AMSC/MATH 420, Spring 2020
First Solo Homework:
Fitting Linear Statistical Models to Data

Due Tuesday, February 4

Problem I

A dataset consisting of the national total numbers of births in the US on each day in 2003 can be found on the course web page as a text file `births2003.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 $\{\chi_{\text{Su}}, \chi_{\text{Mo}}, \chi_{\text{Tu}}, \chi_{\text{We}}, \chi_{\text{Th}}, \chi_{\text{Fr}}, \chi_{\text{Sa}}\}$, where the function χ_{Su} is defined by

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

and the other basis functions are defined similarly. Which days of the week regularly have the smallest numbers of births?

(b) Plot the *residuals* of the fit from part (a). What remains is a sequence of numbers that looks more or less like a curvilinear trend plus "noise" except for relatively few anomalous days. Here "noise" means an apparently patternless sequence of numbers which, either visually or by some other criterion, looks like a sequence of independent, identically distributed values across time.

(c) Identify and examine the anomalous days in (b). Was there anything special about these days that might help account for anomalies?

(d) Add some basis functions to the linear model used in part (a). Use a least squares fit to capture as simply as possible the common curvilinear trend remaining in (b) after adjusting for day-of-week effects and possibly for the "outliers" you found in (c). It is your job to decide on 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)].

(e) For your fit, compute the *residuals*: the original data points (numbers of births) minus the day-of-week adjustment and the trend function you found. 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. (If it isn't then you're not doing the computations correctly). Ideally, 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?

Problem II

Consider the following dataset:

x	0	1	3
y	2	0	2

(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:

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