Homework 6: It's Not the Heat That Gets to You, It's the Conjunction of Sustained Heat with Atmospheric Pollution

36-402, Spring 2024, Section A

Due at 6 pm on Thursday, 14 March 2024

AGENDA: Practice with generalized additive models, including interactions and model-building; practice re-shaping data frames; practice connecting probability theory to modeling choices.

This week's assignment revisits the `chicago` data set, previously seen in HW 1 and HW 5. (You are always welcome to use anything from th solutions to earlier homeworks, but you are particularly encouraged to do so here.)

We begin with some theory problems, which will help guide the data analysis. Note that those problems are all show-that problems, so the conclusions are all given to you.

1. (2) *Binary random variables* T is a binary variable, either `0` or 1. Show that $\mathbb{E}[T] = \mathbb{P}(T = 1)$, and that $\mathrm{Var}[T] = \mathbb{E}[T](1 - \mathbb{E}[T])$.
2. *Binary to binomial* $T_1, T_2, \ldots T_q$ are independent, identically distributed binary variables, where $\mathbb{E}[T_i] = p$. Define $S_q \equiv \sum_{i=1}^{q} T_i$.
   (a) (2) Show that $\mathbb{E}\left[S_q\right] = q\,p$ and $\mathrm{Var}\left[S_q\right] = q\,p(1-p)$.
   (b) (2) Show that $\mathbb{P}\left(S_q = s\right) = \frac{q!}{s!(q-s)!} p^s (1-p)^{q-s}$.
3. *Binomial to Poisson* We continue the setting of Q2, but now we let $q \to \infty$ while $\mu \equiv q\,p$ stays constant, that is, $p = \mu/q$. In the following, all the limits are as $q \to \infty$.
   (a) (2) Show that $\mathbb{E}\left[S_q\right] \to \mu$.
   (b) (2) Show that $\mathrm{Var}\left[S_q\right] \to \mu$.
   (c) (1; harder math than the rest) Show that $\mathbb{P}\left(S_q = s\right) \to \frac{\mu^s e^{-\mu}}{s!}$. *Hints*: (i) you can take as given that $e^x = \lim (1 + x/q)^q$; (ii) you can do this without using Stirling's approximation, but if you *do* use it, you need the form $\log k! \approx k \log k - k$, and not just $\log k! \approx k \log k$.

(d) (3) The random variable Y whose probability mass function is $\mathbb{P}(Y = y; \mu) = \frac{\mu^y e^{-\mu}}{y!}$ is said to have a Poisson distribution, $Y \sim \text{Pois}(\mu)$. Using the earlier parts of this question, show that $\mathbb{E}[Y] = \text{Var}[Y] = \mu$.

(e) (3) Explain the statement "A Poisson distribution totals up a large number of individually-rare events".

4. *Poisson likelihood* $Y_1, Y_2, \ldots Y_n$ are all independent Poisson random variables, with means $\mu_1, \mu_2, \ldots \mu_n$ (which may or may not be equal). The log-likelihood of seeing the data $y_1, y_2, \ldots y_n$ is

$$L(\mu_1, \ldots \mu_n) \equiv \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i; \mu_i) \tag{1}$$

(a) (4) Show that each summand in Eq. 1 is itself a sum of three terms: one term which is a function of y_i but not μ_i, one term which is a function of μ_i but not y_i, and one term which is proportional to $y_i \log \mu_i$.

(b) (1) Assume that each μ_i can be adjusted without affecting any of the other μ_j's. Show that the log-likelihood is maximized at $\hat{\mu}_i = y_i$ for each i.

5. *Theory to modeling*

(a) (4) Using the previous problems, explain why it is reasonable to try modeling the number of people who die each day in a large city as a Poisson-distributed random variable.

(b) (4) Using the previous problems, explain why it is reasonable to try modeling $\log \mu$ as a function of predictor variables. That is, why is log a natural link function for Poisson regression? *Hint:* Think about what we did with logistic regression (see chapter 11).

(c) (2) If each predictor variable makes an additive contribution to $\log \mu$, how will they combine to produce μ?

6. *Time trend* Fit a spline smoothing of `log(death)` on time. (You can use either `smooth.spline` or `gam`.)

(a) (2) Plot the actual values of `death` (not `log(death)`) as a function of time. Add a curve showing the estimates from the smoothing spline.

(b) (3) 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. (Day `0` was 31 December 1993.)

(c) (1) Why do I ask you to smooth `log(death)` against time, but to plot `death` against time?

7. *First GAM* Use `gam` to fit a generalized additive model for `death` on `pm10median`, `o3median`, `so2median`, and `tmpd`. Use spline smoothing for each of these predictor variables. Make sure that you treat `death` as Poisson-distributed, and that

you use a logarithmic link function. *Hint:* Because of some missing-data issues, some plots later may be easier to make if you set the `na.action=na.exclude` option when estimating the model.

(a) (4) Plot the partial response functions, with partial residuals. Describe each partial response functions in words. Carefully describe meaning of the units on the vertical axes of the plots.

(b) (3) Plot the fitted values as a function of time, along with the actual values of `death`. *Hint:* Be careful about the NA values.

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

8. *Time averages* Medically, it makes more sense to suppose that deaths on day t are due conditions over the previous few days, and not just on the conditions on day t. This problem re-shapes the data set to let us model this.

(a) (3) 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 of `lag.mean`?

(b) (3) 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. (*Hint:* you will want to do different things to different columns of `chicago`.) 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?

9. *Second GAM* Fit a generalized additive model, as in Q7, with the time-averaged pollution and temperature variables. (Do not average `time` or `death`.)

(a) (3) Plot the partial response functions and their partial residuals. Describe the shapes of the partial response functions.

(b) (3) Plot the fitted values as a function of time, and the actual values. What has happened to the outliers?

10. *Variable examination and re-specification*

(a) (3) 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) (3) Re-fit the model from problem 9, with an interaction between the two variables you just picked out. Plot the partial response functions. *Hint:* look at examples of interactions in chapter 8.

(c) (3) Plot the fitted values versus time. What has happened to the outliers?

11. *Examining residuals*

(a) (1) Show that if $Y \sim \text{Pois}(\mu)$, then $R = \frac{Y-\mu}{\sqrt{\mu}}$ has $\mathbb{E}[R] = 0$, $\text{Var}[R] = 1$.

(b) (3) For your final estimated model, create plots of the Pearson residuals against the fitted values, and against each of the predictor variables in your model. (If you interacted X_1 and X_2, make a plot of residuals against X_1 and another plot of residuals against X_2, don't try to make a 3D plot of residuals against X_1 *and* X_2.) Does it look like the residuals have mean 0 everywhere? How can you tell? *Should* they have mean 0 everywhere?

(c) (3) Repeat the previous question, but for plots of the squared residuals. Do they look like they have mean 1 everywhere? How can you tell? Should they have mean 1 everywhere?

(d) (3) Explain what Q11a has to do with Q11b and with Q11c.

12. (9) *Conclusions* Pretend that the Chicago city government would like to reduce the number of deaths in the city, and would especially like to reduce the risk of episodes where very large numbers of people die suddenly. Write at least one paragraph, but no more than one page, about what the city should do. For full credit, be as specific as possible in backing your recommendations up with findings from your data analysis.

13. (1) *Timing* How long, roughly, did you spend on this assignment? How much of that time was spent on math, on coding/debugging, and on writing?

PRESENTATION RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All plots and tables are generated by code included in the R Markdown file. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.