PMDS504L: Regression Analysis and Predictive Models

Generalized Linear Models

Dr. Jisha Francis

Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology
Vellore Campus, Vellore - 632 014
India

Table of Contents

1. Introduction to Generalized Linear Model
2. Logistic Regression
3. Estimating Parameters in a Logistic Regression Model
4. Poisson Regression:
5. References

Introduction

- In last session, we explored data transformation for regression models when assumptions of normality and constant variance are violated.
- Transformation is effective for handling response non-normality and variance inequality.
- Weighted least squares is another approach to address non-constant variance.
- This session introduces an alternative: Generalized Linear Models (GLM).

What is a Generalized Linear Model (GLM)?

- GLM unifies linear and nonlinear regression models, incorporating non-normal response distributions. (A non-normal response distribution refers to a situation where the response (dependent) variable in a statistical model does not follow a normal (Gaussian) distribution. In classical linear regression, it is assumed that the response variable is normally distributed with constant variance. However, in many real-world scenarios, this assumption is not valid.)
- Response variable must belong to the exponential family (Normal, Poisson, Binomial, Exponential, Gamma).
- The normal-error linear model is a special case of GLM, making it a powerful tool in empirical modeling.

Generalized Linear Model

- We begin by considering **logistic regression**, where the response variable has only two possible outcomes, typically denoted as 0 (failure) and 1 (success).
- The response is *qualitative*, as the designation of success or failure is arbitrary.
- Next, we explore cases where the response variable represents a **count**, such as:
  - The number of defects in a unit of a product.
  - The number of relatively rare events
- Finally, we discuss how these situations are **unified** under the **Generalized Linear Model (GLM)** framework.

Logistic Regression: Models with a Binary Response Variable

- Consider a regression problem where the response variable takes only two values: 0 or 1.
- These values may represent a qualitative response, e.g., a functional test on a semiconductor device:
    - 1 (Success): The device works properly.
    - 0 (Failure): Due to a short, an open, or another issue.
- Suppose the model has the form:

$$y = x'\beta + \varepsilon$$

where $x' = [1, x_1, x_2, \ldots, x_k]$, $\beta' = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$ and the response variable y takes on the value either 0 or 1.

Binary Response and Bernoulli Distribution

- We will assume that the response variable y_i follows a **Bernoulli distribution** with:

$$P(y_i = 1) = \pi_i, \quad P(y_i = 0) = 1 - \pi_i$$

- The expected value is:

$$E(y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i$$

- This implies:

$$E(y_i) = x'\beta = \pi_i$$

Issues with Linear Regression for Binary Response

- If the response is binary, the error terms ε_i take only two values:

$$\varepsilon_i = 1 - x'\beta \quad (\text{if } y = 1)$$

$$\varepsilon_i = -x'\beta \quad (\text{if } y = 0)$$

- The error terms cannot be normally distributed.
- The variance is not constant:

$$\sigma^2_{y_i} = \pi_i(1 - \pi_i)$$

- The response function must satisfy:

$$0 \leq E(y_i) = \pi_i \leq 1$$

Variance of Binary Response Variable

- The variance of a binary response variable is given by:

$$\sigma^2 = E\{(y - E(y))^2\} \tag{1}$$

- Expanding the expectation:

$$\sigma^2 = (1 - \pi)^2\pi + (0 - \pi)^2(1 - \pi) \tag{2}$$

- Simplifying:

$$\sigma^2 = \pi(1 - \pi) \tag{3}$$

- Since $E(y) = \pi$, this implies:

$$\sigma^2 = E(y)[1 - E(y)] \tag{4}$$

Constraint on Response Function

- The expected response must satisfy:

$$0 \leq E(y_i) = \pi_i \leq 1 \tag{5}$$

- A linear response function can result in predictions outside this range.
- To avoid this issue, a nonlinear response function is preferred.
- Empirical evidence supports a nonlinear, S-shaped response function.

Logistic Response Function

The logistic response function is given by:

$$E(y) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)} = \frac{1}{1 + \exp(-x'\beta)} \tag{6}$$

Since this function is nonlinear, it can be difficult to analyze using traditional linear regression techniques. However, it can be linearized by defining the structural portion of the model in terms of a transformation of the response function mean.

Logistic Response Function

The linear predictor is given by:

$$\eta = x'\beta \tag{7}$$

where η is related to the probability π through the logit transformation:

$$\eta = \ln\left(\frac{\pi}{1-\pi}\right) \tag{8}$$

This transformation, known as the **logit transformation**, maps probabilities (which are constrained between 0 and 1) to an unbounded scale, making the model more suitable for regression analysis.

The ratio $\frac{\pi}{1-\pi}$ in the transformation is called the **odds**, and the logit transformation is often referred to as the **log-odds**.

Estimating Parameters in a Logistic Regression Model

The general form of the logistic regression model is:

$$y_i = E(y_i) + \varepsilon_i \tag{9}$$

where the observations y_i are independent Bernoulli random variables with expected values:

$$E(y_i) = \pi_i = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \tag{10}$$

Estimating Parameters in a Logistic Regression Model

Maximum Likelihood Estimation (MLE) is used to estimate the parameters in the linear predictor $x_i'\beta$. The probability distribution of each observation is:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad i = 1, 2, \ldots, n \tag{11}$$

Since observations are independent, the likelihood function is:

$$L(y_1, y_2, \ldots, y_n, \beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{12}$$

Estimating Parameters in a Logistic Regression Model

Taking the natural logarithm, the log-likelihood function is:

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i x_i' \beta - \sum_{i=1}^{n} \ln[1 + \exp(x_i' \beta)] \tag{13}$$

Incorporating Repeated Observations

Extending to Binomial Responses

- In many experiments, we observe multiple trials at the same level of x.
- Instead of a Bernoulli response, we now have:

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

where:

- Y_i is the number of successes in n_i trials.
- π_i is the probability of success for the i-th observation.

Likelihood Function: $L(y, \beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$

Log-Likelihood: $\ln L(y, \beta) = \sum_{i=1}^{n} \left[y_i \ln \pi_i + (n_i - y_i) \ln(1 - \pi_i) \right]$

Incorporating Repeated Observations

Substituting π_i

$$\pi_i = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$$

Final Log-Likelihood Expression:

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i x_i'\beta - \sum_{i=1}^{n} n_i \ln\left(1 + \exp(x_i'\beta)\right)$$

Example: Vaccine Effectiveness Study

Scenario:

- A clinical trial is conducted to study vaccine effectiveness.
- Participants are grouped based on age.
- The number of infected individuals is recorded for each age group.

Data Representation:

Age Group (x_i)	**Total Participants** (n_i)	**Infected** (y_i)
20–30	100	12
30–40	120	15
40–50	90	20
50–60	80	25

Logistic Regression Model

Modeling the Probability of Infection:

- Let Y_i represent the number of infections in age group i.
- Y_i follows a binomial distribution:

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

- The log-odds of infection is modeled as:

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

Why This is a Case of Repeated Observations?

- Instead of a **single binary outcome (0 or 1) per individual**, we observe **multiple trials** (n_i) at each level of x.
- The response variable (Y_i) represents the **count of "successes" (infections)** rather than individual outcomes.
- This is common in **designed experiments**, where observations are grouped by factors like **age, location, or treatment level**.

MLE and Iteratively Reweighted Least Squares (IRLS)

Finding MLEs: Numerical optimization methods such as **Iteratively Reweighted Least Squares (IRLS)** can be used to estimate β.

Asymptotic Properties of Logistic Regression Estimates

Let $\hat{\beta}$ be the final estimate of the model parameters obtained from the iterative algorithm. If the model assumptions are correct, then asymptotically:

$$E(\hat{\beta}) = \beta, \quad \text{Var}(\hat{\beta}) = (X'VX)^{-1} \tag{14}$$

where V is an $n \times n$ diagonal matrix containing the estimated variance of each observation on the main diagonal. The ith diagonal element of V is:

$$V_{ii} = n_i \hat{\pi}_i (1 - \hat{\pi}_i) \tag{15}$$

Fitted Values in Logistic Regression

The estimated value of the linear predictor is:

$$\hat{\eta}_i = x_i'\hat{\beta} \tag{16}$$

The fitted value of the logistic regression model is:

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} \tag{17}$$

which can also be written as:

$$\hat{\pi}_i = \frac{1}{1 + \exp(-x_i'\hat{\beta})} \tag{18}$$

Interpretation of Parameters in Logistic Regression

Consider a logistic regression model with a single predictor variable. The fitted values for the linear predictor at x_i and $x_i + 1$ are:

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{19}$$

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 (x_i + 1) \tag{20}$$

The difference between the two predicted values is:

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1 \tag{21}$$

Odds Ratio Interpretation

Now, $\hat{\eta}(x_i)$ represents the log-odds when the regressor variable is equal to x_i, and $\hat{\eta}(x_i + 1)$ represents the log-odds when the regressor is equal to $x_i + 1$. Therefore, the difference in the two fitted values is:

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \ln(\text{odds}_{x_i+1}) - \ln(\text{odds}_{x_i}) \tag{22}$$

Odds Ratio

Rearranging, we obtain:

$$\ln\left(\frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}}\right) = \hat{\beta}_1 \tag{23}$$

Taking the antilogarithm, we obtain the odds ratio:

$$\text{OR} = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} = e^{\hat{\beta}_1} \tag{24}$$

Generalizing to Any Change *d*

If the predictor variable increases by d units instead of 1, the new log-odds difference becomes:

$$\hat{\eta}(x_i + d) - \hat{\eta}(x_i) = d\hat{\beta}_1$$

Taking the exponential on both sides:

$$\frac{\text{odds}_{x_i+d}}{\text{odds}_{x_i}} = e^{d\hat{\beta}_1}$$

Thus, the odds ratio associated with a d-unit increase in the predictor variable is:

$$OR = e^{d\hat{\beta}_1}$$

Interpretation of the Odds Ratio

- If $d = 1$, then:

  $$OR = e^{\hat{\beta}_1}$$

  meaning a one-unit increase in x multiplies the odds by $e^{\hat{\beta}_1}$.

- If $d = 2$, then:

  $$OR = e^{2\hat{\beta}_1}$$

  meaning a two-unit increase in x multiplies the odds by $e^{2\hat{\beta}_1}$.

- If $d = -1$, then:

  $$OR = e^{-\hat{\beta}_1}$$

  meaning a one-unit decrease in x divides the odds by $e^{\hat{\beta}_1}$.

Poisson Regression: Regression Modeling for Count Data

We now consider another regression modeling scenario where the response variable of interest is not normally distributed.

In this situation, the response variable represents a **count** of some relatively rare event, such as:

- Defects in a unit of manufactured product
- Errors or “bugs” in software
- Count of particulate matter or other pollutants in the environment

The analyst aims to model the relationship between observed counts and predictor variables.

Example

For example, an engineer may be interested in modeling the relationship between:

- The observed number of **defects** in a unit of product
- The **production conditions** when the unit was manufactured

In such cases, **Poisson regression** is a natural choice for modeling count data.

Poisson Distribution for Count Data

We assume that the response variable y_i is a **count**, such that:

$$y_i = 0, 1, 2, \ldots$$

A reasonable probability model for count data is the **Poisson distribution**:

$$f(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

where the parameter $\mu > 0$ represents the expected count.

Properties of the Poisson Distribution

The Poisson distribution has a key property:

$$E(y) = \mu \quad \text{and} \quad \mathrm{Var}(y) = \mu$$

That is, both the **mean** and **variance** of the Poisson distribution are equal to the parameter μ.

This relationship makes Poisson regression a suitable model for count data.

Poisson Regression Model

The Poisson regression model can be written as:

$$y_i = E(y_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where the expected value of the observed response is:

$$E(y_i) = \mu_i$$

The relationship between the mean response and the linear predictor is given by a function g, known as the **link function:**

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = x_i' \beta$$

Inverse Link Function

The relationship between the mean and the linear predictor is:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i'\beta)$$

Several link functions are commonly used in Poisson regression, including:

- Identity Link: $g(\mu_i) = \mu_i = x_i'\beta$
- Log Link: $g(\mu_i) = \ln(\mu_i) = x_i'\beta$

Log Link Function

The log link function is widely used in Poisson regression:

$$g(\mu_i) = \ln(\mu_i) = x_i'\beta$$

The relationship between the mean response and the linear predictor is:

$$\mu_i = g^{-1}(x_i'\beta) = e^{x_i'\beta}$$

The log link ensures that all predicted values μ_i are nonnegative.

Maximum Likelihood Estimation

The parameters in Poisson regression are estimated using the method of maximum likelihood. Given a random sample of n observations on the response y and the predictors x, the likelihood function is:

$$L(y, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

Simplifying:

$$L(y, \beta) = \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp\left(-\sum_{i=1}^{n} \mu_i\right)}{\prod_{i=1}^{n} y_i!}$$

Once the link function is selected, we maximize the log-likelihood function to estimate β.

Log-Likelihood Function

The log-likelihood function for Poisson regression is:

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!)$$

Maximum Likelihood Estimation

- The **iteratively reweighted least squares (IRLS)** method is used to find the **maximum likelihood estimates** of the parameters in Poisson regression.
- Once the parameter estimates $\hat{\beta}$ are obtained, the fitted Poisson regression model is:

$$\hat{y}_i = g^{-1}(x_i'\hat{\beta})$$

Prediction Equations

The form of the prediction equation depends on the chosen link function:

- Identity Link: $g^{-1}(x_i'\hat{\beta}) = x_i'\hat{\beta}$

$$\hat{y}_i = x_i'\hat{\beta}$$

- Log Link: $g^{-1}(x_i'\hat{\beta}) = \exp(x_i'\hat{\beta})$

$$\hat{y}_i = \exp(x_i'\hat{\beta})$$

Generalized Linear Model (GLM)

- GLM is a unifying approach to regression and experimental design models.
- It includes linear regression, logistic regression, and Poisson regression as special cases.
- The response variable follows a distribution from the **exponential family**.

Exponential Family of Distributions

A key assumption in GLM is that the response variable y follows a distribution from the **exponential family**, which has the general form:

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i, \phi)\right\}$$

where:

- θ_i is the natural location parameter.
- ϕ is a scale parameter.

Mean and Variance in GLM

For members of the exponential family,

- The expected value of the response is:

$$\mu = E(y) = \frac{db(\theta_i)}{d\theta_i}$$

- The variance of y is:

$$\text{Var}(y) = \frac{d^2 b(\theta_i)}{d\theta_i^2} a(\phi) = \frac{du}{d\theta_i} a(\phi)$$

Generalized Linear Model (GLM)

Let

$$\mathrm{Var}(\mu) = \frac{\mathrm{Var}(y)}{a(\phi)} = \frac{d\mu}{d\theta_i} \tag{25}$$

where $Var(\mu)$ denotes the dependence of the variance of the response on its mean. This is a characteristic of all distributions that are a member of the exponential family, except for the normal distribution

From this, we obtain:

$$\frac{d\theta_i}{d\mu} = \frac{1}{\mathrm{Var}(\mu)} \tag{26}$$

This highlights the dependence of the variance on the mean for distributions in the exponential family, except for the normal distribution.

Fundamental Equation in Generalized Linear Models (GLMs)

The equation:

$$\frac{d\theta_i}{d\mu} = \frac{1}{\text{Var}(\mu)} \tag{27}$$

is a fundamental result in Generalized Linear Models (GLMs). It describes how the **natural parameter** θ_i changes with respect to the mean μ.

Significance of the Equation

- Ensures the selection of **appropriate link functions** based on how the variance of the response variable depends on its mean.
- Highlights the difference between **normal regression** and other GLMs:
  - Only the **normal distribution** has a **constant variance** that does not depend on μ.
  - Other distributions in the exponential family (e.g., Poisson, Binomial) have variance that depends on μ.
- Plays a crucial role in defining the **variance function** in GLM formulations.

Link Functions and Linear Predictors

The fundamental idea of a Generalized Linear Model (GLM) is to develop a linear model for a function of the expected value of the response variable. The **linear predictor** is defined as:

$$\eta_i = g[E(y_i)] = g(\mu_i) = x_i'\beta \tag{28}$$

where:

- η_i is the linear predictor.
- g is the **link function**.
- $E(y_i) = \mu_i$ is the expected response.

Inverse Link Function

Since the expected response is related to the linear predictor through the inverse of the link function:

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(x_i'\beta) \tag{29}$$

We call the function g the link function. Recall that we introduced the concept of a link function in our description of Poisson regression. Different choices of g define different types of GLMs.

Canonical Links in GLMs

If we choose

$$\eta_i = \theta_i \tag{30}$$

then η_i is called the **canonical link**. The table below summarizes the canonical links for common distributions:

Distribution	**Canonical Link**
Normal	$\eta_i = \mu_i$ (identity link)
Binomial	$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ (logistic link)
Poisson	$\eta_i = \ln(\lambda_i)$ (log link)
Exponential	$\eta_i = \frac{1}{\lambda_i}$ (reciprocal link)
Gamma	$\eta_i = \frac{1}{\lambda_i}$ (reciprocal link)

Other Common Link Functions

Apart from the canonical links, there are other choices of link functions:

- **Probit Link**: $\eta_i = \Phi^{-1}(E(y_i))$, where Φ is the cumulative standard normal distribution function.
- **Complementary Log-Log Link**:

$$\eta_i = \ln\{\ln[1 - E(y_i)]\} \tag{31}$$

- **Power Family Link**:

$$\eta_i = \begin{cases} E(y_i)^\lambda, & \lambda \neq 0 \\ \ln[E(y_i)], & \lambda = 0 \end{cases} \tag{32}$$

Link Function Selection and Model Fit

- The choice of the link function is similar to choosing a transformation for a linear model.
- Unlike transformations, link functions leverage the natural distribution of the response.
- Poor link function choice can lead to:
  - Misfit of the model.
  - Incorrect parameter estimates.
  - Inefficient predictions.

Parameter Estimation in GLMs

Maximum Likelihood Estimation (MLE) is the theoretical basis for parameter estimation in GLMs. The estimation procedure relies on:

- **Iteratively Reweighted Least Squares (IRLS)**, commonly used in software implementations.
- If $\hat{\beta}$ is the final estimate of regression coefficients, then:

$$E(\hat{b}) = \beta, \quad \text{Var}(\hat{\beta}) = a(\phi)(X'VX)^{-1} \tag{33}$$

where V is a diagonal matrix of variances.

Inference in GLMs

Key points:

- Typically, when experimenters and data analysts use a transformation, they use OLS to actually fit the model in the transformed scale.
- Ordinary Least Squares (OLS) assumes constant variance, but GLMs use **weighted least squares**.
- GLMs provide better estimation when variance is non-constant.
- Similar inference techniques as logistic regression:
  - Model deviance is used for overall model fit.
  - Difference in deviance tests subsets of model parameters.
  - Wald tests construct confidence intervals for individual parameters.

References

This presentation is adapted from:

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis, Fifth Edition. Wiley.
- MTH 416 : Regression Analysis — Shalabh, IIT Kanpur

Thank You!

Thank you for your attention!