# Generalized Linear Models Principles

Søren Lund Pedersen

## 1 Introduction

Generalized Linear Models (GLMs) represent a flexible generalization of ordinary linear regression that allows for the response variable to have a distribution other than the normal distribution. Introduced by John Nelder and Robert Wedderburn in 1972, GLMs have become a cornerstone in statistical modeling, providing a unified framework for modeling various types of data, including binary, count, and continuous responses.

In this article, I aim to provide the motivation, intuition, and framework behind GLMs. We will explore the components that make up GLMs, delve into estimation techniques, and discuss methods for assessing model fit. By the end of this article, you should have a comprehensive understanding of GLMs and how to apply them to diverse data types.

## 2 Motivation

The primary motivation for GLMs is to extend linear models, such as $y = ax + b + \epsilon$, which assume that the data is normally distributed, to accommodate a broader range of response variable distributions. Traditional linear models are limited in their applicability because they rely on the assumption of homoscedasticity (constant variance) and normally distributed errors. When modeling outcomes that fall into binary categories ($y \in \{0, 1\}$), count data, or time-to-event data, the Gaussian assumption becomes inappropriate. Using linear models for such data often results in poor fit and unreliable inference due to the violation of underlying assumptions.

GLMs address these limitations by allowing the response variable to follow any distribution from the exponential family. This flexibility enables the modeling of various types of data while maintaining desirable statistical properties, such as interpretability of coefficients and efficient estimation procedures.

## 3 Intuition

The intuition behind GLMs lies in their three core components:

1. **A random component**: Specifies the probability distribution of the response variable.
2. **A systematic component**: Relates the predictors to a linear predictor through a linear combination.
3. **A link function**: Connects the mean of the response variable to the linear predictor.

This structure allows GLMs to overcome the restrictions of ordinary linear regression and better capture the relationships in non-Gaussian data. By decoupling the mean of the response variable

from its linear predictor through the link function, GLMs provide the necessary flexibility to model different types of data effectively.

# 4 Framework

The framework of GLMs can be broken down into the following components:

## 4.1 A Random Component

Instead of assuming a normal distribution for the response variable, GLMs allow the response to follow any distribution from the exponential family. The exponential family includes a wide range of distributions, such as:

- **Normal Distribution**: For continuous data with constant variance.
- **Binomial Distribution**: For binary or proportion data.
- **Poisson Distribution**: For count data.
- **Gamma Distribution**: For positive continuous data with skewness.

These distributions are characterized by their probability density functions (PDFs) or probability mass functions (PMFs) being expressible in a specific exponential form, which facilitates the derivation of estimation procedures and inference.

## 4.2 Systematic Component

The systematic component relates the predictors to the response variable through a linear predictor. This is typically expressed as:

$$\eta = \mathbf{X}\boldsymbol{\beta}$$

where:

- $\eta$ is the linear predictor.
- $\mathbf{X}$ is the design matrix containing the predictor variables.
- $\boldsymbol{\beta}$ is the vector of coefficients to be estimated.

This linear combination of predictors allows us to model the effect of each predictor on the response variable while maintaining interpretability of the coefficients.

## 4.3 Link Function

The link function connects the mean of the response variable $\mu = \mathbb{E}[Y]$ to the linear predictor $\eta$. Formally, it is defined as:

$$g(\mu) = \eta = \mathbf{X}\boldsymbol{\beta}$$

The choice of link function depends on the distribution of the response variable and the nature of the relationship between the predictors and the mean response. Common examples include:

- **Identity Link**: $g(\mu) = \mu$, used for normal regression where the mean is directly modeled by the linear predictor.
- **Logit Link**: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, used for binary outcomes in logistic regression.

- **Log Link**: $g(\mu) = \log(\mu)$, used for count data in Poisson regression.
- **Inverse Link**: $g(\mu) = \frac{1}{\mu}$, used in certain types of gamma regression.

The link function ensures that the modeled mean $\mu$ remains within the appropriate range for the distribution, such as between 0 and 1 for binary outcomes or positive for count data.

# 5 Maximum Likelihood Estimation (MLE)

GLMs estimate the model parameters $\boldsymbol{\beta}$ using Maximum Likelihood Estimation (MLE). Unlike ordinary least squares (OLS) regression, which has a closed-form solution:

$$\boldsymbol{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

MLE for GLMs generally requires iterative numerical methods due to the lack of closed-form solutions for most distributions.

## 5.1 Estimation Techniques

Common numerical methods for MLE in GLMs include:

- **Newton-Raphson**: An iterative optimization algorithm that uses both the first and second derivatives of the log-likelihood to find the parameter estimates.
- **Iterative Weighted Least Squares (IWLS)**: A specialized version of the Newton-Raphson method tailored for GLMs, which iteratively fits weighted least squares models to update the estimates.
- **Fisher Scoring**: A variant of the Newton-Raphson method that uses the expected information (Fisher information) instead of the observed information.

These methods update the parameter estimates iteratively until convergence criteria are met, typically when changes in the log-likelihood or parameter estimates between iterations fall below a predefined threshold.

## 5.2 Convergence Criteria

To ensure reliable estimation, convergence criteria are employed, such as:

- The change in log-likelihood between iterations is below a specified tolerance.
- The change in parameter estimates between iterations is negligible.
- The gradient of the log-likelihood function approaches zero.

Proper convergence is crucial for obtaining accurate parameter estimates and avoiding issues like overfitting or underfitting the data.

# 6 Likelihood Ratio and Wald Tests

Assessing the significance of predictors and comparing models are essential aspects of GLMs. Two commonly used statistical tests in this context are the Likelihood Ratio (LR) test and the Wald test.

### 6.1 Likelihood Ratio (LR) Test

The LR test compares the goodness-of-fit of two nested models: a full (unrestricted) model and a reduced (restricted) model. The null hypothesis $H_0$ posits that the simpler model fits the data as well as the more complex model.

The LR test statistic is defined as:

$$LR = -2\left[\ell(\text{reduced model}) - \ell(\text{full model})\right],$$

where $\ell(\cdot)$ denotes the log-likelihood of the respective model.

Under the null hypothesis, the $LR$ statistic follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the full and reduced models. Thus, the decision rule is:

- If $LR$ exceeds the critical value from the chi-squared distribution, reject $H_0$ in favor of the full model.
- Otherwise, do not reject $H_0$, indicating that the reduced model is sufficient.

### 6.2 Wald Test

The Wald test assesses whether a specific parameter $\beta_i$ significantly differs from a hypothesized value (commonly zero). It is defined as:

$$W = \frac{\hat{\beta}_i}{\sqrt{\text{Var}(\hat{\beta}_i)}}.$$

Under the null hypothesis $H_0 : \beta_i = 0$, the Wald statistic follows a standard normal distribution.

Confidence intervals for $\beta_i$ can be constructed using the Wald test as follows:

$$\hat{\beta}_i \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{\beta}_i)},$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution for the desired confidence level.

## 7 Goodness of Fit

Evaluating how well a GLM fits the observed data is critical for model validation. Two key measures for assessing goodness of fit are Deviance and the Akaike Information Criterion (AIC).

### 7.1 Deviance

Deviance measures the discrepancy between the fitted model and a saturated model (a model with a parameter for every observation). It is defined as:

$$\text{Deviance} = 2\left[\ell(\text{Sat}) - \ell(\text{MLE})\right],$$

where:

- $\ell(\text{Sat})$ is the log-likelihood of the saturated model.

- $\ell(\text{MLE})$ is the log-likelihood of the fitted model.

Lower deviance values indicate a better fit. Deviance can be used to compare nested models using the Likelihood Ratio test, as a reduction in deviance suggests an improvement in model fit.

### 7.2 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a metric for model selection that balances model fit and complexity. It is calculated as:

$$\text{AIC} = -2\ell(\hat{\beta}) + 2k,$$

where:

- $\ell(\hat{\beta})$ is the log-likelihood of the model evaluated at the maximum likelihood estimates.
- $k$ is the number of parameters in the model.

AIC estimates the relative information loss when a given model is used to represent the true data-generating process. Lower AIC values indicate a better trade-off between goodness of fit and model complexity, thereby avoiding overfitting.

When comparing multiple models, the model with the lowest AIC is typically preferred. However, it's essential to note that AIC does not provide an absolute measure of fit but rather a relative one.

## 8 Residuals

Residuals are the differences between observed and fitted values and are vital for diagnosing the adequacy of a GLM. They provide insights into whether the model assumptions hold and if the model adequately captures the data structure.

### 8.1 Pearson Residuals

Pearson residuals adjust raw residuals for the variance of the response variable. They are defined as:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(y_i)}},$$

where:

- $y_i$ is the observed value.
- $\hat{\mu}_i$ is the fitted value (mean response) for observation $i$.
- $\text{Var}(y_i)$ is the variance of the response variable, often modeled as $\phi V(\mu_i)$, where $V(\mu_i)$ is the variance function and $\phi$ is the dispersion parameter.

Pearson residuals help in identifying patterns that may suggest model inadequacies, such as non-linearity, overdispersion, or heteroscedasticity. They are typically plotted against fitted values or predictors to visually assess the model fit.

### 8.2 Other Types of Residuals

In addition to Pearson residuals, other types of residuals are commonly used in GLMs:

- **Deviance Residuals**: Measure the contribution of each observation to the overall deviance. They are useful for identifying outliers and assessing the fit of individual observations.
- **Response Residuals**: Simply the difference between observed and fitted values, $y_i - \hat{\mu}_i$. They are less informative in the context of GLMs with non-identity link functions.

## 9 Diagnostics and Model Validation

After fitting a GLM, it's crucial to perform diagnostics to ensure that the model adequately captures the underlying data structure. Common diagnostic tools include residual plots, influence measures, and goodness-of-fit tests.

### 9.1 Residual Plots

Residual plots are graphical tools used to assess the fit of a model. By plotting residuals against fitted values or predictors, we can identify patterns that indicate potential model deficiencies:

- **No Pattern**: Indicates a good fit.
- **Patterns or Trends**: Suggests issues like non-linearity, heteroscedasticity, or missing variables.
- **Outliers**: Observations with unusually large residuals may indicate data entry errors or rare events.

### 9.2 Goodness-of-Fit Tests

In addition to deviance and AIC, other goodness-of-fit tests can be employed to assess the adequacy of a GLM:

- **Hosmer-Lemeshow Test**: Commonly used for logistic regression to assess calibration.
- **Pearson Chi-Square Test**: Compares observed and expected frequencies for categorical data.

These tests provide formal assessments of model fit, complementing graphical diagnostics.

## 10 Extensions and Advanced Topics

GLMs serve as a foundation for more complex modeling techniques, such as:

- **Generalized Additive Models (GAMs)**: Extend GLMs by allowing non-linear relationships between predictors and the response variable.
- **Mixed-Effects Models**: Incorporate both fixed and random effects, enabling the modeling of hierarchical or grouped data.
- **Zero-Inflated Models**: Address overdispersion in count data by modeling excess zeros.

These extensions provide additional flexibility to GLMs, enabling them to handle a wider range of data structures and complexities.

# 11 Conclusion

Generalized Linear Models offer a versatile and powerful framework for modeling diverse types of data by extending the principles of ordinary linear regression. By decoupling the mean of the response variable from its linear predictor through the link function and allowing for various distributions from the exponential family, GLMs accommodate binary, count, and continuous data effectively. Understanding the components, estimation techniques, and diagnostic tools associated with GLMs is essential for building robust and interpretable statistical models.

As data becomes increasingly complex and varied, the adaptability of GLMs ensures their continued relevance and utility in statistical analysis and applied research.