Generalized linear models:

General linear models +

Non-normal residuals=

Generalized linear models

Modeling with the Poisson distribution

Reminder: General linear models

- All of the we've discussed models (t-test, linear regression, ANOVA, ANCOVA) are special cases of the ***general linear model***

- General linear models express a continuous response as a linear combination of the effects of discrete (categorical; factors) and/or continuous explanatory variables (covariates)

  Plus a single random contribution from a Normal distribution whose variance is estimated along with the coefficients of all – discrete and continuous – covariates.

Generalized linear models

- Unifying a large number of statistical methods (t-test, regression, ANOVA and ANCOVA) under the umbrella of the ***general linear model*** was a big advance
- Even more significant was the unification of an even wider range of statistical methods within the class of the ***generalized linear model*** or ***GLM*** in 1972 by Nelder and Wedderburn
- A large number of techniques previously thought of as representing quite separate types of analyses, including logistic regression, multinomial regression, Chisquare, log-linear models, general linear model, were all recognized as just a special case of a generalized version of the familiar linear model

Generalized linear models

Main ideas:

1) A *transformation of the expectation* of the response is expressed as a linear combination of covariates rather than the mean response directly.
2) For the random part of the model, *distributions other than the Normal* can be chosen, e.g., Poisson, Binomial or gamma.

Generalized linear models

GLMs are made up of three components:

1) A *statistical distribution* used to describe the random variation in the response *y*; this is the stochastic part of the system description
2) A *linear predictor*, i.e., a linear combination of covariate effects that are thought to make up *E(y)*; this is the systematic or deterministic part of the system description.
3) A *link function* that is applied to the *E(y)*, expectation of the response

Generalized linear models

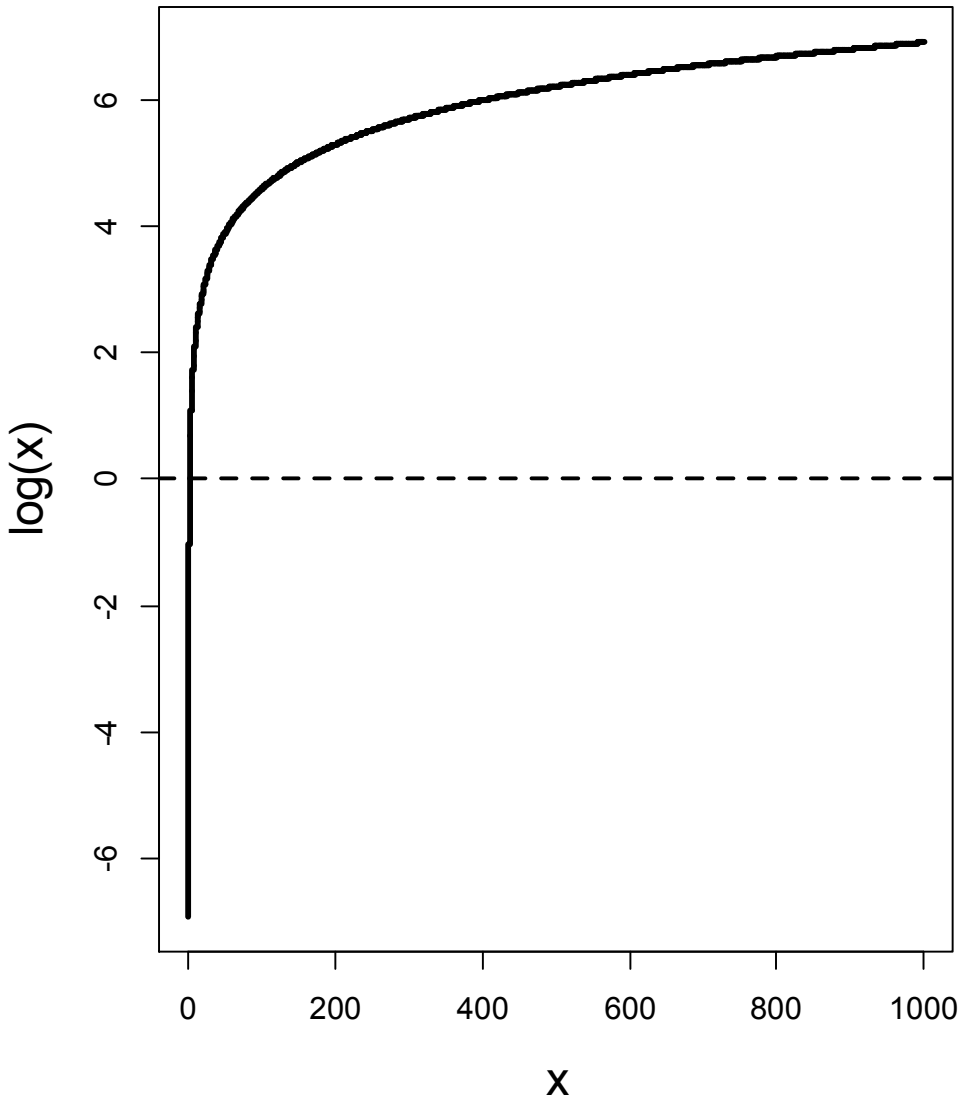
Three most common GLMs (does this look familiar??)

- Normal response:
  - Random part: $y \sim \text{Normal}(\mu, \sigma^2)$
  - (typical) Link function: none ("identity")
  - Systematic part: SLM (some linear model)
- Poisson response:
  - Random part: $y \sim \text{Poisson}(\lambda)$
  - (typical) Link function: $\log(\lambda)$
  - Systematic part: SLM
- Binomial response:
  - Random part: $y \sim \text{Binomial}(p, N) = N * \text{Bernoulli}(p)$
  - (typical) Link function: $\text{logit} = \log(p / (1-p))$
  - Systematic part: SLM

Why use a link function?

- Suppose you were interested in modeling counts (whole numbers)
- Expected count must be greater than or equal to zero
  - What does the log function do? Log(x)
  - Range: $x > 0$
  - Range:

$$-\infty > \log(x) > \infty$$

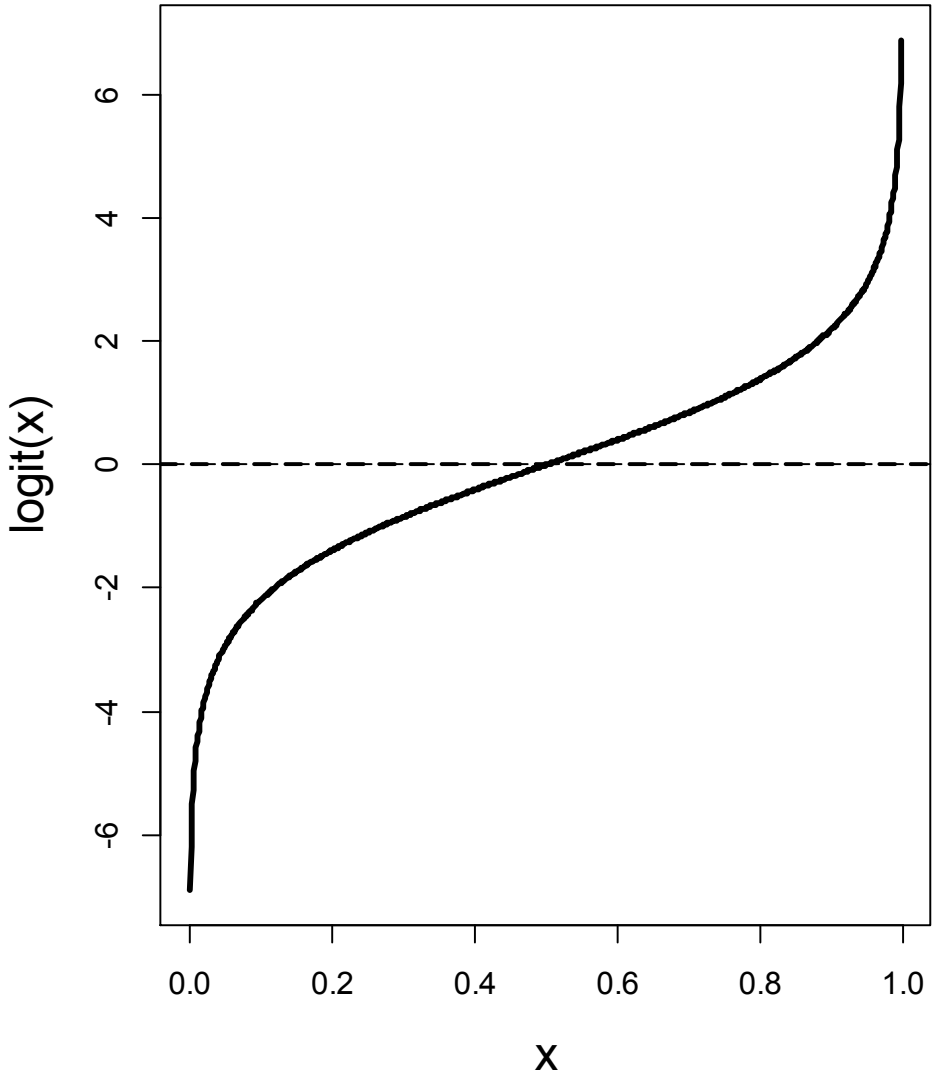


Why use a link function?

- Suppose you were interested in modeling the probability of an event occuring
- Probability must be between zero and one
  - What does the logit function do?
  - Logit(x) =

    log (x) – log (1-x)
  - Range: $0 > x > 1$
  - Range:

    $-\infty > \text{logit}(x) > \infty$



GLMs: stochastic part

Poisson Distribution

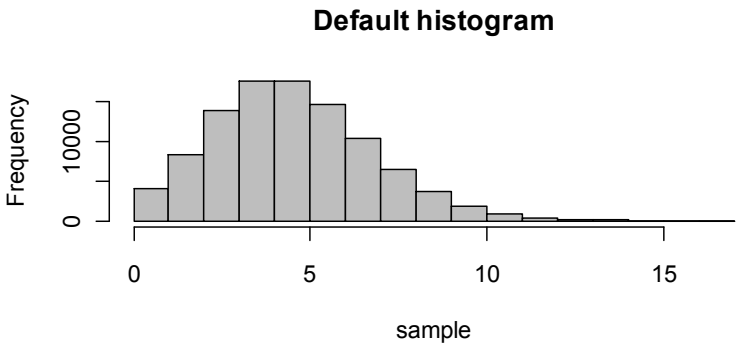
An important *discrete* distribution that is useful for modeling counts

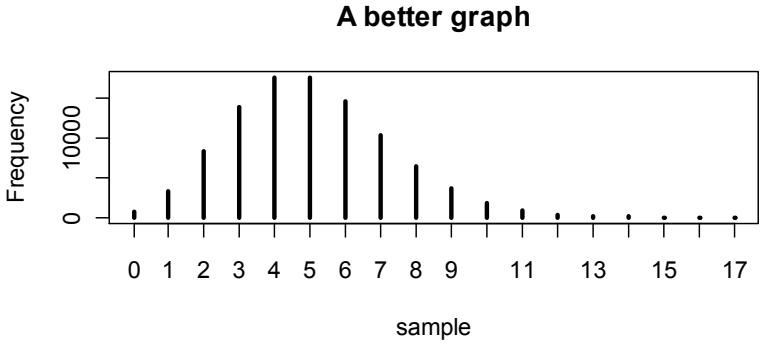
- Denoted: $x \sim Pois(\lambda)$
- Mean: $\lambda > 0$
- Variance: λ
- Support: $x \in \{0,1,2,3, \ldots\}$

GLMs: stochastic part

Poisson Distribution

- *Sampling situation:* Probability of a given number of events occurring in a fixed interval of time/space if events occur with an average rate and are independent
- *Classical examples:* Number of birds that fly over a migration site in 10mins, number of hares per sample quadrat.
- *Mathematical description*: Single parameter, λ, equal to the mean and variance. Variance is not a free parameter.
- *Varieties:* None. But is an approximation to Binomial when N is large and p small and can be approximated by Normal when λ is large, (> 10). The negative binomial dist is an overdispersed version, derived by assuming that λ is a r.v. with a gamma distribution.





GLMs: stochastic part

Poisson Distribution

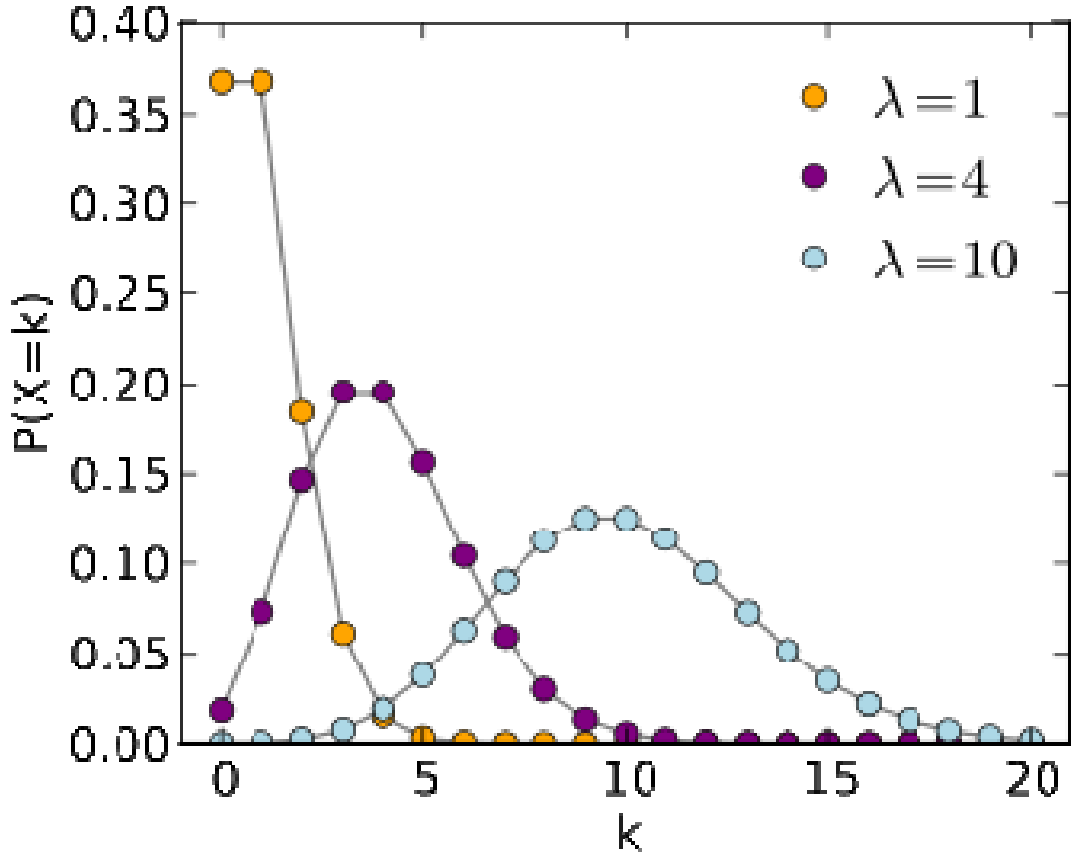
Probability mass function (PMF):

$$p(x = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Likelihood function:

$$L(\lambda \mid X) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

Where $X = \{x_1, x_2, \ldots, x_n\}$



Generalized linear models

Poisson t-test

Assume that we sampled the number of plant species (e.g., through point counts or transect walks) in 50 locations in both OH and MI.

Does point level plant richness differ in OH and MI?

Generalized linear models

Poisson t-test

What is the distribution, link function, and linear predictor we should use?

Distribution:

Link function:

Linear predictor:

Generalized linear models

Poisson t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $C_i \sim Poisson(\lambda_i)$

Link function:

Linear predictor:

Generalized linear models

Poisson t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $C_i \sim Poisson(\lambda_i)$

Link function: $\log(\lambda_i)$

Linear predictor:

Generalized linear models

Poisson t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $C_i \sim Poisson(\lambda_i)$

Link function: $\log(\lambda_i)$

Linear predictor: $\alpha + \beta * x_i$

Generalized linear models

Poisson t-test

Distribution:	$C_i \sim Poisson(\lambda_i)$
Link function:	$\log(\lambda_i) \sim \log(E(C_i))$
Linear predictor:	$\alpha + \beta * x_i$

- Plant species count C_i at point *i* (where *i* exists in either OH or MI) is distributed as a Poisson random variable with mean λ_i.
- The log-transform of λ_i is assumed to be a linear function $\alpha + \beta * x_i$, where α and β are unknown constants and x_i is the value (indicator variable) of an area-specific covariate (i.e., $\log(\lambda_i) = \alpha + \beta * x_i$)
- If x_i is an indicator for MI, then α is the mean species count (log scale) in OH and β, (again on log-scale), is the difference in mean plant count between OH and MI.

Poisson t-test: deterministic part

Design matrix – what does the variable state look like?

$$plant\_counts_i \sim Pois\ (\lambda_i\ )$$
$$log(\lambda_i\ ) = \alpha + \beta * state_i$$

glm(counts ~ state)
model.matrix(~state)

(Intercept)	state2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1

Translates into a set of equations:

$$\log(6) = \alpha * 1 + \beta * 0$$
$$\log(8) = \alpha * 1 + \beta * 0$$
$$\log(5) = \alpha * 1 + \beta * 0$$
$$\log(7) = \alpha * 1 + \beta * 0$$
$$\log(9) = \alpha * 1 + \beta * 1$$
$$\log(9) = \alpha * 1 + \beta * 1$$

Or in matrix notation:

$$\log\begin{pmatrix}6\\8\\5\\7\\9\\9\end{pmatrix} = \begin{pmatrix}1 & 0\\1 & 0\\1 & 0\\1 & 0\\1 & 1\\1 & 1\end{pmatrix} * \begin{pmatrix}\alpha\\\beta\end{pmatrix}$$

Generalized linear models

Poisson t-test

The Poisson t-test is analogous to our standard t-test. The only real difference is that the data are assumed to come from a Poisson rather than from a normal distribution.

We cope with that by modeling the data using a link function, which allows us to create the linear predictor, exactly as we did before.

Generalized linear models

Lab: Poisson t-test

Open script: Poisson t-test