Generalised additive models

## Exercise with solutions

Full resource, see: <https://www.ncrm.ac.uk/resources/online/all/?id=20851>

## Exercise:

The airquality dataset, taken from the datasets R package, contains daily measurements of air quality (Ozone) and weather in New York between May and September 1973. To access this data, install and load the package using the following code within R:

install.packages(“datasets”) # install package (if first time)

library(datasets) # attach package to current session

str(airquality) # view structure of the dataset

Use an appropriate statistical model to investigate whether air quality was related to daily temperature. Perform model checks to ensure the model is valid.

**Note:** The variable Ozone measures the mean ozone parts in the air. The higher this measure is, the worse the air quality.

## Solution:

### Preparing data

To load the airquality dataset, ensure that the datasets package is installed and accessible on your current session of R. We will also require the mgcv package (to fit GAMs), tidyverse (to tidy and plot data), and the marginaleffects package (to extract model results):

# install.packages(c("tidyverse", "mgcv", "marginaleffects", "datasets"))  
  
library(tidyverse)  
library(mgcv)  
library(marginaleffects)  
library(datasets)

Before fitting any models or generating any plots, we must check that the data has been read in correctly and the variables have been correctly specified:

glimpse(airquality)

Rows: 153  
Columns: 6  
$ Ozone <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, 18, 14, …  
$ Solar.R <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256, 290, 27…  
$ Wind <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8.6, 6.9, 9…  
$ Temp <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 68, 58, 64…  
$ Month <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,…  
$ Day <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…

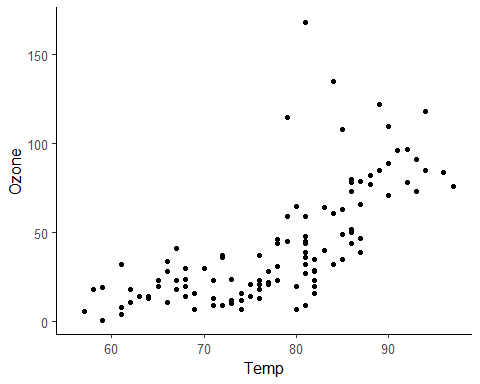
There are 153 observations across 6 variables. All variables are correctly specified as numeric.

### Visualising data

Before fitting any models, it is good practice to visualise the data and inspect the relationships between variables. The aim of our models will be to investigate the relationship between Ozone and Temp so we could visualise this using a scatterplot:

ggplot(data = airquality,  
 mapping = aes(y = Ozone,  
 x = Temp)) +  
 geom\_point() +  
 theme\_classic(base\_size = 12)

Warning: Removed 37 rows containing missing values or values outside the scale range  
(`geom\_point()`).

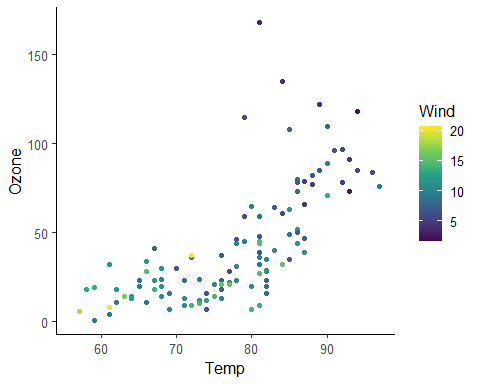


The scatterplot shows a clear positive relationship between temperature and ozone parts, suggesting that air quality is worse in warmer temperatures. We can also clearly see that this relationship is **nonlinear**, suggesting a GAM would be an appropriate model.

There are some outlying values that do not seem to follow the same patterns as other observations. We may wish to include other variables in this scatterplot to gain insight into whether other variables should be included in this model:

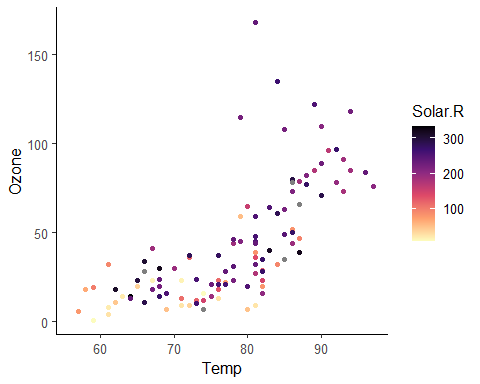
ggplot(data = airquality,  
 mapping = aes(y = Ozone,  
 x = Temp,  
 colour = Wind)) +  
 geom\_point() +  
 scale\_colour\_viridis\_c() +  
 theme\_classic(base\_size = 12)

Warning: Removed 37 rows containing missing values or values outside the scale range  
(`geom\_point()`).



ggplot(data = airquality,  
 mapping = aes(y = Ozone,  
 x = Temp,  
 colour = Solar.R)) +  
 geom\_point() +  
 scale\_colour\_viridis\_c(option = "A", direction = -1) +  
 theme\_classic(base\_size = 12)

Warning: Removed 37 rows containing missing values or values outside the scale range  
(`geom\_point()`).



There appears to be a relationship between wind and air quality: days with very little wind appeared to have worse air quality, regardless of temperature. Therefore, we may consider including wind speed as a covariate into our model.

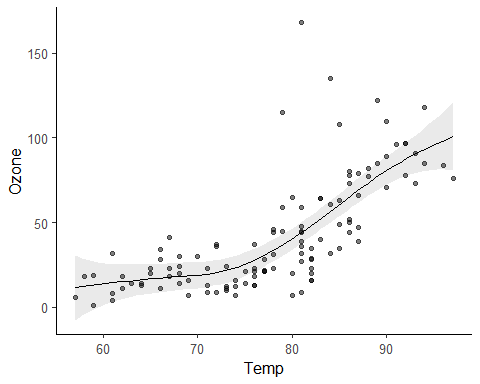
## Fitting a GAM

The first GAM we may consider fitting is one with air quality as the outcome and a smooth function applied to temperature to capture this nonlinear relationship:

gam\_temp <- gam(Ozone ~ s(Temp),  
 data = airquality,  
 method = "REML")  
  
summary(gam\_temp)

Family: gaussian   
Link function: identity   
  
Formula:  
Ozone ~ s(Temp)  
  
Parametric coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 42.129 2.049 20.57 <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(Temp) 3.365 4.201 33.97 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-sq.(adj) = 0.553 Deviance explained = 56.6%  
-REML = 521.6 Scale est. = 486.83 n = 116

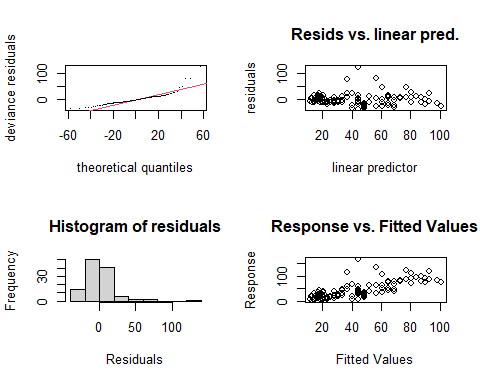
plot\_predictions(gam\_temp,  
 condition = "Temp",  
 points = .5) +   
 theme\_classic(base\_size = 12)



This model appears to show very little, if any, relationship between temperature and air quality until around 75F. Above 75F, it appears as though air quality declined as temperature increased.

Before we can share these findings, we must check the model is appropriate and results can be treated as valid:

par(mfrow = c(2,2))  
gam.check(gam\_temp)



Method: REML Optimizer: outer newton  
full convergence after 5 iterations.  
Gradient range [-1.501701e-06,3.401414e-08]  
(score 521.6017 & scale 486.8272).  
Hessian positive definite, eigenvalue range [0.6349053,57.02481].  
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(Temp) 9.00 3.37 0.99 0.34

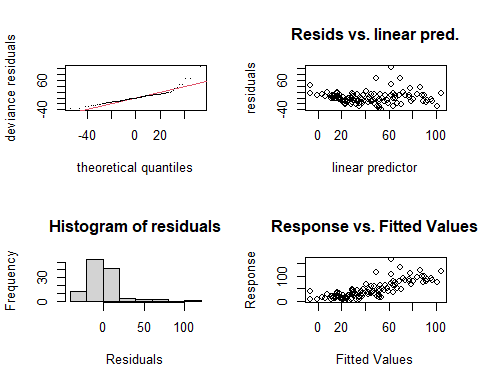
Although the maximum number of knots appears to be appropriate, the diagnostic plots show that there are some concerns about this model’s validity. The residuals follow a skewed distribution, with a longer positive tail, indicating that the model is underestimating a number of points. This is also noticeable in other diagnostic plots and suggests the model could be improved.

Prior checks showed that wind speed may explain some of the outlying points, so the model could be extended to include the wind variable:

gam\_temp\_wind <- gam(Ozone ~  
 s(Temp) + Wind,  
 data = airquality,  
 method = "REML")  
  
summary(gam\_temp\_wind)

Family: gaussian   
Link function: identity   
  
Formula:  
Ozone ~ s(Temp) + Wind  
  
Parametric coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 70.0310 6.4514 10.855 < 2e-16 \*\*\*  
Wind -2.8292 0.6253 -4.525 1.53e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Approximate significance of smooth terms:  
 edf Ref.df F p-value   
s(Temp) 3.237 4.046 19.34 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-sq.(adj) = 0.617 Deviance explained = 63.1%  
-REML = 511.67 Scale est. = 417.1 n = 116

gam.check(gam\_temp\_wind)



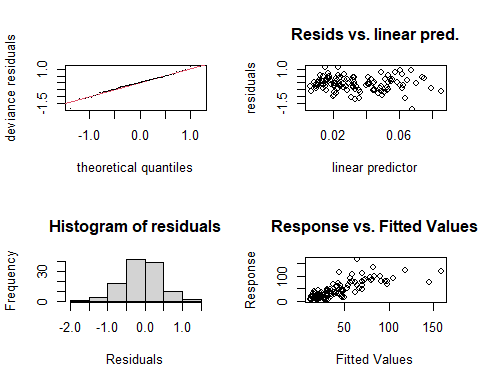
Method: REML Optimizer: outer newton  
full convergence after 5 iterations.  
Gradient range [-1.229657e-07,2.809249e-09]  
(score 511.6701 & scale 417.0997).  
Hessian positive definite, eigenvalue range [0.7249549,56.52243].  
Model rank = 11 / 11   
  
Basis dimension (k) checking results. Low p-value (k-index<1) may  
indicate that k is too low, especially if edf is close to k'.  
  
 k' edf k-index p-value  
s(Temp) 9.00 3.24 0.95 0.22

Model diagnostics show that the residuals were very skewed in this model, and the qq-plot suggests that a normal distributional assumption may not be appropriate. Given the outcome is continuous and non-negative, a Gamma distribution may be more appropriate:

gam\_temp\_gamma <- gam(Ozone ~  
 s(Temp) + Wind,  
 data = airquality,  
 family = Gamma,  
 method = "REML")  
  
summary(gam\_temp\_gamma)

Family: Gamma   
Link function: inverse   
  
Formula:  
Ozone ~ s(Temp) + Wind  
  
Parametric coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.0192060 0.0036713 5.231 8.04e-07 \*\*\*  
Wind 0.0014955 0.0003643 4.105 7.75e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Approximate significance of smooth terms:  
 edf Ref.df F p-value   
s(Temp) 3.3 4.121 14.49 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-sq.(adj) = 0.587 Deviance explained = 60.4%  
-REML = 506.6 Scale est. = 0.24166 n = 116

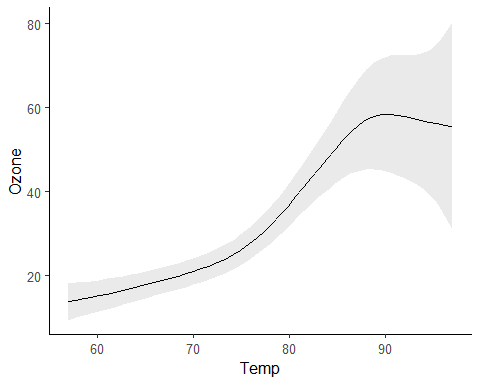
gam.check(gam\_temp\_gamma)



Method: REML Optimizer: outer newton  
full convergence after 5 iterations.  
Gradient range [-3.599286e-07,2.03387e-07]  
(score 506.595 & scale 0.2416603).  
Hessian positive definite, eigenvalue range [0.6962158,61.40889].  
Model rank = 11 / 11   
  
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(Temp) 9.0 3.3 0.88 0.14

This model appears to fit the data far better, the qq-plot shows that the Gamma distribution is a good fit for the residuals, and there are no issues with heteraskedasticity. Therefore, this is the model we will use to answer our original question.

plot\_predictions(gam\_temp\_gamma,  
 condition = "Temp") +   
 theme\_classic(base\_size = 12)



Based on out model, we can determine that air quality decreases as temperature increases, up to around 90 degrees (after adjusting for differences in wind speed). Above 90 degrees, there doesn’t appear to be a strong relationship with air quality.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK