Generalized Linear Models

Statistics · Module 8 — Multivariate Linear Models

Node 8.6

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.6 — Generalized Linear Models.

GLMs extend the normal linear model to responses from any exponential family distribution, replacing the identity link with a general link function. This node depends on the linear model geometry (Node 8.3), exponential family theory (Node 2.6), and M-estimation (Node 7.5). It is the natural generalization of the Gaussian linear model to non-Gaussian responses.

Formal Definition

A generalized linear model specifies three components:

1. **Random component.** $Y_i \mid x_i$ independently follows an exponential family distribution with density

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

where η_i is the canonical parameter, $b(\cdot)$ is the cumulant function, ϕ is the dispersion parameter, and $a(\phi) = \phi / w_i$ for known weights w_i.

2. **Systematic component.** A linear predictor $\eta_i^* = x_i^\top \beta$.

3. **Link function.** A monotone differentiable function g such that $g(\mu_i) = x_i^\top \beta$, where $\mu_i = \mathbb{E}[Y_i \mid x_i] = b'(\eta_i)$.

The **canonical link** sets $g = (b')^{-1}$, so that $\eta_i = x_i^\top \beta$ directly.

Intuition

The normal linear model ties the mean directly to the linear predictor: $\mu_i = x_i^\top \beta$. For binary or count data, this is inappropriate — the mean must be constrained (to $[0, 1]$ or $[0, \infty)$). The link function bridges the gap: it maps the constrained mean space to the unconstrained linear predictor space. The canonical link is special because it makes the sufficient statistics linear in the linear predictor, simplifying the score equations and guaranteeing concavity of the log-likelihood.

Structural Role

GLMs unify regression models for continuous, binary, and count responses under a single framework. The exponential family structure (Node 2.6) determines the variance function $V(\mu) = b''(\eta)$, the canonical link, and the form of the log-likelihood. The IRLS algorithm reduces all GLM fitting to a sequence of weighted least squares problems, specializing to OLS in the Gaussian case. Asymptotic inference follows from M-estimation theory (Node 7.5).

Mathematical Development

Mean-variance relationship. From exponential family theory, $\mu = b'(\eta)$ and $\text{Var}(Y) = a(\phi)b''(\eta) = a(\phi)V(\mu)$, where $V(\mu) = b''((b')^{-1}(\mu))$ is the variance function. The variance function determines the distribution:

Distribution	$b(\eta)$	$V(\mu)$	Canonical link
Normal	$\eta^2/2$	1	Identity: $g(\mu) = \mu$
Bernoulli	$\log(1 + e^\eta)$	$\mu(1-\mu)$	Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$
Poisson	e^η	μ	Log: $g(\mu) = \log \mu$
Gamma	$-\log(-\eta)$	μ^2	Reciprocal: $g(\mu) = 1/\mu$

Log-likelihood and score equations. For the canonical link ($\eta_i = x_i^\top \beta$):

$$\ell(\beta) = \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i x_i^\top \beta - b(x_i^\top \beta) \right] + \text{const.}$$

The score is

$$\nabla_\beta \ell(\beta) = \frac{1}{\phi} \sum_{i=1}^{n} w_i (y_i - \mu_i) x_i = \frac{1}{\phi} X^\top W^{1/2} (y - \mu),$$

where $\mu_i = b'(x_i^\top \beta)$. Setting the score to zero gives the score equations $X^\top (y - \mu(\beta)) = 0$ (up to weights and ϕ), which generalize the normal equations of OLS.

For a general link, the score is

$$\nabla_\beta \ell = \sum_{i=1}^{n} \frac{w_i (y_i - \mu_i)}{a(\phi) V(\mu_i) g'(\mu_i)} x_i.$$

Fisher information. The expected Fisher information is

$$\mathcal{I}(\beta) = \frac{1}{\phi} X^\top W X,$$

where $W = \text{diag}(w_1/(V(\mu_1)[g'(\mu_1)]^2), \ldots, w_n/(V(\mu_n)[g'(\mu_n)]^2))$. For the canonical link, $W = \text{diag}(w_i V(\mu_i))$.

Fisher scoring / IRLS. The MLE is computed iteratively. Fisher scoring updates:

$$\beta^{(t+1)} = \beta^{(t)} + [\mathcal{I}(\beta^{(t)})]^{-1} \nabla_\beta \ell(\beta^{(t)}).$$

This is equivalent to Iteratively Reweighted Least Squares: at each step, solve

$$\hat{\beta}^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)},$$

where the working response is $z_i^{(t)} = \hat{\eta}_i^{(t)} + (y_i - \hat{\mu}_i^{(t)}) g'(\hat{\mu}_i^{(t)})$ and $W^{(t)}$ is the weight matrix evaluated at $\hat{\mu}^{(t)}$.

Convergence. For the canonical link, the log-likelihood is concave (since b is convex), so IRLS converges to the unique global maximum. For non-canonical links, local convergence is guaranteed under smoothness conditions.

Logistic regression (Bernoulli GLM). For binary $Y_i \in \{0, 1\}$ with $\mathbb{P}(Y_i = 1 | x_i) = \mu_i$, the canonical link is the logit: $\log(\mu_i/(1-\mu_i)) = x_i^\top \beta$, so $\mu_i = 1/(1+e^{-x_i^\top \beta})$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - \log(1+e^{x_i^\top \beta})]$. The variance function is $V(\mu) = \mu(1-\mu)$, and the weights are $w_i = \mu_i(1-\mu_i)$.

Poisson regression. For count $Y_i \in \{0, 1, 2, \ldots\}$ with $\mu_i = \mathbb{E}[Y_i | x_i]$, the canonical link is the log: $\log \mu_i = x_i^\top \beta$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - e^{x_i^\top \beta}] + \text{const}$. The variance equals the mean: $V(\mu) = \mu$, and the weights are $w_i = \mu_i$.

Quasi-likelihood. When only the mean-variance relationship $\mathrm{Var}(Y_i) = \phi V(\mu_i)$ is specified (without a full distributional assumption), the quasi-likelihood score equations are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i}{V(\mu_i) g'(\mu_i)} = 0.$$

These coincide with the GLM score equations, so the same IRLS algorithm applies. The quasi-likelihood estimator is consistent and asymptotically normal with the sandwich covariance, even without specifying the full distribution — only the first two conditional moments matter.

Negative binomial regression. For overdispersed count data where $\mathrm{Var}(Y) > \mu$, the negative binomial GLM uses $V(\mu) = \mu + \mu^2/\kappa$ where κ is the overdispersion parameter. This nests Poisson regression ($\kappa \to \infty$) and accommodates extra-Poisson variation.

Probit model. An alternative to logistic regression for binary data uses the link $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the standard normal CDF. The probit model arises naturally from a latent variable formulation: $Y_i = \mathbf{1}(x_i^\top \beta + \varepsilon_i > 0)$ with $\varepsilon_i \sim \mathcal{N}(0, 1)$. Logit and probit give nearly identical fitted probabilities in practice (after rescaling by $\pi/\sqrt{3} \approx 1.81$).

Deviance. The deviance is twice the log-likelihood ratio between the saturated model (one parameter per observation) and the fitted model:

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i(\tilde{\eta}_i - \hat{\eta}_i) - b(\tilde{\eta}_i) + b(\hat{\eta}_i) \right],$$

where $\tilde{\eta}_i$ is the canonical parameter of the saturated model ($\tilde{\mu}_i = y_i$). The deviance plays the role of RSS in the normal linear model. Under regularity conditions, the difference in deviances between nested models follows χ^2_q asymptotically, enabling likelihood ratio tests for model comparison.

Asymptotic inference. By the M-estimation theory of Node 7.5, the MLE satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0,\ \mathcal{I}(\beta)^{-1}).$$

Wald tests use $(\hat{\beta}_j / \mathrm{SE}(\hat{\beta}_j))^2$; Score tests use $\nabla \ell(\beta_0)^\top \mathcal{I}(\beta_0)^{-1} \nabla \ell(\beta_0)$; LR tests use the deviance difference. All three converge to χ^2 under the null.

Worked Examples

Example 1 (Logistic regression). Data: 100 patients, binary outcome (disease/no disease), predictor $x =$ age. Fitted model: $\text{logit}(\hat{\mu}) = -5.0 + 0.08 \cdot \text{age}$. At age 50: $\hat{\mu} = 1/(1 + e^{-(-5+4)}) = 1/(1 + e) \approx 0.269$. The odds ratio per year of age is $e^{0.08} \approx 1.083$: each year increases the odds of disease by 8.3%. The Wald 95% CI for the log-odds-ratio is $0.08 \pm 1.96 \times \text{SE}(\hat{\beta}_1)$.

Example 2 (Poisson regression). Data: counts of insurance claims by age group. Model: $\log \hat{\mu}_i = 1.2 + 0.03 \cdot \text{age}_i$. The rate ratio per year is $e^{0.03} \approx 1.030$. Deviance goodness-of-fit: if the residual deviance is 45.2 on 48 df, the model fits adequately (p-value from χ^2_{48} is ≈ 0.59). If overdispersion is detected ($D/(n-p) \gg 1$), a quasi-Poisson model with estimated $\phi > 1$ inflates standard errors.

Cross-Links

AI. Logistic regression is the canonical binary classifier and the basis for the output layer of neural networks for classification. The cross-entropy loss in deep learning is the negative Bernoulli GLM log-likelihood. Softmax regression extends logistic regression to multinomial responses.

Economics. Probit and logit models are the primary binary choice models in microeconomics. Poisson regression is used for count data in labor economics (number of patents, doctor visits). The quasi-likelihood framework extends GLMs when the full distributional assumption is relaxed.

Linear Algebra. IRLS reduces each GLM iteration to a weighted least squares problem: $(X^\top W X)^{-1} X^\top W z$. The convergence of IRLS is analyzed via the contraction mapping theorem on the Fisher scoring iteration.

Quantum Mechanics. In quantum detector tomography, the detection probabilities follow a multinomial model that can be expressed as a GLM with a log-link on the POVM elements. The Fisher scoring algorithm applies to reconstruct the detector from count data.

Structural Compression

Invariant. The score equations $X^\top(y - \mu(\beta)) = 0$ generalize the normal equations of OLS to any exponential family response; the canonical link aligns the natural parameter with the linear predictor.

Mechanism. The link function maps the constrained mean space to the unconstrained linear predictor; the exponential family structure provides the variance function, score, and Fisher information. IRLS iteratively constructs a locally quadratic approximation via working responses and weights, converging to the MLE.

Cross-System Echo. Logistic regression is maximum entropy classification (AI connection). The deviance is the likelihood ratio statistic comparing fitted and saturated models (Node 5.4). The quasi-likelihood approach relaxes the full distributional assumption to the first two conditional moments, paralleling the Gauss-Markov approach in the linear model.

Exercises

1. Derive the canonical link for the Bernoulli distribution by writing the PMF in exponential family form and identifying $\eta = \log(\mu/(1-\mu))$.
2. Show that the score equations for logistic regression reduce to $X^\top(y - \mu(\beta)) = 0$ and verify that the log-likelihood is concave.
3. For Poisson regression with canonical link, derive the Fisher information matrix and the IRLS weight matrix.

4. Prove that the deviance for the normal linear model equals RSS/σ^2, recovering the residual sum of squares as a special case.

5. Fit a logistic regression to the data $(x, y) = \{(1,0),(2,0),(3,1),(4,0),(5,1),(6,1),(7,1)\}$ by writing the log-likelihood and performing one Newton-Raphson iteration starting from $\beta = (0,0)$.

6. Explain the concept of overdispersion in the Poisson GLM, derive the quasi-Poisson variance $\mathrm{Var}(Y_i) = \phi\mu_i$, and describe how standard errors are adjusted.

7. Argue, structurally, why the canonical link occupies a privileged position among all possible link functions, connecting its role to sufficiency, concavity of the log-likelihood, and the information matrix equality.

Statistics · Module 8 — Multivariate Linear Models · Node 8.6
Concepts: exponential-families, maximum-likelihood-estimation, convergence
Prerequisites: 8.3, 2.6, 7.5

hbar.university — own.your.cognition