GLeMa: Maximum Likelihood Estimation for Generalized Linear Models

Stochastic Batman

December 2025

1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is fundamentally a technique that operates in the reverse direction of traditional probability. While probability functions calculate the likelihood of observing data given specified parameters, MLE seeks to identify the optimal parameters that maximize the likelihood of observing the given data. This method is essential for many statistical procedures and forms the backbone of various algorithms used in data science and machine learning, such as the Naive Bayes classifier [1], Gaussian mixture models [2], and more. By efficiently identifying parameters, MLE enhances model performance and predictive accuracy, allowing for robust analysis and decision-making based on empirical data. This section is based on the works from [3].

Intuitive Explanation

The basic intuition behind Maximum Likelihood Estimation is grounded in probability. Given a random sample of data, $X_1, X_2, \ldots, X_n$, and assume this data comes from a specific probability distribution (like a Normal or Bernoulli distribution). This distribution depends on some unknown parameter, denoted by θ.

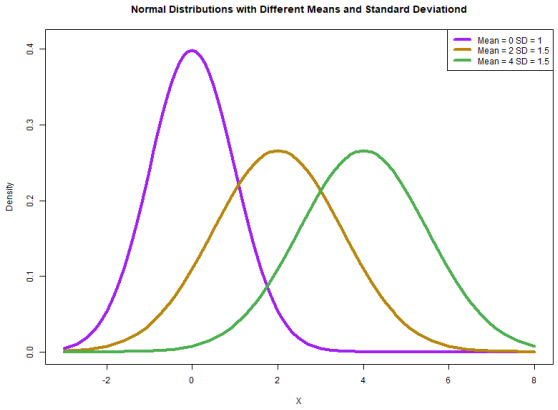

Figure 1: Normal distributions with different means and standard deviations

The method suggests that the optimal parameter value is the specific θ that maximizes the probability (or likelihood) of having generated the data that was actually collected.

In a nutshell: We treat the probability of our data as a function of the unknown parameter. We then ask, "Which value of the parameter makes our observed data most likely to have happened?"

The Likelihood Function

Definition: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \ldots, \theta_m)$. Suppose that the parameter vector $(\theta_1, \ldots, \theta_m)$ is restricted to a given parameter space Ω.

The joint probability density (or mass) function of $X_1, X_2, \ldots, X_n$, when regarded as a function of $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_m$, is called the **Likelihood Function**, denoted by L:

$$L(\boldsymbol{\theta}) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

The first equality is merely the definition of the joint probability mass function. The second equality stems from the fact that it is a random sample, which inherently means that the samples are independent.

The function is defined for all parameter vectors $(\theta_1, \ldots, \theta_m)$ in the parameter space Ω.

Example: Likelihood for Bernoulli Data

Given a random sample of n independent coin flips $X_1, \ldots, X_n$, where $X_i \in \{0, 1\}$, and the probability of success is p, the probability mass function is $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$. The likelihood function $L(p)$ is the product of the individual probabilities:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

The Necessity of Log-Likelihood

For large sample sizes, the Likelihood Function $L(\theta)$, which is a product of many probability values (each between 0 and 1), quickly becomes an extremely small number. This causes two main problems:

1. **Numerical Underflow:** Computational systems may fail to store these tiny numbers accurately, leading to a value of zero (underflow).
2. **Calculus Complexity:** Products are difficult to differentiate. By taking the logarithm, the product simplifies to a sum, which is mathematically much easier to optimize.

Mathematical Justification for Log-Likelihood

The use of the natural logarithm is justified because it is a *strictly monotonically increasing function*. This property ensures that the maximum point of the Likelihood function $L(\theta)$ is exactly the same as the maximum point of the Log-Likelihood function $\ln L(\theta)$.

Consider any two parameter values θ_a and θ_b:

$$L(\theta_a) > L(\theta_b) \iff \ln(L(\theta_a)) > \ln(L(\theta_b))$$

Therefore, maximizing the simpler function, $\ln L(\theta)$, yields the identical maximum likelihood solution, $\hat{\theta}$. The optimization procedure involves differentiating the Log-Likelihood, setting it to zero, and solving for θ:

$$\frac{\partial \ln L(\theta)}{\partial \theta} \overset{\text{set}}{=} 0$$

Example: Log-Likelihood for Bernoulli Data

Applying the natural logarithm to the Bernoulli likelihood function $L(p)$ from the example above:

$$\ln L(p) = \ln \left(p^{\sum x_i} (1-p)^{n - \sum x_i} \right)$$

Using the logarithm property $\ln(a^b c^d) = b \ln(a) + d \ln(c)$:

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i \right) \ln(p) + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p)$$

Thus (remember, the necessary condition for a maximum is that the derivative equals 0):

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \overset{\text{set}}{=} 0$$

Then:

$$\left(\sum x_i \right) (1-p) - \left(n - \sum x_i \right) p = 0 \implies \sum x_i - np = 0$$

Solving for p (hat on the parameter $\hat{p}$ indicates that it is an estimate) to get an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ or, alternatively, an estimator: } \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimators vs. Estimates

A common source of confusion in mathematical statistics is the distinction between an *estimator* and an *estimate*. While used interchangeably in casual conversation, they represent distinct mathematical concepts.

The Recipe vs. The Meal

Before defining formally, to understand the difference intuitively, consider the analogy of cooking:

1. **The Estimator (The Recipe):** Think of an estimator as a *rule* or a *formula*. It tells you how to process ingredients (data) to get a result. Before you actually cook, the recipe exists as a procedure. In statistics, the estimator is the formula we choose to use (e.g., "I will calculate the average of whatever data I collect"). Because we haven't collected the data yet, the result of the estimator is unknown and random.

2. **The Estimate (The Meal):** Think of the estimate as the specific result you get after you have finished cooking. It is a concrete, fixed product. In statistics, once you collect your specific data points and plug them into the formula, you get a single number (e.g., "The average is 5.2"). This number is the estimate.

The Statistic

A **statistic** is formally defined as any function of the observable random variables in a sample that does not depend on any unknown parameters.

Let $X_1, X_2, \ldots, X_n$ be a random sample. A statistic T is any function:

$$T = g(X_1, X_2, \ldots, X_n)$$

Since T is a function of random variables, the statistic T itself is a random variable and has a probability distribution (called its sampling distribution).

The Estimator

An **estimator** is a statistic, which is a function of the random sample $X_1, X_2, \ldots, X_n$. Because the sample variables X_i are random variables (before the experiment is performed), the estimator is also a random variable.

Let Θ be the parameter of interest. An estimator $\hat{\Theta}$ is defined as:

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

The function h is the mathematical **rule or procedure** that maps the values of the random sample space (the observed data) to a value in the parameter space (the estimate). It represents the specific formula chosen for estimation (e.g., the formula for the sample mean or the sample variance). The choice of h determines the properties of the estimator $\hat{\Theta}$.

The Estimate

An **estimate** is a specific realization of the estimator. It is calculated using the observed values $x_1, x_2, \ldots, x_n$ (after the experiment is performed).

Let $x_1, \ldots, x_n$ be the observed data. The estimate $\hat{\theta}$ is:

$$\hat{\theta} = h(x_1, x_2, \ldots, x_n)$$

Feature	**Estimator**	**Estimate**
Symbol	$\hat{\Theta}$ or $\bar{X}$ (Capital)	$\hat{\theta}$ or $\bar{x}$ (Lowercase)
Mathematical Nature	Random Variable (Function)	Constant (Real Number)
Input	Random Sample $X_1, \ldots, X_n$	Observed Data $x_1, \ldots, x_n$
Timing	Pre-data collection	Post-data collection
Properties	Bias, Variance, Consistency	Accuracy (for a specific case)

Table 1: Comparison of Estimators and Estimates

2 Generalized Linear Models

A generalized linear model (GLM) is a statistical framework for analyzing relationships between variables where the outcomes, referred to as response variables, can follow various patterns - not just a straight-line relationship. The response variable is the specific outcome being predicted or explained, such as the number of successes in a series of trials or the probability of an event occurring. GLMs enable the linking of these response variables to one or more influencing factors, accommodating scenarios where the outcomes are counts, proportions, or other types of data rather than just averages.

Understanding traditional linear models is essential before exploring the complexities of GLMs. Traditional linear models establish a direct relationship between a response variable and one or more predictor variables, providing the foundation for the more flexible approach offered by GLMs.

Linear Models

The simplest and most intuitive way to model the relationship between variables is to assume a straight-line connection. Imagine trying to predict someone's weight based on their height - we might reasonably expect that taller people tend to weigh more, and this relationship could be approximated by a line. Linear models capture this idea mathematically, providing a framework for understanding how changes in one variable correspond to changes in another.

In a linear model, the response variable is expressed as a weighted combination of predictor variables plus some random noise that accounts for natural variability and measurement error. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the discrepancies between our predictions and the actual observed values.

The Linear Model Framework

Formally, a linear model expresses the relationship between a response variable Y and predictor variables $X_1, X_2, \ldots, X_p$ as:

$$Y_i = \beta_0 + \left(\sum_{j=1}^{p} \beta_j X_{i_j} \right) + \epsilon_i$$

where:

- Y_i is the i-th observation of the response variable
- X_{i_j} represents the j-th predictor variable for the i-th observation
- β_0 is the intercept term (the expected value of Y when all predictors are zero)
- $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients (slopes) that quantify the effect of each predictor
- ϵ_i is the random error term for observation i assumed to be independent and identically distributed (i.i.d.) such that $\mathbb{E}[\epsilon_i] = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ (where σ^2 is a constant). Typically, it is assumed that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In matrix notation, the linear model for all n observations can be written compactly as (with the first column of $\mathbf{X}$ typically being all ones for the intercept):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ with } \mathbf{Y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times(p+1)}, \boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}, \boldsymbol{\epsilon} \in \mathbb{R}^n$$

Maximum Likelihood Estimation for Linear Models

Under the normality assumption for the errors, the response variable Y_i follows a normal distribution:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{where} \quad \mu_i = \beta_0 + \left(\sum_{j=1}^{p} \beta_j X_{i_j}\right)$$

The probability density function for observation Y_i is:

$$f(y_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

Following the procedure outlined in Section 1, the likelihood function for all n independent observations is:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

Taking the natural logarithm yields the log-likelihood function:

$$\begin{aligned}
\ln L(\boldsymbol{\beta}, \sigma^2) &= \ln\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(\exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)\right)\right] \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right] \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2
\end{aligned}$$

Substituting $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$ (where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$ as a column vector):

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

To find the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$, differentiate with respect to $\boldsymbol{\beta}$ and set equal to zero:

$$\frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{x}_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) = \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{\text{set}}{=} \mathbf{0}$$

Solving this equation yields the "normal equations":

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}$$

Assuming $\mathbf{X}^T \mathbf{X}$ is invertible, the maximum likelihood estimator for $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

This is the ordinary least squares (OLS) estimator. Remarkably, under the normality of the error term assumption, the maximum likelihood estimator and the least squares estimator coincide.

For the variance parameter, differentiating the log-likelihood with respect to σ^2 and setting to zero yields:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Limitations of Linear Models

While linear models are powerful and widely applicable, they impose several restrictive assumptions that may not hold in many practical scenarios:

1. **Normality of Errors:** The assumption that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ may be violated when the response variable is binary, count-based, or highly skewed.

2. **Constant Variance (Homoscedasticity):** Linear models assume that the variance of the error terms is constant across all levels of the predictors. In reality, variance often depends on the mean (e.g., count data typically exhibit variance proportional to the mean).

3. **Unbounded Predictions:** The linear predictor $\mathbf{X}\boldsymbol{\beta}$ can take any real value from $-\infty$ to $+\infty$. This is problematic when modeling probabilities (which must lie in $[0, 1]$), counts (which must be non-negative), or other constrained quantities.

4. **Linear Relationship:** The model assumes that the expected value of Y is a linear function of the predictors. Non-linear relationships cannot be captured without transformation.

These limitations motivate the development of Generalized Linear Models, which relax many of these assumptions while retaining the interpretability and computational advantages of the linear framework. GLMs achieve this flexibility by introducing a link function that connects the linear predictor to the expected value of the response through a non-linear transformation, and by allowing the response to follow distributions beyond the normal distribution - including binomial, Poisson, gamma, and others from the exponential family.

Structure of Generalized Linear Models

Generalized Linear Models extend the linear modeling framework by relaxing the restrictive assumptions imposed by ordinary linear models. While linear models are limited to normally distributed responses with constant variance, GLMs accommodate a much broader class of response distributions and allow the variance to depend on the mean in a natural way.

Exponential Family

Most commonly used statistical distributions belong to the exponential family, whose probability density (or mass) functions can be written in the canonical from [4]:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi + c(y,\phi)}\right)$$

where:

- θ is the canonical parameter (also called the natural parameter)
- ϕ is the dispersion parameter
- $b(\theta)$ and $c(y,\phi)$ are specific functions that define the distribution

A remarkable property of the exponential family is that the mean and variance can be derived directly from the function $b(\theta)$:

$$\mathbb{E}[Y] = b'(\theta) = \mu$$
$$\text{var}(Y) = \phi b''(\theta) = \phi V(\mu)$$

where $b'(\theta)$ and $b''(\theta)$ denote the first and second derivatives of b with respect to θ.

The Three Components of a GLM

A GLM consists of three components that work together to model the relationship between predictors and response:

1. **The Random Component:** This specifies the probability distribution of the response variable Y_i. In a GLM, it is assumed that Y_i follows a distribution from the *exponential family*, which includes:
   - Normal distribution (for continuous data)
   - Binomial distribution (for binary or proportion data)
   - Poisson distribution (for count data)
   - Gamma distribution (for positive continuous data)
2. **The Systematic Component:** This is the linear predictor, which combines the predictor variables linearly:

   $$\mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_p x_{i_p} = \sum_{j=0}^{p} \beta_j x_{ij}$$

   where $x_{i_0} = 1$ for all i to account for the intercept term. The linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ can take any real value in $\mathbb{R}$.

3. **The Link Function:** This function $g(\cdot)$ connects the mean of the response variable $\mu_i = \mathbb{E}(Y_i)$ to the linear predictor:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

The link function transforms the expected value of the response (which may be constrained to a specific range) to the scale of the linear predictor (which is unrestricted). The inverse link function is denoted as g^{-1}, so that:

$$\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

The Variance Function

In addition to these three components, GLMs specify how the variance of the response depends on its mean through a variance function:

$$\text{var}(Y_i) = \phi V(\mu_i)$$

where ϕ is the dispersion parameter (a constant) and $V(\mu_i)$ is the variance function that describes how variance changes with the mean. For the normal distribution, $V(\mu_i) = 1$ (constant variance), but for other distributions, the variance typically depends on μ_i.

Canonical Links

For any exponential family distribution, there exists a special link function called the **canonical link**. The canonical link is defined as:

$$g(\mu_i) = g(b'(\theta_i)) = (b'(\theta_i))^{-1} = \beta_0 + \beta_1 x_{i_1} + \cdots + \beta_p x_{i_p} = \theta_i$$

In other words, the canonical link function sets the linear predictor equal to the canonical parameter. Canonical links lead to simpler mathematical forms for the likelihood equations and maximum likelihood estimation becomes numerically more stable. However, the canonical link is not always the most appropriate choice for modeling. While it offers computational and theoretical advantages, the choice of link function should ultimately be guided by the scientific context and the interpretability of the model parameters.

Common GLM Families and Their Properties

Every probability and statistics course covers these distributions, but one might not be very familiar with the distributions in the exponential family form.

Normal Distribution (Gaussian Family)

For the normal distribution with mean μ and variance σ^2:

Canonical form:

$$f(y; \mu, \sigma^2) = \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right)$$

Properties:

- Canonical parameter: $\theta = \mu$
- $b(\theta) = \frac{1}{2}\theta^2$
- Dispersion parameter: $\phi = \sigma^2$
- Variance function: $V(\mu) = 1$
- Canonical link: $g(\mu) = \mu$ (identity link)
- Mean-variance relationship: $\text{var}(Y) = \sigma^2$ (constant)

This is precisely the ordinary linear model, showing that linear regression is a special case of GLMs.

Binomial Distribution from [4]

Suppose $Y_i \sim \text{Binomial}(n_i, p_i)$, where n_i is the number of trials and p_i is the probability of success. The proportions Y_i/n_i are typically modeled.

Properties:

- Mean: $\mathbb{E}[Y_i/n_i] = p_i$
- Variance: $\text{var}\left(Y_i/n_i\right) = \dfrac{1}{n_i} p_i(1-p_i)$
- Canonical parameter: $\theta = \log\left(\frac{p}{1-p}\right)$ (log-odds)
- Dispersion parameter: $\phi = 1$ (known)
- Variance function: $V(\mu_i) = \mu_i(1-\mu_i)$
- Canonical link: $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ (logit link)

The logit link function maps probabilities from the interval $(0, 1)$ to the entire real line $(-\infty, \infty)$, making it appropriate for the linear predictor. The inverse logit (logistic function) is:

$$\mu_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}$$

Alternative link functions for binomial data include:

- Probit link: $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the standard normal CDF
- Complementary log-log link: $g(\mu) = \log(-\log(1-\mu))$

Poisson Distribution

Suppose $Y_i \sim \text{Poisson}(\lambda_i)$, typically used for modeling count data.

Properties:

- Mean: $\mathbb{E}(Y_i) = \lambda_i$
- Variance: $\text{var}(Y_i) = \lambda_i$
- Canonical parameter: $\theta = \log(\lambda)$
- Dispersion parameter: $\phi = 1$ (known)
- Variance function: $V(\mu_i) = \mu_i$
- Canonical link: $g(\mu_i) = \log(\mu_i)$ (log link)

The log link ensures that the predicted mean is always positive, which is necessary for count data. The inverse link is:

$$\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

This means that the effects of predictors are multiplicative on the original scale:

$$\mathbb{E}[Y_i] = \exp(\beta_0 + \beta_1 x_{i_1} + \cdots + \beta_p x_{i_p})$$

A one-unit increase in predictor x_j multiplies the expected count by a factor of $\exp(\beta_j)$.

Maximum Likelihood Estimation for GLMs

Parameter estimation in GLMs is performed using maximum likelihood. For a sample $y_1, y_2, \ldots, y_n$ from an exponential family distribution, the log-likelihood is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi_i + c(y_i, \phi_i)} \right\}$$

Following [4], assume that the dispersion parameters can be written as:

$$\phi_i = \frac{\phi}{a_i}$$

where ϕ is a single dispersion parameter and a_i are known prior weights. For example, binomial proportions with known index n_i have $\phi = 1$ and $a_i = n_i$.

The maximum likelihood estimates are obtained by solving the score equations:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = \sum_{i=1}^{n} \frac{a_i (y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0$$

for each parameter β_j, where $g'(\mu_i)$ is the derivative of the link function with respect to μ_i. Note that these estimating equations do not depend on ϕ, which may be unknown.

Iteratively (Re-)Weighted Least Squares (IRLS)

Unlike ordinary linear models where a closed-form solution exists, GLMs generally require iterative numerical methods. The standard algorithm is Iteratively (Re-)Weighted Least Squares (IRLS), also known as Fisher's Method of Scoring. For models using the canonical link, IRLS simplifies to the Newton-Raphson method, as the expected and observed information matrices coincide.

Algorithm 1 Iteratively (Re-)Weighted Least Squares (IRLS) for GLMs

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, response vector $\mathbf{y} \in \mathbb{R}^n$, link function $g(\cdot)$, variance function $V(\cdot)$, prior weights $\mathbf{a} \in \mathbb{R}^n$, tolerance $\epsilon > 0$, maximum iterations M

Ensure: Parameter estimates $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$

Initialize: Set $\boldsymbol{\beta}^{(0)} \leftarrow \mathbf{0}$ (or other suitable starting values)
Compute $\boldsymbol{\eta}^{(0)} \leftarrow \mathbf{X}\boldsymbol{\beta}^{(0)}$
Compute $\boldsymbol{\mu}^{(0)} \leftarrow g^{-1}(\boldsymbol{\eta}^{(0)})$ ▷ Apply inverse link elementwise
Set $r \leftarrow 0$ ▷ For termination
repeat
    // Step 1: Compute working responses
    for $i = 1$ to n **do**
        $z_i^{(r)} \leftarrow \mathbf{x}_i^T \boldsymbol{\beta}^{(r)} + (y_i - \mu_i^{(r)}) \cdot g'(\mu_i^{(r)})$
    end for
    // Step 2: Compute iterative weights
    for $i = 1$ to n **do**
        $w_i^{(r)} \leftarrow \dfrac{a_i}{V(\mu_i^{(r)}) \cdot (g'(\mu_i^{(r)}))^2}$
    end for
    $\mathbf{W}^{(r)} \leftarrow \mathrm{diag}(w_1^{(r)}, w_2^{(r)}, \ldots, w_n^{(r)})$
    // Step 3: Weighted least squares update
    $\boldsymbol{\beta}^{(r+1)} \leftarrow (\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{z}^{(r)}$
    // Step 4: Update fitted values
    $\boldsymbol{\eta}^{(r+1)} \leftarrow \mathbf{X}\boldsymbol{\beta}^{(r+1)}$
    $\boldsymbol{\mu}^{(r+1)} \leftarrow g^{-1}(\boldsymbol{\eta}^{(r+1)})$
    // Check convergence
    $\delta \leftarrow \|\boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)}\|_2$
    $r \leftarrow r + 1$
until $\delta < \epsilon$ **or** $r \geq M$
return $\boldsymbol{\beta}^{(r)}$

3 Model Selection and Evaluation (For Fun)

Once a GLM has been fitted, it is essential to evaluate its adequacy and compare it with alternative models. This section covers the primary tools for model selection and evaluation in the GLM framework, including deviance, the Akaike Information Criterion, and residual analysis.

Deviance

The deviance is a key measure of model fit in GLMs, generalizing the residual sum of squares from linear regression. It quantifies how well the fitted model explains the data by comparing it to a saturated model. The saturated model is one in which the number of parameters equals the number of observations, yielding perfect fit with $\hat{y}_i = y_i$ for all observations. The deviance measures twice the difference in log-likelihood between the saturated model and the fitted model, scaled by the dispersion parameter.

Deviance

The deviance is a key measure of model fit in GLMs, generalizing the residual sum of squares from linear regression. It quantifies how well the fitted model explains the data by comparing it to a saturated model.

Definition

The deviance of a fitted model is defined as:

$$D = 2\phi(L_{\text{sat}} - L_{\text{mod}})$$

where:

- L_{mod} is the log-likelihood of the fitted model
- L_{sat} is the log-likelihood of the saturated model
- ϕ is the dispersion parameter

The saturated model is one in which the number of parameters equals the number of observations, yielding perfect fit with $\hat{y}_i = y_i$ for all observations.

Example: Saturated Model

Consider a simple dataset with three observations: $y_1 = 2$, $y_2 = 5$, $y_3 = 3$. In a saturated model, there are 3 parameters (one per observation), resulting in the perfect fit:

$$\hat{\mu}_1 = 2, \quad \hat{\mu}_2 = 5, \quad \hat{\mu}_3 = 3$$

This model always achieves $\hat{y}_i = y_i$, representing the best possible fit. Any simpler model with fewer parameters (e.g., $\mu_i = \beta_0 + \beta_1 x_i$ with only two parameters) will generally fit less well, and the deviance measures this loss of fit.

Properties and Interpretation

- **Scale:** The deviance is always non-negative, with $D = 0$ indicating perfect fit (the fitted model equals the saturated model)
- **Degrees of Freedom:** The deviance has $n - p$ degrees of freedom, where n is the number of observations and p is the number of parameters (including the intercept)
- **Comparison:** For nested models, the difference in deviances follows approximately a chi-squared distribution under certain regularity conditions, enabling likelihood ratio tests. Two models are **nested** when one model (the reduced model) is a special case of the other (the full model), obtained by setting some parameters to zero or imposing constraints. For example, the model $\mu_i = \beta_0 + \beta_1 x_{i_1}$ is nested within $\mu_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}$.

  Specifically, if Model 0 (reduced) has p_0 parameters with deviance D_0, and Model 1 (full) has $p_1 > p_0$ parameters with deviance D_1, then under the null hypothesis that the reduced model is adequate:

$$\chi^2 = D_0 - D_1 \sim \chi^2_{p_1 - p_0} = \sum_{i=1}^{p_1 - p_0} Z_i \text{ where } Z_i \sim \mathcal{N}(0,1) \text{ (i.i.d.)}$$

  where $\chi^2_{p_1 - p_0}$ denotes a chi-squared distribution with $p_1 - p_0$ degrees of freedom.

Akaike Information Criterion (AIC)

The Akaike Information Criterion is a widely used measure for model selection that balances goodness of fit against model complexity. Unlike deviance-based tests, AIC can be used to compare non-nested models.

Definition

The AIC is defined as ($L_{\text{mod}} \leftarrow$ the log-likelihood of the fitted model; $p \leftarrow$ the number of parameters):

$$\text{AIC} = -2L_{\text{mod}} + 2p$$

Interpretation and Use

- **Penalty Term:** The term $2p$ penalizes model complexity, discouraging overfitting
- **Comparison:** Lower AIC values indicate better models. AIC does not provide an absolute measure of fit, only a relative comparison between models
- **Non-nested Models:** AIC can compare models that are not nested
- **Trade-off:** AIC explicitly trades off goodness of fit (measured by $-2L_{\text{mod}}$) against model parsimony (measured by $2p$)

Note: AIC values are only comparable when models are fitted to the same data and use the same probability distribution. Constants in the log-likelihood function are sometimes omitted in computational implementations, so AIC values for models with different error distributions (e.g., normal vs. gamma) may not be directly comparable.

Likelihood Ratio Tests

For nested models (where one model is a special case of another), the difference in deviances can be used to test whether the additional parameters significantly improve the fit.

Test Statistic

Given two nested models:

- Model 0 (reduced): p_0 parameters, deviance D_0
- Model 1 (full): p_1 parameters, deviance D_1, where $p_1 > p_0$

The likelihood ratio test statistic is:

$$\Lambda = D_0 - D_1 = 2(L_1 - L_0)$$

Under the null hypothesis that the reduced model is adequate, Λ follows approximately a chi-squared distribution with $p_1 - p_0$ degrees of freedom:

$$\Lambda \sim \chi^2_{p_1 - p_0}$$

Decision Rule

The null hypothesis is rejected (and favor the more complex model) if:

$$\Lambda > \chi^2_{p_1 - p_0, \alpha}$$

where $\chi^2_{p_1 - p_0, \alpha}$ is the critical value of the chi-squared distribution at significance level α.

Model Selection Strategy

A systematic approach to model selection typically involves:

1. **Start with a Full Model:** Include all potentially relevant predictors and interactions
2. **Compare Nested Models:** Use likelihood ratio tests (deviance differences) to test whether terms can be removed
3. **Use AIC for Non-nested Comparisons:** When comparing models that are not nested (e.g., different link functions or variance structures), use AIC
4. **Consider Scientific Context:** Statistical criteria should be balanced with subject-matter knowledge and interpretability

4 Calculating MLE for Binomial GLM

This section demonstrates the complete maximum likelihood estimation procedure for a binomial GLM with the canonical logit link. We will derive the log-likelihood, score equations, and show how the IRLS algorithm applies to this specific case.

Model Setup

Suppose we have N independent observations $Y_1, Y_2, \ldots, Y_N$ where:

$$Y_i \sim \text{Binomial}(n_i, p_i)$$

Here, n_i is the known number of trials for observation i, and p_i is the unknown probability of success we wish to model. The expected value is $\mathbb{E}[Y_i] = n_i p_i$ and variance is $\text{var}(Y_i) = n_i p_i (1-p_i)$.

We model the probability p_i using the logit link (the canonical link for binomial data):

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i_1} + \cdots + \beta_p x_{i_p}$$

The inverse link gives us the *sigmoid* function:

$$p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}$$

Likelihood Function

The probability mass function for a single observation Y_i is:

$$f(y_i; p_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

Since the observations are independent, the likelihood function is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

Log-Likelihood Function

Taking the natural logarithm:

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\ln \binom{n_i}{y_i} + y_i \ln(p_i) + (n_i - y_i) \ln(1-p_i) \right]$$

Substituting $p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$ and $1 - p_i = \frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$:

$$\ln(p_i) = \ln\left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right) = \mathbf{x}_i^T \boldsymbol{\beta} - \ln(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}})$$

$$\ln(1-p_i) = \ln\left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right) = -\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})$$

Substituting these into the log-likelihood:

$$\begin{aligned}
\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{N}\left[\ln\binom{n_i}{y_i} + y_i(\mathbf{x}_i^T\boldsymbol{\beta} - \ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})) - (n_i - y_i)\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right] \\
&= \sum_{i=1}^{N}\left[\ln\binom{n_i}{y_i} + y_i\mathbf{x}_i^T\boldsymbol{\beta} - n_i\cdot\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right] \\
&= \sum_{i=1}^{N}\ln\binom{n_i}{y_i} + \sum_{i=1}^{N}\left[y_i\mathbf{x}_i^T\boldsymbol{\beta} - n_i\cdot\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right]
\end{aligned}$$

The first term does not depend on $\boldsymbol{\beta}$ and can be ignored during optimization. Thus, the log-likelihood to maximize is:

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left[y_i\mathbf{x}_i^T\boldsymbol{\beta} - n_i\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right]$$

Score Equations

To find the maximum likelihood estimates, we compute the score function (the gradient of the log-likelihood with respect to $\boldsymbol{\beta}$):

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial\beta_j} = \sum_{i=1}^{N}\left[y_i\frac{\partial\mathbf{x}_i^T\boldsymbol{\beta}}{\partial\beta_j} - n_i\cdot\frac{\partial}{\partial\beta_j}\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right]$$

Since $\mathbf{x}_i^T\boldsymbol{\beta} = \sum_{k=0}^{p}\beta_k x_{i_k}$ (with $x_{i0}=1$), we have $\frac{\partial\mathbf{x}_i^T\boldsymbol{\beta}}{\partial\beta_j} = x_{i_j}$.

For the second term:

$$n_i\cdot\frac{\partial}{\partial\beta_j}\ln(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}) = n_i\cdot\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\cdot\frac{\partial\mathbf{x}_i^T\boldsymbol{\beta}}{\partial\beta_j} = n_i\cdot p_i\cdot x_{i_j}$$

Therefore:

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial\beta_j} = \sum_{i=1}^{N}x_{i_j}(y_i - n_i\cdot p_i) = \sum_{i=1}^{N}x_{i_j}(y_i-\mu_i)$$

where $\mu_i = n_i\cdot p_i = \mathbb{E}[Y_i]$ is the expected value.

Setting the score equations to zero:

$$\sum_{i=1}^{N}x_{i_j}(y_i-\mu_i) = 0 \quad \text{for } j = 0,1,\ldots,p$$

In matrix form:

$$\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}$$

These equations do not have a closed-form solution and must be solved iteratively using IRLS.

IRLS Components for Binomial GLM

For the binomial GLM with logit link, we need to specify the components of the IRLS algorithm (Algorithm 1):

1. Link function and its derivative:

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}$$

$$g'(p_i) = \frac{\partial}{\partial p_i}\log\left(\frac{p_i}{1-p_i}\right) = \frac{1}{p_i(1-p_i)}$$

2. Inverse link function:

$$p_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$$

3. Variance function:
For binomial data with $Y_i \sim \text{Binomial}(n_i, p_i)$, we have:

$$\text{var}(Y_i) = n_i p_i(1-p_i) = n_i \cdot \frac{\mu_i}{n_i}\left(1-\frac{\mu_i}{n_i}\right) = \mu_i\left(1-\frac{\mu_i}{n_i}\right)$$

where $\mu_i = n_i p_i$. Expanding further:

$$\text{var}(Y_i) = \mu_i\left(1-\frac{\mu_i}{n_i}\right) = \mu_i - \frac{\mu_i^2}{n_i} = \frac{n_i\mu_i - \mu_i^2}{n_i} = \frac{\mu_i(n_i-\mu_i)}{n_i}$$

The variance function $V(\mu_i)$ is defined such that $\text{var}(Y_i) = \phi V(\mu_i)$ with $\phi = 1$ for binomial data. Therefore:

$$V(\mu_i) = \frac{\mu_i(n_i-\mu_i)}{n_i}$$

Equivalently, in terms of $p_i = \mu_i/n_i$:

$$V(\mu_i) = n_i p_i(1-p_i)$$

4. Working responses:

$$z_i^{(r)} = \mathbf{x}_i^T\boldsymbol{\beta}^{(r)} + (y_i - \mu_i^{(r)})\cdot g'(\mu_i^{(r)}/n_i) = \mathbf{x}_i^T\boldsymbol{\beta}^{(r)} + \frac{(y_i-\mu_i^{(r)})n_i}{\mu_i^{(r)}(n_i-\mu_i^{(r)})}$$

5. Iterative weights (with prior weights $a_i = n_i$):

$$w_i^{(r)} = \frac{n_i}{V(\mu_i^{(r)})\cdot(g'(\mu_i^{(r)}/n_i))^2} = \frac{n_i}{\frac{\mu_i^{(r)}(n_i-\mu_i^{(r)})}{n_i}\cdot\frac{n_i^2}{(\mu_i^{(r)})^2(n_i-\mu_i^{(r)})^2}} = \frac{\mu_i^{(r)}(n_i-\mu_i^{(r)})}{n_i}$$

Equivalently, in terms of $p_i^{(r)} = \mu_i^{(r)}/n_i$:

$$w_i^{(r)} = n_i p_i^{(r)}(1-p_i^{(r)})$$

Algorithm 2 IRLS for Binomial GLM with Logit Link

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{N \times (p+1)}$, response vector $\mathbf{y} \in \mathbb{R}^N$, trial counts $\mathbf{n} \in \mathbb{R}^N$, tolerance $\epsilon > 0$, maximum iterations M

Ensure: Parameter estimates $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$

Initialize: Set $\boldsymbol{\beta}^{(0)} \leftarrow \mathbf{0}$ (or use starting values from linear probability model)
Set $r \leftarrow 0$
repeat
  // **Step 1: Compute linear predictors**
  for $i = 1$ to N **do**
    $\eta_i^{(r)} \leftarrow \mathbf{x}_i^T \boldsymbol{\beta}^{(r)}$
  end for
  // **Step 2: Compute fitted probabilities and means**
  for $i = 1$ to N **do**
    $p_i^{(r)} \leftarrow \dfrac{e^{\eta_i^{(r)}}}{1 + e^{\eta_i^{(r)}}}$ ▷ Inverse logit
    $\mu_i^{(r)} \leftarrow n_i \cdot p_i^{(r)}$
  end for
  // **Step 3: Compute working responses**
  for $i = 1$ to N **do**
    $z_i^{(r)} \leftarrow \eta_i^{(r)} + \dfrac{y_i - \mu_i^{(r)}}{n_i p_i^{(r)}(1 - p_i^{(r)})}$
  end for
  // **Step 4: Compute iterative weights**
  for $i = 1$ to N **do**
    $w_i^{(r)} \leftarrow n_i p_i^{(r)}(1 - p_i^{(r)})$
  end for
  $\mathbf{W}^{(r)} \leftarrow \text{diag}(w_1^{(r)}, w_2^{(r)}, \ldots, w_N^{(r)})$
  // **Step 5: Weighted least squares update**
  $\boldsymbol{\beta}^{(r+1)} \leftarrow (\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{z}^{(r)}$
  // **Step 6: Check convergence**
  $\delta \leftarrow \|\boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)}\|_2$
  $r \leftarrow r + 1$
until $\delta < \epsilon$ **or** $r \geq M$
return $\boldsymbol{\beta}^{(r)}$

Special Case: Logistic Regression

When $n_i = 1$ for all observations (Bernoulli trials), this becomes standard logistic regression:

- $\mu_i = p_i$ (the expected value equals the probability)
- $V(\mu_i) = \mu_i(1 - \mu_i)$
- $w_i^{(r)} = p_i^{(r)}(1 - p_i^{(r)})$
- $z_i^{(r)} = \eta_i^{(r)} + \frac{y_i - p_i^{(r)}}{p_i^{(r)}(1 - p_i^{(r)})}$

The interpretation of the coefficients is in terms of log-odds: a one-unit increase in predictor x_j changes the log-odds of success by β_j, or equivalently, multiplies the odds by $\exp(\beta_j)$.

5 The Simulation

The following results are from running the code in [5].

References

[1] Langley, P., & Sage, S. (1994). **The importance of naive Bayes**. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 333-337). Morgan Kaufmann.

[2] McLachlan, G. J., & Basford, K. E. (1988). **Mixture Models: Inference and Applications to Clustering**. Wiley.

[3] **PennState - Eberly College of Science.** STAT 415 - Introduction to Mathematical Statistics, 1.2 - Maximum Likelihood Estimation. From `https://online.stat.psu.edu/stat415/lesson/1/1.2`

[4] Turner, H. **Introduction to Generalized Linear Models**. ESRC National Centre for Research Methods, UK, and Department of Statistics, University of Warwick, UK. Retrieved from `https://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf`

[5] **GLeMa GitHub Repository.** MLE meets GLM. `https://github.com/Stochastic-Batman/GLeMa`