Unit 4: Generalized linear models: General theory

Eugene Katsevich

October 27, 2021

Units 1-3 focused on the most common class of models used in applications: linear models. Despite their versatility, linear models do not apply in all situations. In particular, they are not designed to deal with binary or count responses. In Unit 4, we introduced *generalized linear models* (GLMs), a generalization of linear models that encompasses a wide variety of incredibly useful models including logistic regression and Poisson regression.

We'll start Unit 4 by introducing exponential family models (Section 1), a generalization of the Gaussian distribution that serves as the backbone of GLMs. Then we formally define a GLM, demonstrating logistic regression and Poisson regression as special cases (Section 2). Next we discuss maximum likelihood inference in GLMs (Section 3). Finally, we discuss how to carry out statistical inference in GLMs (Section 4).

1 Exponential family distributions

Definition and examples. Let's start with the Gaussian distribution, taking variance $\sigma^2 = 1$ for simplicity. If $y \sim N(\mu, 1)$, then it has density

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) = \exp\left(\mu y - \frac{1}{2}\mu^2\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right). \tag{1}$$

Here is a way of generalizing this density:

$$f_\eta(y) = \exp(\eta y - \psi(\eta))h(y). \tag{2}$$

Here η is called the *natural parameter*, ψ is called the *log-partition function*, and h is called the *base measure*. The distribution with density f_η is called a *one-parameter natural exponential family*. Therefore, $y \sim N(\mu, 1)$ is in the exponential family with

$$\eta = \mu, \quad \psi(\eta) = -\frac{1}{2}\eta^2, \quad h(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right). \tag{3}$$

Several other well-known distributions are in the exponential family as well. For example, consider $y \sim \text{Ber}(\pi)$. Then, we have

$$f(y) = \pi^y(1-\pi)^{1-y} = \exp\left(y \log\frac{\pi}{1-\pi} + \log(1-\pi)\right). \tag{4}$$

Therefore, we have $\eta = \log\frac{\pi}{1-\pi}$, so that $\log(1-\pi) = -\log(1+e^\eta)$. It follows that

$$\eta = \log\frac{\pi}{1-\pi}, \quad \psi(\eta) = \log(1+e^\eta), \quad h(y) = 1. \tag{5}$$

As another example, consider the Poisson distribution $y \sim \text{Poi}(\mu)$. We have

$$f(y) = e^{-\mu}\frac{\mu^y}{y!} = \exp(y \log \mu - \mu)\frac{1}{y!}. \tag{6}$$

Therefore, we have $\eta = \log \mu$, so that $\mu = e^\eta$. It follows that

$$\eta = \log \mu, \quad \psi(\eta) = e^\eta, \quad h(y) = \frac{1}{y!}. \tag{7}$$

Moments of exponential family distributions. It turns out that the derivatives of the log-partition function ψ give the moments of y. Indeed, let's start with the relationship

$$\int f_\eta(y)dy = \int \exp(\eta y - \psi(\eta))h(y)dy = 1. \tag{8}$$

Differentiating in η and interchanging the derivative and the integral, we obtain

$$0 = \frac{d}{d\eta}\int f_\eta(y)dy = \int (y - \dot{\psi}(\eta))f_\eta(y)dy, \tag{9}$$

from which it follows that

$$\dot{\psi}(\eta) = \int \dot{\psi}(\eta)f_\eta(y)dy = \int y f_\eta(y)dy = \mathbb{E}_\eta[y] \equiv \mu_\eta. \tag{10}$$

Thus, the first derivative of the log partition function is the mean of y. Differentiating again, we get

$$\ddot{\psi}(\eta) = \int y(y - \dot{\psi}(\eta))f_\eta(y)dy = \int y(y - \mu_\eta)f_\eta(y)dy = \int (y - \mu_\eta)^2 f_\eta(y)dy = \text{Var}_\eta[y]. \tag{11}$$

Thus, the second derivative of the log-partition function is the variance of y.

Relationship between mean and natural parameter. The log-partition function ψ induces a connection (10) between the natural parameter η and the mean μ. Because

$$\frac{d\mu}{d\eta} = \frac{d}{d\eta}\dot{\psi}(\eta) = \ddot{\psi}(\eta) = \text{Var}_\eta[y] > 0, \tag{12}$$

it follows that μ is a strictly increasing function of η, so in particular the mapping between μ and η is bijective. Therefore, we can think of equivalently parameterizing the distribution via μ or η. In the context of GLMs (see Section 2), the mean-variance relationship is quantified in terms of the *canonical link function* g, which maps the mean to the natural parameter:

$$\eta = \dot{\psi}^{-1}(\mu) \equiv g(\mu). \tag{13}$$

Relationship between mean and variance. Note that the mean of an exponential family distribution determines its variance (since it determines the natural parameter η). For example, a Poisson random variable with mean μ has variance μ and a Bernoulli random variable with mean μ has variance $\mu(1 - \mu)$. The mean-variance relationship turns out to characterize the exponential family distribution, i.e. an exponential family distribution with mean equal to its variance is the Poisson distribution.

2 Generalized linear models and examples

In this class, the focus is on building models that tie a vector of predictors ($\boldsymbol{x}_{i*}$) to a response y_i. For linear regression, the mean of y was modeled as a linear combination of the predictors $\boldsymbol{x}_{i*}^T\boldsymbol{\beta}$: $\mu = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}$. Typically, the "right" thing to do is to model the response linearly on the scale of the natural parameter η rather than on the scale of the mean parameter μ. It just happens for linear models (where the underlying distribution is Gaussian) that these two parameters coincide.

Definition. We define $\{(y_i, \boldsymbol{x}_{i*})\}_{i=1}^n$ as following a generalized linear model based on the exponential family f_η if

$$y_i \overset{\text{ind}}{\sim} f_{\eta_i}, \quad \eta_i = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}. \tag{14}$$

GLMs are often written in terms of their link functions g, which relate the mean of y to the linear predictor $\boldsymbol{x}_{i*}^T\boldsymbol{\beta}$. When modeling the natural parameter as a linear function in the predictors, as in the definition (14), we get a GLM with *canonical link function* $g = \dot{\psi}^{-1}$:

$$g(\mathbb{E}[y_i]) = \dot{\psi}^{-1}(\mathbb{E}[y_i]) = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}. \tag{15}$$

Examples. For example, *logistic regression* is the GLM based on the Bernoulli distribution:

$$y_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); \quad \eta_i = \log\frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}. \tag{16}$$

Thus the canonical link function for logistic regression is the *logistic link function* $g(\mu) = \log\frac{\mu}{1-\mu}$. As another example, *Poisson regression* is the GLM based on the Poisson distribution:

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i); \quad \eta_i = \log\mu_i = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}. \tag{17}$$

Thus the canonical link function for Poisson regression is the *log link function* $g(\mu) = \log\mu$.

3 Maximum likelihood estimation in GLMs

GLM normal equations. Recall that the least squares estimate $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate. For general GLMs, we also estimate $\boldsymbol{\beta}$ via maximum likelihood. To derive this estimates, let's write down the GLM likelihood and then take a derivative. The GLM likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^n f_{\eta_i}(y_i) = \prod_{i=1}^n \exp(\eta_i y_i - \psi(\eta_i))h(y_i). \tag{18}$$

Taking a logarithm, we have

$$\log\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^n(\eta_i y_i - \psi(\eta_i)) + \sum_{i=1}^n \log h(y_i) = \sum_{i=1}^n(\boldsymbol{x}_{i*}^T\boldsymbol{\beta}\, y_i - \psi(\boldsymbol{x}_{i*}^T\boldsymbol{\beta})) + \sum_{i=1}^n \log h(y_i). \tag{19}$$

Taking a gradient in $\boldsymbol{\beta}$, we get

$$\nabla_\beta \log\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^n(\boldsymbol{x}_{i*}y_i - \boldsymbol{x}_{i*}\dot{\psi}(\boldsymbol{x}_{i*}^T\boldsymbol{\beta})) = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \tag{20}$$

Setting this expression to zero, we get the normal equations:

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = 0. \tag{21}$$

Recall that, for least squares, we got the same equation, with $\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. We can interpret the normal equations as stating that $\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$ is a projection of $\boldsymbol{y}$ onto the model "space"

$$C_\mu(\boldsymbol{X}) \equiv \{\boldsymbol{\mu} = \dot{\psi}(\boldsymbol{\eta}) = \dot{\psi}(\boldsymbol{X}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}. \tag{22}$$

parallel to the columns of $\boldsymbol{X}$. Note that the subscript μ on $C_\mu(\boldsymbol{X})$ indicates that we are considering the "space" (actually, *set*) of possible $\boldsymbol{\mu}$ as opposed to the space $C_\eta(\boldsymbol{X})$ of possible $\boldsymbol{\eta}$, which we denoted in Unit 1 simply as $C(\boldsymbol{X})$. For linear models, it is the case that $C_\mu(\boldsymbol{X}) = C_\eta(\boldsymbol{X})$, but in general, these two are different. Note that $C_\mu(\boldsymbol{X})$ in general is a manifold as opposed to a linear subspace of $\mathbb{R}^n$, while $C_\eta(\boldsymbol{X})$ is always a linear subspace.

Log-concavity of GLM likelihood. Unlike linear regression, in general GLMs the function $\boldsymbol{\mu}(\boldsymbol{\beta})$ is nonlinear. Therefore, there is in general no closed-form solution to the GLM normal equations (21). We must instead iteratively compute the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$. Before talking about the computation of the MLE $\widehat{\boldsymbol{\beta}}$, we state the important fact that $\log \mathcal{L}(\boldsymbol{\beta})$ is a concave function of $\boldsymbol{\beta}$, which implies that this function is "easy to optimize", i.e. has no local maxima.

Proposition 3.1. *The function* $\log \mathcal{L}(\boldsymbol{\beta})$ *defined in* (19) *is concave in* $\boldsymbol{\beta}$.

Proof. We claim it suffices to show that ψ is a convex function. Indeed, then $\log \mathcal{L}(\boldsymbol{\beta})$ would be the sum of a linear function of $\boldsymbol{\beta}$ and the composition of a concave function with a linear function. To verify that ψ is convex, it suffices to recall that $\ddot{\psi}(\eta) = \mathrm{Var}_\eta[y] \geq 0$. □

Proposition (3.1) gives us confidence that an iterative algorithm will converge to the global maximum of the likelihood. We present such an iterative algorithm next.

Newton-Raphson. We can solve the equation (21) using the Newton Raphson algorithm, which involves the gradient and Hessian of the function we'd like to maximize. We already computed the gradient in equation (20). To compute the Hessian, we take another gradient in $\boldsymbol{\beta}$. We have

$$\begin{aligned} \nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}}(\boldsymbol{X}^T(\boldsymbol{y} - \dot{\psi}(\boldsymbol{X}\boldsymbol{\beta}))) = -\nabla_{\boldsymbol{\beta}}\boldsymbol{X}^T\dot{\psi}(\boldsymbol{X}\boldsymbol{\beta}) \\ &= -\boldsymbol{X}^T \mathrm{diag}(\ddot{\psi}(\boldsymbol{X}\boldsymbol{\beta}))\boldsymbol{X} \equiv -\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}. \end{aligned} \tag{23}$$

Here, $\dot{\psi}$ and $\ddot{\psi}$ applied to vectors are interpreted element-wise and $\boldsymbol{W}(\boldsymbol{\beta}) \in \mathbb{R}^{n\times n}$ is the diagonal matrix such that

$$W_{ii}(\boldsymbol{\beta}) = \mathrm{Var}_{\boldsymbol{\beta}}[y_i]. \tag{24}$$

The Newton-Raphson iteration is therefore

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - (\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\widehat{\boldsymbol{\beta}}^{(t)}))^{-1}\nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\widehat{\boldsymbol{\beta}}^{(t)}) = \widehat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(t)})). \tag{25}$$

Iteratively reweighted least squares (IRLS). A nice interpretation of the Newton-Raphson algorithm is as a sequence of weighted least squares fits, known as the iteratively reweighted least squares (IRLS) algorithm. Suppose that we have a current estimate $\widehat{\boldsymbol{\beta}}^{(t)}$, and suppose we are looking for a vector $\boldsymbol{\beta}$ near $\widehat{\boldsymbol{\beta}}^{(t)}$ that fits the model even better. We have

$$\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{y}] = \dot{\psi}(\boldsymbol{X}\boldsymbol{\beta}) \approx \dot{\psi}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)}) + \mathrm{diag}(\ddot{\psi}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)}))(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)}) = \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(t)}) + \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)}).$$

and

$$\mathrm{Var}_{\boldsymbol{\beta}}[\boldsymbol{y}] \approx \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)}).$$

Thus, up to the first two moments, near $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(t)}$ the distribution of $\boldsymbol{y}$ is approximately

$$\boldsymbol{y} = \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(t)}) + \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})), \tag{26}$$

or, equivalently,

$$\boldsymbol{z}^{(t)} \equiv \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(t)})) + \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(t)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}', \quad \boldsymbol{\epsilon}' \sim N(\boldsymbol{0}, \boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})^{-1}). \tag{27}$$

The regression of the *adjusted response variable* $\boldsymbol{z}^{(t)}$ on $\boldsymbol{X}$ leaves us with a weighted linear regression, whose maximum likelihood estimate is

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}}^{(t)})\boldsymbol{z}^{(t)}, \tag{28}$$

which we define as our next iterate. It's easy to verify that the IRLS iteration (28) is equivalent to the Newton-Raphson iteration (25).

4 Inference in GLMs

Asymptotic normality and GLM standard errors. Inference in GLMs is based on asymptotic likelihood theory. Using the Hessian computation (23), we can compute the Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbb{E}_{\boldsymbol{\beta}}[\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta})] = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}, \tag{29}$$

recalling the definition of $\boldsymbol{W}$ in equation (24). Therefore, likelihood theory tells us that, as the sample size n grows, we have

$$\widehat{\boldsymbol{\beta}} \mathrel{\dot\sim} N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X})^{-1}). \tag{30}$$

Using the plug-in variance estimate, we typically construct GLM standard errors based on

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] \equiv (\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}. \tag{31}$$

Confidence intervals. A confidence interval for each coordinate β_j can be obtained via

$$\mathrm{CI}(\widehat{\beta}_j) \equiv \widehat{\beta}_j \pm 2 \cdot \mathrm{SE}(\widehat{\beta}_j), \quad \text{where} \quad \mathrm{SE}(\beta_j) \equiv \sqrt{(\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}_{jj}}. \tag{32}$$

A confidence interval for $\eta_i = \boldsymbol{x}_{i*}^T\boldsymbol{\beta}$ can be obtained via

$$\mathrm{CI}(\widehat{\eta}_i) \equiv \boldsymbol{x}_{i*}^T\widehat{\boldsymbol{\beta}} \pm 2 \cdot \mathrm{SE}(\eta_i), \quad \text{where} \quad \mathrm{SE}(\widehat{\eta}_i) \equiv \sqrt{\boldsymbol{x}_{i*}^T(\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}\boldsymbol{x}_{i*}}. \tag{33}$$

A confidence interval for $\mu_i \equiv \mathbb{E}_{\boldsymbol{\beta}}[y_i] = \dot\psi(\eta_i)$ can be obtained by applying the strictly increasing function $\dot\psi$ to the endpoints of the confidence interval for η_i. Note that the resulting confidence interval may be asymmetric.

Testing a single coordinate. We can invert the confidence interval (32) to get a test of the hypothesis $H_0 : \beta_j = \beta_j^0$ for any $\beta_j^0 \in \mathbb{R}$:

$$\phi(\boldsymbol{X}, \boldsymbol{y}) = \mathbb{1}(|z(\boldsymbol{X}, \boldsymbol{y})| > z_{1-\alpha/2}), \quad \text{where} \quad z(\boldsymbol{X}, \boldsymbol{y}) \equiv \frac{\widehat{\beta}_j}{\mathrm{SE}(\widehat{\beta}_j)}. \tag{34}$$

This is the analog of the t-test for a linear regression.

Testing multiple coordinates. Likelihood ratio test, deviances, goodness of fit testing. (TBD)

5 Bells and whistles

Offsets.

Exponential dispersion families.

Non-canonical links.