Generalized linear models:

General linear models +

Non-normal residuals=

Generalized linear models

Logistic Regression

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

Modeling with the binomial distribution

Binomial response

- Random part: y ~ Binomial(p, N) = N*Bernoulli(p)
- (typical) Link function: logit = log (p / (1-p))
- Systematic part: Some linear model (e.g., t-test, regression)

Logistic regression

Used with binary outcomes – a Bernoulli trial

- Heads or tails of a coin
- Presence or absence of a species
- The success or failure of breeding
- Occurrence of a color morph
- Detection or non detection of an individual

Logistic regression

Estimating a binomial proportion (commonly called a logistic regression) is analogous to summing up the number of successes in a fixed number of trials

- Flip a coin N times. Sum up the number of successes (heads) in the N trials
- Pr (k heads given N trials) -> depends on the probability (p) of success

- Binomial distribution is bounded by N, different from the Poisson, which is unbounded
- The Bernoulli distribution is a special case when N=1.

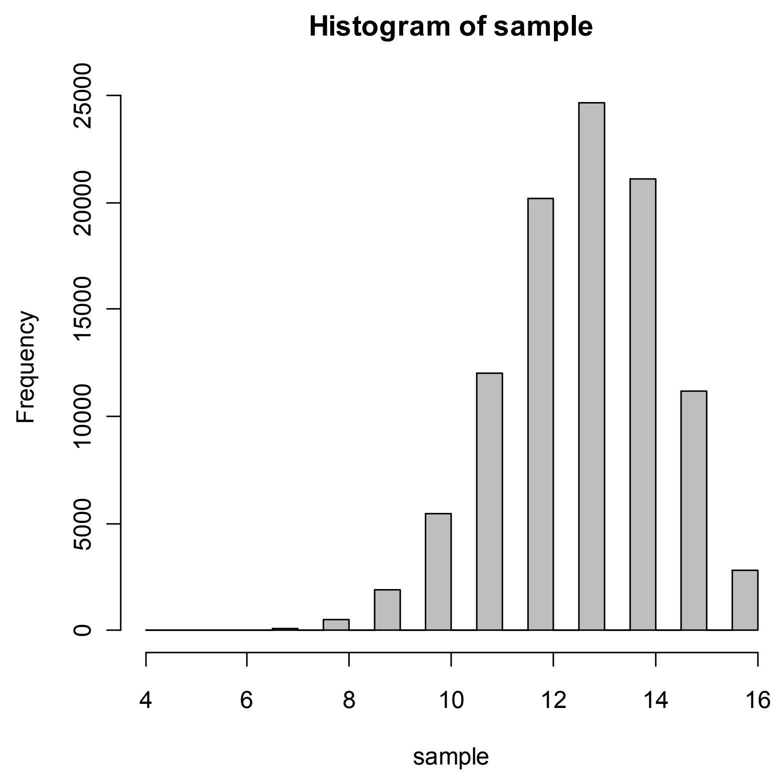
GLMs: stochastic part
Binomial Distribution

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

- Denoted: $x \sim Bin(p, N)$
- Mean: Np (for $0 > p > 1$)
- Variance: $Np(1 - p)$
- Support: $x \in \{0,1,2,3, \dots, N\}$

GLMs: stochastic part
Binomial Distribution

- *Sampling situation:* N things that have the same probability p of making it into a sample (e.g., being counted or dead)
- *Classical examples:* Number of males in a clutch of size N; Number of individuals among all present that are observed.
- *Varieties:* Bernoulli distribution is a single coin flip and has only a single parameter, p (e.g., a Binomial is a sum of N Bernoullis)
- *Mathematical description*: 2 parameters: success probability p, and “binomial total” or “size”, N. N represents a ceiling to the binomial counts; Usually is observed and therefore is not a parameter.



GLMs: stochastic part

Binomial Distribution

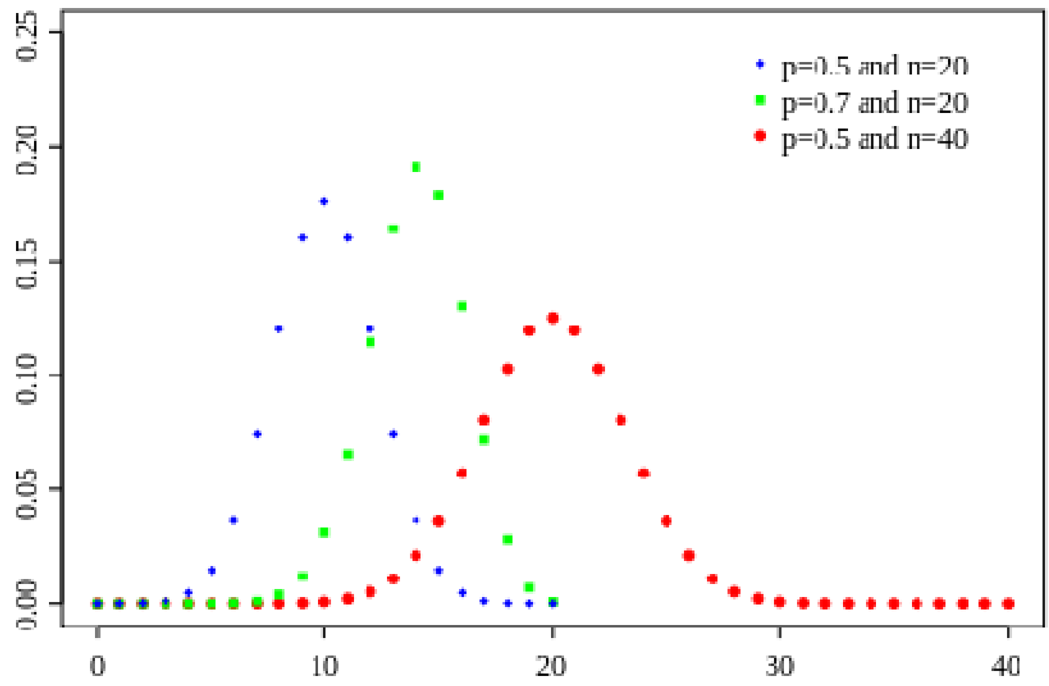
Probability mass function (PMF):

$$p(x = k \mid p, N) = \binom{N}{k} p^k (1-p)^{N-k}$$

Likelihood function:

$$L(p, N \mid X) = \prod_{i=1}^{n} \binom{N}{k_i} p^{k_i} (1-p)^{N-k_i}$$

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



Logistic regression

AKA: Binomial regression

- Consider an inventory of adder snakes, which have two color morphs: all black and zigzag.
- You hypothesize that the black color confers thermal advantages -> more black adders in cooler and wetter locations.

Question: Is color morph related to temperature and average precipitation?

Logistic regression

AKA: Binomial regression

- Data collection:
  - C_i the number of black morphs out of N_i total number of observed adder snakes at location *i*.
  - $temp_i$ = the average summer temperature at *i.*
  - $prec_i$ = the total amount of annual rainfall at *i.*

Want to estimate whether the proportion of black morphs is higher in locations with higher temp and prec values, a relationship that could change (a possible interaction between the variables).

Logistic regression

Binomial distribution

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

Distribution:

Link function:

Linear predictor:

Logistic regression

Binomial distribution

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

Distribution: $C_i \sim Bin(p_i, N_i)$

Link function:

Linear predictor:

Logistic regression

Binomial distribution

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

Distribution: $C_i \sim Bin(p_i, N_i)$

Link function: $\text{logit}\,(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$

Linear predictor:

Logistic regression

Binomial distribution

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

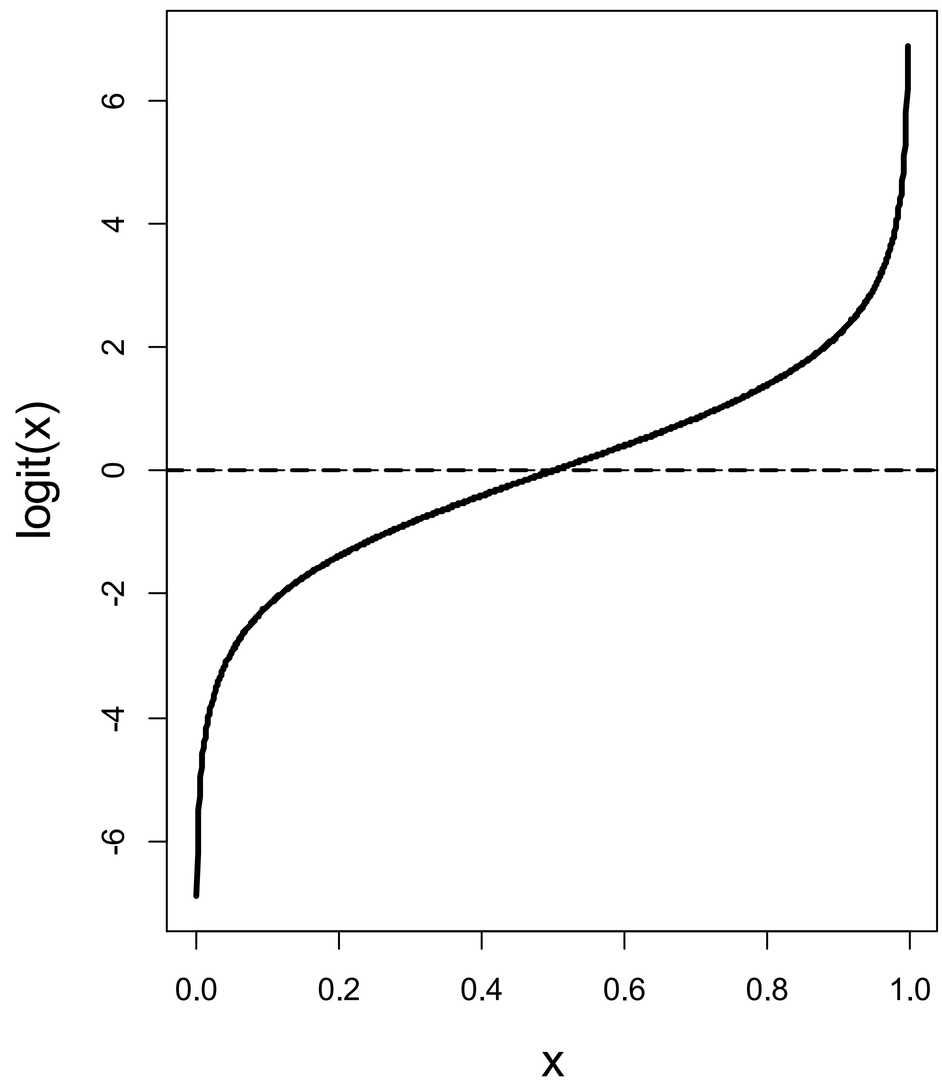
Distribution: $C_i \sim Bin(p_i, N_i)$

Link function: $\text{logit}\,(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$

Linear predictor: $\alpha + \beta 1 * temp_i + \beta 2 * prec_i + \beta 3 * temp_i * prec_i$

Logit link function for probabilities

- What to model the probability that each snake is black (or the proportion of N total snakes)
- 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:
    $-\infty > \text{logit}(x) > \infty$

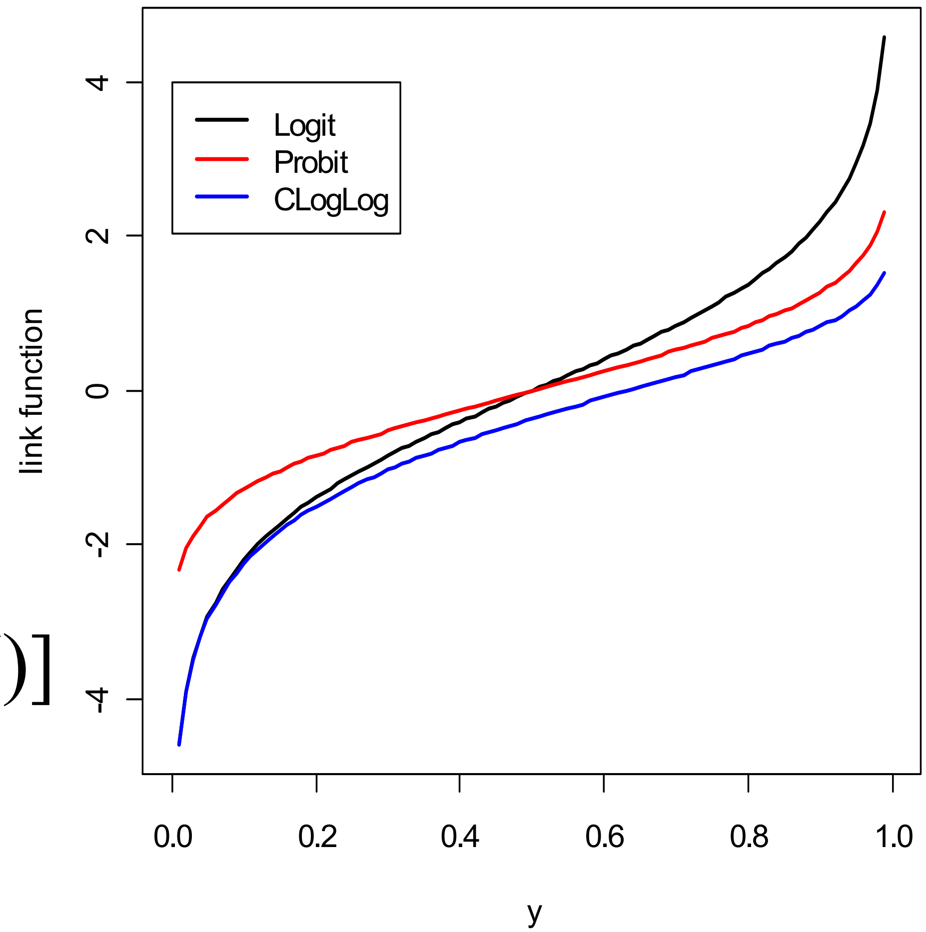


Other link function for probabilities

$$\text{logit}(Y) = \log\left(\frac{Y}{1-Y}\right)$$

$$\text{probit}(Y) = \Phi^{-1}(Y)$$

$$c\log\log(Y) = \log[-\log(1-Y)]$$



Logistic regression

Distribution: $C_i \sim Bin(p_i, N_i)$

Link function: $\text{logit}\ (p_i)$

Linear predictor: $\alpha + \beta 1 * temp_i + \beta 2 * prec_i + \beta 3 * temp_i * prec_i$

- Count of black color morphs C_i out of N_i observed adder snakes is distributed as a binomial random variable with mean $p_i N_i$.

- The logit-transform of p_i is assumed to be a linear function of the intercept and the covariate values.

$$\text{logit}\ (p_i) = \alpha + \beta 1 * temp_i + \beta 2 * prec_i + \beta 3 * temp_i * prec_i$$

Logistic regression

Question: Does the proportion of black morphs vary with weather?

Example with 10 data points:

Location	N	C	temp	precip
1	20	15	64	6
2	17	13	67	8
3	18	12	68	9
4	25	12	72	7
5	22	14	72	5
6	16	8	75	7
7	27	13	77	9
8	24	10	80	8
9	24	11	82	6
10	21	9	85	6

Logistic regression

Translates into a set of equations:

$$C_i \sim Bin(p_i, N_i)$$
$$\text{logit}\ (p_i) = \text{linear predictor}$$

First data point:

$$15 \sim Bin(p_1, 20)$$
$$logit(p_1) = \alpha * 1 + \beta_1 * 64 + \beta_2 * 6 + \beta_2 * 64 * 6$$

Second data point:

$$13 \sim Bin(p_2, 17)$$
$$logit(p_2) = \alpha * 1 + \beta_1 * 67 + \beta_2 * 8 + \beta_2 * 67 * 8$$

Logistic regression

Or in matrix notation:

$$C_i \sim Bin(p_i, N_i)$$

$$logit\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \\ p_9 \\ p_{10} \end{pmatrix} = \begin{pmatrix} 1 & 64 & 6 & 64*6 \\ 1 & 67 & 8 & 67*8 \\ 1 & 68 & 9 & 68*9 \\ 1 & 72 & 7 & 72*7 \\ 1 & 72 & 5 & 72*5 \\ 1 & 75 & 7 & 75*7 \\ 1 & 77 & 9 & 77*9 \\ 1 & 80 & 8 & 80*8 \\ 1 & 82 & 6 & 82*6 \\ 1 & 85 & 6 & 85*6 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Logistic regression

Assumptions

- The probability of success is independent for each sample.
- The max number of successful outcomes for each trial is N_i

- Covariates should not be collinear. This should always be true!

Standardizing covariates

What is standardizing?

- Rescaling regression coefficients by subtracting the mean and dividing by the standard deviation:

$$standardized.temp = \frac{temp - mean(temp)}{sd(temp)}$$

- So that:

$$\text{mean}(standardized.temp) = 0$$
$$\text{sd}(standardized.temp) = 1$$

Standardizing covariates

Why standardize?

- Improves interpretation of parameter estimates
  - What is the interpretation of an intercept term when we regress a variable against year (2000-2014)?
    - Effect of the variable when year = 0
  - What about if year where standardized?
    - Effect of the variable when year is at its average value (2007)
  - Now the slope can be interpreted in units of standard deviations with respect to the corresponding predictor

Standardizing covariates

Why standardize?

- Improves interpretation of parameter estimates
  - Especially true when there are many regression coefficients and interactions.
  - When regression coefficients are on hugely different scales, interpretation is difficult

Standardizing covariates

Why standardize?

- Improves numerical stability
  - Many situations where it makes estimation easier
  - Especially true in Bayesian analyses. Must pretty much always standardize.

Generalized linear models

Lab: Logistic regression – ANCOVA style