AMSC/MATH 420, Spring 2021
Solo Homework 1:
Fitting Linear Statistical Models to Data

Due Tuesday, February 2

1. A dataset that gives the total number of births in the US on each day of 1996 is posted on the course website under “Homework” as a text file `births1996.txt`. Using these data:

(a) Show that there is an important day-of-the-week effect on the way these numbers of births turn out. This is done by using a least squares fit the linear model generated by the basis $\{\chi_{\mathrm{Su}}, \chi_{\mathrm{Mo}}, \chi_{\mathrm{Tu}}, \chi_{\mathrm{We}}, \chi_{\mathrm{Th}}, \chi_{\mathrm{Fr}}, \chi_{\mathrm{Sa}}\}$, where the function χ_{Su} is defined by

$$\chi_{\mathrm{Su}}(t) = \begin{cases} 1 & \text{if day } t \text{ is Sunday,} \\ 0 & \text{if day } t \text{ is not Sunday,} \end{cases}$$

and the other basis functions are defined similarly. These are the seven *indicator functions* for days of the week. Which days of the week regularly have the smallest numbers of births?

(b) Plot the *residuals* of the fit from part (a). Recall from lecture that if the constant functions are in the span of your basis functions then the mean of the *residuals* should be zero. Because the indicator functions sum to the constant function 1, that should be the case here. (If it isn’t then you’re not doing the computations correctly). The residuals should be a sequence of numbers that looks more or less like

- a curvilinear trend
- plus “noise”
- except for relatively few “outlier” days.

Here “noise” means an apparently patternless sequence of numbers which looks like a sequence of independent, identically distributed values across time.

(c) Identify and examine the “outlier” days in (b). Was there anything special about these days that might help account for why they are outliers?

(d) Add some basis functions to the linear model used in part (a) to acount for

- the “outliers”
- the curvilinear trend.

Use a least squares fit to capture these phenomena as simply as possible. Your job to choose a suitable set of basis functions. There is no “right” basis, but some are more suitable than others – see, in particular, the comments in (e). Justify your choice.

(e) For your new fit, compute the *residuals*. Ideally, they should look like pure noise. There should not be an obvious trend in the residuals; such a trend may suggest something “missing” from your basis functions.

(f) Discuss the function you fitted in (d) in relation to real-life factors that vary over the course of year. Is there significant seasonal variation, and why or why not?

2. Consider the following dataset:

x	0	1	4
y	10	7	10

(g) Find by hand a quadratic function $x \mapsto f(x) = ax^2 + bx + c$ that best fits this dataset, in the least-squares sense.

(h) For the function f computed at (g), find its extreme values (minimum and maximum) and where those extreme values are achieved. Assume the domain of definition for variable x is $[0, 10]$. In other words, solve:

$$\underset{x \in [0,10]}{\text{minimum}} \quad f(x) \qquad , \qquad \underset{x \in [0,10]}{\text{maximum}} \quad f(x)$$

(i) For the same dataset at part (g), find by hand the linear function $x \mapsto g(x) = dx + e$ that best fits the data, in the least-squares sense.

(j) For the function g computed at (i), find its extreme values (minimum and maximum) and where those extreme values are achieved. Assume the domain of definition for variable x is $[0, 10]$. In other words, solve:

$$\underset{x \in [0,10]}{\text{minimum}} \quad g(x) \qquad , \qquad \underset{x \in [0,10]}{\text{maximum}} \quad g(x)$$