# Lecture 3
# Generalized Linear Model

**Chao Song**

College of Ecology

Lanzhou University

July 17, 2025

# Classic linear models

Recall that a classic linear model is of the following form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Or in the matrix notation as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

This means that $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. That is, the response variable $Y$ follows a normal distribution and its mean is predicted by a linear predictor $\boldsymbol{X}\boldsymbol{\beta}$.

# Limitations of classic linear models

While classic linear models is a very versatile tool, it has the following limitation:

- $E(\boldsymbol{Y}) = \boldsymbol{X}\beta$ is unlimited in range, but in many problems, the range of $\boldsymbol{Y}$ is restricted.
- Most inferences assumes a normal distribution of errors;
- The errors are additive.

# Limitations of classic linear models

**Example**: In toxicology research, we often want to study how the concentration of a pollutant or the dose of a chemical influence deathl rate of organisms. The response we collect is a binary, i.e., live (0) or die (1).

Clearly, using a linear model here is not adequate because

- The response is binary and does not follow a normal distribution;
- A linear model can produce predictions that are out of the [0,1] range

To address these issues, we may consider

- Use an alternative distribution of model the response, e.g, Bernoulli.
- Map the response into permissible range via a nonlinear function.

This extends the classic linear model to **generalized linear model.**

# Generalized linear model

A Generalized linear model has three essential components:

- Systematic component: the linear predictor $\eta = \boldsymbol{X}\boldsymbol{\beta}$
- Random component: the distribution of the response variables;
- Link function: map $\mu$ to the range of $\eta$.

**Example**: To investigate how dose of a chemical influence the death rate of fish, we recorded 100 fishes at various dose levels. Our response $Y$ is binary, e.g., death (1) or live (0). We know that $Y$ has a Bernoulli distribution with probability $p$ of being equal to 1. A GLM for this analysis could be

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

This type of model is called **logistic regression** and is a type of GLM.

# Common distributions for GLM

While GLM can be applied to any univariate distributions, we typically encounter the following

- Normal
- Bernoulli
- Binomial
- Poisson
- Negative Binomial

Understanding the meaning of these distribution is a critical first step in correctly using GLM in data analysis.

# Bernoulli distribution

A **Bernoulli trial** is a random experiment, the outcome of which can be classified in one of the two mutually exclusive and exhaustive ways–say, success of failure. Let $X$ be a random variable associated with a Bernoulli trial such that $X = 1$ for success and $X = 0$ for failure, $X$ follows a **Bernoulli distribution**.

**Example**: Suppose that the probability of germination of a beet seed is 0.8 and the germination of a seed is called a success. If we plant 10 seeds and can assume that the germination of one seed is independent of the germination of another seed. This would correspond to 10 Bernoulli trials with $p = 0.8$.

# Bernoulli distribution

The probability mass function of $X$ following a Bernoulli distribution is

$$f(x) = \begin{cases} p, & X = 1 \\ 1 - p, & X = 0 \end{cases},$$

Or more concisely, $f(x) = p^x(1-p)^{1-x}$.

The mean and variance of a Bernoulli distribution is

- $E(X) = 1 \times p + 0 \times (1-p) = p$
- $Var(X) = (1-p)^2 p + (0-p)^2(1-p) = p(1-p)$

# Binomial distribution

In a sequence of Bernoulli trials, we are often interested in the total number of successes, but not the actual order of their occurrences. Let random variable $X$ equal the number of observed successes in $n$ Bernoulli trials.

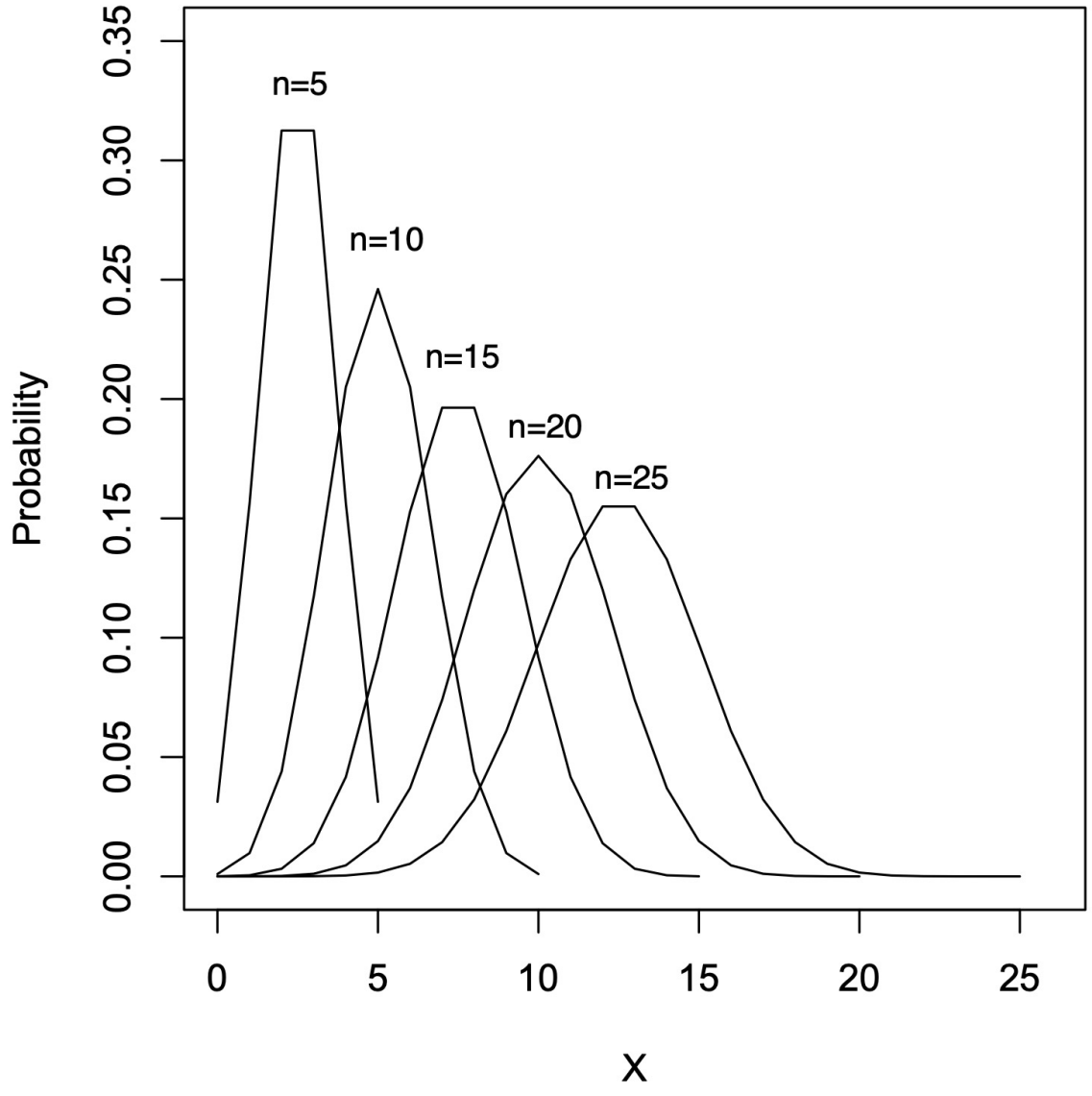
**Binomial distribution**: If a random variable $X$ denotes the number of successes in $n$ independent Bernoulli trials, $X$ follows a binomial distribution and its PMF is

$$P(X = k) = \mathbf{C}_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots,$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$

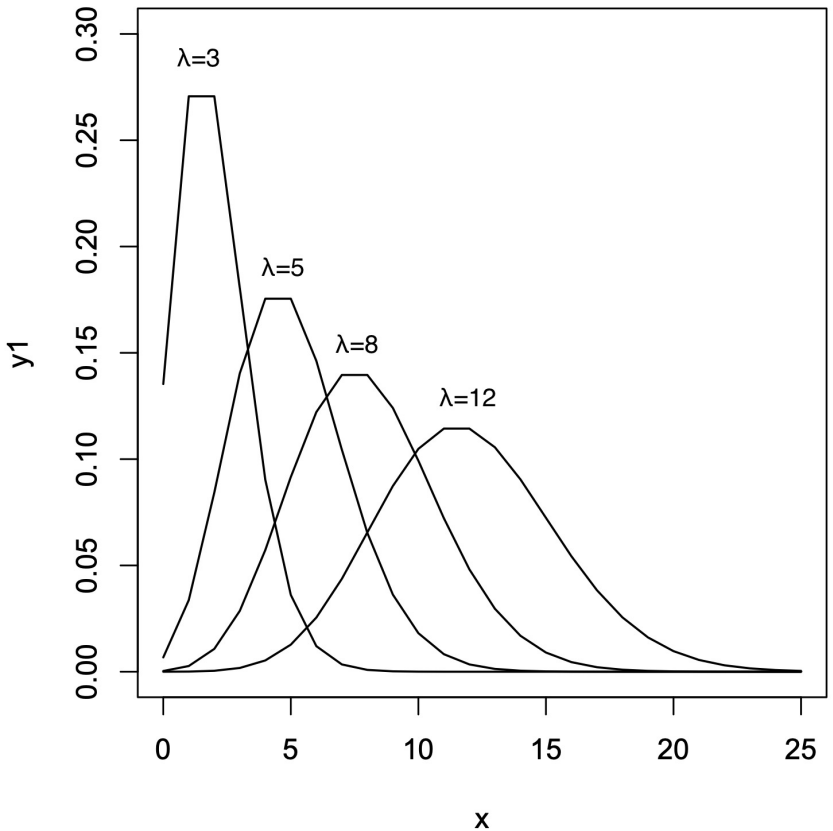
# Binomial distribution



# Poisson distribution

**Poisson distribution**: Let $\lambda$ be a positive number. A random variable is said to have a Poisson distribution if its probability mass function is

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \ k = 0, 1, 2, \ldots$$



# Poisson distribution

**What does a Poisson distributed variable model?**

Poisson distribution models the number of events in a time interval $t$.

- Divide $t$ into $n$ segments such that at most one event occur within a segment;
- Probability of occurrence is $\mu t/n$;
- Number of occurrence is modeled with a binomial distribution.

$$
\begin{aligned}
P(X=k) &= \lim_{n\to\infty} \mathbf{C}_n^k p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} (\frac{\mu t}{n})^k (1-\frac{\mu t}{n})^{n-k} \\
&= \lim_{n\to\infty} \frac{(\mu t)^k}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} (1-\frac{\mu t}{n})^{-k} (1-\frac{\mu t}{n})^{n} \\
&= \frac{(\mu t)^k}{k!} e^{-\mu t}
\end{aligned}
$$

# Poisson distribution

Poisson distribution is a limiting case of a binomial distribution. Here, $\lambda = \mu t$ is often referred to as the rate parameter of the Poisson distribution.

This derivation gives us a mechanistic insights into when we can use Poisson distribution. When some events occur at a constant rate, we can model the count of event with a Poisson distribution.

An important property of Poisson distribution is that both its mean and variance are $\lambda$. This is often used in practice to see if Poisson distribution is adequate.

# Negative binomial distribution

**Negative binomial distribution**: In a sequence of independent Bernoulli trials with success probability $p$, let $X$ be the number of failure until $r$ successes. Then $X$ has a negative binomial distribution with probability mass function

$$P(X = k) = \mathbf{C}_{k+r-1}^{k}(1-p)^k p^r$$

**Why is this called a negative binomial distribution?**

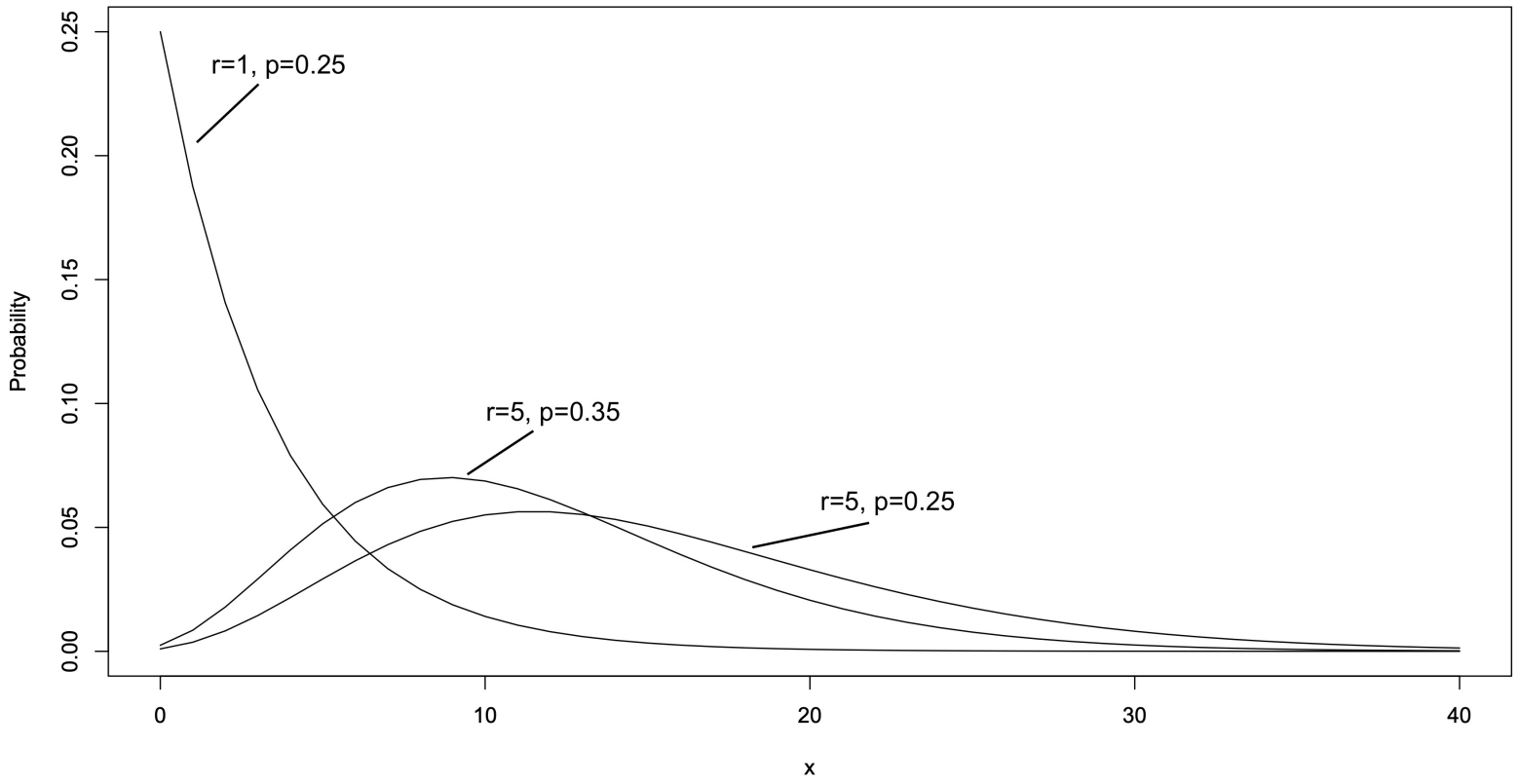
Let $q = 1 - p$ and $h(q) = (1-q)^{-r}$. Using Taylor expansion at $q = 0$

$$h(q) = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} q^k = \sum_{k=0}^{\infty} \mathbf{C}_{k+r-1}^{r-1} q^k = \sum_{k=0}^{\infty} \mathbf{C}_{k+r-1}^{k} q^k$$

Thus, we can see that the PMF of a negative binomial distribution is the summand of $p^r p^{-r}$

# Negative binomial distribution

The negative binomial distribution can take on a variety of shapes, depending on the parameters *r* and *p*. An important feature of negative binomial distribution is that its variance is larger than the mean.



# Link function

Now we know how to choose a distribution for the response variable, we now need to choose a link function that maps the mean of the response to the linear predictor.

While many plausible link function exist, we typically use what is referred to as the canonical link function statistical convenience.

| Distribution | Canonical link |
|---|---|
| Normal | $g(\mu) = \mu$ |
| Poisson | $g(\mu) = \log(\mu)$ |
| Binomial | $g(\mu) = \log(\frac{\mu}{1-\mu})$ |
| Gamma | $g(\mu) = \mu^{-1}$ |

# Fitting GLM

Unlike classic linear models that are typically fit by least squares, generalized linear models are fit using the method of **maximum likelihood.** To understand how GLMs are fit, it is essential to understand the method of maximum likelihood.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with PDF or PMF $f(x|\theta_1, \theta_2, \ldots, \theta_k)$, the joint PDF or PMF regarded as a function the $\theta_1, \theta_2, \ldots, \theta_k$ is called the **likelihood** function

$$\begin{aligned} L(\theta|X) &= L(\theta_1, \theta_2, \ldots, \theta_k|X_1, X_2, \ldots, X_n) \\ &= f(X_1, X_2, \ldots, X_n|\theta_1, \theta_2, \ldots, \theta_k) \end{aligned}$$

# Maximum likelihood estimators

Suppose we flipped a coin 3 times and observed heads, heads, and tail. What is the probability of observing such a result if $p = 0.5$ or $p = 0.6$?

The result of a coin flipping follows a Bernoulli distribution. Thus, the probability of observing heads, heads, and tail is

$$P(HHT) = p \times p \times (1 - p)$$

Thus, we have

$$P(HHT) = 0.5 \times 0.5 \times (1 - 0.5) = 0.125 \quad \textit{if } p = 0.5$$
$$P(HHT) = 0.6 \times 0.6 \times (1 - 0.6) = 0.144 \quad \textit{if } p = 0.6$$

In this case, if we do not know the probability of success and want to estimate it from observation, what would be the best estimates?

# Maximum likelihood estimators

**Definition**: For a particular sample, let $\hat{\theta}$ be the parameter value at which $L(\theta|X)$ attains its maximum as a function of $\theta$, with $X$ held fixed. $\hat{\theta}$ is called the **maximum likelihood estimator (MLE)** of the parameter $\theta$ based on the sample $X$.

The MLEs possess several useful properties that allowed us to use it as a universal way for hypothesis testing:

- The maximum likelihood estimate is approximately normal
- Likelihood ratio of nested models is the most powerful test

# Likelihood ratio test

Let $\Omega$ be the set of all possible values of parameter $\theta$ given by either $H_0$ or $H_a$. Let $\omega$ be a subset of $\Omega$ and $\omega'$ be its complement. The null and alternative hypotheses can be stated as

$$H_0 : \theta \in \omega, \quad H_a : \theta \in \omega'$$

Let $L(\hat{\omega})$ be the maximum of the likelihood function with respect to $\theta$ when $\theta \in \omega$ and $L(\hat{\Omega})$ be the maximum of the likelihood function with respect to $\theta$ when $\theta \in \Omega$. To test $H_0$ against $H_a$, the critical region is the set of points in the sample space for which

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \leqslant k,$$

where $0 < k < 1$ and $k$ is selected so that the test has a desired significance level $\alpha$.

# Likelihood ratio test

The likelihood ratio method does not always produce a test statistic with a known probability distribution. How do we use likelihood ratio test then?

**Wilk's theorem**: Let $r_0$ and $r$ be the number of free parameters under $\omega$ and $\Omega$, respectively. Under regularity conditions, $-2\ln(\lambda)$ asymptotically approaches $\chi^2(r - r_0)$ as sample size approaches $\infty$.

- The theorem gives us a general way of hypothesis testing. When sample size is large, we compare $-2\ln(\lambda)$ to a chi-square distribution with appropriate degrees of freedom. We reject the null hypothesis if the test statistic $-2\ln(\lambda)$ exceeds the critical value.
- The regularity conditions mainly involve the existence of derivatives of the likelihood function with respect to the parameters and the condition that the region over which the likelihood function is positive does not depend on unknown parameters. These conditions are satisfied for almost all distributions we discussed in this class.