Generalized linear models

Data analytics

Jerzy Baranowski

Linear models are simple

And why we would want anything else?

$$y_i \sim \mathrm{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$

This is nice, friendly, and usually analytical (up to solving system of linear equations)

There may be some reasons

Outliers

$$y_i \sim \mathrm{t}_\nu(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$

With Student's-t distribution we get outlier rejection, but principles stay the same.

But what to do if our data are not real numbers?

Problems with integers and constraints

- Normal and student-t distributions are good for real valued data
- We have some distributions that are integer valued
  - Bernouli - binary results
  - Binomial - sequence of binary results
  - Poisson - integers
  - …
- There are also zero constrained distributions
  - Exponential
  - Gamma
  - LogNormal

Lets focus on integers

Maybe we just switch the likelihood?

$$y_i \sim \text{Bernouli}(\theta_i)$$
$$\theta_i = \alpha + \beta x_i$$

Wrong!

Parameters of integer valued distributions are constrained

- For Binomial or Bernouli $\theta \in [0,1]$
- For Poisson $\lambda > 0$

- And $\alpha + \beta x_i$ is generally unbounded so we have problems
- Constraining is not an option, as it would computationally screw us

Link functions are the solution!

Its where the “generalized” in GLM comes from

- In general we ‘link’ our linear model with distribution by a function

$$y_i \sim \text{Distribution}(\theta_i)$$
$$f(\theta_i) = \alpha + \beta x_i$$

- That would mean

$$y_i \sim \text{Distribution}(f^{-1}(\alpha + \beta x_i))$$

- And function $f^{-1} : \mathbb{R} \to [a, b]$, where a and b are distribution dependent constraints

Examples

Exponential (or logarithmic link)

- In Poisson distribution λ has to be a positive, it can be ensured by

$$\lambda_i = \exp(\alpha + \beta x_i)$$

- Or equivalently

$$y_i \sim \text{Poisson}(\lambda_i)$$
$$\log \lambda_i = \alpha + \beta x_i$$

Logit link

Logistic regression

- In any case where we want to estimate probability of an event Bernouli distribution is useful.

- In order to constraint our linear expression to $[0,1]$ we can use logit function

$$\operatorname{logit}(\theta_i) = \log \frac{\theta_i}{1-\theta_i}$$

- We get

$$\begin{aligned} y_i &\sim \operatorname{Bernouli}(\theta_i) \\ \operatorname{logit}(\theta_i) &= \alpha + \beta x_i \end{aligned} \qquad \theta_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Probit link

Probit regression (same but not the same as logit)

- Probit function is based on Cumulative Distribution Function of Normal distribution i.e.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

- Probit regression is formulated as

$$y_i \sim \mathrm{Bernouli}(\theta_i)$$
$$\Phi^{-1}(\theta_i) = \alpha + \beta x_i$$

- Difference from logit is in the tails (larger values of $\alpha + \beta x_i$ are faster approaching zero or one)

Time for the examples