STA303 Assignment 2

Talia Fabregas Kasandra Tworzynaski

2025-03-28

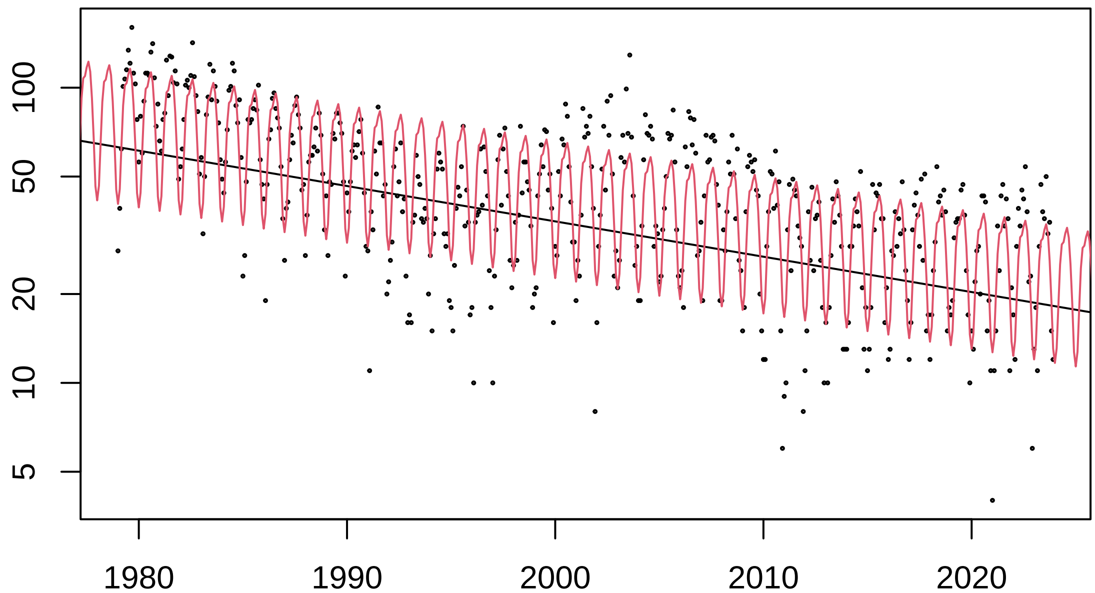

Figure 1: Provided on the assignment 2 handout

Question 1: Motorcycle Accidents

Part 1

Write down, in equations not R code, a generalized additive model suitable for this problem. Explain each of the parts of the model and give a rationale for them (i.e. "The response variable is Gamma distributed because the number of deaths must be positive"). (4 points)

A generalized additive model (GAM) suited for this problem is the Negative Binomial GAM. We chose a Negative Binomial because our response variable, Y_i (number of motorcycle deaths in month i) is a non-negative count variable. Unlike a Poisson, a Negative Binomial can account for over-dispersion (flexible enough to account for high or very low over-dispersion).

The GAM that we will use for this problem is as follows:

$$Y_i \sim \text{NegBinom}(D_i\mu_i, \tau)$$

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{12} \beta_j \mathbb{I}(\text{month}_i = j) + f(t_i; \alpha)$$

where:

- Y_i is the number of motorcycle deaths in month i
- μ_i is the expected number of motorcycle deaths in month i. We use the log link function to
- D_i is the number of days in month i. We use `logMonthDays` as an offset in our model to account for differences in exposure time.
- β_0 is the intercept.
- $\beta_1, ..., \beta_{12}$ are the fixed effects for month, where month is categorical.
- f is the smooth function over time (dateInt) and we use $k = 50$ knots.
- α is the smoothing coefficient.

Part 2

Show R code which fits this model using the mgcv package. (2 points)

```
# fit the model
motorcycleGAM <- mgcv::gam(killed ~ month + offset(logMonthDays) + s(dateInt, k=50),
                    data=x,
                    family=nb,
                    method='ML'
                    )

# show the formula (just for reference)
motorcycleGAM$formula
```

```
## killed ~ month + offset(logMonthDays) + s(dateInt, k = 50)
```

Part 3

Question 2: Heat

The formulas for `res1` , `res2`, `res3` :

```
## Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")), k = 100) +
##     s(yearFac, bs = "re") + sinpi(dateInt/182.625) + cospi(dateInt/182.625) +
##     sinpi(dateInt/91.3125) + cospi(dateInt/91.3125)
```

```
## Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")), k = 4) +
##     s(yearFac, bs = "re") + sinpi(dateInt/182.625) + cospi(dateInt/182.625) +
##     sinpi(dateInt/91.3125) + cospi(dateInt/91.3125)
```

```
## Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")), k = 100) +
##     sinpi(dateInt/182.625) + cospi(dateInt/182.625) + sinpi(dateInt/91.3125) +
##     cospi(dateInt/91.3125)
```

Part 1

Equations for `res1`

$$
\begin{aligned}
&Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \\
&\mu_i = \beta_0 + \beta_1 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_2 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_3 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) + \beta_4 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) \\
&\quad + f_1(\text{dateInt}_i) + f_2(\text{yearFac}_i)
\end{aligned}
$$

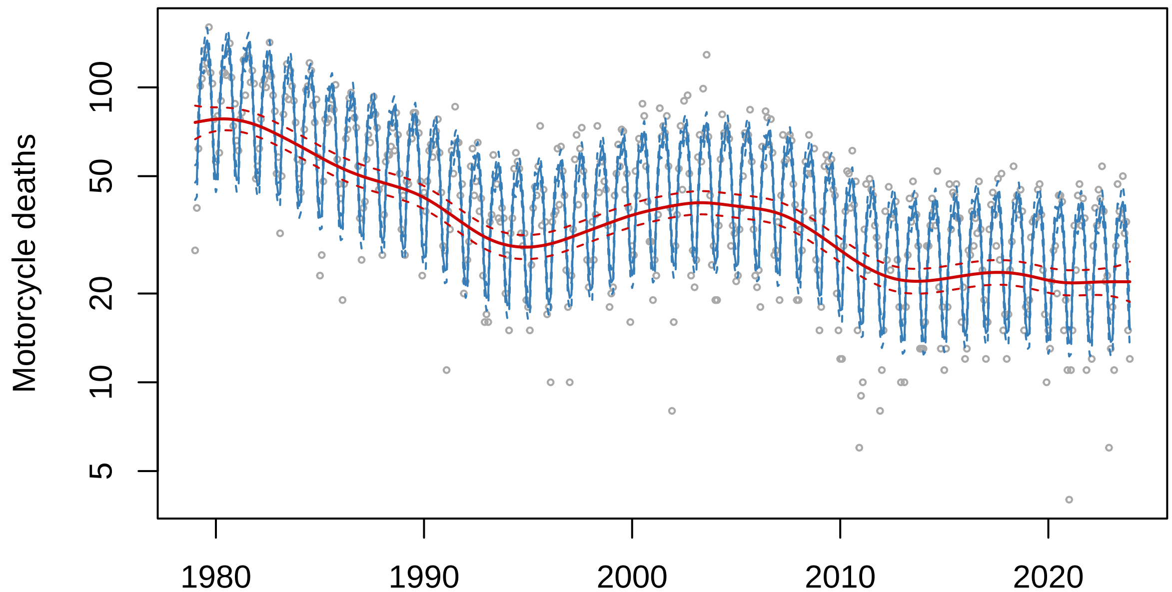

Figure 2: The figure displays motorcycle death data over time (dark gray points) with a log scale on the y-axis. The blue lines represent the seasonal effect, with the 95% prediction interval. The red lines illustrate the trend over time, with a 95% prediction interval shown by the red dotted lines. The data suggest a seasonal fluctuation in motorcycle deaths, along with a clear trend over time, as modeled by a generalized additive model (GAM)

where: - Y_i is the Max Temp recorded in month i

- β_1, β_2 are the coefficients of the cosine and sine terms of the 12-month cycle
- β_3 and β_4 are the coefficients of the cosine and sine terms of the 6-month cycle
- f_1 is the smooth trend over time (dateInt), modeled using a spline with k=100 knots
- f_2 is the year random effect which accounts for variability between years. The use of `bs = "re"` in the code tells `mgcv::gam` to treat it as a random effect, and not a smoothing term.

Equations for `res2`

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_2 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_3 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) + \beta_4 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) \\
&\quad + f_1(\text{dateInt}_i) + f_2(\text{yearFac}_i)
\end{aligned}
$$

where: - Y_i is the Max Temp recorded in month i

- β_1, β_2 are the coefficients of the cosine and sine terms of the 12-month cycle
- β_3 and β_4 are the coefficients of the cosine and sine terms of the 6-month cycle
- f_1 is the smooth trend over time (dateInt), modeled using a spline with k=4 knots. The number of knots, k is the key difference between `res1` and `res2`
- f_2 is the year random effect which accounts for variability between years. The use of `bs = "re"` in the code tells `mgcv::gam` to treat it as a random effect, and not a smoothing term.

Equations for `res3`

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_2 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{365.25}\right) + \beta_3 \cdot \cos\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) + \beta_4 \cdot \sin\left(\frac{2\pi \text{dateInt}_i}{182.625}\right) + f(\text{dateInt}_i)$$

where: - Y_i is the Max Temp recorded in month i

- β_1, β_2 are the coefficients of the cosine and sine terms of the 12-month cycle
- β_3 and β_4 are the coefficients of the cosine and sine terms of the 6-month cycle
- f is the smooth trend over time (dateInt), modeled using a spline with k=100 knots.

Differences between the three models The models defined in `res1`, `res2`, and `res3` have a couple of key differences. Firstly, the models in `res1` and `res2` have one key difference: the number of knots, k. The `res1` model uses $k = 100$ knots which allows for a more detailed and flexible smoothing function, whereas the `res2` model uses $k = 4$ knots which allows for a "smoother", less detailed smoothing function. Larger k is generally better because it fits the data better; the model in `res2` with $k = 4$ probably has too few knots to fit the data well or capture details. The model in `res3` has $k = 100$ knots just like the model in `res1`, but it omits the `yearFac` random effect and only includes a smooth term for `dateInt`. The `res3` model is the simplest but it does not capture unobserved temperature variability between years.

Part 2

Part 3

Part 4

These results hint that there may have been excess warming between 1996 and 2025, but the 95% prediction intervals (shown by the purple dotted lines) fall both above and below the actual max temperature data, so it's inconclusive.

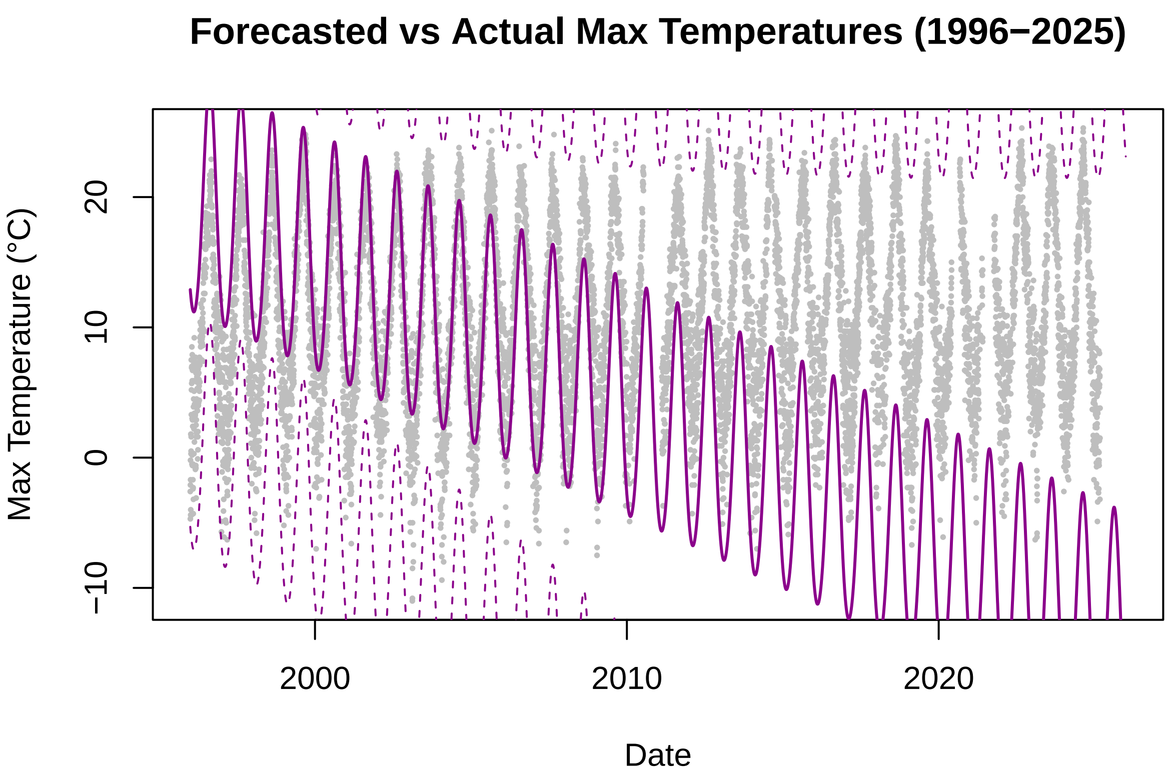

Figure 3: write the caption here. 2-3 sentences