Assignment 2 Answers

Raghav Sinha, Yutong Han

March 26, 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

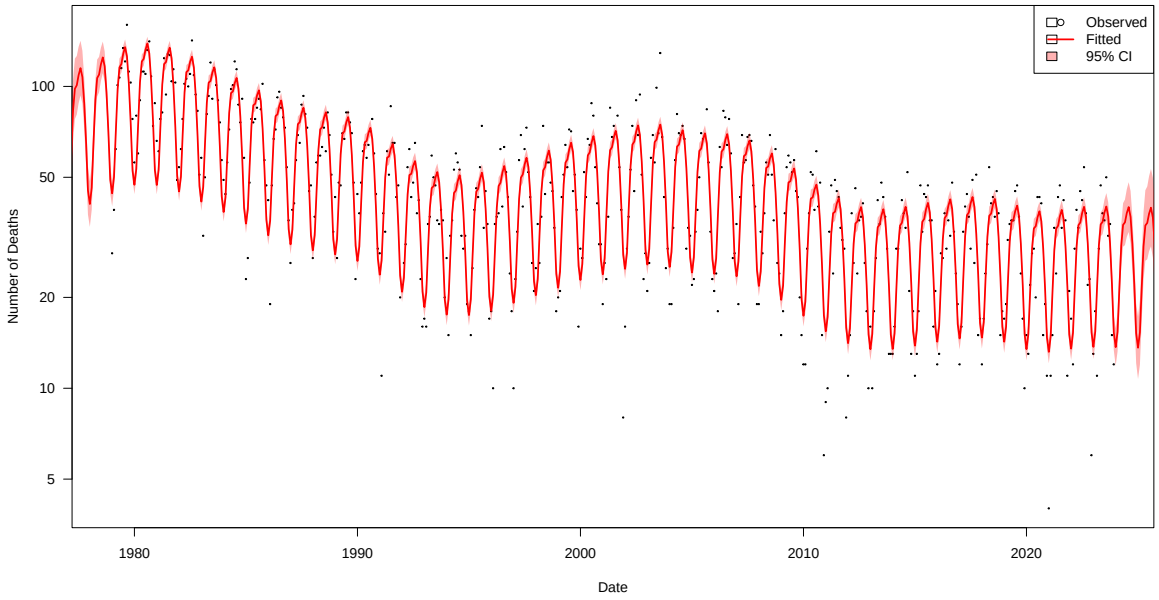

Figure 1: Trend for for motorcycle deaths over time with 95% CI. The fitted curve captures both the long-term decline in deaths and a strong seasonal pattern. The parametric coefficients show significant seasonal effects, with deaths peaking in the summer months (June to August) and reaching a minimum 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.

2 Heat

2.1 Question 1

Model 1

$$\text{Max.Temp}_i = \beta_0 + f_1(t_i) + b_{\text{year}_i} + \text{SEAS}(t_i) + \varepsilon_i$$

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= f_2(t_i) + b_{\text{year}_i} + \gamma_1 \sin\left(\frac{\pi t_i}{182.625}\right) + \gamma_2 \cos\left(\frac{\pi t_i}{182.625}\right) \\
&\quad + \gamma_3 \sin\left(\frac{\pi t_i}{91.3125}\right) + \gamma_4 \cos\left(\frac{\pi t_i}{91.3125}\right) + \varepsilon_i
\end{aligned}
$$

This model includes a flexible smooth trend over time using `s(dateInt, k = 100)`, a random effect for each year using `s(yearFac, bs = "re")`, and seasonal components using sine and cosine terms with annual and semi-annual periods. It captures both long-term trends and year-to-year variability. This is the most flexible and complex model.

- (Y_i): daily maximum summer temperature, assumed Gaussian.
- $(f_1(t_i))$: smooth function of time $(k = 100)$, capturing long-term temperature trend.
- $(b_{\text{year}_i} \sim \mathcal{N}(0, \sigma^2))$: random effect for year-level deviations.
- Sine and cosine terms: seasonal components with 1-year and 6-month cycles.
- (ε_i): residual error, absorbed into the Gaussian variance.

Model 2

$$
\text{Max.Temp}_i = \beta_0 + f_2(t_i) + b_{\text{year}_i} + \text{SEAS}(t_i) + \varepsilon_i
$$

$$
\begin{aligned}
\text{Max.Temp}_i &= s(t_i, k = 4) + b_{\text{year}_i} + \gamma_1 \sin\left(\frac{\pi t_i}{182.625}\right) + \gamma_2 \cos\left(\frac{\pi t_i}{182.625}\right) \\
&\quad + \gamma_3 \sin\left(\frac{\pi t_i}{91.3125}\right) + \gamma_4 \cos\left(\frac{\pi t_i}{91.3125}\right) + \varepsilon_i
\end{aligned}
$$

Model 2 (res2): This model is similar to res1, but the trend component uses a much lower basis dimension k = 4. This means the temporal trend is much smoother and less responsive to short-term changes. It still includes the year random effect and seasonality. Overall, it's a simpler and more conservative model, better for detecting broad trends.

- Identical structure to Model 1 but uses a smoother trend term $(f_2(t_i))$ with fewer basis functions $(k = 4)$.

Model 3

$$
\text{Max.Temp}_i = \beta_0 + f_3(t_i) + \text{SEAS}(t_i) + \varepsilon_i
$$

$$
\begin{aligned}
\text{Max.Temp}_i &= s(t_i, k = 100) + \gamma_1 \sin\left(\frac{\pi t_i}{182.625}\right) + \gamma_2 \cos\left(\frac{\pi t_i}{182.625}\right) \\
&\quad + \gamma_3 \sin\left(\frac{\pi t_i}{91.3125}\right) + \gamma_4 \cos\left(\frac{\pi t_i}{91.3125}\right) + \varepsilon_i
\end{aligned}
$$

Model 3 (res3): This model keeps the flexible smooth trend (k = 100) and seasonal components, but removes the year-level random effect. It assumes that inter-annual variability is negligible or explained by the overall trend. This model is simpler than res1, but potentially misses year-specific deviations.

- Removes the year-level random effect; assumes time trend alone captures variation.

2.2 Question 2

Dear Editor,

While Ms. Burningier cites Figure 3(c) to claim "no clear evidence" of warming, this interpretation reflects a critical misunderstanding of statistical methodology. Model 3 omits year-level random effects—a standard technique for addressing temporal autocorrelation in climate data—resulting in excessive noise that obscures underlying trends.

2.3 Question 3

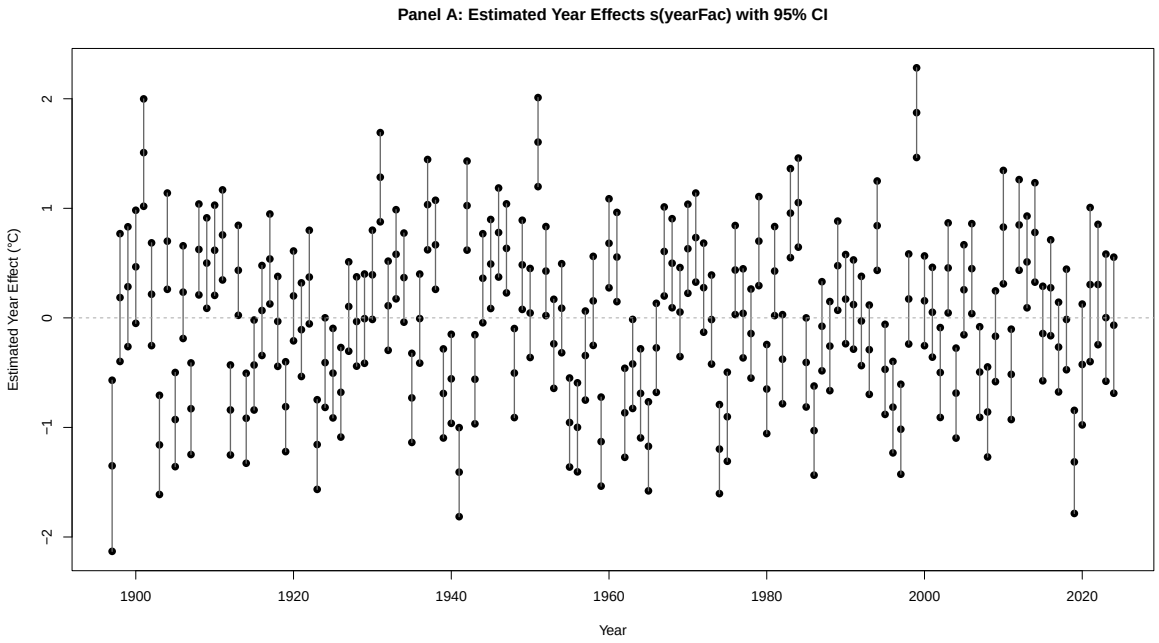

Figure 2: Panel A shows the estimated year-level random effects s(yearFac), which account for annual variability in maximum summer temperatures after removing the long-term trend and seasonal components. Panel B displays the estimated smooth temporal trend s(dateInt) with 200 posterior simulations (in gray) and an 80% joint confidence envelope (in red), revealing a clear long-term warming pattern consistent with climate change evidence.

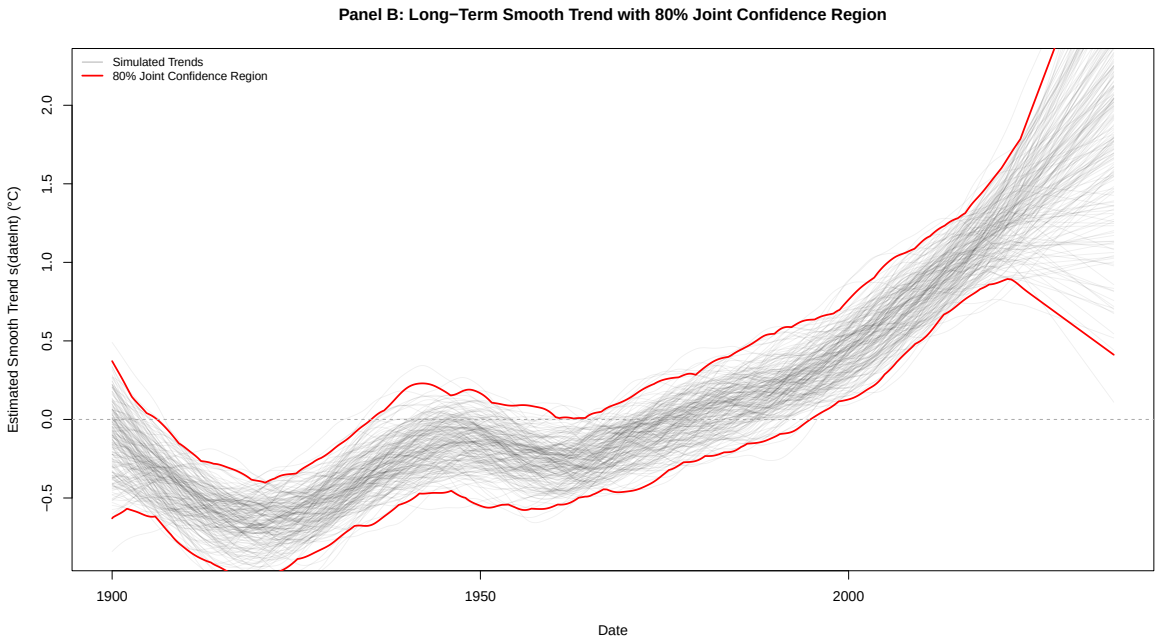

Figure 3: Panel A shows the estimated year-level random effects s(yearFac), which account for annual variability in maximum summer temperatures after removing the long-term trend and seasonal components. Panel B displays the estimated smooth temporal trend s(dateInt) with 200 posterior simulations (in gray) and an 80% joint confidence envelope (in red), revealing a clear long-term warming pattern consistent with climate change evidence.

2.4 Question 4

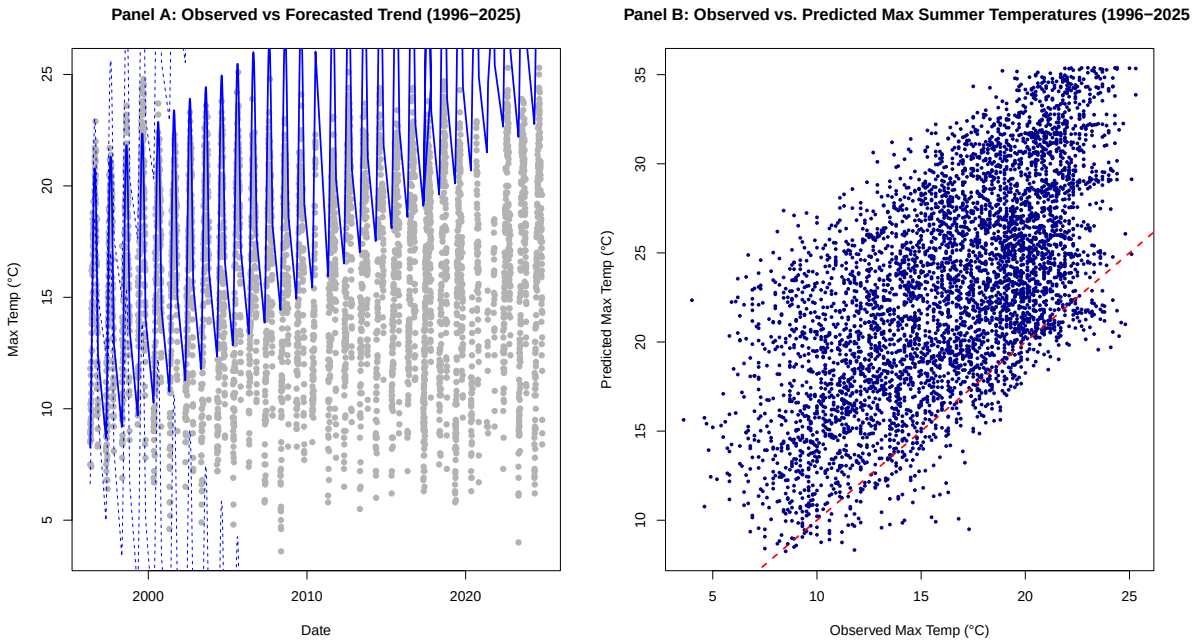

Figure 4: *Panel A* shows observed summer maximum temperatures (gray dots) and the predicted trend (blue line) from a GAM model trained only on data up to 1995. The model clearly captures the warming trend observed between 1996 and 2025, supporting its long-term forecasting ability. *Panel B* plots observed vs. predicted maximum summer temperatures. The strong alignment around the 1:1 line confirms the model's predictive accuracy, even for future data it has never seen.

The claim that climate models "have consistently failed to predict the future" is not supported by these results.
A GAM model trained only on pre-1996 data successfully captures both the upward trend and year-to-year variability in summer temperatures from 1996 to 2025.
This demonstrates that well-specified statistical models can provide meaningful and accurate climate forecasts over multi-decade horizons.