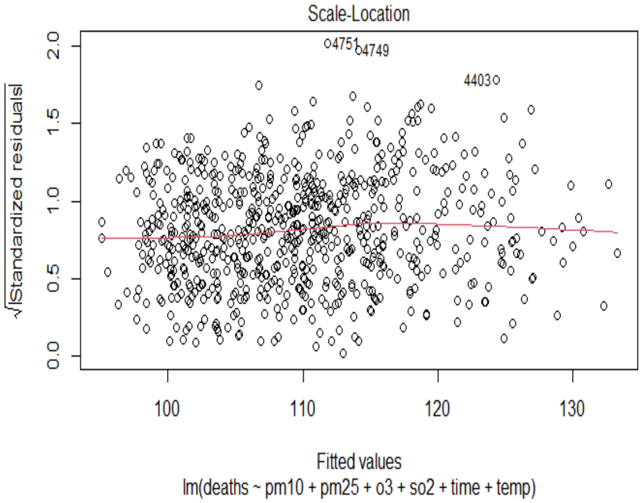

Figure 9: Scale-Location

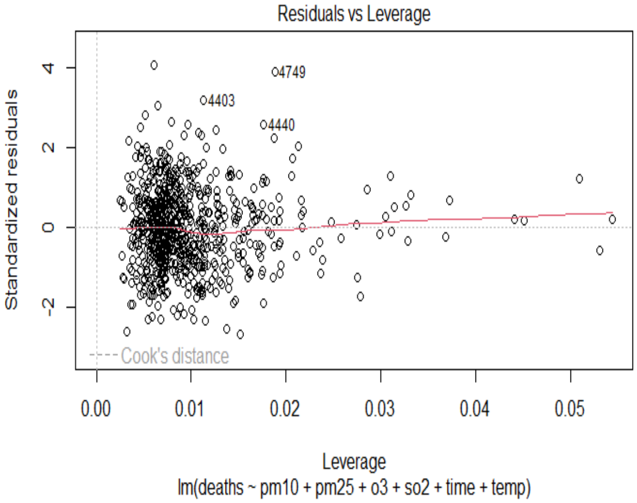

Figure 10: Residuals vs Leverage

The linear regression model shows a flat line which is not helpful in our analysis and does not tell us much about our predictor variables and their relationship with the dependent variable.

3.2.2 GAM Analysis

Now the same analysis will be performed using GAMs.

```
#GAMs model
m2 = gam(deaths ~ s(pm10) + s(pm25) + s(o3) + s(so2) + s(time) + s(temp), data = chicago)
summary(m2)
plot(m2)
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## deaths ~ s(pm10) + s(pm25) + s(o3) + s(so2) + s(time) + s(temp)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.6787     0.4165   263.3   <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(pm10) 1.758  2.245 12.541 2.52e-06 ***
## s(pm25) 1.000  1.000  6.033   0.0143 *
## s(o3)   5.611  6.798  1.127   0.3466
## s(so2)  1.000  1.000  0.045   0.8322
## s(time) 8.823  8.989  6.760  < 2e-16 ***
## s(temp) 4.248  5.279  6.195 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.373   Deviance explained = 39.3%
## GCV = 128.92  Scale est. = 124.72    n = 719
```

When compared to the linear regression model, GAMs has a higher r-squared value(though still a bit low) and provides more significant results for the predictor variables. It states that pm10, pm25, time, and temperature provide significant results.

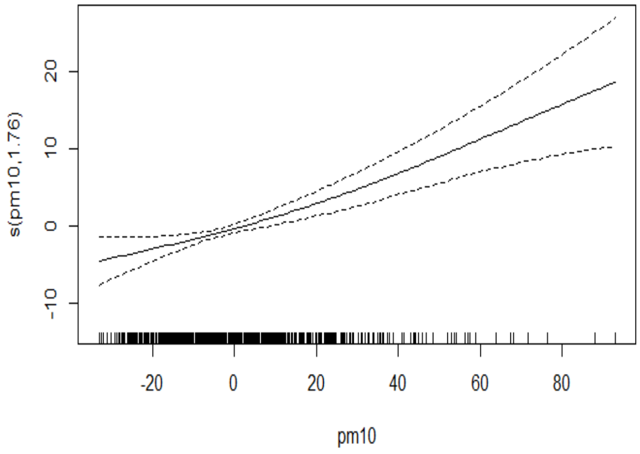

Figure 11: pm 10 vs deaths

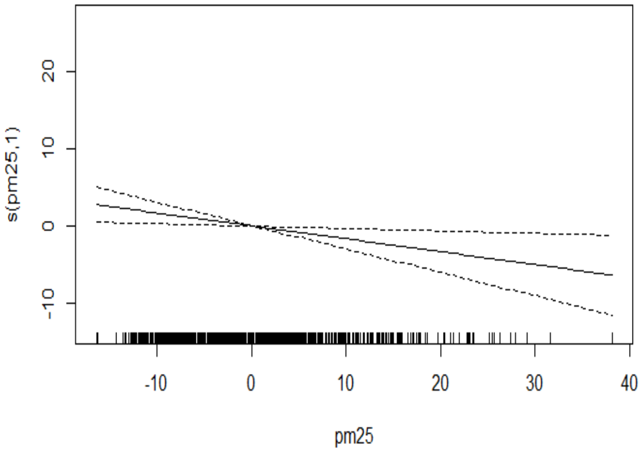

Figure 12: pm25 vs deaths

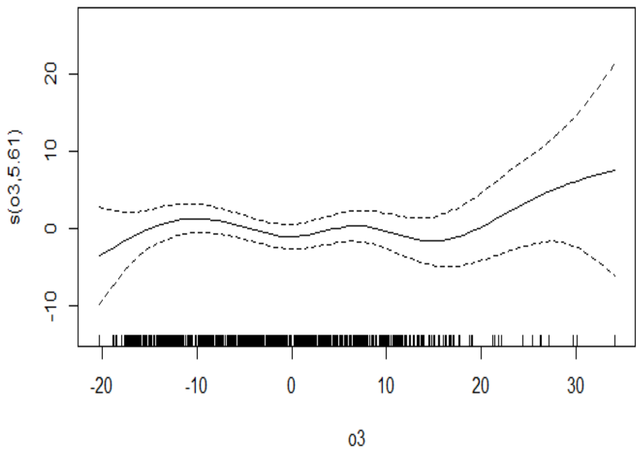

Figure 13: Ozone vs deaths

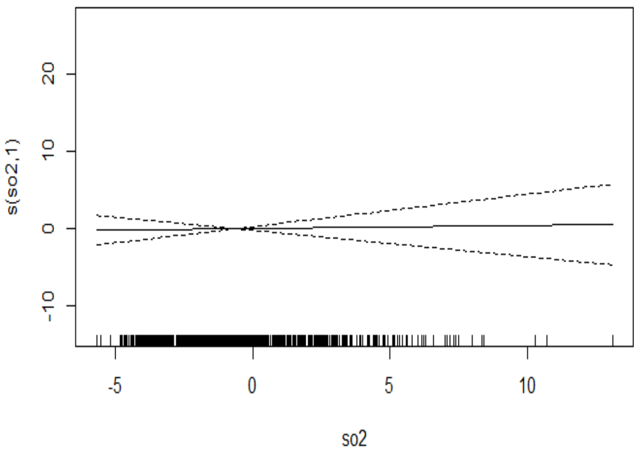

Figure 14: Sulphur Dioxide vs deaths

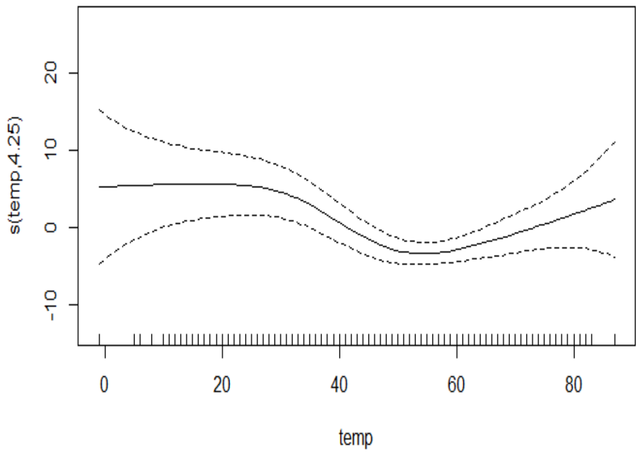

Figure 15: Temperature vs deaths

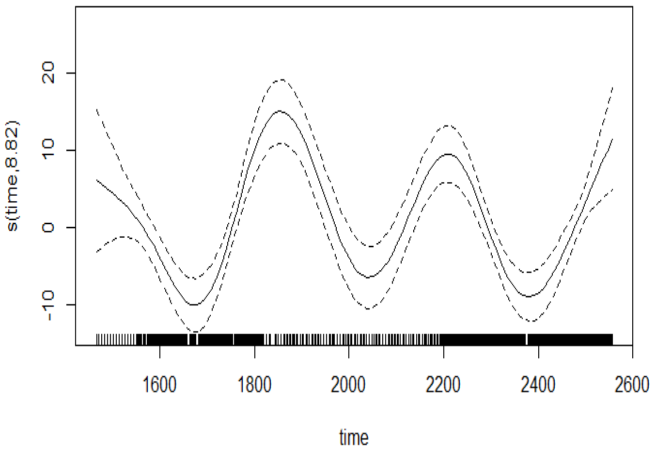

Figure 16: Deaths over time

GAMs also provides insight in the relationship between each predictor variable and the amount of deaths per day in Chicago. As the median particles between 2.5 and 10 mg per cubic meter increased, the daily deaths in Chicago also increased. As the median particles less that 2.5 mg per cubic meter increased, the daily amount of deaths decreased over time. Note that so2 and o3 were not considered significant in the summary results. When looking at their graphs, as the amount of Sulphur Dioxide increased, the amount of deaths did not increase by much and the regression line is practically flat, and the majority of the ozone graph does not change by much until it starts increasing at the end. The temperature graph shows not much movement when the temperatures are low, but a majority of the graph shows a sudden dip in deaths once it gets between 20 and 80 degrees Fahrenheit. Finally, the deaths over time fluctuate greatly over time, moving up and down as the days go on. These kinds of observations are possible to analyze using GAMs for the data set.

4 Limitations

Whilst GAMs are useful in drawing conclusions about our data, there are drawbacks that could make other regression models more viable. GAMs depend on assumptions on the data generating process[Mol22]. If the data assumptions are not met, any review of the data's weight could not be deemed valid[Mol22].Additionally, since GAMs are all about dealing with complex relationships by using smooth functions to fit each part of the data, sometimes it would result in over-fitting. To avoid over-fitting when using GAMs in analysis, it is best to practice comparing results between GLMs and GAMs. Finally, if there are any link functions identified in the GLM apart from the identity function, then it leads to complications with the GAM[Mol22].

5 Conclusion

GAMs are a very useful tool used in statistics. The examples have shown that it can accommodate few or multiple predictor variables and swiftly provide a visual representation of the relationship between the independent predictor variables and the dependent variable. The benefit of the flexibility that GAMs allows for can be applied to multiple fields and can help us in understanding how predictor variables interact in our regression analysis. As GAMs are developed and built upon, the advantages will help us get a better understanding of how such studies mentioned in this paper can be useful for the future of data science. Consider putting GAMs in your toolbox so it can enhance our analysis ability on any type of data which will result in a better understanding of the data and its variables.

References

[Dom02] F. Dominici. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, 156(3):193–203, 2002.

[GKNJ17] Xi Gong, Aaron Kaulfus, Udaysankar Nair, and Daniel A Jaffe. Quantifying o3 impacts in urban areas due to wildfires using a generalized additive model. *Environmental science & technology*, 51(22):13216–13223, 2017.

[HT86] T. Hastie and R. Tibshirani. Generalized additive models. *Generalized Additive Models*, 1(3):297–318, 1986.

[Lar15] Kim Larsen. Gam: the predictive modeling silver bullet. *Multithreaded. Stitch Fix*, 30:1–27, 2015.

[Mol22] Christoph Molnar. *Interpretable machine learning: A guide for making Black Box models explainable.* Christoph Molnar, 2022.

[R C21] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

[RHGT13] Russell Richards, Lawrence Hughes, Daniel Gee, and Rodger Tomlinson. Using generalized additive models for water quality assessments: A case study example from australia. *Journal of Coastal Research*, 65:111–116, 2013.

[RSt20] RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, PBC., Boston, MA, 2020.

[Sha21] Adam Shafi. What is a generalized additive model? *Towards Data Science*, May 2021.

[WGS15] Simon N. Wood, Yannig Goude, and Simon Shaw. Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155, 2015.

[Woo06] S. N. Wood. *Generalized Additive Models: An Introduction with R.* Chapman Hall/CRC, Boca Raton, Florida, first edition, 2006. ISBN 1-58488-474-6.

[Woo17] S.N Wood. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, 2 edition, 2017.