CSC 121, Spring 2017 — Small Assignment #3

Worth 5% of the course grade. Due by the start of class on March 3, to be handed in using MarkUs. This assignment may be handed in late, with a 20% penalty, by start of class on March 7. Assignments will not usually be accepted after that. Contact the instructor as soon as possible if you have a legitimate excuse (such as documented illness) for handing in the assignment late (without penalty).

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you shouldn't leave a discussion with someone else with any written notes (either paper or electronic).

In this assignment, you will write and test an R function filling in missing observations in a vector of sequential observations, and apply it in an R script file for reading and modifying data in an R data frame.

As discussed in the week 7 lectures, a "missing" observation can be indicate in R by a special NA value, which can occur anywhere a number, string, or logical value would otherwise appear. Since missing observations are very common in practice, handling them in some way is an important part of statistical analysis. One approach is to replace each missing observation with some value found from the other observations. A simple approach, for example, is to replace them with the average of all the non-missing observations, but this will often be too simple, producing misleading conclusions.

In this assignment you will implement another approach, that is applicable when the observations come in a sequence (eg, are a time series), for which order is meaningful. In this situation, one might decide to replace a missing observation by the average of the observation before and the observation after. However, either or both of those might be missing as well. So more generally, the method is to fill in a missing observation by linearly interpolating between the nearest non-missing observations that come before and after the missing observation. Missing observations at the beginning of the sequence (with no non-missing observations before) are filled in with the first non-missing observation, and similarly for missing observations at the end of the sequence. If all observations are missing, they remain missing, since there is no data at all to use to fill them in.

You should write a function called `na_interpolate` that takes as its only argument a numeric vector, and returns as its value the result of replacing missing values (if any) in this vector according to the method described above.

Here is an example call of this function:

```
> na_interpolate (c (4,NA,5,NA,NA,NA,8,10,NA,NA))
 [1]  4.00  4.50  5.00  5.75  6.50  7.25  8.00 10.00 10.00 10.00
```

The general formula for filling in a missing value by linear interpolation is as follows:

$$\frac{d_1 \times x_2 + d_2 \times x_1}{d_1 + d_2}$$

Here, d_1 is the distance to the closest non-missing observation before the one to be filled in, x_1 is the value of this observation, d_2 is the distance to the closest non-missing observation after the one to be filled in, and x_2 is the value of this observation. So, for example, the missing observation after the 5 in the example above is filled in as

$$\frac{1 \times 8 + 3 \times 5}{1 + 3}$$

You should create a test script with examples such as the one above in order to test your `na_interpolate` function.

Once your `na_interpolate` function is working, you should create a script to apply it to a made-up data set recording weather hour-by-hour on two days, which you can read (as a data frame) from the course web page, as follows:

```
    read.table("http://www.cs.utoronto.ca/~radford/csc121/sma3-data",head=TRUE)
```

This data frame has several variables (columns), including `temperature` and `pressure`. Your script should fill in missing values (separately) in these two variables using your `na_interplate` function, and then print the modified data frame.

You should hand in three script files, one with only the definition of your `na_interpolate` function, one with your tests of this function, and one that uses this function to fill in missing values in the weather data described above. You should also hand in the output of the last two scripts, as two text files.

Here is a suggested approach to writing the `na_interpolate` function. Start by creating a vector the same length as the argument vector, which at position i contains the index in the argument vector of the last non-missing observation at or before i (or zero if there is none). Similarly, create a vector that at index i contains the index of the earliest non-missing observation at or after i. Then use these two vectors to modify elements of the argument vector by filling in interpolated missing values.

You should not use features that we have not covered yet when doing this assignment (except for minor features that don't affect the overall method used). In particular, you should not try to use some R package for filling in missing values that may already implement this method!