Generalized linear models

Katie Pollard

BMI 206
docpollard.org/bmi206
October 5, 2016

Generalized linear model (GLM)

If outcome is not quantitative, the linear model framework can be extended via data transformations, called link functions.

- Binary: logit (alternatives: probit, log-log)
- Counts: log (also known as log-linear model)

The covariates are still a linear combination.

The parameters are estimated by numerical methods (e.g., Newton-Raphson).

But the error has a different distribution.

Link functions in GLMs

Link function systematically relates expected value of outcome ($E[Y] = \mu$) to a linear combination of covariates (X):

$$g(\mu)=ß'X$$

- Identity link: $g(\mu) = \mu$
- Log link: $g(\mu) = \log(\mu)$
- Logit link: $g(\mu) = \log(\mu/(1-\mu))$
- Log-log link: $g(\mu) = \log(-\log(1-\mu))$
- Probit link: $g(\mu) = \mathrm{Phi}^{-1}(\mu)$

Error distributions in GLMs

Different types of outcome variables require different error distributions, e.g.,

- Continuous (link=identity): Gaussian
- Binary (link=logit): Binomial
- Counts (link=log): Poisson

These are the random components.

They are examples of the exponential family:

$$f(y) = a(\mu)b(y)\exp\{g(\mu)y\}$$

Logistic regression parameters

Consider:

$$\text{logit}(\pi) = \beta_0 + \beta_1 X$$

Interpretation of β_1 is the expected change in logit for a unit increase in X. What is this?

If X is binary (e.g., 0=wild-type vs. 1=mutant):

$$\text{odds}(X=0) = \exp\{\beta_0\}, \text{odds}(X=1) = \exp\{\beta_0\}\exp\{\beta_1\}$$

Odds increase multiplicatively by $\exp\{\beta_1\}$ per unit X.

$$\text{Odds ratio} = \text{odds}(X=1)/\text{odds}(X=0) = \exp\{\beta_1\}$$

Logistic regression model

What is the distribution function?

Can we write it as an exponential family?

What is the canonical link?

What is the systematic component?

Poisson regression parameters

Consider:

$$\log(\mu) = \beta_0 + \beta_1 X$$

Interpretation of β_1 is the expected change in log count for a unit increase in X.

Exponentiate to get back to count scale.

If X is binary (e.g., 0=wild-type vs. 1=mutant):

$$\mu(X{=}0) = \exp\{\beta_0\} \text{ and } \mu(X{=}1) = \exp\{\beta_0 + \beta_1\}$$

$$\text{Relative risk} = \mu(X{=}1)/\mu(X{=}0) = \exp\{\beta_1\}$$

Poisson regression model

What is the distribution function?

Can we write it as an exponential family?

What is the canonical link?

What is the systematic component?

Over-dispersion

The Poisson distribution has the variance equal to the mean. Count data in bioinformatics frequently violates this assumption, e.g.,

- Gene expression via RNA-seq (read counts/transcript)
- Taxon abundance in metagenomics (reads counts/taxa)

Variance > mean is called “over-dispersion”.

The negative binomial distribution is a good alternative:

$$\text{mean} = \mu,\ \text{variance} = \mu + \mu^2/k$$

Linear model as a GLM

What is the distribution function?

Can we write it as an exponential family?

What is the canonical link?

What is the systematic component?