Birth Data

Laurie Davies
Faculty of Mathematics
University of Duisburg-Essen, 45117 Essen, Federal Republic of Germany
e-mail:laurie.davies@uni-due.de

1 Description of the analysis

The sample size is $n = 7305$. The data are shown in Figure 1

In a first step a trend was calculated using a polynomial of order 7. The result is shown in Figure 2. Aki Vehtari pointed out that this step is separate from the remainder of the analysis. This seems to me to be perfectly reasonable. I prefer to separate the trend, mostly due to the number of births, from the periodic components. The choice of trend is somewhat ad hoc and should be done in collaboration with a demographer. Failing this the calculated trend seems quite reasonable.

The trend was subtracted and the residuals were analysed using a simple robust linear regression with an intercept and the covariates

$$xs_j(i) = \sin(\pi ji/n) \text{ and } xc_j(i) = \cos(\pi ji/n),\ i, j = 1, ..., n,$$

giving $2n + 1 = 14611$ covariates in all. The xs and xc were treated in pairs xs_j and xc_j for no particular reason. The selection of the covariates was done as described in [Davies, 2016] using Huber's ψ-function with tuning constant $cnt = 1$ (page 174 of [Huber and Ronchetti, 2009]). The cut-off p-value was set to $p = 0.01$. This resulted in 109 pairs being included in the regression. Finally a robust regression was performed using all 218 covariates plus an intercept. The running time was 1072 seconds using Fortran 77. An R programme is very simple to write (as is indeed the Fortran programme) but is much slower. This so to speak terminates the analysis. The remainder consists of showing and interpreting the results.

2 The results

Figure 3 shows the complete regression function: Figure 4 the first 28 days in black and the days 7274:7301 in red, the increase in amplitude is apparent. The first year is shown in Fig 5. The first four (main) periodicities are 7, 3.5, 365.25 and 182.625 days.

The residuals are shown in Figure 6. There is a tendency for the negative residuals to increase in number and size with time.

There are 266 observations whose residuals are less than -400. This is 3.6% of the data set. Table 1 gives those days with at least two outliers.

1st	January	16	13th	June	2	24th	November	3
2nd	January	8	3rd	July	2	25th	November	5
9th	January	2	4th	July	14	26th	November	4
13th	February	4	5th	July	4	27th	November	6
15th	February	2	13th	August	2	28th	November	3
29th	February	3	1st	September	4	29th	November	2
1st	April	6	2nd	September	2	18th	December	4
13th	May	2	3rd	September	3	19th	December	4
17th	May	2	4th	September	2	20th	December	3
25th	May	2	5th	September	3	22nd	December	3
26th	May	3	6th	September	3	23rd	December	4
27th	May	2	7th	September	3	24th	December	14
28th	May	3	13th	October	3	25th	December	15
29th	May	2	31st	October	2	26th	December	8
30th	May	4	22nd	November	3	28th	December	2
31st	May	3	23rd	November	5	31st	December	2
						13th	All months	19

Table 1: Days of the year where number of outliers $<$ -400 is at least two followed by the actual number of outliers.

There are 162 observations whose residuals are greater than 400. This is 2.2% of the data set. Table 2 gives those days with at least two outliers.

Figure 7 shows the mean of the regression function for each of the 365 days of the year plus the 29th February. It may be compared with the figure shown in

(*) http://andrewgelman.com/2012/06/12/simple-graph-win-the-example-of-birthday-frequencies/.

Figure 8 shows the mean of the residuals for each of the 365 days of the year plus the 29th February. It may be compared with the figure shown in

(**) http://andrewgelman.com/2012/06/14/cool-ass-signal-processing-using-gaussian-processes/.

2.1 Valentine's Day, Halloween, Bastille Day and the 13th

Is there anything special about Valentine's Day, the 14th of February? The answer is clearly yes. In Table 2 it is listed with 6 outliers of size greater than 400. The previous day, the 13th February, is listed in Table 1 with 4 outliers of size less than -400. Both days are immediately apparent in Figure 8. In Figure 7 the order is reversed showing that the relying on the regression function rather than the residuals can lead to false conclusions.

Is there anything special about Halloween, the 31st of October? The fact that it is located at a local minimum of the averaged regression function (Fig-

3rd	January	5	7th	July	4	15th	December	2
4th	January	2	14th	July	2	18th	December	3
14th	February	6	8th	August	4	21st	December	3
27th	May	2	3rd	September	2	24th	December	2
29th	May	5	5th	September	4	27th	December	6
31st	May	2	9th	September	2	28th	December	6
12th	June	2	10th	September	3	29th	December	3
2nd	July	5	22nd	September	2	30th	December	13
3rd	July	4	10th	October	2	31st	December	8
6th	July	3	25th	November	2	13th	All months	0

Table 2: Days of the year where number of outliers > 400 is at least two followed by the actual number of such outliers.

ure 7) makes it no more special than any of the other days located at local minima, for example the 26th of April. Tables 1 and 2 do not help as there are two large positive outliers and two large negative outliers. The best evidence is provided by Figure 8 which shows a local minimum of -118 at this point. However if the Halloween is regarded as special then the 13th of October is even more so. There are three large negative outliers not compensated for by positive outliers and the local minimum of the averaged residuals at this point is -146. Moreover, to descend into the realm of pure speculation, people avoid giving birth on the 13th of October by giving birth three days earlier on the 10th of October with a local maximum of 109.0. The conclusion is that there is nothing special about Halloween.

Is there anything special about Bastille Day, the 14th of July? In favour is that there are two large positive outliers on this date, Table 2, not compensated for by negative outliers. Note that there are many more negative outliers, 266, than positive outliers, 162, so that perhaps positive outliers are more special. The special nature of Bastille Day is also supported by Figure 8 where the average residual function has a local maximum of 98.9. The conclusion is that there is slightly more evidence for Bastille Day than for Halloween. Maybe there are just sufficient Americans who regard it as chic, the mot juste, to have children born on Bastille Day for this to have a slight impact on the birth data.

Do parents avoid the 13th? This is clearly so. There are 19 large negative outliers as against 0 positive outliers. Of the 240 13ths 156 have negative residuals and the mean is -377.

3 Acknowledgment

I thank Aki Vehtari for comments on earlier versions of this paper.

References

[Davies, 2016] Davies, L. (2016). Stepwise choice of covariates in high dimensional regression. arXiv:1610.05131 [math.ST].

[Huber and Ronchetti, 2009] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, New Jersey, second edition.

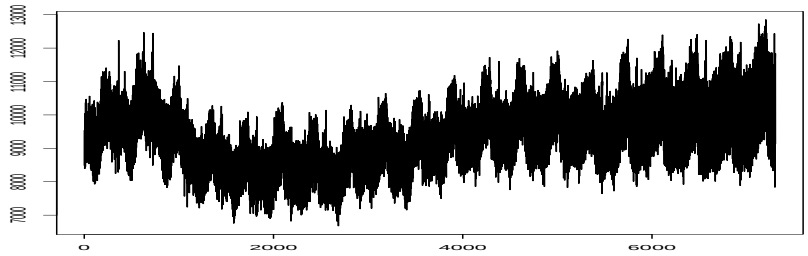

Figure 1: The birth data.

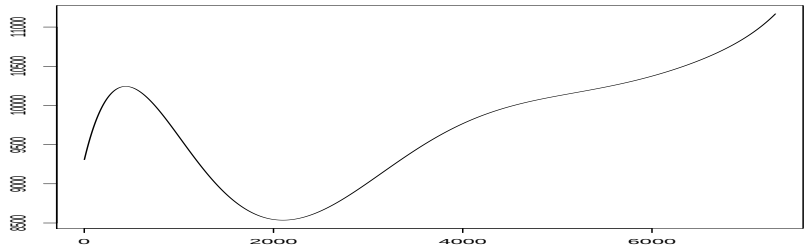

Figure 2: The trend.

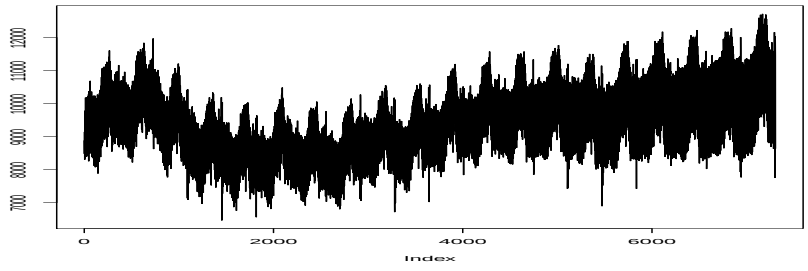

Figure 3: The robust regression function.

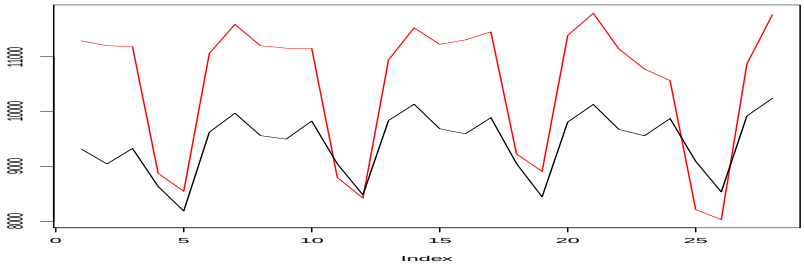

Figure 4: The regression function evaluated at the days 1:28 in black and 7274:7301 in red.

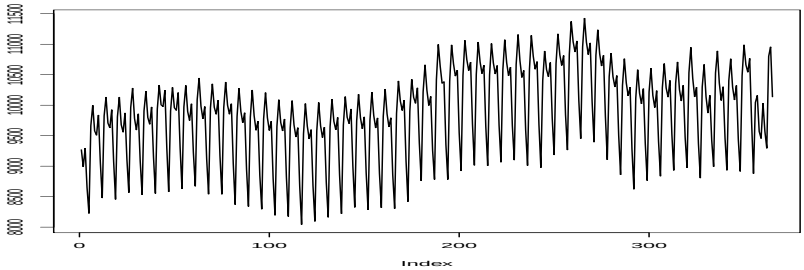

Figure 5: The regression function evaluated for the first year.

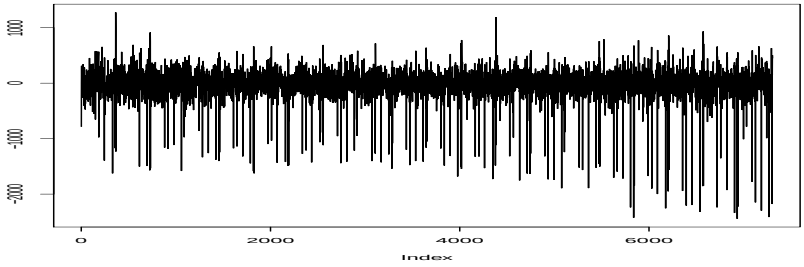

Figure 6: The residuals.

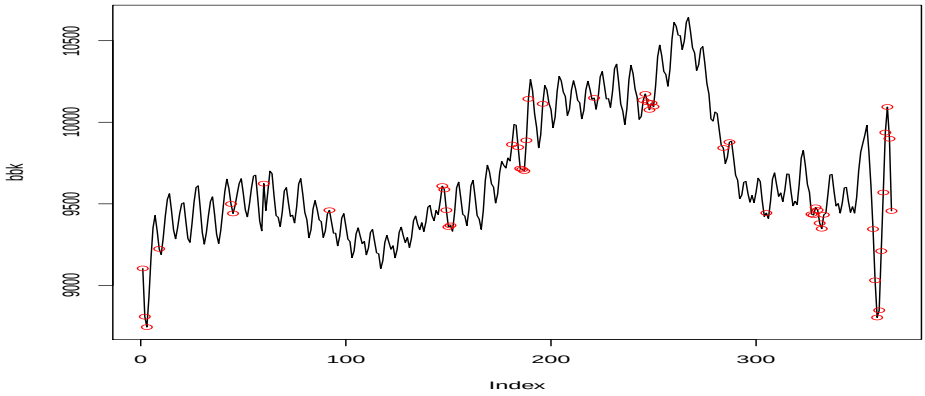

Figure 7: Mean of regression function plotted against day of the year. The marked observations are (i) New Year, (ii) 13th and 14th of February, (iii) 29th February, (iv) 1st of April, (v) end of May, (vi) 2nd-7th July, (vii) 14th July, (viii) 8th August, (ix) beginning of September, (x) 10th October, (xi) 13th October, (xii) 31st October, (xiii) end of November and (xiv) Christmas.

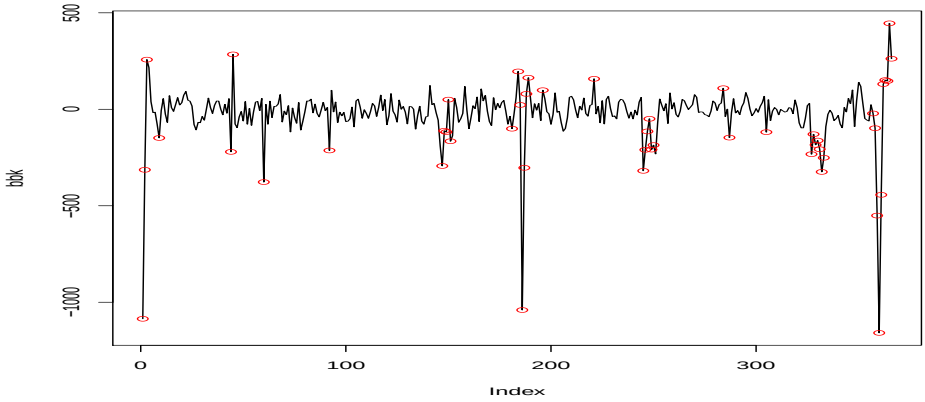

Figure 8: Mean of residuals plotted against day of the year. The marked observations are as in Figure 8