Generalized Linear Models

AMS 274

Fall 2016

This is a graduate-level course on the theory, methods and applications of Generalized Linear Models (GLMs). Emphasis will be placed on statistical modeling, building from standard normal linear models, extending to GLMs, and briefly covering more specialized topics. With regard to inference, prediction, and model assessment, we will study both likelihood and Bayesian methods. In particular, within the Bayesian modeling framework, we will discuss practically important hierarchical extensions of the standard GLM setting.

1 September 22, 2016: Introduction and Exponential Family

I missed this class due to a conference, but I've pieced the content together based on Tatiana's slides:

- Model building
- Extending linear modeling to get around assumption violations
- Exponential family

2 September 27, 2016:

GLM definition: Responses are $y_i \in \mathcal{S} \subseteq \mathbb{R}$ (finite, countable, or uncountable) with a support that does not depend on the parameters. A vector of (fixed) covariates are associated with each response $x_i = (x_{i1}, \ldots, x_{ip})^T$ for $i = 1, \ldots, n$

1. Random component

   The assumption is that the responses (y_i) are independent from the **exponential dispersion** (ED) family with parameters (θ_i, ϕ) with a pdf or pmf

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

   with $\theta_i \in \Theta$ and $\phi > 0$. This also has the location characteristic $\mu_i = E(y_i)$. Note that there can be a value in extending the systematic component to be represented as $\eta_i = h(x_1) + \ldots + h(x_p)$, which is not completely general (could go nonparametric), but it is easier to implement.

2. Systematic component

   $\eta_i = x_i^T\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$

3. Link function

   We have a transformation $g(\mu_i) = \eta_i = x_i^T\boldsymbol{\beta}$. This function is assumed to be fully specified

Exponential Family

$\theta = (\theta_1, \ldots, \theta_k) \in \Theta$

We have the distribution of the responses

$$f(y|\theta) = \exp\left\{c(\theta) + d(y) + \sum_{i=1}^{k} a_i(y)b_i(\theta)\right\}$$

This decomposition is **absolutely essential** to using generalized linear models. We have to be able to separate a function dependent on y that does not depend on parameters. Once we have this form, we can reparameterize using η (different from the systematic component).

$$f(y|\eta) = \exp\left\{c^*(\eta) + d^*(y) + \sum_{i=1}^{k} \eta_i a_i^*(y)\right\}$$

The important thing to notice here is that the function $b_i(\theta)$ becomes characterized by η_i

Example

We have $\mathrm{N}(\mu, \sigma^2)$

$$b_1(\mu, \sigma^2) = \frac{1}{\sigma^2} \qquad a_1(y) = -\frac{y^2}{2}$$
$$b_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2} \qquad a_2(y) = y$$

So the **natural characterization** is

$$\eta_1 = \frac{1}{\sigma^2}$$
$$\eta_2 = \frac{\mu}{\sigma^2}$$

Let's simplify our view to the one-parameter exponential family. Now we have

$$f(y|\theta) = \exp\{c(\theta) + d(y) + a(y)b(\theta)\}$$

So now the ED is $ED(\theta, \phi)$, where ϕ is fixed. This highlights the important characteristic that we cannot have a b_j that depends both on y and ϕ.

Examples

1. If you think of the normal distribution $\mathrm{N}(\mu, \sigma^2)$, you can write it in the form of the general exponential family. $\theta = \mu$, and $a(\phi) = \sigma^2 = \phi$
2. Another example is the gamma distribution $Gamma(\nu, \mu)$ where the parameter ν controls the shape and μ characterizes the location. In this case, $\theta = -\frac{1}{\mu}$, and $a(\phi) = \frac{1}{\nu}$.
3. Poisson(λ), $\theta = \log(\lambda)$
4. Binomial with n known, $\theta = \log\left(\frac{p}{1-p}\right)$

Something to take home: The reason we use generalized linear models is because η_i in the systematic component has support on $\mathbb{R}$, but the expected value of each response y_i is not always supported on $\mathbb{R}$.

If you are more concerned with the distribution theory, check out *Jorgensen (1987), JRSS-B, vol 49, p. 127-162*. This spends a little time talking about the **cumulant generating function**

$$k(\theta) = \log \int \exp(x^T\theta)g(x)dx$$

where $g(x)$ is the "density" of $x \in \mathbb{R}^k$, and $\theta \in \Theta = \left\{\theta \in \mathbb{R}^k : k(\theta) < \infty\right\}$ with $x \in \mathbb{R}$. For a univariate distribution, we have

$$k(t) = \log \int e^{tx} g(x) dx = \log(E(e^{tx}))$$

Note: in a special case of a univariate distribution, $b(\theta)$ is the cumulant generating function.

Properties for the ED family

$y|\theta, \phi \sim ED(\theta, \phi)$

What is the expectation and variance given θ and ϕ? Let's use some standard results relating to the score function:

$$\begin{aligned} \ell(\theta, \phi; y) &= \log L(\theta, \phi; y) \\ &= \log f(y|\theta, \phi) \\ E\left(\frac{\delta \ell}{\delta \theta}\right) &= 0 \end{aligned} \tag{1}$$

under **regularity conditions** (in order to switch the order of differentiation and integration). This was somewhat covered in 205B (covered in Casella & Berger). Additionally,

$$E\left(\frac{\delta^2 \ell}{\delta \theta^2}\right) + E\left[\left(\frac{\delta \ell}{\delta \theta}\right)^2\right] = 0 \tag{2}$$

This gives us a way to find the expectation and variance for the ED family.

$$\begin{aligned} \ell(\theta, \phi; y) &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \\ \frac{\delta \ell}{\delta \theta} &= \frac{y - b'(\theta)}{a(\phi)} \\ \frac{\delta^2 \ell}{\delta \theta^2} &= -\frac{b''(\theta)}{a(\phi)} \end{aligned}$$

From (1),

$$E(y|\theta, \phi) = b'(\theta)$$

and from (2),

$$\text{Var}(y|\theta, \phi) = a(\phi) b''(\theta),$$

so $b''(\theta)$ is used to define the "variance function", along with ϕ. Next, the variance-mean relationship is defined as

$$\text{Var}(y|\theta, \phi) = a(\phi) \left(\frac{dE(y|\theta, \phi)}{d\theta}\right).$$

In the normal distribution, there is no mean-variance relationship, but once you leave the comfort of the normal, all bets are off. Thanos suggests going through the earlier examples to see the mean-variance relations. If you go check out Jorgensen, you may see the notation $V(\mu)$, which signifies that the variance is somehow a function of the mean. We may use this notation, too. If you see variance functions of this type:

$$V(\mu) = \mu^p,$$

these power functions are particularly attractive within the context of GLMs. Jorgensen goes over different values of p that are "allowed". If you have a different value of p, you may leave the ED family, and your time in the realm of GLMs is up.

Convolution Property

$$y_i \overset{ind}{\sim} ED(\mu, \phi/w_i),\ i = 1, \ldots, n$$

Then we know that $\mu = b'(\theta)$ and $a_i(\phi) = \frac{\phi}{w_i}$ for fixed weights $w_i > 0$. Therefore,

$$\frac{1}{w}\sum_{i=1}^{n} w_i y_i \sim ED(\mu, \phi/w),$$

where $w = \sum_{i=1}^{n} w_i$. This means that the weighted average stays within the ED family with the same mean. If, instead, you assume that the responses are iid,

$$y_i \overset{iid}{\sim} ED(\mu, \phi),$$

then

$$\frac{1}{n}\sum_{i=1}^{n} y_i \sim ED(\mu, \phi/n)$$

Asymptotic Normality

In the context of

$$y \sim ED(\mu, \phi),$$

a result follows such that

$$\frac{y-\mu}{\sqrt{\phi}} \xrightarrow{d} \mathrm{N}(0, V(\mu)).$$

If $a(\phi) = \phi$, then this resulting distribution will have $\phi \to 0$.

Score Function

If we're in the context of

$$\ell(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi),$$

then the **score function**

$$U = \frac{\delta \ell}{\delta \theta}.$$

Note that in the two-dimensional case, this is actually a matrix, so thinking in this way is not, in Thanos terms, "legal", but this helps understand the concept. A key property (under regularity conditions) is that

$$E(U) = 0.$$

Another important quantity is the variance of U, which is often referred to as the expected Fisher information:

$$J = \mathrm{Var}(U) = -E\left(\frac{\delta^2 \ell}{\delta \theta^2}\right) = E\left(-\frac{\delta U}{\delta \theta}\right).$$

In our earlier notation, this is simplified to

$$U = \frac{y - b'(\theta)}{a(\phi)},$$

and

$$J = \mathrm{Var}(U) = \cdots = \frac{b''(\theta)}{a(\phi)} = \frac{\mathrm{Var}(y)}{a^2(\phi)}$$

3 September 29, 2016:

Notes were provided by Thanos and are on the website: `https://ams274-fall16-01.courses.soe.ucsc.edu/system/files/attachments/notes_0.pdf`.

Quick Remark on the Link Function

We have established that $g(\mu_i) = \mathbf{X}_i^T\boldsymbol{\beta}$ and $\mu_i = b'(\theta_i)$. First remark, for all standard GLMs, without loss of generality, g is invertible. Therefore,

$$E(y_i) = \mu_i = g^{-1}(\mathbf{X}_i^T\boldsymbol{\beta}).$$

This is **not** the same thing as the non-linear function $h(\mathbf{X}_i^T, \boldsymbol{\beta})$ mentioned in the notes. Note from our two equations:

$$b'(\theta_i) = g^{-1}(\mathbf{X}_i^T\boldsymbol{\beta}).$$

If we want to find θ_i, we could find θ_i by applying the inverse of b' to $g^{-1}(\mathbf{X}_i^T\boldsymbol{\beta})$. It helps to see this relationship between the θs and the βs. At some point in the scoring algorithm, we will need to use the relation

$$\left(f^{-1}\right)'(y_0) = \frac{1}{f'(x_0)}$$
$$y_0 = f(x_0)$$

Note that it is an important concept that θ_i is a function $\boldsymbol{\beta}$, but not for ϕ. Another important note is that if the scoring algorithm can be simplified such that it removes dependence on ϕ, then one can quickly find the MLE for $\hat{\boldsymbol{\beta}}$, and then find the maximization for ϕ from $\ell(\hat{\boldsymbol{\beta}}, \phi)$.

4 October 4, 2016

Some key results from last time:

- $y_i \overset{ind}{\sim} f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$ is the form of the ED family with $E(y_i) = \mu_i = b'(\theta_i)$, $\text{Var}(y_i) = a(\phi)b''(\theta_i) = V(\mu_i)$, and the link $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} \Rightarrow \mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$
- MLE estimation for $\boldsymbol{\beta}$ (Scoring method): $U = (U_1, \ldots, U_p)$ is the *score vector* corresponding to the regression coefficients where $U_j \equiv U_j(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^n \frac{(y_i - \mu_i)}{\text{Var}(y_i)} \frac{d\mu_i}{d\eta_i} x_{ij}$. Note that ϕ only enters this equation via the variance. The only other thing we needed we needed was the expected Fisher Information matrix $J_{kj} \equiv (J(\boldsymbol{\beta}))_{kj} = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{\text{Var}(y_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2$
- $b^{m+1} = b^m + (J(b^m))^{-1}U(b^m, y)$

The algorithm that we discussed for finding the MLE for $\boldsymbol{\beta}$ starts with $\ell(\boldsymbol{\beta}, \phi; y, x)$ in which we fix ϕ. Then

$$\hat{\boldsymbol{\beta}}_\phi \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \phi),$$

since $\hat{\boldsymbol{\beta}}_\phi$ is free of ϕ, then $\hat{\boldsymbol{\beta}}_\phi \equiv \hat{\boldsymbol{\beta}}$, which is the MLE for $\boldsymbol{\beta}$. Then we use $\ell(\phi) = \ell(\hat{\boldsymbol{\beta}}, \phi)$ to find that

$$\hat{\phi} = \arg\max_{\phi} \ell(\hat{\boldsymbol{\beta}}, \phi)$$

Inference for $\boldsymbol{\beta}$ through $\hat{\boldsymbol{\beta}}$

Next, we would like to perform additional inference, such as confidence intervals/hypothesis testing/what have you. Let's start with the final result:

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}_p(\boldsymbol{\beta}, J^{-1}(\hat{\boldsymbol{\beta}})).$$

If ϕ is present, then the covariance is $J^{-1}(\hat{\boldsymbol{\beta}}, \tilde{\phi})$, where $\tilde{\phi}$ is a consistent estimator of ϕ.One source for all of the details and theory, check out *Fahrmeir & Kaufman (1985) Annals of Statistics, pp. 342 - 368.*

Beginning with the score vector U: What is the expectation of U_j? It is zero, clearly, as $E(y_i) = \mu_i$, which will cancel out the term in the numerator of each summand, bringing $E(U) = 0$. The covariance of U is $J(\boldsymbol{\beta})$. In the univariate case, we could find the covariance of U as $E(-U') = J(\boldsymbol{\beta})$. Therefore, asymptotically,

$$U \dot{\sim} \mathrm{N}_p(\mathbf{0}, J(\boldsymbol{\beta})).$$

From this, we can also get that

$$U^T J^{-1}(\boldsymbol{\beta}) U \dot{\sim} \chi^2_p.$$

Back to the inference for $\boldsymbol{\beta}$. Look at the first-order Taylor series expansion:

$$U(\boldsymbol{\beta}) \simeq U(\hat{\boldsymbol{\beta}}) - I(\boldsymbol{\beta}, \mathbf{y})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

where I is the observed information matrix. We can (and should) replace I with J, the expected information matrix. This is actually better because then this expansion is not a function of $\mathbf{y}$. That is the practicality behind using the expected information matrix. Under the assumption that the expansion is being done on the parameter space, $U(\hat{\boldsymbol{\beta}}) = 0$ because we'd be taking the derivative with respect to the parameter, not its estimate. This gets us

$$\Rightarrow \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq J^{-1}(\boldsymbol{\beta}) U(\boldsymbol{\beta}),$$

which, after taking the expectation, gives us

$$E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \simeq 0.$$

For the covariance, we have the result

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &\simeq E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right) \\ &\simeq E\left(J^{-1}(\boldsymbol{\beta}) U(\boldsymbol{\beta}) U^T(\boldsymbol{\beta}) J^{-1}(\boldsymbol{\beta})\right) \\ &\simeq J^{-1}(\boldsymbol{\beta}) E\left[U(\boldsymbol{\beta}, \mathbf{y}) U^T(\boldsymbol{\beta}, \mathbf{y})\right] J^{-1}(\boldsymbol{\beta}) \\ &\simeq J^{-1}(\boldsymbol{\beta}) \end{aligned}$$

Try to verify this in your R output in the first homework.

Normal Linear Model (identity link function)

We have that $\theta = \mu$ and $b(\theta) = \frac{\theta^2}{2}$ and $\phi = \sigma^2$. The link function is the identity function, so

$$g(\mu_i) = \mu_i = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Therefore, the score function is found as

$$U_j = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$$

and the expected information matrix is given as

$$J_{kj} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik}.$$

If we want the covariance estimation, we can start with the typical linear regression formulation $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I_n)$. So the MLE for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}.$$

Taking the expectation with respect to $\mathbf{y}$, we have that

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T E(\mathbf{y}) = \boldsymbol{\beta},$$

and the covariance is given as

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &= E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right) \\ &= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1} = J^{-1} \end{aligned}$$

Classical Semiparametric Estimation

We're in the setting with y_i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, as we want to estimate a relationship between y_i and $\mathbf{x}_i$. The estimation approach here will rely on the first and second moments.

$$\begin{aligned} E(y_i) &= \mu_i \equiv \mu_i(\boldsymbol{\beta}) \text{ which comes in the form ofa regression equation} \\ Var(y_i) &= \phi V_i(\mu_i),\ \phi > 0 \end{aligned}$$

Note here that the variance includes the function V_i, but also the additional ϕ term. Typically, $Var(y_i) = V_i(\mu_i)$. So this is referred to as **quasi-likelihood estimation**. Now consider that the quasi-score function would appear as

$$U(\mu, \phi; y) = \frac{y - \mu}{\phi V(\mu)} \left(= \frac{y - E(y)}{Var(y)}\right)$$

Recall that $E(U) = 0$ and $Var(U) = \frac{1}{\phi V(\mu)}$, as

$$E\left(-\frac{\delta U}{\delta \mu}\right) = \frac{1}{\phi V(\mu)} = Var(U).$$

So U behaves like $\frac{\delta \ell(\mu, \phi; y)}{\delta \mu}$. The expression for the log quasi-likelihood for μ arises as

$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\phi V(t)} dt.$$

Where does this come from? If we reverse-engineer a little, we can find that

$$\frac{\delta Q(\mu; y)}{\delta \mu} = \frac{y - \mu}{\phi V(\mu)} \Rightarrow Q(\mu; y) = Q(\mu_0; y) + \int_{\mu_0}^\mu \frac{y - t}{\phi V(t)} dt,$$

where μ_0 is in the interval of values for μ. A reasonable choice for μ_0 is y.

5 October 6, 2016

Quasi-Likelihood Estimation

We're in the regression setting with responses y_i and covariates $\mathbf{x}_i$, leading to the estimation of coefficients $\boldsymbol{\beta}$. Instead of requiring a likelihood, we instead make assumptions about the moments:

- $E(y_i) = \mu_i \equiv \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)$, which in the context of GLM, $\mu_i = g^{-1}(\mathbf{x}_i^T, \boldsymbol{\beta})$
- $Var(y_i) = \phi V_i(\mu_i)$, which does not need to correspond to the variance term of a proper parametric distribution. Typically, $V_i(\mu_i) = V(\mu_i)$.

Quasi-score Function

Set a score such that

$$U_i = U(\mu_i, \phi_i; y_i) = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

Based on this formualtion, one can easily chekc that $E(U) = 0$ and that

$$Var(U) = \frac{1}{\phi V(\mu)} = E\left(-\frac{\delta U}{\delta \mu}\right) = Var(U).$$

This U behaves (as far as the first and second moments) like the partial derivative of the log-likelihood with respect to μ. That is,

$$U \sim \frac{\delta \ell(\mu, \phi; y)}{\delta \mu}.$$

Log quasi-likelihood for μ

$$Q(\mu, \phi; y) = \int_y^\mu \frac{y - t}{\phi V(t)} dt,$$

and after taking a derivative, we have that

$$\frac{\delta Q(\mu, \phi; y)}{\delta \mu} = \frac{y - \mu}{\phi V(\mu)},$$

the solution of which gives us that

$$Q = \int_y^\mu \frac{y - t}{\phi V(\mu)} dt$$

Now say we take the data $\mathbf{y} = (y_1, \ldots, y_n)$ with corresponding $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Then

$$Q(\mu, \phi; y) = \sum_{i=1}^n Q(\mu_i, \phi; y).$$

In theory, the quasi likelihood could functionally depend on the observation, but in practice, this is rarely done.

Examples (for $V(\mu)$)

1. $V(\mu) = 1$ and $y \in \mathbb{R}$. Then
$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\phi} dt = -\frac{1}{2\phi}(y - \mu)^2$$

2. Now say that $y \in \mathbb{N}$ and $V(\mu) = \mu$). Then
$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\phi t} dt = \frac{1}{\phi} \underbrace{(y \log(\mu) - \mu)}_{\text{Poisson}} + \frac{1}{\phi}(y - y \log(y)).$$

Note that this does not agree with the Poisson, or any other proper underlying distribution.

Quasi-Likelihood Estimates $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$?

Use the scoring method

$$U = (U_1(\boldsymbol{\beta}; y), \ldots, U_p(\boldsymbol{\beta}, y)),$$

where

$$U_j(\boldsymbol{\beta}; y) = \frac{\delta Q(\mu(\boldsymbol{\beta}), \phi; y}{\delta \beta_j} = \sum_{i=1}^n \frac{\delta Q(\mu_i(\boldsymbol{\beta}), \phi; y)}{\delta \beta_j} = \sum_{i=1}^n \frac{\delta Q(\mu_i, \phi_i; y_i)}{\delta \mu_i} \frac{\delta \mu_i}{\delta \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\delta \mu_i}{\delta \beta_j}.$$

Another equivalent form for the U vector is

$$U(\boldsymbol{\beta}, \phi; y) = \frac{1}{\phi} D^T V^{-1} (y - \mu),$$

where $V = diag(V(\mu_1), \ldots, V(\mu_n))$ and $D_{ij} = \frac{\delta \mu_i}{\delta \beta_j}$. This result is exactly the same as for the GLM. Now the expected information matrix is found to be

$$\begin{aligned} J(\boldsymbol{\beta}, \phi) &= -E\left(-\frac{\delta^2 Q(\mu(\boldsymbol{\beta}), \phi; y)}{\delta \beta_k \delta \beta_j}\right) \\ &= Cov(U(\beta, \phi; y)) \\ &= \left(\frac{1}{\phi} D^T V^{-1}\right) Cov(\mathbf{y}) \left(\frac{1}{\phi} V^{-1} D\right) \\ &= \frac{1}{\phi^2} D^T V^{-1} (\phi V) V^{-1} D \\ &= \frac{1}{\phi} D^T V^{-1} D \end{aligned}$$

So then, if b^m is the m-th iteration for $\boldsymbol{\beta}$, then

$$b^{m+1} = b^m + \left(D^T V^{-1} D\right)^{-1} D^T V^{-1} (y - \mu).$$

The point here is that this algorithm does not depend on ϕ. Another important concept is that this algorithm is identical to that corresponding to the standard GLM estimates. However, differences will arise when doing something like creating confidence intervals for $\boldsymbol{\beta}$.

CIs for $\boldsymbol{\beta}$

We find that

$$\tilde{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, J^{-1}(\tilde{\boldsymbol{\beta}}, \tilde{\phi})\right),$$

where $\tilde{\phi}$ must be a consistent estimator for ϕ. So what we'll find is that

$$Var(\tilde{\boldsymbol{\beta}}) = \tilde{\phi} Var(\hat{\boldsymbol{\beta}})$$

Estimate for ϕ under the QL Approach

$$\tilde{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \tilde{\mu}_i)^2}{V(\tilde{\mu}_i)}$$

Model Assessment/Checking/Comparisons

We will discuss

- Deviances
- Residuals

We're back in the GLM setting. Remember, we have some choice to make that may govern how well we fit our data:

- ED family distribution
- Link function (g)
- Linear predictor

How can we compare models in which we made different choices? A good part of our discussion will center around choices of linear predictors. To set the stage, we have our responses y_i and covariates x_i, and the fitted values $\hat{\mu}_i$ based on the particular GLM. Recall that $\hat{\mu}_i = g^{-1}\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)$. Something we might use is the deviance, which corresponds to the likelihood ratio test statistic. We compare the full (saturated) model with m parameters to our specific model with $p < m$ parameters. The full model basically has the most parameters given the dataset. We refer to the parameter vector in this scenario as $\hat{\boldsymbol{\beta}}_{max}$. We find the likelihood $L(\hat{\boldsymbol{\beta}}; y)$ and/or the log-likelihood $\ell(\hat{\boldsymbol{\beta}}; y)$. The idea is that we want to compare $L(\hat{\boldsymbol{\beta}}; y)$ and $L(\hat{\boldsymbol{\beta}}_{max}; y)$ with the same ED family distribution and link function. Note that

$$L(\hat{\boldsymbol{\beta}}; y) \leq L(\hat{\boldsymbol{\beta}}_{max}; y).$$

The LRT statistic is

$$\lambda = \frac{L(\hat{\boldsymbol{\beta}}; y)}{L(\hat{\boldsymbol{\beta}}_{max}; y)}.$$

Looking at this on the log scale gets us

$$\log\lambda = \ell(\hat{\boldsymbol{\beta}}; y) - \ell(\hat{\boldsymbol{\beta}}_{max}; y).$$

The **scaled deviance** is found as

$$D^* = -2\log(\lambda) = -2\left[\ell(\hat{\boldsymbol{\beta}}; y) - \ell(\hat{\boldsymbol{\beta}}_{max}; y)\right],$$

which is a function of $\hat{\boldsymbol{\beta}}$, ϕ, and the data.

In the context of GLMs, we find $\hat{\boldsymbol{\beta}}$ through $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$. The natural parameter in this case is $\hat{\theta}_i = (b')^{-1}(\hat{\mu}_i)$. In the context of the full model, we'd use $\tilde{\boldsymbol{\beta}}$ and $\tilde{(}\mu_i)$. Remember that under the full model, $\tilde{\mu}_i = y_i$. So then the scaled deviance is

$$\begin{aligned} D^* = -2\left[\ell(\hat{\theta}, \phi; y) - \ell(\tilde{\theta}, \phi; y)\right] &= -2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)} \\ &= \frac{2}{\phi}\sum_{i=1}^{n} w_i\left\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right\} \\ &= \frac{1}{\phi}D(\hat{\mu}; y) \end{aligned}$$

which has a term with nothing to do with ϕ.