Assignment 2 Answers

Raghav Sinha, Yutong Han

March 14, 2025

Table of contents

1 Motorcycle Deaths

1.1 Question 1

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$log(\mu_t) = \beta_0 + f(t) + \text{offset}(\log(\text{MonthDays}_t)) + g(\text{month}_t) + \epsilon_t$$

- Y_t represents the weekly number of deaths at time t. It is assumed to follow a Poisson distribution.
- The Poisson distribution is suitable as number of deaths is a count variable.
- The log link function is used to connect the mean to the linear predictor. It ensures that the mean μ_t remains positive.
- β_0 represents the baseline level of deaths when all other terms are zero.
- $f(t)$ is a smooth function of time, it captures the non-linear trend in deaths over time.
- $\log(\text{MonthDays}_t)$ is an offset that accounts for the varying number of days in each month.
- $g(\text{month}_t)$ is a function that captures the seasonal effect of the month on the number of deaths (eg. more accidents in the summer due to increased riding.)

1.2 Question 2

```r
library(mgcv)
library(Hmisc)

x$dateInt <- as.integer(x$date)
x$logMonthDays <- log(Hmisc::monthDays(x$date))
x$month <- factor(format(x$date, "%b"),
                  levels = format(ISOdate(2000, 1:12, 1), "%b"))

# Fit the GAM
gam_model <- gam(killed ~ s(dateInt, bs = "cr", k = 50)
                 + offset(logMonthDays) + month,
                 data = x,
                 family = poisson(link = "log"),
                 method = "REML")

# Summary of the model
summary(gam_model)
```

```
Family: poisson
Link function: log

Formula:
killed ~ s(dateInt, bs = "cr", k = 50) + offset(logMonthDays) +
    month

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33370    0.03041 -10.975  < 2e-16 ***
monthFeb     0.11785    0.04258   2.768  0.00564 **
monthMar     0.44270    0.03883  11.400  < 2e-16 ***
monthApr     0.77497    0.03683  21.043  < 2e-16 ***
monthMay     0.93406    0.03579  26.101  < 2e-16 ***
monthJun     0.95391    0.03587  26.594  < 2e-16 ***
monthJul     1.02488    0.03537  28.972  < 2e-16 ***
monthAug     1.07111    0.03518  30.448  < 2e-16 ***
monthSep     0.99581    0.03571  27.885  < 2e-16 ***
monthOct     0.78972    0.03666  21.540  < 2e-16 ***
monthNov     0.47967    0.03899  12.303  < 2e-16 ***
```

```
monthDec     0.11103    0.04198   2.645  0.00817 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq p-value
s(dateInt) 17.79  22.03   4129  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.866   Deviance explained = 86.8%
-REML = 2105.8  Scale est. = 1         n = 540
```

1.3 Question 3

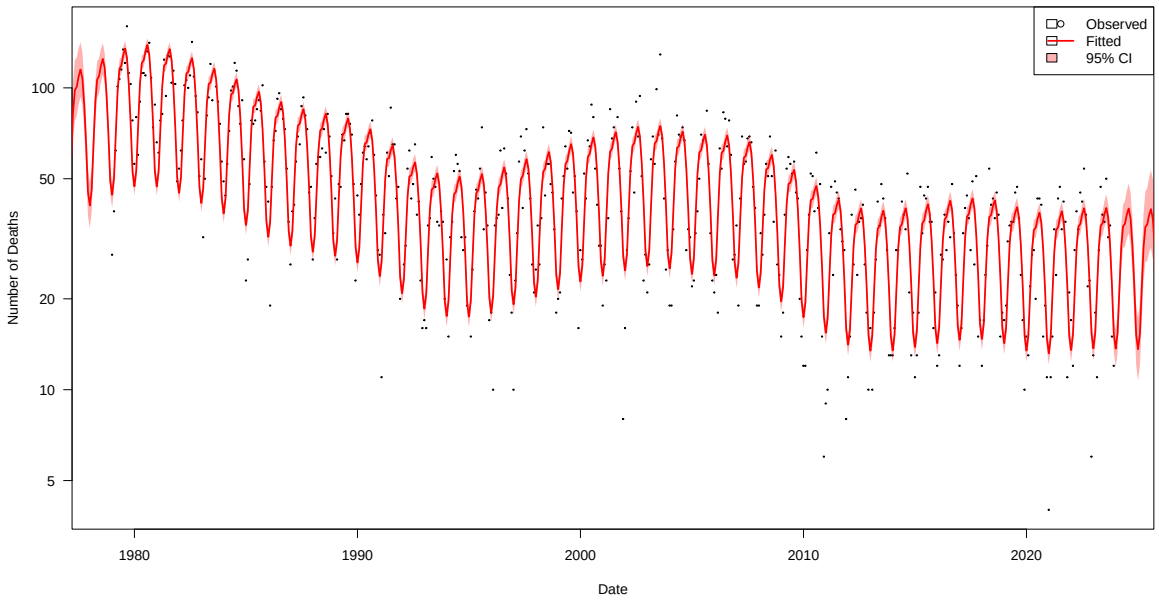

Figure 1: Trend for for motorcycle deaths over time with 95% CI. The fitted curve captures both the long-term decline in deaths and a strong seasonal pattern. The parametric coefficients show significant seasonal effects, with deaths peaking in the summer months (June to August) and reaching a minimum in winter. For example, compared to January, deaths increase by 107% in August (exp(1.071) 2.92), while December shows no significant difference from January.