

Generalised Linear Models

Aayush Bajaj

August 31, 2026

a generalised linear model is what remains of linear regression once you stop pretending every response is gaussian. nelder and wedderburn's 1972 observation was that the gaussian, binomial, poisson, exponential and gamma models all run on the *same* three-part engine — a linear predictor, a link, and an exponential-family response — so estimation (fisher scoring), inference (deviance) and diagnostics (standardised residuals) can be built *once* and inherited by every member.

*this page distils my math5806
(applied regression analysis)
notes, which follow dobson and
barnett's textbook
parametrisation throughout*

this page is the general theory. logistic regression instantiates it for bernoulli/binomial responses; the machinery for maximum likelihood itself lives on the estimation and inference page.

Contents

1 the three-part anatomy	2
2 the exponential family	2
2.1 canonical form	2
2.2 the members	3
3 link functions	3
4 fitting: fisher scoring, alias irls	3
4.1 the score and the information	3
4.2 the update is weighted least squares	4
4.3 why lm is a one-lap special case	4
5 the deviance	5
5.1 definition and use	5
6 poisson regression, worked	5
6.1 the count glm	5
6.2 by-hand iterative least squares	5
6.3 link choice as a modelling decision	6
7 log-linear models for contingency tables	6
8 beyond one mean: the rest of the family	7
9 quasi-likelihood and overdispersion	7
10 diagnostics	7
11 results	8
12 see also	8

1 the three-part anatomy

every glm consists of:

1. a **random component**: independent responses $Y_1, \dots, Y_N$ drawn from the same exponential-family distribution, whose means μ_i may differ;
2. a **systematic component**: a linear predictor $\eta_i = \mathbf{x}_i^\top \beta$ built from covariates — linear *in the parameters*, so polynomial and dummy-coded terms are welcome;
3. a **link function** g , monotone and differentiable, joining the two: $g(\mu_i) = \eta_i$.

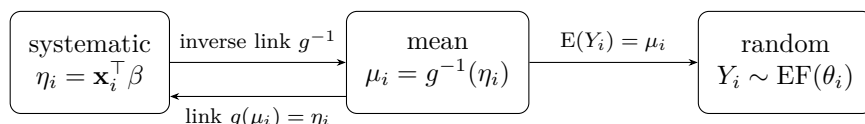


Figure 1: the glm engine: only the random component and the link change between members; the linear predictor and the fitting machinery are shared.

the course's opening taxonomy is worth pinning down: the *scale of the response* chooses the model, and most of classical statistics turns out to be a glm wearing a costume.

response	covariates	method
continuous	continuous	multiple regression
continuous	categorical	anova
continuous	categorical & continuous	ancova
binary	categorical & continuous	logistic regression
nominal, more than 2 categories	categorical & continuous	nominal logistic regression
ordinal	categorical & continuous	ordinal logistic regression
counts	categorical & continuous	poisson regression
counts (cross-classified)	categorical	log-linear models
survival times	categorical & continuous	survival analysis (exponential)

anova and ancova are linear gaussian models with dummy-coded factors; contingency-table analysis is poisson log-linear regression

2 the exponential family

2.1 canonical form

a distribution belongs to the exponential family if its density or mass function factors as

$$f(y | \theta) = \exp[a(y) b(\theta) + c(\theta) + d(y)], \quad (1)$$

this is dobson and barnett's parametrisation; other books write the same family as $\exp[(y \theta - b(\theta))/a(\theta) + c(y, \theta)]$ — the identities below translate

and the form is **canonical** when $a(y) = y$. the function $b(\theta)$ is then called the **natural parameter**. everything the fitting algorithm needs is carried by b and c :

$$\mathbb{E}(Y) = -\frac{c'(\theta)}{b'(\theta)}, \quad \text{Var}(Y) = \frac{b''(\theta) c'(\theta) - c''(\theta) b'(\theta)}{[b'(\theta)]^3}, \quad (2)$$

both obtained from the same trick: differentiate $\int f dy = 1$ under the integral sign, once for the mean and twice for the variance (proofs in the [results](#) below).

2.2 the members

running the factorisation on each classical distribution:

distribution	natural parameter $b(\theta)$	canonical link	Var(Y) as a function of μ
normal $N(\mu, \sigma^2)$	μ/σ^2	identity	constant σ^2
bernoulli / binomial $B(n, \pi)$	$\log[\pi/(1 - \pi)]$	logit	$n\pi(1 - \pi)$
poisson $P(\lambda)$	$\log \lambda$	log	λ
exponential / gamma	$-1/\mu$ (up to scale)	negative inverse	μ^2/ν

the variance column is the family’s fingerprint: a glm believes not just in a mean structure but in a specific **mean–variance relationship**. counts spread out as their mean grows; proportions are most variable at one half; gaussian noise alone ignores its mean. least squares gets away with a single constant weight precisely because of that last row — and only there.

3 link functions

the link’s job is geometric: carry the mean’s natural range onto the whole real line, where a linear predictor can roam freely.

- **identity** $g(\mu) = \mu$: gaussian responses, range already $\mathbb{R}$;
- **log** $g(\mu) = \log \mu$: positive means (counts, rates) — effects become *multiplicative*;
- **logit** $g(\pi) = \log[\pi/(1 - \pi)]$: probabilities in $(0, 1)$ — effects become odds ratios.

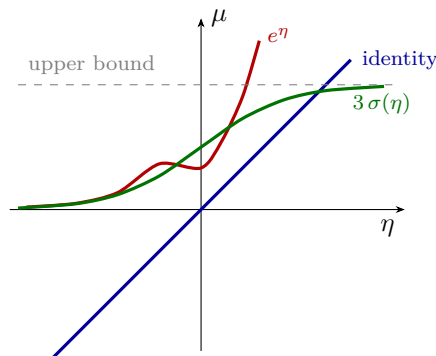


Figure 2: three inverse links carrying the linear predictor back to the mean scale: identity (gaussian), exponential (poisson), sigmoid (bernoulli). each respects its response’s natural range.

the **canonical link** is the choice $g = b$: set the linear predictor equal to the natural parameter itself. it is not compulsory — poisson regression with an identity link is perfectly legal, and fitted below — but it buys real simplifications: sufficient statistics $\mathbf{X}^\top \mathbf{y}$, score equations of the clean form $\mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$, and a hessian that equals its own expectation, making newton–raphson and fisher scoring the same algorithm.

4 fitting: fisher scoring, alias irls

4.1 the score and the information

the log-likelihood of a glm is $\ell(\beta) = \sum_i [y_i b(\theta_i) + c(\theta_i) + d(y_i)]$, and the chain rule $\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$ collapses it to a formula that mentions the distribution only through its mean and variance:

mismatched links can also break: a log link on a bernoulli response maps probabilities to the non-positive half-line while the linear predictor roams all of $\mathbb{R}$ — the mle routinely fails to exist

$$U_j = \sum_{i=1}^N \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right), \quad \mathcal{I}_{jk} = \sum_{i=1}^N \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2. \quad (3)$$

in matrix form $\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$ with diagonal weights $w_{ii} = [d\mu_i/d\eta_i]^2 / \text{Var}(Y_i)$. the score equations $\mathbf{U} = \mathbf{0}$ are nonlinear in β for every member except the gaussian, so there is no closed form and we iterate.

4.2 the update is weighted least squares

the **method of scoring** replaces the hessian in newton–raphson with its expectation, the information:

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \mathbf{U}^{(m-1)}. \quad (4)$$

multiply through by $\mathcal{I}^{(m-1)}$ and the right-hand side rearranges into $\mathbf{X}^\top \mathbf{W} \mathbf{z}$, where

$$z_i = \eta_i + (y_i - \mu_i) \left(\frac{d\eta_i}{d\mu_i} \right) \quad (5)$$

is the **working response** — a first-order linearisation of $g(y_i)$ about the current mean. each iteration is therefore an ordinary weighted least-squares solve,

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{X} \hat{\beta}^{(m)} = \mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}, \quad (6)$$

with $\mathbf{W}$ and $\mathbf{z}$ refreshed from the current fit: **iteratively reweighted least squares**. this is what R’s `glm()` runs — hence the **Number of Fisher Scoring iterations** line in every summary.

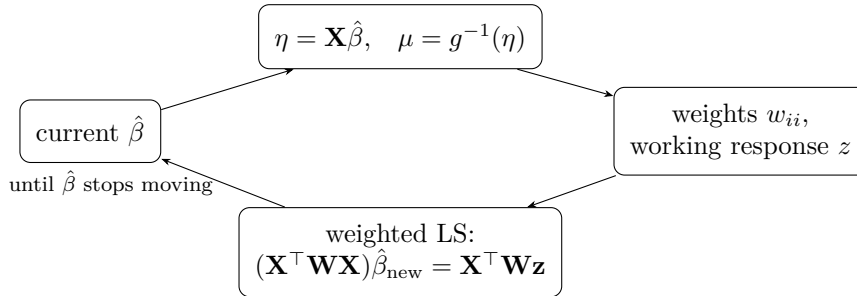


Figure 3: one lap of irls. convergence is checked on the coefficients; at the fixed point the inverse information is the asymptotic covariance of the estimates, free of charge.

4.3 why lm is a one-lap special case

take the gaussian family with the identity link: $\mu = \eta$ makes $d\mu/d\eta = 1$, so the working response is $z = \eta + (y - \mu) = y$, and constant variance makes $\mathbf{W} \propto \mathbf{I}$. the update becomes $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ — the ordinary least squares formula — and since neither $\mathbf{z}$ nor $\mathbf{W}$ depends on β any more, irls converges in a single iteration. that collapse is also the honest accounting of what generalisation costs: the linear gaussian model has exact small-sample t and F inference, while everything else on this page is asymptotic — wald z -values, deviances against χ^2 — which is the

this is why `glm(..., family = gaussian)` reproduces `lm` exactly; R still reports two scoring iterations because it takes one more lap to notice nothing moved

deeper reason a `glm()` summary shows deviance and aic where `lm()` showed R^2 and an F -statistic (see estimation and inference).

irls can also fail where least squares cannot: perfectly separated classes in logistic regression drive $\|\hat{\beta}\| \rightarrow \infty$ and the mle does not exist. the only failure mode ols has is exact collinearity.

5 the deviance

5.1 definition and use

fix the data and ask: how much log-likelihood do we forfeit by describing N observations with p parameters instead of N ? the **saturated model** fits one parameter per observation (or per covariate pattern) and attains the maximum achievable likelihood $\ell(\hat{\theta}_{\text{sat}})$. the **deviance** of a fitted model is twice the gap:

$$D = 2 \left[\ell(\hat{\theta}_{\text{sat}}; \mathbf{y}) - \ell(\hat{\theta}; \mathbf{y}) \right]. \quad (7)$$

under the hypothesis that the model is correct, $D \sim \chi^2(N - p)$ approximately. its close cousin is the pearson statistic $P^2 = \sum_i (y_i - \hat{\mu}_i)^2 / \text{Var}(\hat{Y}_i)$, asymptotically equivalent to D ; the course's practical guidance is that P^2 is often the better-behaved of the two because D is unduly influenced by very small frequencies — and both approximations fail when expected counts are small, most flagrantly for ungrouped binary data.

for gaussian responses the deviance is just the residual sum of squares scaled by the variance — the saturated model interpolates the data exactly

two model-comparison workhorses fall out:

- **nested models:** for $M_0 \subset M_1$, the offending saturated terms cancel in $\Delta D = D_0 - D_1 \sim \chi^2(p_1 - p_0)$ — trustworthy even where the absolute deviances are not. this is the glm generalisation of the F -test.
- **non-nested models:** penalised scores $\text{AIC} = -2\ell(\hat{\theta}) + 2p$ and $\text{BIC} = -2\ell(\hat{\theta}) + p \log N$ trade fit against complexity.

6 poisson regression, worked

6.1 the count glm

for counts $Y_i \sim \text{Poisson}(\lambda_i)$, the canonical form is $f(y | \lambda) = \exp[y \log \lambda - \lambda - \log y!]$: natural parameter $\log \lambda$, so the canonical link is the log and the model is $\log \lambda_i = \mathbf{x}_i^\top \beta$. exponentiating a coefficient gives a **rate ratio** — the multiplicative change in expected count per unit of covariate — the poisson sibling of the logistic odds ratio.

6.2 by-hand iterative least squares

the course fits dobson's toy count data ($N = 9$, a single covariate with values $-1, 0, 1$) with an *identity* link, both by hand and via `glm()`. the by-hand version is five laps of the loop above, with weights $w_{ii} = 1/\lambda_i = 1/(\beta_1 + \beta_2 x_i)$:

```
1 library(dobson)
2 data("poisson")
3 x <- poisson$x; y <- poisson$y
4
5 ILS <- function(niter, b, x, y) {
6   N <- length(y)
```

```

7  z <- as.matrix(y)                                # identity link: working response is
  ↪ y itself
8  X <- matrix(c(rep(1, N), x), N, 2)
9  for (i in 1:niter) {
10     w <- 1 / (b[1] + b[2] * x)                    # 1 / Var(Y_i), and dmu/deta = 1
11     W <- diag(w, N, N)
12     b <- solve(t(X) %*% W %*% X) %*% t(X) %*% W %*% z
13  }
14  inv.inform <- solve(t(X) %*% W %*% X)
15  list(b = b, sd = sqrt(diag(inv.inform)))
16 }
17 ILS(5, b = c(7, 5), x = x, y = y)

```

```

iteration:   b1      b2
  1         7.451389  4.937500
  2         7.451632  4.935314
  3         7.451633  4.935300
  4         7.451633  4.935300      <- converged
standard errors: 0.884  1.089

```

from the crude start (7, 5) the estimates settle to $\hat{\beta} = (7.4516, 4.9353)$ in three effective iterations, and the identity-link `glm()` fit reproduces both the coefficients and the standard errors (0.884, 1.089) exactly — the standard errors being square roots of the diagonal of $\mathcal{I}^{-1} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$. wald 95% intervals are then $\hat{\beta}_j \pm 1.96 \text{ se}$, giving (5.72, 9.18) and (2.80, 7.07).

6.3 link choice as a modelling decision

refitting with the canonical log link gives $\hat{\beta} = (1.889, 0.670)$ and a rate ratio $e^{0.670} = 1.95$: each unit of x roughly doubles the expected count. goodness of fit for the log-link model: pearson $P^2 = 2.90$ and deviance $D = 2.94$ on 7 degrees of freedom, far below $\chi^2_{7,0.95} = 14.07$ — no evidence of misfit; and the deviance difference against the intercept-only model, $\Delta D = 15.48 \gg \chi^2_{1,0.95} = 3.84$, says the covariate earns its keep. note the identity-link fit had the *smaller* residual deviance here (1.89 vs 2.94): the canonical link is a default, not a decree, and the deviance is precisely the instrument that lets the data adjudicate.

7 log-linear models for contingency tables

cross-classified counts are poisson regression on factor covariates. for the course's melanoma table (3 tumour sites $\times$ 4 histological types, $N = 12$ cells), model the expected cell count as $\log \mu_{jk} = \text{const} + \text{site}_j + \text{type}_k$ (+ interaction $_{jk}$) and the classical test of independence becomes an analysis of deviance:

model	deviance	df	reading
minimal (intercept only)	295.20	11	all cells share one mean — hopeless
additive (site + type)	51.80	6	the <i>independence</i> hypothesis
saturated (site * type)	0.00	0	one parameter per cell — perfect by fiat

the additive model's deviance 51.80 on 6 degrees of freedom sits far beyond $\chi^2_{6,0.95} = 12.59$: site and tumour type are associated, and the interaction terms are needed. the saturated fit's deviance of zero is the definition of the yardstick, not an achievement.

with a poisson likelihood the cell totals are random; conditioning on the grand total recovers the multinomial, which is why log-linear deviance tests agree with the classical chi-squared test of independence

8 beyond one mean: the rest of the family

- **binomial and multinomial:** the bernoulli/binomial member is developed in full on the logistic regression page, including nominal (baseline-category) and ordinal (proportional-odds) extensions for responses with more than two categories.
- **gamma and exponential:** positive continuous responses with variance growing as μ^2 — waiting times, claim sizes, survival without censoring. usually fitted with a log link for interpretability rather than the canonical negative inverse.
- **linear discriminant analysis**, the course’s week-5 companion to logistic regression, is *not* a glm — it models $f(\mathbf{x} \mid Y = k)$ generatively and inverts with bayes’ rule — but it answers the same classification question, remains stable under perfectly separated classes (where the logistic mle diverges), and handles many classes gracefully.

9 quasi-likelihood and overdispersion

real counts are routinely noisier than the poisson’s $\text{Var}(Y) = \mu$ permits — omitted covariates, clustering, contagion. the symptom is a residual deviance far exceeding its degrees of freedom. wedderburn’s **quasi-likelihood** observation: fisher scoring only ever touched the distribution through μ_i and $\text{Var}(Y_i)$, so specify *just those two*

$$\text{Var}(Y_i) = \phi V(\mu_i), \quad (8)$$

a variance function V plus a free dispersion ϕ — and run the same algorithm. the point estimates are unchanged; standard errors inflate by $\sqrt{\hat{\phi}}$, with $\hat{\phi} = P^2/(N-p)$. in R this is `family = quasipoisson` or `quasibinomial`. the price: no true likelihood, so aic is undefined and tests use F -ratios of quasi-deviances. for ungrouped binary data overdispersion is not even identifiable — one more reason to fit grouped covariate patterns when the design allows.

10 diagnostics

the residual toolkit generalises componentwise, standardising each observation by its own model variance:

- **pearson residuals** $P_i = (y_i - \hat{\mu}_i) / \sqrt{\text{Var}(\hat{Y}_i)}$, whose squares sum to P^2 ;
- **deviance residuals** $d_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{D_i}$, whose squares sum to D ;
- both are standardised by leverage, dividing by $\sqrt{1 - h_{ii}}$ with h_{ii} taken from the hat matrix of the final irls solve.

plot them against fitted values, against each covariate, and in measurement order; superimpose a smoother before trusting your eyes, since discrete responses stripe the plots into curves that mislead the naked eye. the fuller story — including why binary-response residual plots degenerate — is on the logistic page.

11 results

a curated crib of the statements this page leans on.

definition 11.1 (exponential family, canonical form). $f(y \mid \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$. the form is canonical when $a(y) = y$, and $b(\theta)$ is then the natural parameter.

theorem 11.2 (mean and variance identities). *for the canonical form,*

$$\mathbb{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}, \quad \text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}.$$

definition 11.3 (generalised linear model). independent $Y_1, \dots, Y_N$ from one exponential-family distribution, a linear predictor $\eta_i = \mathbf{x}_i^\top \beta$, and a monotone differentiable link with $g(\mu_i) = \eta_i$, where $\mu_i = \mathbb{E}(Y_i)$.

definition 11.4 (canonical link). the choice $g = b$, equating the linear predictor with the natural parameter. under it the score equations reduce to $\mathbf{X}^\top(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$ and fisher scoring coincides with newton–raphson.

theorem 11.5 (score and information of a glm).

$$U_j = \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \frac{d\mu_i}{d\eta_i}, \quad \mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X}, \quad w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2,$$

and $\mathcal{I} = \mathbb{E}[\mathbf{U}\mathbf{U}^\top] = -\mathbb{E}[\partial \mathbf{U} / \partial \beta]$.

theorem 11.6 (fisher scoring is irls). *the scoring update $\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \mathbf{U}^{(m-1)}$ is equivalent to the weighted least-squares solve $\mathbf{X}^\top \mathbf{W} \mathbf{X} \hat{\beta}^{(m)} = \mathbf{X}^\top \mathbf{W} \mathbf{z}$ with working response $z_i = \eta_i + (y_i - \mu_i) d\eta_i / d\mu_i$.*

corollary 11.7 (one-step collapse to ols). *for the gaussian family with identity link, $z = \mathbf{y}$ and $\mathbf{W} \propto \mathbf{I}$, so irls converges to $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ in one iteration.*

definition 11.8 (deviance). $D = 2[\ell(\hat{\boldsymbol{\theta}}_{\text{sat}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})]$, the log-likelihood forfeited relative to the saturated model. approximately $\chi^2(N - p)$ when the model is correct and expected counts are not small.

theorem 11.9 (analysis of deviance). *for nested models $M_0 \subset M_1$ with $p_0 < p_1$ parameters, $D_0 - D_1 \sim \chi^2(p_1 - p_0)$ approximately under M_0 — valid even when the individual deviances are poorly approximated.*

definition 11.10 (quasi-likelihood). specify only $\mathbb{E}(Y_i) = \mu_i(\beta)$ and $\text{Var}(Y_i) = \phi V(\mu_i)$; run fisher scoring unchanged. estimates coincide with the glm's when V matches a real family; standard errors scale by $\sqrt{\hat{\phi}}$, $\hat{\phi} = P^2 / (N - p)$.

theorem 11.11 (asymptotic distribution of the estimator). *under regularity conditions $\hat{\beta}$ is asymptotically $N(\beta, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1})$; wald intervals and z -tests follow.*

12 see also

- estimation and inference — score, information, and the wald/score/likelihood-ratio trio in general
- logistic regression — the binomial member, in full
- ordinary least squares — the gaussian member, and the one-lap special case
- regularised regression — penalised likelihood on top of any of this
- dobson & barnett, *an introduction to generalized linear models* — the math5806 text; my worked solutions live there