Homework 5: It's Not the Heat that Gets to You, It's the Sustained Heat with Pollution

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 18 February 2013

The data set `chicago`, in the package `gamair`, contains data on the relationship between air pollution and the death rate in Chicago from 1 January 1987 to 31 December 2000. The seven variables are: the total number of (non-accidental) deaths each day (`death`); the median density over the city of large pollutant particles (`pm10median`); the median density of smaller pollutant particles (`pm25median`); the median concentration of ozone (O_3) in the air (`o3median`); the median concentration of sulfur dioxide (SO_2) in the air (`so2median`); the time in days (`time`); and the mean daily temperature (`tmpd`).

We will model how the death rate changes with pollution and temperature. Epidemiologists tell us that risk factors usually multiply together rather than adding, so we will fit additive models to the logarithm of the number of deaths. These problems *can* all be done using either the `gam` or the `mgcv` packages for fitting additive models, but you will probably find it easier to use `mgcv`.

Warning: The bootstrapping in Problem 7g might be time-consuming; don't wait to the last minute.

1. (5) Load the data set and run `summary` on it.
   (a) (1) Is temperature given in degrees Fahrenheit or degrees Celsius?
   (b) (2) The pollution variables are negative at least half the time. What might this mean?
   (c) (2) We will ignore the `pm25median` variable in the rest of this problem set. Why is this reasonable?
2. (10) Fit a spline smoothing of `log(death)` on time. (You can use either `smooth.spline` or `gam`.)
   (a) (3) Plot the smoothing spline along with the actual values.
   (b) (4) There should be four large outliers, right next to each other in time. When are they? For full credit, give calendar dates, not day numbers. (*Hint:* day 0 was 31 December 1993.)
   (c) (3) How many degrees of freedom did your smoothing spline have? Add curves to the plot which would result from using 10, 50, 100 and 2000

degrees of freedom. (Make sure these differ in color and/or line-style.) What happens to the spline curves as you change the degrees of freedom?

3. (15) Use `gam` to fit an additive model for `log(death)` on `pm10median`, `o3median`, `so2median`, `tmpd` and `time`. Use spline smoothing for each of these predictor variables.

   (a) (7) Plot the partial response functions, with partial residuals. Describe the partial response functions in words.

   (b) (4) Plot the fitted values as a function of time, along with the actual values of `log(death)`.

   (c) (4) Are the outliers still there? Are they any better?

4. (15) It is medically implausible to supposed that deaths on day t are only due to heat or pollution on that day, and not on earlier ones.

   (a) (8) Suppose that on any given day, we want to know the average value of some variable over today and the previous k days. Explain how the following code computes that.

```
lag.mean <- function(x, window) {
  n <- length(x)
  y <- rep(0,n-window)
  for (t in 0:window) {
    y <- y + x[(t+1):(n-window+t)]
  }
  return(y/(window+1))
}
```

   In particular, how is k related to the arguments?

   (b) (7) Create a new data frame with the same column names as `chicago`, but where, on each day, the value of the pollution concentrations and temperature is the average of that day's value with the previous three days. How many rows should this data frame have? Make sure that the `time` and `death` columns are properly aligned with the new, time-average predictor variables. How can you check that this is working properly?

5. (10) Fit an additive model, as in problem 3, with the time-averaged pollution and temperature variables. (Do not average `time` or `death`.)

   (a) (5) Plot the partial response functions and their partial residuals.

   (b) (5) Plot the fitted values as a function of time, and the actual values. What has happened to the outliers?

6. (15) *Variable examination*

   (a) (4) Find the rows in the data frame (with the time-averaged values) corresponding to the large-death outliers. Look at all variables for them, and

for three days on either side. Now compare this to the same stretch of time a year earlier. Which two variables, aside from `death`, are unusually high or low around the outliers?

(b) (7) Re-fit the model from problem 5, with an interaction between the two variables you just picked out. Plot the partial response functions.

(c) (4) Plot the fitted values versus time. What has happened to the outliers?

7. (25) Using the last model you fit, we will consider the predicted impact of a 2° Celsius increase in temperature on `log(death)`, taking the last full year of the data as a baseline.[1].

   (a) (1) Prepare a data frame containing only the last full year of the data. What is the average predicted value of `log(deaths)`?

   (b) (1) Modify this data frame to increase all temperatures by 2°C.

   (c) (3) Find the new average *change* in the predicted values of `log(deaths)` associated with a 2°C warming.

   (d) (5) Find a standard error for this average predicted change, using the standard errors for the prediction on each day, and assuming no correlation among them. Also give the corresponding Gaussian 95% confidence interval.

   (e) (5) Find the predicted change in the number of deaths (not change in `log(death)` from a 2°C warming over the course of a whole year. *Hint:* remember that $\overline{e^x} \neq e^{\overline{x}}$.

   (f) (5) Explain how you could use bootstrapping to give a 95% confidence interval for the average increase in `log(death)` over the year. More credit will be given for more precise, complete and clear explanations.

   (g) (5) Implement your bootstrapping scheme and give the confidence interval.

8. (5) Explain at least one reason that this estimate of what would happen if Chicago warmed by 2°C might be systematically flawed. (Do not repeat the problems mentioned in the footnote. Doubts that such warming will happen do not count.) For full credit, suggest ways of improving the estimates.

[1] 2°C is in the middle range of current projections for the global average effect of climate change by the end of this century (`http://www.ipcc.ch/publications_and_data/ar4/wg1/en/contents.html`)q. Of course it's unrealistic to suppose that would be an even shift throughout the year, or for that matter that Chicago would necessarily warm by the average amount. In fact, some of the models (`http://www.ipcc.ch/publications_and_data/ar4/wg1/en/ch11s11-5-3.html`, Figure 11.11) have 4°C of warming in the middle of their prediction intervals for central North America.