Generalized Linear Additive Models

Emily Corcoran & Kathryn Shore

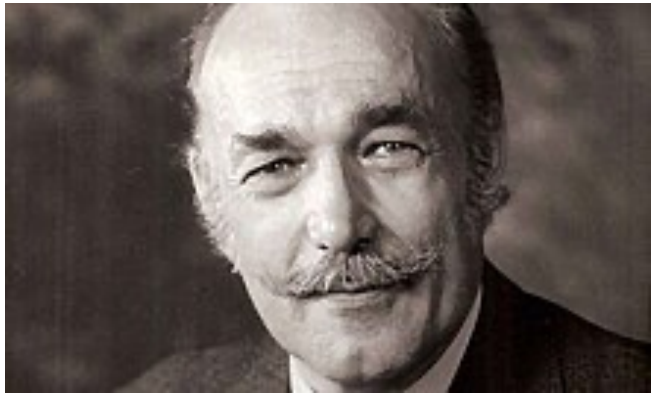
What are Generalized Linear Models?

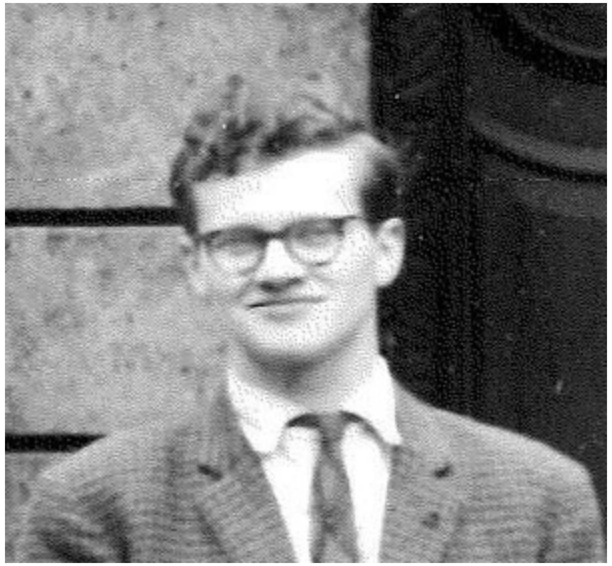
Linear Models

- Gauss 1809

Generalized Linear Models

- Nelder & Wedderburn 1972





What are Generalized Linear Models?

A generalized linear model is a flexible generalization of ordinary linear regression

- GLMs allow the extension of linear modelling ideas to more response types including count data or binary responses
- Includes linear regression, logistic regression, and Poisson regression
- Normally we assume residuals normally distributed; for GLM's it does not have to be

What are Generalized Linear Models?

A generalized linear model is a flexible generalization of ordinary linear regression

- We can use GLM’s (instead of LMs or MLR) when…
  - residuals not normally distributed
  - data is heteroscedastic
  - data is non-linear

What are Generalized Linear Models?

A generalized linear model is a flexible generalization of ordinary linear regression

1. Systematic component
   (the function that links the predictor to the outcome) $\beta_0 + \beta_1 x$

2. Link function
   (function that "bends the line") $(\beta_0 + \beta_1 x)^2$, $e^{\beta_0 + \beta_1 x}$, $\log(\beta_0 + \beta_1 x)$

3. Random component
   (epsilon does not have to be normally distributed) Normal, Poisson, Gamma, Binomial

What are Generalized Linear Models?

An ordinary linear model is a special case of a GLM

1. Systematic component $\beta_0 + \beta_1 x$
2. Link function Identity ($g(y) = y$)
3. Random component Normally distributed

What are Generalized Linear Models?

An ordinary linear model is a special case of a GLM

- In R...
  - `lm()` fits models of the form $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
  - `glm()` fits models of the form $g(Y) = X\beta + \epsilon$ where $g()$ and the distribution of ϵ need to be specified.
  - The default link function for `glm()` is the identity function and the default error distribution is Normal

With these defaults, `lm()` and `glm()` fit the same model

Example Scenarios for GLMs

- Predicting # of times people go to therapy (non-negative count data)
- Predicting death from heart disease (binary data)
- Predicting the grade someone will get in a class (ordinal data)

Pros and Cons of GLMS

- The response does not have to be normally distributed
- Able to deal with categorical predictors
- Modelling is interpretable
- Flexibility
- Makes strict assumptions about shape
- Can be prone to overfitting
- Can be sensitive to outliers

CODE DEMO

Generalized Linear Additive Models

Emily Corcoran & Kathryn Shore

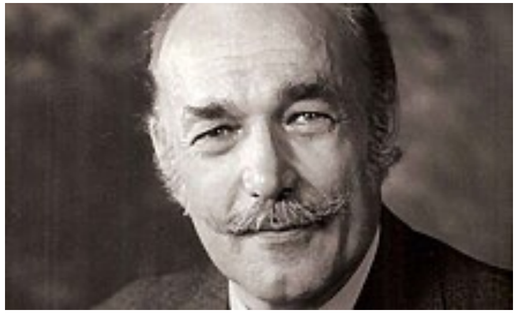
What are Generalized Additive Models?

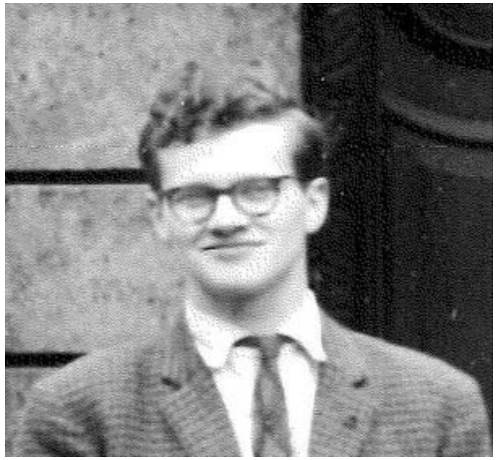
Linear Models

- Gauss 1809

Generalized Linear Models

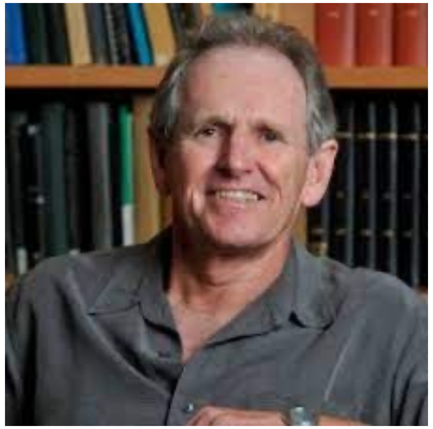
- Nelder & Wedderburn 1972





Generalized Additive Models

- Hastie & Tibshirani 1986



What are Generalized Additive Models?

A GAM is a GLM in which the response variable depends linearly on smooth functions of predictor variables

- GLMs are extensions of multiple linear regression: the problem of predicting Y on the basis of several predictors $X_1, X_2, \ldots, X_p$.
- allow us to extend a linear model to allow non-linear functions while maintaining additivity
- provide a compromise between linear and fully nonparametric models

What are Generalized Additive Models?

A GAM is a GLM in which the response variable depends linearly on smooth functions of predictor variables

- The main difference is GLMs assume a fixed form of the relationship between the dependent variable and the covariates, but GAMs do not assume a specific form *a priori*
- In GLMs we have a weighted sum of the covariates, in GAMs we have a sum of smooth functions
- GAMs more flexible

GAM Example

MLR is just a special case of a GAM

Multiple linear regression model:

$$y_i = \beta_0 + \beta_{1x_{i1}} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

We can extend this model to allow non-linear relationships by replacing each linear component ($\beta_j x_{ij}$) with a smooth non-linear function $f_j(x_{ij})$:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

e.g., $\mathrm{y_i} = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \mathrm{sqrt}(x_{i2}) + \beta_3 \log(x_{i3}) + \epsilon_i$

GAM Example

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

This is an additive model because we calculate a separate f_j for each X_j and add together all of their contributions.

GAM Example

- CORIS survey – Coronary Risk Factor Study survey
- Generalized Additive Models: Some Applications by Hastie and Tibshirani

GAM Example

- CORIS survey
- Investigating intensity of ischemic heart disease risk factors in rural areas of South Africa
- Risk factors included
    - Systolic blood pressure
    - Cumulative tobacco
    - Cholesterol ratio
    - “Type A” (measure or psychosocial stress on Bortner Scale)
    - Age
    - Total energy
    - Family history (binary)
- Fitted nonparametric logistic regression

Pros and Cons of GAMS

- GAMs allow us to fit a non-linear f_j to each X_j so we can automatically model non-linear relationships that standard linear regression will miss

  ➡ We do not need to manually try out many different transformations on each variable individually

- The non-linearity can potentially make more accurate predictions
- Since the model is additive, we can examine the effect of each X_j on Y individually while holding all the other variables fixed
- The smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom

Pros and Cons of GAMs

- The model is restricted to be additive

  ➡ important interactions can be missed

However...

We can manually add interaction terms

➡ i.e. predictors of the form $X_j \times X_k$ or interaction functions of the form $f_{jk}(X_j, X_k)$

CODE DEMO