Unit 4: Generalized linear models: General theory

Eugene Katsevich

November 8, 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

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

Here is a way of generalizing this density:

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

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

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

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

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

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

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

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 $\theta = \log \mu$, so that $\mu = e^\theta$. It follows that

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

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

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

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

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

from which it follows that

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

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

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

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 θ and the mean μ. Because

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

it follows that μ is a strictly increasing function of θ, so in particular the mapping between μ and θ is bijective. Therefore, we can think of equivalently parameterizing the distribution via μ or θ. 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:

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

Relationship between mean and variance. Note that the mean of an exponential family distribution determines its variance (since it determines the natural parameter θ). 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 θ 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_θ if

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

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:

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

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:

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

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

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

Taking a logarithm, we have

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

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"

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

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 $C_\theta(\boldsymbol{X})$ of possible $\boldsymbol{\theta}$, which we denoted in Unit 1 simply as $C(\boldsymbol{X})$. For linear models, it is the case that $C_\mu(\boldsymbol{X}) = C_\theta(\boldsymbol{X})$, 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 $C_\theta(\boldsymbol{X})$ 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 $\ddot{\psi}(\theta) = \mathrm{Var}_\theta[y] \geq 0$. □

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).

Deviance (definition). Suppose that

$$y_i \overset{\text{ind}}{\sim} f_{\theta_i} \tag{29}$$

for some vector $\boldsymbol{\theta} \in \mathbb{R}^n$. Then, the log likelihood, expressed as a function of $\boldsymbol{\mu} \in \mathbb{R}^n$, is

$$L(\boldsymbol{y}; \boldsymbol{\mu}) \equiv \sum_{i=1}^n \theta_i y_i - \psi(\theta_i) + \sum_{i=1}^n \log h(y_i) = \sum_{i=1}^n g(\mu_i) y_i - \psi(g(\mu_i)) + \sum_{i=1}^n \log h(y_i). \tag{30}$$

When we fit a GLM, we choose

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{y}; \boldsymbol{\mu}(\boldsymbol{\beta})) \quad \Longleftrightarrow \quad \widehat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu} \in C_\mu(\boldsymbol{X})} L(\boldsymbol{y}; \boldsymbol{\mu}). \tag{31}$$

Thus a GLM can be viewed as a constrained optimization problem over $\boldsymbol{\mu} \in C_\mu(\boldsymbol{X}) \subset \mathbb{R}^n$. What if we were to maximize $L(\boldsymbol{y}; \boldsymbol{\mu})$ over all $\boldsymbol{\mu} \in \mathbb{R}^d$? It is easy to see that the $\boldsymbol{\mu}$ we would obtain is $\boldsymbol{\mu} = \boldsymbol{y}$. This model is called the *saturated model*. Inspired by this fact, we define the *deviance* statistic

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \equiv 2(L(\boldsymbol{y}; \boldsymbol{y}) - L(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})) = 2\left(\max_{\boldsymbol{\mu} \in \mathbb{R}^d} L(\boldsymbol{y}; \boldsymbol{\mu}) - \max_{\boldsymbol{\mu} \in C_\mu(\boldsymbol{X})} L(\boldsymbol{y}; \boldsymbol{\mu})\right). \tag{32}$$

We can view $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \geq 0$ as a measure of the *lack of fit* of a GLM. We could in principle define the deviance for any pair $(\boldsymbol{y}, \boldsymbol{\mu})$ via

$$D(\boldsymbol{y}; \boldsymbol{\mu}) \equiv 2(L(\boldsymbol{y}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{\mu})) = 2\left(\sum_{i=1}^n (g(y_i) - g(\mu_i))y_i - (\psi(g(y_i)) - \psi(g(\mu_i)))\right). \tag{33}$$

Then, it is clear that maximizing the likelihood is equivalent to minimizing the deviance:

$$\widehat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu} \in C_\mu(\boldsymbol{X})} L(\boldsymbol{y}; \boldsymbol{\mu}) = \arg\min_{\boldsymbol{\mu} \in C_\mu(\boldsymbol{X})} D(\boldsymbol{y}; \boldsymbol{\mu}). \tag{34}$$

Deviance (examples). Let's first compute the deviance for $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$. We have $L(\boldsymbol{y}, \boldsymbol{\mu}) = -\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 - \frac{n}{2}\log 2\mu$ and $L(\boldsymbol{y}, \boldsymbol{y}) = -\frac{n}{2}\log 2\mu$, so

$$D(\boldsymbol{y}; \boldsymbol{\mu}) = \|\boldsymbol{y} - \boldsymbol{\mu}\|^2, \tag{35}$$

which we recognize as the familiar residual sum of squares (RSS). Therefore, the deviance is a generalization of the RSS. Let's compute the deviance for a Poisson regression, where $\psi(\theta) = e^\theta$ and $g(\mu) = \log(\mu)$. We have

$$D(\boldsymbol{y}; \boldsymbol{\mu}) = 2\left(\sum_{i=1}^n (g(y_i) - g(\mu_i))y_i - (\psi(g(y_i)) - \psi(g(\mu_i)))\right) = 2\left(\sum_{i=1}^n y_i \log\frac{y_i}{\mu_i} - (y_i - \mu_i)\right). \tag{36}$$

Now, if $\widehat{\boldsymbol{\mu}}$ is the maximum likelihood mean vector for a Poisson regression including an intercept, the normal equations tell us that $\mathbf{1}^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = 0$, so

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^n y_i \log\frac{y_i}{\widehat{\mu}_i}. \tag{37}$$

This is the lack-of-fit measure that a Poisson regression seeks to minimize.

4 Inference in GLMs

Inferential goals. There are two types of inferential goals: hypothesis testing and confidence interval construction. Within hypothesis testing, we can test $H_0 : \beta_j = 0$ (importance of a single coefficient), $H_0 : \boldsymbol{\beta}_S = \mathbf{0}$ for some $S \subset \{0, \ldots, p-1\}$ (importance of a group of coefficients), or $H_0 : \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$ (goodness of fit). Within confidence intervals, we may want to construct intervals for the coefficients β_j or for fitted values θ_i or μ_i.

Inferential tools. Inference in GLMs is based on asymptotic likelihood theory. Hypothesis tests (and, by inversion, confidence intervals) can be constructed in three asymptotically equivalent ways: Wald tests, likelihood ratio tests (LRT), and score tests. Despite their asymptotic equivalence, in finite samples some tests may be preferable to others. We will discuss the most commonly applied methods for each inferential task, though others are possible as well.

4.1 Wald tests and confidence intervals

Asymptotic normality and Wald standard errors. Wald tests and confidence intervals are based on the large-sample distribution of the MLE, with covariance matrix equal to the Fisher information. Using the Hessian computation (23), we can compute the Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbb{E}_\beta[\nabla^2_\beta \log \mathcal{L}(\boldsymbol{\beta})] = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}, \tag{38}$$

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

$$\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X})^{-1}). \tag{39}$$

Using the plug-in variance estimate, we can construct Wald standard errors based on

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] \equiv (\boldsymbol{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}. \tag{40}$$

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

$$\text{CI}(\widehat{\beta}_j) \equiv \widehat{\beta}_j \pm 2 \cdot \text{SE}(\widehat{\beta}_j), \quad \text{where} \quad \text{SE}(\widehat{\beta}_j) \equiv \sqrt{(\boldsymbol{X}^T \boldsymbol{W}(\widehat{\boldsymbol{\beta}}) \boldsymbol{X})^{-1}_{jj}}. \tag{41}$$

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

$$\text{CI}(\widehat{\theta}_i) \equiv \boldsymbol{x}_{i*}^T \widehat{\boldsymbol{\beta}} \pm 2 \cdot \text{SE}(\theta_i), \quad \text{where} \quad \text{SE}(\widehat{\theta}_i) \equiv \sqrt{\boldsymbol{x}_{i*}^T (\boldsymbol{X}^T \boldsymbol{W}(\widehat{\boldsymbol{\beta}}) \boldsymbol{X})^{-1} \boldsymbol{x}_{i*}}. \tag{42}$$

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

Wald test for a single coefficient. We can invert the confidence interval (41) to get a test of the hypothesis $H_0 : \beta_j = \beta_j^0$ for any $\beta_j^0 \in \mathbb{R}$:

$$\phi(\boldsymbol{X}, \boldsymbol{y}) = \mathbb{1}(|z(\boldsymbol{X}, \boldsymbol{y})| > z_{1-\alpha/2}), \quad \text{where} \quad z(\boldsymbol{X}, \boldsymbol{y}) \equiv \frac{\widehat{\beta}_j - \beta_j^0}{\text{SE}(\widehat{\beta}_j)}. \tag{43}$$

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

4.2 Likelihood ratio tests and confidence intervals

Testing one or more coefficients. Suppose that $S \subset \{0, 1, \ldots, p-1\}$ and we wish to test the null hypothesis $H_0 : \boldsymbol{\beta}_S = \boldsymbol{0}$. For linear regression, we used an F-test for this purpose. In Homework 2, we saw that an F-test is related to a likelihood ratio test. The likelihood ratio test can be defined for arbitrary GLMs, and is usually how we test multiple coordinates. To define the likelihood ratio test, let $\widehat{\boldsymbol{\mu}}_{\text{-}S} \in \mathcal{R}^n$ the maximum likelihood mean vector under the null hypothesis, and let $\widehat{\boldsymbol{\mu}}$ denote the maximum likelihood mean vector without restrictions on $\boldsymbol{\beta}$. Then, the likelihood ratio test statistic is

$$T^{\text{LRT}} \equiv 2(L(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) - L(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}_{\text{-}S})), \tag{44}$$

and

$$\text{under } H_0, \quad T^{\text{LRT}} \xrightarrow{d} \chi^2_{|S|}. \tag{45}$$

Note that the LRT test statistic can also be expressed as a difference in deviances:

$$T^{\text{LRT}} = D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}_{\text{-}S}) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}). \tag{46}$$

We see the connection with the F-test, whose numerator is the difference in the RSSs of the partial and full models.

LRT-based confidence intervals. Sometimes, Wald confidence intervals do not work very well in finite samples, e.g. if $\widehat{\boldsymbol{\beta}} \to \infty$. In these cases, the LRT can be inverted to get more reliable confidence intervals, though this is less straightforward conceptually and computationally.

Goodness of fit tests. In some cases, we want to compare a GLM model to a *saturated model*. In this case, we can use a likelihood ratio test similar to that applied to test multiple coefficients. It turns out that $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ is exactly the likelihood ratio statistic we want. Under *small dispersion asymptotics*, we can expect it to have a χ^2_{n-p} distribution under the null.

4.3 Score tests

Goodness of fit tests. Score tests are primarily used as alternatives to likelihood ratio tests for testing goodness of fit in GLMs. Score tests are based on the fact that

$$\text{under } H_0, \quad \nabla_\theta \log \mathcal{L}(\widehat{\boldsymbol{\theta}}_{\mathbf{0}}) I^{-1}(\widehat{\boldsymbol{\theta}}_{\mathbf{0}}) \nabla_\theta \log \mathcal{L}(\widehat{\boldsymbol{\theta}}_{\mathbf{0}}) \to \chi^2_{n-p}, \tag{47}$$

where $\widehat{\boldsymbol{\theta}}_{\mathbf{0}}$ is the maximum likelihood estimate under the null hypothesis. For GLMs, note that

$$\nabla_\theta \log \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \quad \text{and} \quad I(\boldsymbol{\theta}) = \operatorname{diag}(\ddot{\psi}(\boldsymbol{\theta})). \tag{48}$$

Therefore, we arrive at the statistic

$$X^2 \equiv \nabla_\theta \log \mathcal{L}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) I^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \nabla_\theta \log \mathcal{L}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{W}^{-1} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\operatorname{var}(\widehat{\mu}_i)}. \tag{49}$$

This is Pearson's famous chi-squared statistic, which he proposed in 1900. It was only pointed out that this is a score test many decades later.

5 Further generalizations

The definitions and theory of GLMs introduced in the previous sections were simplified in several ways for the sake of exposition. Here we discuss a more general definition of GLMs that accounts for (1) a dispersion parameter, (2) offsets, and (3) non-canonical links. These elements will be introduced below.

5.1 Exponential dispersion models (EDMs)

Definition. An EDM is a generalization of exponential family models that includes a *dispersion parameter*. An EDM $f_{\theta,\phi}$ is parameterized by a natural parameter $\theta \in \mathbb{R}$ and a dispersion parameter $\phi > 0$:

$$f_{\theta,\phi}(y) = \exp\left(\frac{\theta y - \psi(\theta)}{\phi}\right) h(y, \phi). \tag{50}$$

Sometimes, we parameterize this distribution using its mean and dispersion, writing

$$y \sim \operatorname{EDM}(\mu, \phi). \tag{51}$$

Examples. For example, the distribution $N(\mu, \sigma^2)$ falls into this class:

$$f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) = \exp\left(\frac{\mu y - \frac{1}{2}\mu^2}{\sigma^2}\right) \cdot \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} y^2\right). \tag{52}$$

Therefore, we have

$$\theta = \mu; \quad \psi(\theta) = \frac{1}{2}\theta^2; \quad \phi = \sigma^2; \quad h(y, \phi) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} y^2\right). \tag{53}$$

The Bernoulli and Poisson distributions are special cases with $\phi = 1$, and θ and $\psi(\theta)$ as derived before. Binomial proportions y such that $my \sim \operatorname{Bin}(m, \mu)$ also have EDM distributions:

$$f(y) = \binom{m}{my} \mu^{my} (1-\mu)^{m(1-y)} = \exp\left(m\left(y \log \frac{\mu}{1-\mu} + \log(1-\mu)\right)\right) \binom{m}{my}, \tag{54}$$

so

$$\theta = \log \frac{\mu}{1-\mu}; \quad \psi(\theta) = \frac{e^\theta}{1+e^\theta}; \quad \phi = 1/m; \quad h(y,\phi) = \binom{m}{my}. \tag{55}$$

Many other examples fall into this class, including the negative binomial, gamma, and inverse-Gaussian distributions.

Mean and variance. We can employ similar tricks as before to derive the mean and variance of an EDM:

$$\mu = \mathbb{E}_\theta[y] = \dot{\psi}(\theta); \quad \text{Var}_\theta[y] = \phi \cdot \ddot{\psi}(\theta). \tag{56}$$

There are the same relationships we found before, except the variance function has an extra factor of ϕ.

5.2 GLMs based on EDMs

Definition. We define a GLM based on an EDM as follows:

$$y_i \overset{\text{ind}}{\sim} \text{EDM}(\mu_i, \phi/w_i), \quad \eta_i \equiv g(\mu_i) = o_i + \boldsymbol{x}_{i*}^T \boldsymbol{\beta}. \tag{57}$$

Here, w_i are known *observation weights*, g is the *link function*, η_i is the *linear predictor*, and o_i are *offsets* (known terms contributing additively to the linear predictor). The parameters $\boldsymbol{\beta}$ are unknown, and ϕ might or might not be known. For example, in Poisson regression ϕ is known to be 1 but in linear regression $\phi = \sigma^2$ is unknown. For example, consider logistic regression with *grouped data*:

$$n_i y_i \sim \text{Bin}(n_i, \mu_i); \quad \eta_i = \log \frac{\mu_i}{1-\mu_i} = o_i + \boldsymbol{x}_{i*}^T \boldsymbol{\beta}. \tag{58}$$

Here, $\phi = 1$ and $w_i = n_i$.

Deviance. The log-likelihood of a GLM is

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\theta_i y_i - \psi(\theta_i)}{\phi/w_i} + \sum_{i=1}^n \log h(y_i, \phi/w_i). \tag{59}$$

Expressing this in terms of $\boldsymbol{\mu}$, we have

$$L(\boldsymbol{y}; \boldsymbol{\mu}) = \sum_{i=1}^n \frac{\dot{\psi}^{-1}(\mu_i) y_i - \psi(\dot{\psi}^{-1}(\mu_i))}{\phi/w_i} + \sum_{i=1}^n \log h(y_i, \phi/w_i). \tag{60}$$

We define the deviance $D(\boldsymbol{y}; \boldsymbol{\mu})$ via

$$2(L(\boldsymbol{y}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{\mu})) = \frac{1}{\phi} \sum_{i=1}^n w_i \left((\dot{\psi}^{-1}(y_i) - \dot{\psi}^{-1}(\mu_i)) y_i - (\psi(\dot{\psi}^{-1}(y_i)) - \psi(\dot{\psi}^{-1}(\mu_i))) \right) \equiv \frac{1}{\phi} D(\boldsymbol{y}; \boldsymbol{\mu}). \tag{61}$$

Estimation of $\boldsymbol{\beta}$. Taking a gradient in $\boldsymbol{\beta}$ using the chain rule, we obtain:

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = (\boldsymbol{y} - \boldsymbol{\mu})^T \text{diag}(\phi/w_i)^{-1} \cdot \text{diag}(\ddot{\psi}(\theta_i))^{-1} \cdot \text{diag}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \cdot \boldsymbol{X}. \quad (62)$$

Transposing and setting to zero, we get the normal equations

$$0 = \left(\frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T = \boldsymbol{X}^T \text{diag}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \text{diag}\left(\frac{\phi}{w_i}\ddot{\psi}(\theta_i)\right)^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \equiv \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}). \quad (63)$$

Here, $\boldsymbol{D} = \text{diag}(\partial \mu_i / \partial \eta_i)$ and $\boldsymbol{V} = \text{diag}\left(\frac{\phi}{w_i}\ddot{\psi}(\boldsymbol{\theta})\right) = \text{diag}(\text{Var}[y_i])$. We can solve these normal equations using a generalized version of iteratively reweighted least squares. Notably, the dispersion parameter ϕ cancels from the normal equations, so estimation of ϕ is not required to estimate $\boldsymbol{\beta}$.

Estimation of ϕ. While sometimes the parameter ϕ is known (e.g. for binomial or Poisson GLMs), in other cases ϕ must be estimated (e.g. for the normal linear model). It turns out that we can generalize the linear model estimator $\widehat{\sigma}^2 = \frac{1}{n-p}\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2$ to

$$\widehat{\phi} = \frac{1}{n-p} D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}). \quad (64)$$

This estimator performs decently well.

Wald inference. Let's first compute the Fisher information matrix:

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\beta}) &= \text{Var}\left[\left(\frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T\right] \\ &= \text{Var}[\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})] \\ &= \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} \text{Var}[\boldsymbol{y}] \boldsymbol{V}^{-1} \boldsymbol{D} \boldsymbol{X} \\ &= \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} \boldsymbol{V} \boldsymbol{V}^{-1} \boldsymbol{D} \boldsymbol{X} \\ &= \boldsymbol{X}^T \boldsymbol{D}^2 \boldsymbol{V}^{-1} \boldsymbol{X} \\ &\equiv \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}. \end{aligned} \quad (65)$$

Here,

$$\boldsymbol{W} = \text{diag}\left(\frac{(\partial \mu_i / \partial \eta_i)^2}{\text{Var}[y_i]}\right). \quad (66)$$

Therefore, once again we have

$$\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}). \quad (67)$$

Using the plug-in principle (including plugging in an estimator of ϕ if this parameter is unknown), we define

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}] \equiv (\boldsymbol{X}^T \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1}, \quad (68)$$

based on which we can conduct Wald tests and construct Wald confidence intervals. If a plug-in estimate is used for ϕ, then in small samples t_{n-p} is a better approximation of the null distribution than $N(0, 1)$.

Likelihood ratio test inference. Suppose we want to test $H_0 : \boldsymbol{\beta}_S = \mathbf{0}$. Then, asymptotic theory tells us that under the null,

$$2(L(y;\widehat{\boldsymbol{\mu}}) - L(y;\widehat{\boldsymbol{\mu}}_{\text{-}S})) = \frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}_{\text{-}S}) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{\phi} \to \chi^2_{|S|}. \tag{69}$$

If ϕ is known, then we can construct a chi-square test directly based on the above asymptotic null distribution. If ϕ is unknown, we can estimate it as discussed above, and construct an F-statistic as follows:

$$F \equiv \frac{(D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}_{\text{-}S}) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}))/|S|}{\widehat{\phi}} = \frac{(D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}_{\text{-}S}) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}))/|S|}{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)}. \tag{70}$$

In normal linear model theory, the null distribution of F is *exactly* $F_{|S|,n-p}$. For GLMs, the null distribution of F is *approximately* $F_{|S|,n-p}$. For ϕ known, we can also construct a goodness of fit test: This includes comparing the GLM to a saturated model, to get a goodness of fit test via

$$\frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{\phi} \to \chi^2_{n-p}, \tag{71}$$

assuming the saturated model can be estimated relatively well (small dispersion asymptotics).

Score test inference. By the same exact logic as in Section 4.3, we get that

$$X^2 \equiv \nabla_\theta \log \mathcal{L}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) I^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \nabla_\theta \log \mathcal{L}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T \text{diag}(\ddot{\psi}(\boldsymbol{\theta}))^{-1} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^n \frac{(y_i - \widehat{\mu}_i)^2}{\frac{1}{\phi}\text{var}(\widehat{\mu}_i)}.$$

the one difference being the extra factor of ϕ. Under small-dispersion asymptotics, this test statistic has null distribution χ^2_{n-p}.

6 R demo

Let's revisit the crime data from Homework 2, this time fitting a logistic regression to it.

```
# read crime data
crime_data = read_tsv("../data/Statewide_crime.dat")

# read and transform population data
population_data = read_csv("../data/state-populations.csv")
population_data = population_data %>%
  filter(State != "Puerto Rico") %>%
  select(State, Pop) %>%
  rename(state_name = State, state_pop = Pop)

# collate state abbreviations
state_abbreviations = tibble(state_name = state.name,
                             state_abbrev = state.abb) %>%
  add_row(state_name = "District of Columbia", state_abbrev = "DC")

# add CrimeRate to crime_data
crime_data = crime_data %>%
```

```
  mutate(STATE = ifelse(STATE == "IO", "IA", STATE)) %>%
  rename(state_abbrev = STATE) %>%
  filter(state_abbrev != "DC") %>%       # remove outlier
  left_join(state_abbreviations, by = "state_abbrev") %>%
  left_join(population_data, by = "state_name") %>%
  mutate(CrimeRate = Violent/state_pop) %>%
  select(state_abbrev, CrimeRate, Metro, HighSchool, Poverty, state_pop)

crime_data
```

```
## # A tibble: 50 x 6
##    state_abbrev CrimeRate Metro HighSchool Poverty state_pop
##    <chr>            <dbl> <dbl>      <dbl>   <dbl>     <dbl>
##  1 AK           0.000819   65.6       90.2     8       724357
##  2 AL           0.0000871  55.4       82.4    13.7    4934193
##  3 AR           0.000150   52.5       79.2    12.1    3033946
##  4 AZ           0.0000682  88.2       84.4    11.9    7520103
##  5 CA           0.0000146  94.4       81.3    10.5   39613493
##  6 CO           0.0000585  84.5       88.3     7.3    5893634
##  7 CT           0.0000867  87.7       88.8     6.4    3552821
##  8 DE           0.000664   80.1       86.5     5.8     990334
##  9 FL           0.0000333  89.3       85.9     9.7   21944577
## 10 GA           0.0000419  71.6       85.2    10.8   10830007
## # ... with 40 more rows
```

We can fit a GLM using the `glm` command, specifying as additional arguments the observation weights as well as the exponential dispersion model. In this case, the weights are the state populations and the family is binomial:

```
glm_fit = glm(CrimeRate ~ Metro + HighSchool + Poverty,
              weights = state_pop,
              family = "binomial",
              data = crime_data)
```

We can print the summary table as usual:

```
summary(glm_fit)
```

```
##
## Call:
## glm(formula = CrimeRate ~ Metro + HighSchool + Poverty, family = "binomial",
##     data = crime_data, weights = state_pop)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -21.043   -9.176    0.418    9.053   47.174
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.609e+01  3.520e-01  -45.72   <2e-16 ***
## Metro       -2.586e-02  5.727e-04  -45.15   <2e-16 ***
## HighSchool   9.106e-02  3.450e-03   26.39   <2e-16 ***
## Poverty      6.077e-02  4.852e-03   12.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15590  on 49  degrees of freedom
## Residual deviance: 11742  on 46  degrees of freedom
## AIC: 12136
##
## Number of Fisher Scoring iterations: 5
```

Amazingly, everything is very significant! This is because the weights for each observation (the state populations) are very high, effectively making the sample size very high.

We can test individual coefficients or groups of coefficients using the likelihood ratio test, via `anova`. For example, let's take a look at the p-value for `Metro`:

```
glm_fit_partial = glm(CrimeRate ~ HighSchool + Poverty,
                      weights = state_pop,
                      family = "binomial",
                      data = crime_data)

anova_fit = anova(glm_fit_partial, glm_fit, test = "LRT")
anova_fit
```

```
## Analysis of Deviance Table
##
## Model 1: CrimeRate ~ HighSchool + Poverty
## Model 2: CrimeRate ~ Metro + HighSchool + Poverty
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        47      13649
## 2        46      11742  1   1907.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can manually carry out the LRT as a sanity check:

```
deviance_partial = deviance(glm_fit_partial)
deviance_full = deviance(glm_fit)
lrt_stat = deviance_partial - deviance_full
p_value = pchisq(lrt_stat, df = 1, lower.tail = FALSE)
tibble(lrt_stat, p_value)
```

```
## # A tibble: 1 x 2
##   lrt_stat p_value
##      <dbl>   <dbl>
```

```
## 1    1907.       0
```

We can get Wald confidence intervals for the coefficients using `confint`:

```
confint(glm_fit)

## Waiting for profiling to be done...

##                     2.5 %       97.5 %
## (Intercept) -16.78344072 -15.40360346
## Metro        -0.02697681  -0.02473192
## HighSchool    0.08430723   0.09783210
## Poverty       0.05125776   0.07027803
```

Or for the fitted values on the log-odds (natural parameter) scale using `predict`:

```
ci_log_odds = predict(glm_fit,
                      newdata = crime_data %>%
                        column_to_rownames(var = "state_abbrev"),
                      se.fit = TRUE) %>%
  as.data.frame() %>%
  rownames_to_column(var = "state") %>%
  as_tibble() %>%
  select(state, fit, se.fit)
ci_log_odds

## # A tibble: 50 x 3
##    state    fit  se.fit
##    <chr>  <dbl>   <dbl>
##  1 AK     -9.09 0.0124
##  2 AL     -9.19 0.0149
##  3 AR     -9.50 0.0221
##  4 AZ     -9.96 0.0144
##  5 CA    -10.5  0.0162
##  6 CO     -9.79 0.0104
##  7 CT     -9.88 0.0125
##  8 DE     -9.93 0.0175
##  9 FL     -9.99 0.0112
## 10 GA     -9.53 0.00788
## # ... with 40 more rows
```

Or for the fitted values on the probability scale by applying the logistic transformation to the endpoints of the above intervals:

```
logistic = function(x)(exp(x)/(1+exp(x)))
ci_probability = ci_log_odds %>%
  mutate(lower = logistic(fit-2*se.fit),
         upper = logistic(fit + 2*se.fit)) %>%
  select(state, lower, upper)
ci_probability
```

```
## # A tibble: 50 x 3
##    state     lower     upper
##    <chr>     <dbl>     <dbl>
##  1 AK    0.000110  0.000116
##  2 AL    0.0000991 0.000105
##  3 AR    0.0000714 0.0000780
##  4 AZ    0.0000457 0.0000484
##  5 CA    0.0000269 0.0000287
##  6 CO    0.0000547 0.0000570
##  7 CT    0.0000497 0.0000522
##  8 DE    0.0000468 0.0000502
##  9 FL    0.0000448 0.0000469
## 10 GA    0.0000716 0.0000739
## # ... with 40 more rows
```

R code for goodness of fit testing will be provided in Unit 5.