AMSC/MATH 420, Spring 2023
First Solo Homework:
Fitting Linear Statistical Models to Data

Problem I

A dataset consisting of the total electrical energy demand in the Mid-Atlantic region of the US on each day in 2020 can be found on the course web page as a text file `TotalEnergyDemand_MIDA_2020.txt`[1]. First number if the daily energy demand on January 1, 2020, expressed in Megawatt hour (MWh). Note there are 366 numbers, one number for each day, and the year 2020 was a leap year. Using these data:

(a) Determine whether there is an important day-of-the-week effect on the daily energy demand. 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. Which days of the week regularly have the smallest energy demand?

(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.

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

(c) 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 (b). 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 (d)].

(d) For your fit, compute the *residuals*: the original data points (daily energy demand) 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.

(e) 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?

---

[1] See: https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48 Click on DownloadData and then choose Balancing Authority/Region Files. E.G., https://www.eia.gov/electricity/gridmonitor/knownissues/xls/Region_MIDA.xlsx

Problem II Consider the following Matlab code:

```
f1 = 0;
f2 = 1;
h = 0.1;
y = 1;
for t=0:h:1-h
    f1 = 2*t-y^4;
    f2 = 2*(t+h)+(y+h*f1)^2;
    y = y + f1*h;
end
% The output of this code is y
```

The output of this code is y, and the time step is h.

(a) What initial value problem does this Matlab code solve? What method does it use?

(b) If the error with respect to the exact solution at t=1 is 10-2 what should the step size be to achieve an accuracy of 10-4 (that is an error less than 10-4)?

Problem III
Consider the following dataset:

x	0	1	4
y	10	7	10

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

(b) For the function f computed at (a), 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:

$$\begin{array}{ll} \text{minimum} & f(x) \\ x \in [0,10] & \end{array} \quad , \quad \begin{array}{ll} \text{maximum} & f(x) \\ x \in [0,10] & \end{array}$$

(c) Repeat the same problems (a) and (b) for the class of linear functions. This means:
(c1) For the same dataset, find by hand the linear function $x \mapsto g(x) = dx + e$ that best fits in the least-squares sense.
(c2) For the function g computed at (c), 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:

$$\begin{array}{ll} \text{minimum} & g(x) \\ x \in [0,10] & \end{array} \quad , \quad \begin{array}{ll} \text{maximum} & g(x) \\ x \in [0,10] & \end{array}$$