Kericho

greyhypotheses

The set of external functions used thus far - relative to, therefore based in, GitHub repository premodelling/time - are

```
sys.source(file = 'R/kericho/StudyData.R',
           envir = knitr::knit_global())

sys.source(file = 'R/kericho/problems/explore/ExplorationGraph.R',
           envir = knitr::knit_global())
sys.source(file = 'R/kericho/problems/rainfall/Graphs.R',
         envir = knitr::knit_global())
sys.source(file = 'R/kericho/problems/specified/PredictionsGraph.R',
           envir = knitr::knit_global())

sys.source(file = 'R/functions/TimeDependentLag.R',
           envir = knitr::knit_global())

sys.source(file = 'docs/programme/mathematics/auxiliary_function.R',
           envir = knitr::knit_global())
```

Data Set-up

The original data set, with appended time dependent variables, is

```
'data.frame': 310 obs. of  11 variables:
 $ Year   : int  1979 1979 1979 1979 1979 1979 1979 1979 1979 1979 ...
 $ Month  : Ord.factor w/ 12 levels "Jan"<"Feb"<"Mar"<..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Cases  : int  25 25 20 30 18 18 15 15 10 20 ...
 $ Rain   : num  3.7 3.2 5.6 8.3 8.1 5.4 5.5 6.1 5.7 5.6 ...
 $ minT   : num  11.8 11.3 10.9 12 10.9 11.4 10.2 10.1 10.2 11.1 ...
 $ maxT   : num  24 23.5 25.1 23.6 22.9 22.1 22.1 23 23.9 25.2 ...
 $ VCAP   : num  78.5 56.6 131.9 467.6 277 ...
 $ CasesLN: num  3.22 3.22 3 3.4 2.89 ...
 $ datestr: chr  "1979-01" "1979-02" "1979-03" "1979-04" ...
 $ date   : Date, format: "1979-01-01" "1979-02-01" ...
 $ time   : num  0 1 2 3 4 5 6 7 8 9 ...
```

Explore

An exploration of the relationship between ln(cases) and maximum temperature, minimum temperature, and rain.

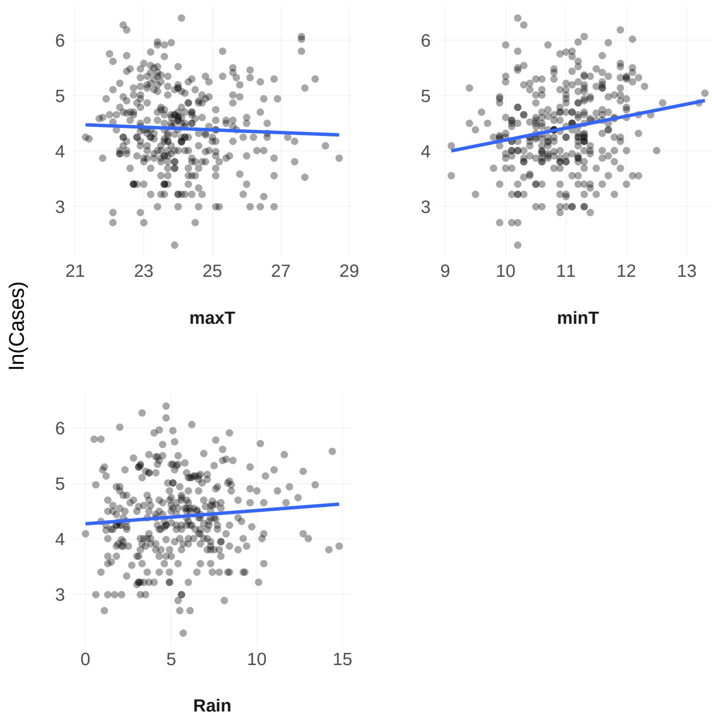

Figure 1: the relationship between ln(cases) and maximum temperature, minimum temperature, and rain

Rainfall

The function *TimeDependentLag()* creates lagged fields. Hence, expression

```
dataset <- TimeDependentLag(
  frame = instances, frame.date = 'date', frame.date.granularity = 'month',
  variables = 'Rain', lags = seq(from = 0, to = 4) )
```

creates lagged rainfall series; appended to the original data set.

```
# A tibble: 6 x 16
   Year Month Cases  Rain  minT  maxT  VCAP CasesLN datestr date        time
  <int> <ord> <int> <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>   <date>     <dbl>
1  1979 Jan      25   3.7  11.8  24    78.5    3.22 1979-01 1979-01-01     0
2  1979 Feb      25   3.2  11.3  23.5  56.6    3.22 1979-02 1979-02-01     1
3  1979 Mar      20   5.6  10.9  25.1 132.     3.00 1979-03 1979-03-01     2
4  1979 Apr      30   8.3  12    23.6 468.     3.40 1979-04 1979-04-01     3
5  1979 May      18   8.1  10.9  22.9 277.     2.89 1979-05 1979-05-01     4
6  1979 Jun      18   5.4  11.4  22.1 132.     2.89 1979-06 1979-06-01     5
# ... with 5 more variables: rain_lag_0 <dbl>, rain_lag_1 <dbl>,
#   rain_lag_2 <dbl>, rain_lag_3 <dbl>, rain_lag_4 <dbl>
```

The graphs of *fig. 2* illustrate the relationship between ln(cases) and each lagged rainfall series. The numeric suffix of each graph's title denotes the rain series lag, in months.

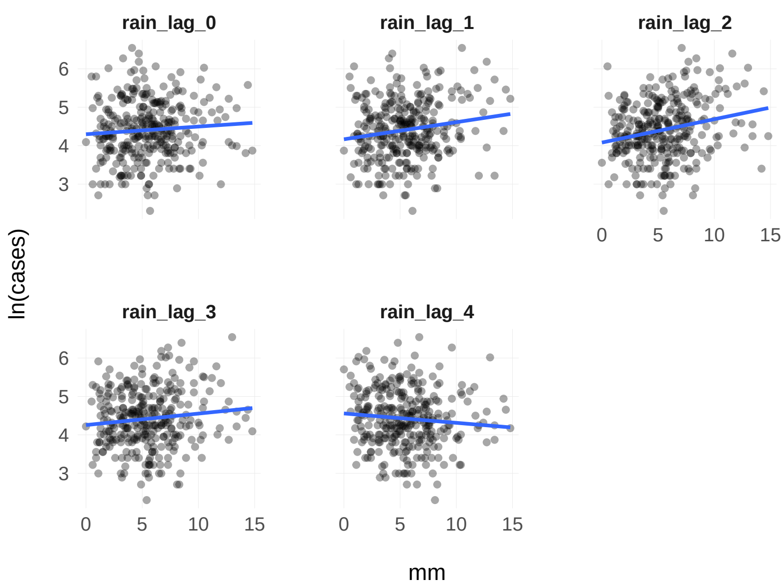

Figure 2: ln(cases) and the lagged rainfall series. the lags range from 0 to 4 months.

The degree of correlation between ln(cases) and each lagged rainfall series is quantifiable via the Pearson Correlation Coefficient. For each pairing the correlation values are:

```
          rain_lag_0 rain_lag_1 rain_lag_2 rain_lag_3  rain_lag_4
ln(cases) 0.07433887  0.1662839  0.2280889  0.1124358 -0.09247881
```

Trends

Is about . . .

Specified

Considering the time series model

$$\begin{aligned} Y(t) = \beta_0 + \beta_1 t + \beta_2 I(pmax(t - 50, 0)) + \beta_3 I(t > 225) \\ + \beta_4 minT(t - k) + \beta_5 maxT(t - k) + \beta_6 Rain(t - k) \\ + \mathcal{W}(t) + Z(t) \end{aligned} \tag{1}$$

for the Kericho malaria cases data, wherein

variable	description
t	time (months)
$minT$	mininum temperature
$maxT$	maximum temperature
$Rain$	rainfall (millimetres)
k	lag; $k = 2$ months
$\mathcal{W}(t)$	A Matern processwhereby $\kappa = 2.5$
$Z(t)$	Gaussian noise

The function *TimeDependentLag()* creates lagged fields. Hence, the lagged minimum temperature, maximum temperature, and rain fields:

```
variables <- c('minT', 'maxT', 'Rain')

T <- TimeDependentLag(
  frame = instances, frame.date = 'date', frame.date.granularity = 'month',
  variables = variables, lags = seq(from = 2, to = 2))
data <- T$frame
```

```
'data.frame': 310 obs. of  14 variables:
 $ Year      : int  1979 1979 1979 1979 1979 1979 1979 1979 1979 1979 ...
 $ Month     : Ord.factor w/ 12 levels "Jan"<"Feb"<"Mar"<..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Cases     : int  25 25 20 30 18 18 15 15 10 20 ...
 $ Rain      : num  3.7 3.2 5.6 8.3 8.1 5.4 5.5 6.1 5.7 5.6 ...
 $ minT      : num  11.8 11.3 10.9 12 10.9 11.4 10.2 10.1 10.2 11.1 ...
 $ maxT      : num  24 23.5 25.1 23.6 22.9 22.1 22.1 23 23.9 25.2 ...
 $ VCAP      : num  78.5 56.6 131.9 467.6 277 ...
 $ CasesLN   : num  3.22 3.22 3 3.4 2.89 ...
 $ datestr   : chr  "1979-01" "1979-02" "1979-03" "1979-04" ...
 $ date      : Date, format: "1979-01-01" "1979-02-01" ...
 $ time      : num  0 1 2 3 4 5 6 7 8 9 ...
 $ mint_lag_2: num  NaN NaN 11.8 11.3 10.9 12 10.9 11.4 10.2 10.1 ...
 $ maxt_lag_2: num  NaN NaN 24 23.5 25.1 23.6 22.9 22.1 22.1 23 ...
 $ rain_lag_2: num  NaN NaN 3.7 3.2 5.6 8.3 8.1 5.4 5.5 6.1 ...
```

Exercise 1: Model Fitting

Prior to fitting *Eq. 1*, records that have `NaN` values …

```
condition <- !is.na(instances$rain_lag_2) | !is.na(instances$mint_lag_2) |
  !is.na(instances$maxt_lag_2)
excerpt <- data[condition, ]
```

```
str(excerpt)
```

```
'data.frame': 0 obs. of  14 variables:
 $ Year      : int
 $ Month     : Ord.factor w/ 12 levels "Jan"<"Feb"<"Mar"<..:
 $ Cases     : int
 $ Rain      : num
 $ minT      : num
 $ maxT      : num
 $ VCAP      : num
 $ CasesLN   : num
 $ datestr   : chr
 $ date      : 'Date' num(0)
 $ time      : num
 $ mint_lag_2: num
 $ maxt_lag_2: num
 $ rain_lag_2: num
```