Generalized linear models

Katie Pollard

BMI 206

In this unit we will learn…

- How to formulate generalized linear models (GLMs) with outcomes that are not Normally distributed (e.g., binary, counts)
- The main components of GLMs
- Interpretation of parameters in GLMs such as logistic and Poisson regression
- The "exponential family" of distributions
- How to fit and interpret LMs and GLMs in R

Relating Different Data Types

Covariate (independent variable)

Outcome (dependent variable)

	Continuous or Both	Categorical
Continuous	Linear Regression / ANCOVA	ANOVA
Categorical	Generalized Linear Model Regression	Contingency Tables / Log-linear Model Regression

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

The link function, denoted g(), systematically relates expected value of outcome (E[Y] = μ) to a linear combination of covariates (X):

$$g(\mu)=\beta' X$$

- 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 (Normal)
- Binary (link=logit): Binomial
- Counts (link=log): Poisson

These are the random components.

The systematic component is the mean ß'X, e.g.:

$$ß_0 + ß_1 X$$

Logistic regression parameters

Consider: Y binary with $E(Y) = Pr(Y=1) = \pi$

$$\text{logit}(\pi) = \log(\pi/(1-\pi)) = \beta_0 + \beta_1 X$$

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):

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

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

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

Poisson regression parameters

Consider: Y counts with $E(Y) = \mu$

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

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):

$\mu \mid X=0 = \exp\{\beta_0\}$ and $\mu \mid X=1 = \exp\{\beta_0 + \beta_1\}$

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

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:

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

GLMs and Exponential Family

The common error distributions in GLMs (Gaussian, Binomial, Poisson) are all members of the exponential family of distributions which can be written:

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

where g() is the link function.

If you can arrange the error distribution into this form, the result gives you the canonical link function.

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?

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 model

What is the distribution function?

Can we write it as an exponential family?

What is the canonical link?

What is the systematic component?