Generalized Linear Models II

Today you use GLMs to explore a data set on ozone pollution: https://is.gd/YeptW1

People chronically exposed to ground-level ozone (not ozone "holes" in the troposphere) are more likely to suffer from asthma, bronchitis, and cardiopulmonary problems. In principle, ozone levels are increased by sunlight (here radiation, or "rad") and warmer temperatures, but might be diluted by winds.

Your Mission: Use the techniques from the GLMs I class and prior classes to obtain the most plausible model of ozone levels, where the assumption of the statistical distribution is also most legitimate. You have two jobs: *most plausible* AND *most legit.* Plausible comes from AIC comparisons. Legit comes from distribution family and collinearity reduction.

Here is an approximate sequence of steps / tricks:

1. **`pairs`** to squint at all variables in a data set. This can help you decide if you should use:
   (a) interactive terms (e.g., temp*wind), or
   (b) simple (y = a + bx), or
   (c) quadratic (e.g., quadratic; $y = a + bx + cx^2$) functions for predictors. And yes, a quadratic function is a linear model, because it's all about the coefficients, and x is not raised here to some solved-for coefficient (e.g., $y = ax^Z$).
2. use **`gamlss`** or **glm** or **glm.nb** in MASS) to try different families
3. use **`scale`** to make all predictors in units of SD (and thus more comparable), despite having different units and ranges
4. identify your most efficient model based on **`AICc`** (e.g., with the bbmle package)
5. evaluate assumptions and collinearity for your favored model – using tools appropriate to the GLM package you used

Having obtained your most plausible and legit model, here is a nice tool to visually compare predictors. A **classification and regression tree** (aka CART) recursively partitions the response variable into subsets based on its relationship to predictor variables. The predictor variable at the first split yields the greatest change in explained deviance (like minimizing SSE in an ANOVA).

Install and load the **`tree`** package and then run this command on your final, "best" model (*where you fill in the `predictors`*):

```
mytree <- tree(ozone ~ predictors)
plot(mytree)
text(mytree)
```

Does this make sense when compared to the scaled coefficients in your output? A limitation: a CART cannot represent interactive terms.