Chapter 10:Scatterplot Smoothers and Generalized Additive Models

An Online Course

Sponsored by The Georgia R School

Presented by Geoffrey S. Hubona

The Men’s Olympic 1500m

- Modern Olympics began in 1896 in Greece.
- Men’s 1500 meter foot race has always been the star track event.
- Winning times continue to decline 1896-2004.
- ***Can we use these winning times as the basis of a statistical model to predict winning times in future Olympics?***

Air Pollution in U.S. Cities

- Data on Air Pollution in 41 U.S. cities.
    - Annual mean concentration of **sulphur dioxide** (SO2), in micrograms per cubic meter.
- ***Which aspects of climate and human ecology determine pollution?***
    - **`temp:`** average annual temperature (F)
    - **`manu:`** number of manufacturers with > 20 employees
    - **`popul:`** population size
    - **`wind:`** average wind speed
    - **`precip:`** average annual precipitation
    - **`predays:`** average annual number of days with precipitation

Risk Factors for Kyphosis

- 81 Children Undergoing Corrective Surgery of the Spine
- **Kyphosis** is a medical condition in children characterized by an outward curvature of the spine.
- ***What are risk factors for kyphosis following surgery?***
  - `Age:` age in months
  - `Start:` starting vertebral level of the surgery
  - `Number:` number of vertebrae involved

Smoothers and GAMs

- How could we let the functional form of the relationship between the response variable and the predictor variables be estimated by the data?
- The secret is to ***replace the global estimates from the regression models with local estimates***.
  - Statistical dependency between two variables is described not with a single global parameter like a regression coefficient, but with a series of local estimates.
- This approach **is useful when**:
  - Relationship between variables is complex and not easily fitted by standard linear or non-linear models.
  - No *a priori* reason to use a particular model.
  - We would like the data to suggest the appropriate functional form of the relationship.

Smoothers (Everitt and Hothorn)

- Non-parametric ‘smoothers’ summarize the relationship between two variables with a line drawing.
- The simplest smoother is a ***local weighted regression*** or ***lowess*** fit:

$$y_i = g\left(x_i\right) + \varepsilon_i, \quad \text{where} \quad i = 1,\ldots,n$$

- Two parameters control the shape of a **Lowess** curve:
  - **Smoothing parameter**, α, the ***span***, or width of the local neighborhood; and
  - **Lambda**, λ, the ***degree of the polynomials*** that are fitted by this method.
- Selecting values for these parameters requires judgment and, often, trial and error.

Generalized Additive Models (E&H)

- More general, semi-parametric approach to modeling scenarios ***with more than one explanatory variable*** (like US air pollution data).
- Can model relationship between response variable and each explanatory variable using:
  - **Linear** coefficient (parametric)
  - **Lowess** smoothers (non-parametric)
  - **Cubic splines** smoothers (parametric)
- GAMs are a type of GLM in that the expectation of the value of the response variable is modeled as a sum of (parametric and non-parametric) functions.
  - Each explanatory variable can have its own unique parametric or non-parametric form.

Variable Selection and Model Choice

- Quantifying the influence of covariates goes beyond estimating a coefficient
  - Careful implementation of **variable selection**: *what subset of covariates enter the model?*
  - Careful **model choice**: *Linear*? *Non-Linear*?
- Two general approaches:
  - Fit models using a ***target function with a penalty term*** that increases in severity as model complexity increases.
  - Iteratively fit ***simple, univariate models which sum*** to a more complex generalized additive model.
    - Known as ***boosting***.
    - Need a ***stop criterion*** for the iterative model-fitting algorithm.