Generalized Linear Models

Introduction

Demetris Athienitis

UNIVERSITY of FLORIDA

Introduction to GLMs

The rest of the class we will focusing on fitting models similar to the multiple regression models from your previous classes....which is a special case of GLMs

Models come with many nice advantages:

- Automatically handle continuous explanatory variables
- Can handle large numbers of explanatory variables
- Constructing confidence intervals is relatively straightforward
- Interpretable parameters

Introduction to GLMs

- GLMs relate response variable Y to a set of predictors $\boldsymbol{X}$
  - Understand associations between predictors and outcome
  - Predict the outcome for given predictor levels
- Throughout, we will be treating $\boldsymbol{X}$ as a fixed quantity
- Inference will proceed by looking at the conditional distribution of Y given $\boldsymbol{X}$

Introduction to GLMs

- A GLM has three key components
  1. Random component
  2. Systematic component
  3. Link function
- There are a number of options for each component and we must choose one for each
- There are default choices for the random and link components

Random component

- The random component specifies the probability distribution for Y
- The three most common choices (adequate for most situations) are:
  - Normal distribution for continuous data
  - Binomial distribution for binary data
  - Poisson distribution for count data
- Many other probability distributions work with GLMs (exponential family)

Systematic component

Specifies how the explanatory variables are related to the response.

Typically we use a linear function, such as

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- may include interaction terms, e.g. $x_1 x_2$ and polynomial terms, e.g. x_1^3
- will involve model building techniques such as those developed for multiple linear regression

Link function

The link function specifies the functional relationship between the linear predictor and the mean of the outcome.

Let $\mu = E(Y)$, then we specify

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$g()$ is the link function used.

Link function

- Every probability distribution that can be used for GLMs has a special function of the mean called the *natural parameter*
- The link function that uses the natural parameter as the link function is called the *canonical link* which has some benefits
  - Won't cover these in class
  - We will typically use the canonical link

Link function

The canonical links we will use are

- For the normal distribution, **identity link**

$$g(\mu) = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- For the Poisson distribution, **log link**

$$g(\mu) = \log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- For the Bernoulli distribution, **logit link**

$$g(\mu) = log\left[\frac{\mu}{1-\mu}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Remark

When

- Random component is normal
- Systematic component is linear (will always be)
- Link is the identity link

this yields ordinary multiple regression via maximum likelihood estimation which almost equivalent to least squares estimation (Chapter 1 of STA 4210).

We learned

The 3 components of a Generalized Linear Model, of which ordinary regression is a special case.