STA303, Homework 2

Patrick Brown, University of Toronto

Due Tuesday 25 March

As before you may work in pairs and submit a joint project. Submit on Quercus.

1 Motorcycle deaths (10 points)

The dataset below is a subset of the data from www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents, with all of the road traffic accidents in the UK from 1979 to 2023. The data below consist of all injuries involved in motorcycle accidents with either fatal, moderate, or slight injuries.

Below is code to retreive the data

```
> theUrl = "http://pbrown.ca/teaching/appliedstats/data/motorcycle.rds"
> theFile = basename(theUrl)
> if (!file.exists(theFile)) download.file(theUrl, theFile)
> x = readRDS(theFile)
```

The data are weekly numbers of deaths (numbers killed, seriously injured, or slightly injured). Below is code to fit a Poisson GLM and plot the fitted values

```
> 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"))
> res = glm(killed ~ offset(logMonthDays) + dateInt + month, data = x,
+   family = poisson(link = "log"))
> newdata = data.frame(date = seq(as.Date("1975/1/1"), as.Date("2030/1/1"),
+   by = "month"))
> newdata$dateInt = as.integer(newdata$date)
> newdata$logMonthDays = log(30)
> newdata$month = "Mar"
> pred1 = predict(res, newdata)
> newdata$month = format(newdata$date, "%b")
> pred2 = predict(res, newdata)
```

```
> plot(x$date, x$killed, cex = 0.2, log = "y", xlab = "", ylab = "")
> matlines(newdata$date, exp(cbind(pred1, pred2)), lty = 1)
```

Figure fig. 1 shows that there has been a clear downward trend in fatal motorcycle accidents over time. Using a Generalized Additive Model it is expected that a more accurate and informative estimate of the trend over time can be discerned.

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

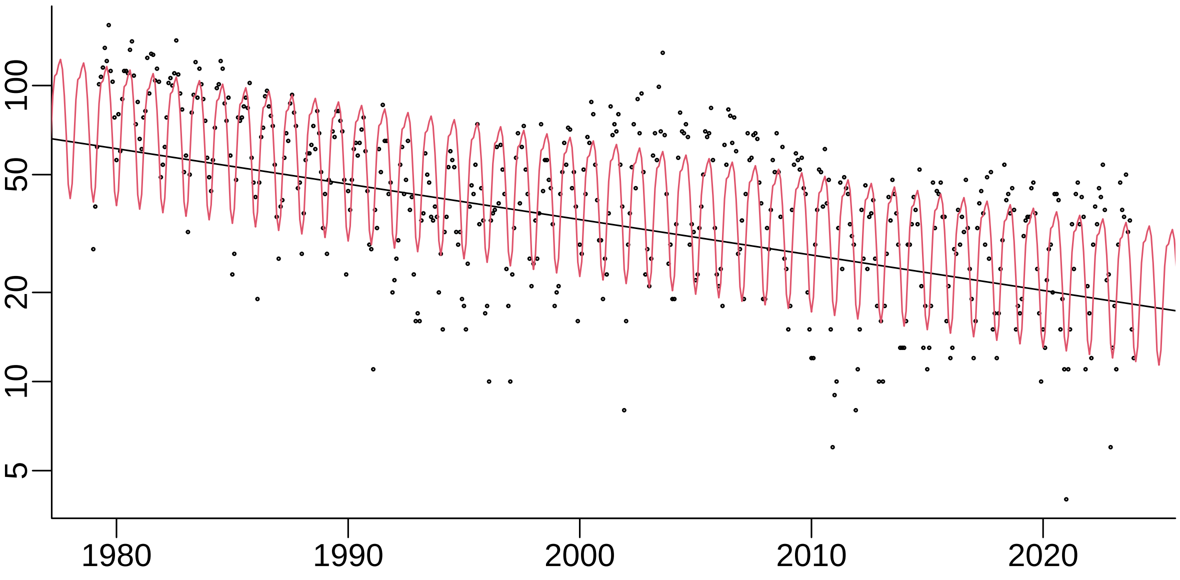


Figure 1: Estimated trend (black), trend with seasonal effect (red) and data (dots) for motorcycle fatalities.

3. Produce a figure similar to Figure fig. 1 which is able to visualize the trend estimated from the motorcycle data. You're marked on the figure looking professional (with clear labels and a caption) as well as conveying the important statistical information (prediction intervals as well as point predictions). (4 points)

2 Heat (10 points)

Figure fig. 2 a shows daily maximum temperature data recorded on Sable Island, off the coast of Nova Scotia. Figure 2 b shows the period from 2020 to the present, with summer months (May to October inclusive) in red and winter in black. In 1996 (when I was a 4th year undergraduate) I attended a talk by David Thompson (now at Queen's University) where looked for evidence of global warming in these data. Most long time series of temperatures are from weather stations which were initially in rural areas but over time became urban as cities expanded, and the worry is that any temperature increases are a result of an "urban heat island" effect rather than climate change. The Sable Island station does not have this problem.

Notice that the winter temperatures are more variable than summer temperatures, and you can assume you have been advised by a reliable environmental scientist to consider only summer temperatures when modelling historical temperature time series (since winter temperatures are governed by a different and much more complex physical process).

Writing in 2019, the IPCC stated

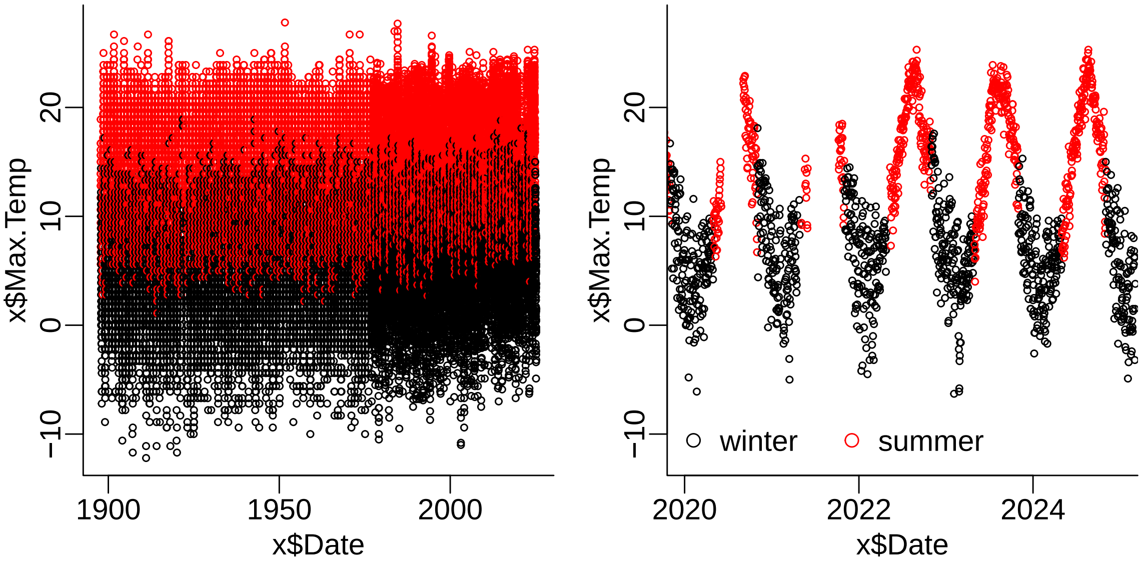
> Human activities are estimated to have caused approximately 1.0°C of global warming above pre-industrial levels, with a likely range of 0.8°C to 1.2°C. Global warming is likely to reach 1.5°C between 2030 and 2052 if it continues to increase at the current rate. (high confidence)

see www.ipcc.ch/sr15/resources/headline-statements

The People's Party of Canada states

> There is however no scientific consensus on the theory that CO2 produced by human activity is causing dangerous global warming today or will in the future, and that the world is facing environmental catastrophes unless these emissions are drastically reduced. Many renowned scientists continue to challenge this theory.

https://www.peoplespartyofcanada.ca/issues/environment



(a) all

(b) 2020-2025

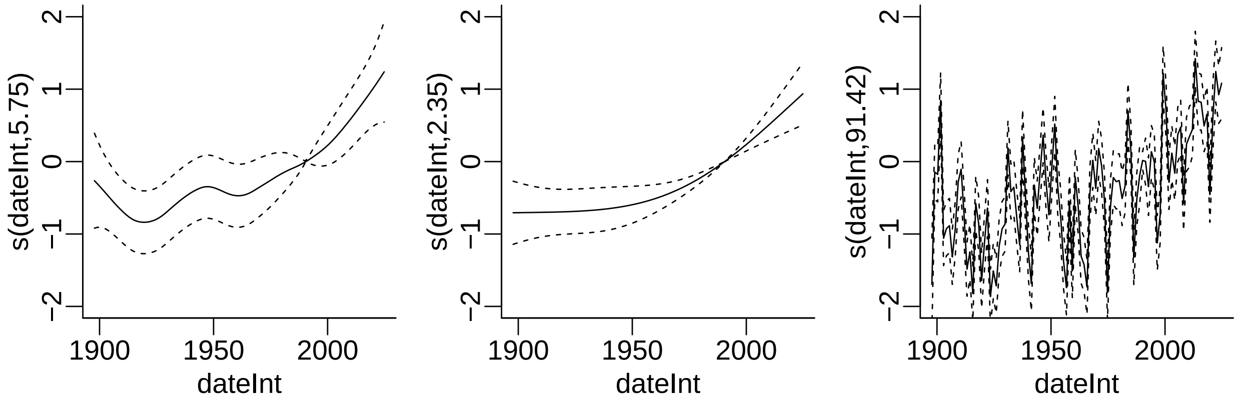
Figure 2: Sable island maximum daily temperature

Below is code for three generalized additive models

```
> x[100, ]

           Date Max.Temp month summer
2565 1898-01-08      4.4     1  FALSE

> x$dateInt = as.integer(x$Date)
> x$yearFac = factor(format(x$Date, "%Y"))
> xSub = x[x$summer & !is.na(x$Max.Temp), ]
> library("mgcv")

> res1 = gam(update.formula(Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")),
+   k = 100) + s(yearFac, bs = "re"), Pmisc::seasonalFormula(period = 365.25,
+   harmonics = 1:2, var = "dateInt")), data = xSub, method = "ML",
+   optimizer = "efs")

> res2 = gam(update.formula(Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")),
+   k = 4) + s(yearFac, bs = "re"), Pmisc::seasonalFormula(period = 365.25,
+   harmonics = 1:2, var = "dateInt")), data = xSub, method = "ML",
+   optimizer = "efs")

> res3 = gam(update.formula(Max.Temp ~ s(dateInt, pc = as.integer(as.Date("1990/7/1")),
+   k = 100), Pmisc::seasonalFormula(period = 365.25, harmonics = 1:2,
+   var = "dateInt")), data = xSub, method = "ML", optimizer = "efs")
```

- The Pmisc::seasonalFormula bit in the code creates sines and cosines with periods 1 year and 6 months, evident in the formula below

```
> res1$formula

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)
```



(a) res1

(b) res2

(c) res3

Figure 3: predicted trend

- The component `s(yearFac, bs='re')` is an independent random effect, equivalent to doing `(1|yearFac)` in `glmmTMB`.

- `optimizer='efs'` uses the "extended Fellner Schall method of Wood and Fasiolo (2017)" as the numerical optimizer. The model had problems fitting with the default optimizer.

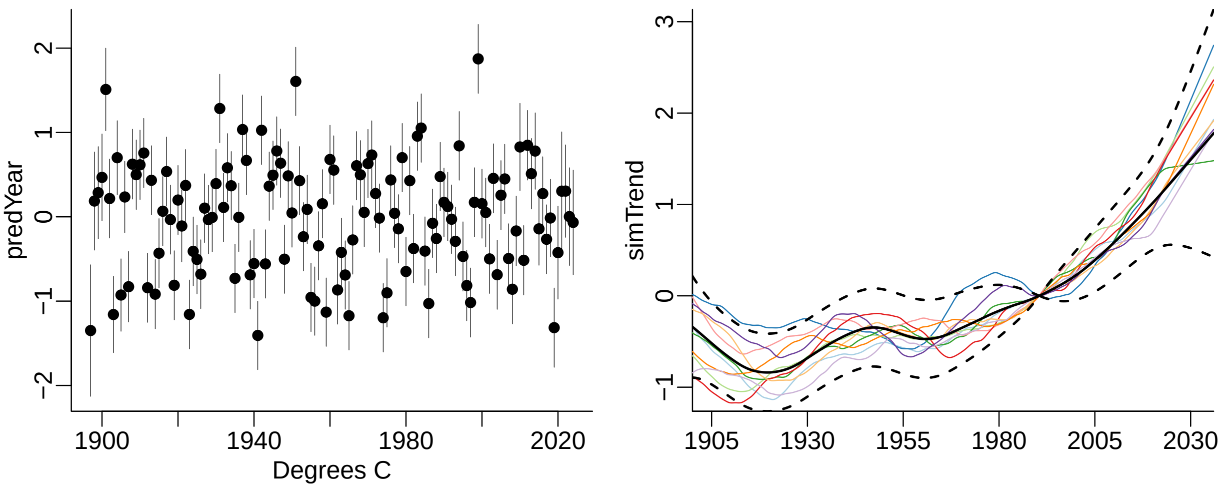
Plots of the fitted trends appear in fig. 3, the produced by executing `plot(resX, select=1)`, with `X` taking values 1,2, and 3. f

Below is code where some predictions are made for the model in `res1`

```
> Syear = unique(xSub$yearFac)
> predYear = do.call(cbind, predict(res1, newdata = data.frame(yearFac = Syear,
+   dateInt = 0), type = "terms", terms = "s(yearFac)", se.fit = TRUE)) %*%
+   Pmisc::ciMat()
> newdat = data.frame(Date = seq(as.Date("1900/1/1"), as.Date("2035/12/31"),
+   by = "2 weeks"), yearFac = Syear[1])
> newdat$dateInt = as.integer(newdat$Date)
> predTrend = do.call(cbind, predict(res1, newdat, type = "terms", terms = "s(dateInt)",
+   se.fit = TRUE)) %*% Pmisc::ciMat()
> newX = predict(res1, newdata = newdat, type = "lpmatrix")
> simCoef <- rmvn(10, coef(res1), vcov(res1))
> isTrend = grep("s[(]dateInt", colnames(newX))
> simTrend = tcrossprod(newX[, isTrend], simCoef[, isTrend])
```

fig. 4 is produced with the following code.

```
> Syear = as.numeric(as.character(Syear))
> matplot(Syear, predYear, xlab = "Degrees C", cex = c(1, 0, 0), pch = 16,
+   col = "black")
> segments(Syear, predYear[, 2], Syear, predYear[, 3], lwd = 0.5)
> matplot(newdat$Date, simTrend, type = "l", lty = 1, col = RColorBrewer::brewer.pal(ncol(simTrend),
+   "Paired"), xaxt = "n", xaxs = "i", yaxs = "i", ylim = range(predTrend),
+   xlab = "")
> matlines(newdat$Date, predTrend, lty = c(1, 2, 2), col = "black",
+   lwd = 2)
> forX = as.Date(ISOdate(seq(1880, 2050, by = 25), 1, 1))
> axis(1, forX, format(forX, "%Y"))
```



(a) one

(b) two

Figure 4: A Figure

1. Write down a set of equations (or several sets of equations) which corresponds to the three calls to `gam`. Explain the differences between the models in `res1`, `res2` and `res3`? (2 points)
2. Maxine Burningier, leader of the People's Party of Canada, sees fig. 3 c and proclaims in a post on Truth Social that “There is no clear evidence of global temperature increase! Model 3 is the best model and it shows these data are only noise!’’ Write a short (3 or 4 sentence) letter to the editor of your local tabloid newspaper either providing scientific justification for the Esteemed Leader's statement or explaining clearly why you disagree. (2 points)
3. Create a nicer version of fig. 4, with axis labels and an informative caption. You can refer to variables given in the equations from Question 1. The caption should be two or three sentences, a reader familiar with generalized additive models should understand everything about the figure (even if they have not read this homework assignment) (3 points)
4. David Thompson found a clear evidence of global warming in these data in 1996. Re-fit the model using only data from 1995 and earlier and compare that model's forecasts to the observed data from 1996-2025. Is the People's Party of Canada correct in saying “Climate change alarmism is based on flawed models that have consistently failed at correctly predicting the future.” (3 points)

Warning: I haven't tried this, so there might be problems fitting the models. If you get extremely rough or perfectly linear estimated trends it could signify a problem with optimizing.

A Additional Code

Download temperature data

```
> heatUrl = "http://pbrown.ca/teaching/appliedstats/data/sableIsland.rds"
> dir.create("cache", showWarnings = FALSE)
> heatFile = file.path("cache", basename(heatUrl))
> if (!file.exists(heatFile)) download.file(heatUrl, heatFile)
> x = readRDS(heatFile)
```

```
> names(x) = gsub("[.]+C[.]", "", names(x))
> x$Date = as.Date(x$Date)
> x$month = as.numeric(format(x$Date, "%m"))
> x$summer = x$month %in% 5:10
```