Module X: Introduction to GLM's

Rebecca C. Steorts

Agenda

What should you learn?

- XXX

Notation

- $$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} : \quad \text{response variable}$$
- $x_{1:p} = (x_1, \ldots, x_p)$: explanatory variables

Linear Regression Model

Two components of regression model:

1. random and
2. systematic components.

Random Component

The random component assumes the response variables y_i ($i = 1, \ldots, n$) have constant variance σ^2.[1]

[1] Or it assumes $var[y_i] = \sigma^2 / w_i$ for known positive weights w_i.

Systematic Component

The systematic component assumes

$$E[y_i] = \mu_i$$

is linearly related to the explanatory variables x_j (for all j) such that:

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}.$$

Linear regression

Combining both components, a linear regression model has the form:

$$var[y_i] = \sigma^2 / w_i \tag{1}$$

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji},$$

where $E[y_i] = \mu_i$ and the prior weights are both known.

The regression parameters $(\beta_0, \beta_1, \ldots, \beta_p)$ and the error variance σ^2 must be estimated from the data.

Linear regression

- β_0: intercept (value of y when all explanatory variables are 0.)
- $(\beta_1, \ldots, \beta_p)$: slopes for the corresponding explanatory variables.

Linear regression

1. Suppose $p = 1$, then $\mu = \beta_0 + \beta_1 x_1$ is known as a simple linear regression model.
2. When $w_i = 1$ (for $i = 1, \ldots, n$) refers to an ordinary linear regression model, which we contrast with a weighted linear regression model.
3. If $p > 1$, we refer to this as a multiple regression model.

Linear Regression Assumptions

The assumptions for establishing equation 1 are

1. Suitability: The same regression model is appropriate for all observations.
2. Linearity: The true relationship between μ and each quantitative explanatory variable is linear
3. Constant variance: The unknown part of the variance of the responses, σ^2 is constant
4. Independence: The responses y are independent of each other

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming the above conditions are all met.
- **Generalized linear models** require different assumptions and can accommodate violations above
  - Response variable comes from a general family of distributions, called an exponential family
  - Relationship between response and predictor(s) can be nonlinear
  - Variance in response can differ at each level of predictor(s)
  - **The independence assumption still must hold!**
- **Multilevel models** are used to model data that violate the independence assumption, i.e. correlated observations

All Models are Wrong but Some Are Useful

Box and Draper [2, p.424] stated "all models are wrong, but some are useful. However, the approximate nature of the model must alawys be borne in mind."

- Models are not an exact representation of reality.
- Models are useful approximations for trying to understand data.
- In this class, we will try and understand what types of models are appropriate for certain types of data.