GLM (Generalized Linear Model) #1 (version 9)

Paul Johnson

March 10, 2006

This handout describes the basics of estimating the Generalized Linear Model: the exponential distribution, familiar examples, the maximum likelihood estimation process, and iterative re-weighted least squares. The first version of this handout was prepared as lecture notes on Jeff Gill's handy book, *Generalized Linear Models: A Unified Approach* (Sage, 2001). In 2004 and 2006, I revised in light of

- Myers, Montgomery and Vining, Generalized Linear Models with *Applications in Engineering and the Sciences* (Wiley, 2002),
- Annette J. Dobson, *An Introduction to Generalized Linear Models, 2ed*, (Chapman and Hall, 2002)
- Wm. Venables and Brian Ripley, *Modern Applied Statistics with S, 4ed (2004)*
- P. McCullagh and J.A. Nelder, *Generalized Linear Models, 2ed* (Chapman & Hall, 1989).
- D. Firth, "Generalized Linear Models," in D.V. Hinkley, N. Reid and E.J. Snell, Eds, *Statistical Theory and Modeling* (Chapman and Hall, 1991).

This version has some things that remain only to explain (to myself and others) the approach in Gill, but most topics are now presented with a more general strategy.

Contents

1 Motivating idea. 3
1.1 Consider anecdotes . . . 3
1.1.1 linear model . . . 3
1.1.2 logit/probit model . . . 3
1.1.3 Poisson . . . 4
1.2 The exponential family unites these examples . . . 4
1.3 Simplest Exponential Family: one parameter . . . 4
1.4 More General Exponential Family: introducing a dispersion parameter . . . 5

2 GLM: the linkage between θ_i and $X_i b$ 5
2.1 Terms . . . 6
2.2 Think of a sequence . . . 6

3 About the Canonical link **6**
3.1 The canonical link is chosen so that $\theta_i = \eta_i$. 6
3.2 Declare $\theta_i = q(\mu_i)$. 7
3.3 How can you find $q(\mu_i)$? 7
3.4 Maybe you think I'm presumptuous to name that thing $q(\mu_i)$. 8
3.5 Canonical means convenient, not mandatory 8

4 Examples of Exponential distributions (and associated canonical links) **8**
4.1 Normal 8
4.1.1 Massage $N(\mu_i, \sigma_i^2)$ into canonical form of the exponential family 9
4.1.2 The Canonical Link function is the identity link 9
4.1.3 We reproduced the same old OLS, WLS, and GLS 10
4.2 Poisson. 10
4.2.1 Massage the Poisson into canonical form 11
4.2.2 The canonical link is the log 11
4.2.3 Before you knew about the GLM 11
4.3 Binomial 12
4.3.1 Massage the Binomial into the canonical exponential form 12
4.3.2 Suppose $n_i = 1$ 13
4.3.3 The canonical link is the logit 13
4.3.4 Before you knew about the GLM, you already knew this 13
4.4 Gamma 13
4.5 Other distributions (may fill these in later). 14

5 Maximum likelihood. How do we estimate these beasts? **14**
5.1 Likelihood basics 14
5.2 Bring the "regression coefficients" back in: θ_i depends on b 15

6 Background information necessary to simplify the score equations **15**
6.1 Review of ML handout #2 15
6.2 GLM Facts 1 and 2 16
6.2.1 GLM Fact #1: 16
6.2.2 GLM Fact #2: The Variance function appears! 17
6.2.3 Variance functions for various distributions 18

7 Score equations for the model with a noncanonical link. **18**
7.1 How to simplify expression 72. 18
7.2 The Big Result in matrix form 20

8 Score equations for the Canonical Link **20**
8.1 Proof strategy 1 21
8.2 Proof Strategy 2. 22
8.3 The Matrix form of the Score Equation for the Canonical Link 22

9 ML Estimation: Newton-based algorithms. 23
9.1 Recall Newton's method . . . 23
9.2 Newton-Raphson method with several parameters . . . 24
9.3 Fisher Scoring . . . 25
9.3.1 General blithering about Fisher Scoring and ML . . . 27

10 Alternative description of the iterative approach: IWLS 27
10.1 Start by remembering OLS and GLS . . . 28
10.2 Think of GLM as a least squares problem . . . 28
10.3 The IWLS algorithm. . . . 29
10.4 Iterate until convergence . . . 30
10.5 The Variance of $\hat{b}$. . . . 30
10.6 The weights in step 3. . . . 30
10.7 Big insight from the First Order Condition . . . 30

1 Motivating idea.

1.1 Consider anecdotes

1.1.1 linear model

We already know

$$y_i = X_i b + e_i \tag{1}$$

where X_i is a row vector of observations and b is a column vector of coefficients. The error term is assumed Normal with a mean of 0 and variance σ^2.

Now think of it a different way. Suppose that instead of predicting y_i, we predict the mean of y_i. So put the "linear predictor" $X_i b$ into the Normal distribution where you ordinarily expect to find the mean:

$$y_i \sim N(X_i b, \sigma^2) \tag{2}$$

In the GLM literature, they often tire of writing $X_i b$ or such, and they start using (for "shorthand") the Greek letter "eta."

$$\eta_i = X_i b \tag{3}$$

1.1.2 logit/probit model

We already studied logit and probit models. We know that we can think of the $X_i b$ part, the systematic part, as the input into a probability process that determines whether the observation is 1 or 0. We might write

$$y_i \sim B(\Phi(X_i b)) \tag{4}$$

where Φ stands for a cumulative distribution of some probability model, such as a cumulative Normal or cumulative Logistic model, and B is some stupid notation I just threw in to denote a

"binary" process, such as a Bernoulli distribution (or a Binomial distribution). Use the Greek letter π_i to represent the probability that y_i equals 1, $\pi_i = \Phi(X_i b)$. For the Logistic model,

$$\pi_i = \frac{1}{1+exp(-X_i b)} = \frac{exp(X_i b)}{1+e(X_i b)} \tag{5}$$

which is the same thing as:

$$ln\left[\frac{\pi_i}{1-\pi_i}\right] = X_i b \tag{6}$$

1.1.3 Poisson

We also now know about the Poisson model, where we can calculate the impact of $X_i b$ on the probability of a certain outcome according to the Poisson model. Check back on my Poisson notes if you can't remember, but it was basically the same thing, where we said we are taking the impact of the input to be $exp(X_i b)$, or $e^{X_i b}$, and then we assert:

$$y_i \sim Poisson(e^{X_i b}) \tag{7}$$

1.2 The exponential family unites these examples

In these examples, there is a general pattern. We want to predict something about y, and we use $X_i b$ to predict it. We estimate the b's and try to interpret them. How far can a person stretch this metaphor?

According to Firth (1991), Nelder and Wedderburn (1972) were the first to identify a general scheme, the Generalized Linear Model (GLM). The GLM ties together a bunch of modeling anecdotes.

1.3 Simplest Exponential Family: one parameter

The big idea of the GLM is that $y_i$is drawn from one of the many members of the "exponential family of distributions." In that family, one finds the Normal, Poisson, Negative Binomial, exponential, extreme value distribution, and many others.

Depending on which book you read, this is described either in a simple, medium, or really general form. The simple form is this (as in Dobson, 2000, p. 44). Note the property that the effects of y and θ are very SEPARABLE:

$$prob(y_i|\theta_i) = f(y_i|\theta_i) = s(\theta_i) \cdot t(y_i) \cdot exp\{a(y_i) \cdot b(\theta_i)\} \tag{8}$$

Dobson says that if $a(y) = y$, then we have the so-called **canonical form** (canonical means "standard"):

$$prob(y_i|\theta_i) = f(y_i|\theta_i) = s(\theta_i) \cdot t(y_i) \cdot exp\{y_i \cdot b(\theta_i)\} \tag{9}$$

Note we've got y_i (the thing we observe), and an unknown parameter for each observation, θ_i.

$b(\theta)$ in Dobson, the function $b(\theta)$ is called the "natural parameter".

θ In other treatments, they assume $b(\theta) = \theta$, so θ itself is the "natural parameter" (see comments below).

$s(\theta)$ a function that depends on θ but NOT on y

$t(y)$ a function that depends on y but NOT on θ.

This is the simplest representation.

Following Dobson, rearrange like so:

$$f(y_i|\theta_i) = exp\{y \cdot b(\theta_i) + ln[s(\theta_i)] + ln[t(y_i)]\} \tag{10}$$

And re-label again:

$$f(y_i|\theta_i) = exp\{y_i \cdot b(\theta_i) - c(\theta_i) + d(y_i)\} \tag{11}$$

Most treatments assume that $b(\theta_i) = \theta_i$. So one finds

$$f(y_i|\theta_i) = exp\{y \cdot \theta_i - c(\theta_i) + d(y_i) \tag{12}$$

Many books relabel of these functions $b(\theta)$, $c(\theta)$, and $d(y_i)$. It is a serious source of confusion.

1.4 More General Exponential Family: introducing a dispersion parameter

We want to include a parameter similar to the "standard deviation," an indicator of "dispersion" or "scale". Consider this reformulation of the exponential family (as in Firth (1991)):

$$f(y_i) = exp\left\{\frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i)\right\} \tag{13}$$

The new element here is ϕ_i (the Greek letter "phi", pronounced "fee"). That is a "dispersion parameter." If $\phi_i = 1$, then this formula reduces to the canonical form seen earlier in 9.

The critical thing, once again, is that the natural parameter θ_i and the observation y_i appear in a particular way.

2 GLM: the linkage between θ_i and $X_i b$

In all these models, one writes that the observed y_i depends on a probability process, and the probability process takes an argument θ_i which (through a chain of reasoning) reflects the input $X_i b$.

2.1 Terms

In order to qualify as a generalized linear model, the probability process has to have a certain kind of formula.

Most sources on the GLM use this set of symbols:

$\eta_i = X_i b$ η_i is the "linear predictor"

$\mu_i = E(y_i)$ the expected value (or mean) of y_i.

$\eta_i = g(\mu_i)$ the link function, $g()$, translates between the linear predictor and the mean of y_i. It is assumed to be monotonically increasing in μ_i.

$g^{-1}(\eta_i) = \mu_i$ the inverse link function, $g^{-1}()$, gives the mean as a function of the linear predictor

The only difficult problem is to connect these ideas to the observed y_i.

2.2 Think of a sequence

- η_i depends on X_i and b (think of that as a function also called $\eta_i(X_i, b)$)
- μ_i depends on η_i (via the inverse of the link function, $\partial\mu_i/\partial\eta_i = g^{-1}(\eta_i)$
- θ_i depends on μ_i (via some linkage that is distribution specific; but below I will call it $q()$)
- $f(y_i|\theta_i)$ depends on θ_i

In this sequence, the only "free choice" for the modeler is $g()$. The others are specified by the assumption of linearity in $X_i b$ and the mathematical formula of the probability distribution.

3 About the Canonical link

Recall that the link translates between μ_i (the mean of y_i) and the linear predictor.

$$\eta_i = g(\mu_i) \tag{14}$$

3.1 The canonical link is chosen so that $\theta_i = \eta_i$.

Use of the canonical link simplifies the maximum likelihood analysis (by which we find estimates of the b's).

If you consider the idea of the sequence that translates $X_i b$ into θ_i as described in section 2.2, notice that a very significant simplifying benefit is obtained by the canonical link.

But the simplification is costly. If we are to end up with $\theta_i = \eta_i$, that means there is a major restriction on the selection of the link function, because whatever happens in $g^{-1}(\eta_i)$ to translate η_i into μ_i has to be exactly undone by the translation from μ_i into θ_i.

McCullagh and Nelder state that the canonical link should not be used if it contradicts the substantive ideas that motivate a research project. In their experience, however, the canonical link is often substantively acceptable.

3.2 Declare $\theta_i = q(\mu_i)$.

I will invent a piece of terminology now! Henceforth, this is Johnson's q. (Only kidding. You can just call it q.) Consider $\theta_i = q(\mu_i)$. (I chose the letter q because, as far as I know, it is not used for anything important.)

The function $q(\mu_i)$ is not something we choose in the modeling process. Rather, it is a feature dictated by the probability distribution.

If we had the canonical link, $q(\mu_i)$ would have the exact opposite effect of $g^{-1}(\eta_i)$. That is to say, the link $g()$ must be chosen so that:

$$\eta_i = q(g^{-1}(\eta_i)) \tag{15}$$

One can reach that conclusion by taking the following steps.

- By definition, $\mu_i = g^{-1}(\eta_i)$, which is the same as saying $\eta_i = g(\mu_i)$.
- The GLM framework assumes that a mapping exists between θ_i and μ_i. I'm just calling it by a name, q: $\theta_i = q(\mu_i)$
- If you connect the dots, you find:

$$\eta_i = g(\mu_i) = \theta_i = q(\mu_i) \tag{16}$$

  In other words, the function $g()$ is the same as the function $q()$ if the link is canonical. So if you find $q()$ you have found the canonical link.

3.3 How can you find $q(\mu_i)$?

For convenience of the reader, here's the canonical family definition:

$$f(y_i) = exp\left\{\frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i,\phi_i)\right\} \tag{17}$$

A result detailed below in section 6.2 is that the average of y_i is equal to the derivative of $c()$. This is the finding I label "GLM Fact #1":

$$\mu_i = \frac{dc(\theta_i)}{d\theta_i} \tag{18}$$

That means that the "mysterious function" $q(\mu_i)$ is the value you find by calculating $dc/d\theta_i$ and then re-arranging so that θ_i is a function of μ_i.

Perhaps notation gets in the way here. If you think of $dc/d\theta_i$ as just an ordinary function, using notation like $\mu_i = c'(\theta_i)$, then it is (perhaps) easier to think of the inverse (or opposite, denoted by $^{-1}$) transformation, applied to μ_i:

$$\theta_i \quad = \quad [c']^{-1}(\mu_i) \tag{19}$$

The conclusion, then, is that $q(\mu_i) = [c']^{-1}(\mu_i)$ and the canonical link are the same.

Just to re-state the obvious, if you have the canonical family representation of a distribution, the canonical link can be found by calculating $c'(\theta_i)$ and then solving for θ_i.

3.4 Maybe you think I'm presumptuous to name that thing $q(\mu_i)$.

There is probably some reason why the textbooks don't see the need to define $\theta_i = q(\mu_i)$ and find the canonical link through my approach. Perhaps it makes sense only to me (not unlikely).

Perhaps you prefer the approach mentioned by Firth (p.62). Start with expression 18. Apply $g()$ to both sides:

$$g(\mu_i) = g\left[\frac{dc(\theta_i)}{d\theta_i}\right] \tag{20}$$

And from the definition of the canonical link function,

$$\theta_i = \eta_i = g(\mu_i) = g\left[\frac{dc(\theta_i)}{d\theta_i}\right] \tag{21}$$

As Firth observes, for a canonical link, one must choose $g()$ so that it is the inverse of $\frac{dc(\theta_i)}{d\theta_i}$. That is, whatever $dc(\theta_i)/d\theta_i$ does to the input variable θ_i, the link $g()$ has to exactly reverse it, because we get θ_i as the end result. If g "undoes" the derivative, then it will be true that

$$\eta_i = g(\mu_i) = \theta_i \tag{22}$$

That's the canonical link, of course. Eta equals theta.

3.5 Canonical means convenient, not mandatory

In many books, one finds that the authors are emphatic about the need to use the canonical link. Myers, Montgomery, and Vining are a agnostic on the matter, acknowledging the simplifying advantage of the canonical link but also working out the maximum likelihood results for non-canonical links (p. 165).

4 Examples of Exponential distributions (and associated canonical links)

We want to end up with a dependent variable that has a distribution that fits within the framework of 13.

4.1 Normal

The Normal is impossibly easy! The canonical link

$$g(\mu_i) = \mu_i$$

Its especially easy because the Normal distribution has a parameter named μ_i which happens also to equal the expected value. So, if you have the formula for the Normal, there's no need to do any calculation to find its mean. It is equal to the parameter μ_i.

4.1.1 Massage $N(\mu_i, \sigma_i^2)$ into canonical form of the exponential family

Recall the Normal is

$$f(y_i;\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(y_i-\mu_i)^2} \tag{23}$$

$$= \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma^2}(y_i^2+\mu_i^2-2y_i\mu_i)} \tag{24}$$

$$= \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\left[\frac{y_i\mu_i}{\sigma^2}-\frac{\mu_i^2}{2\sigma^2}-\frac{y_i^2}{2\sigma^2}\right]} \tag{25}$$

Next recall that

$$ln\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\right] = ln\left[(2\pi\sigma_i^2)^{-1/2}\right] \tag{26}$$

$$= -\frac{1}{2} ln(2\pi\sigma_i^2)$$

In addition, recall that

$$A \cdot e^x = e^{x+ln(A)} \tag{27}$$

So that means 25 can be reduced to:

$$exp\left[\frac{y_i \cdot \mu_i}{\sigma_i^2} - \frac{\mu_i^2}{2\sigma_i^2} - \frac{y_i^2}{2\sigma_i^2} - \frac{1}{2} ln(2\pi\sigma_i^2)\right] \tag{28}$$

That fits easily into the general form of the exponential family in 13. The dispersion parameter $\phi_i = \sigma_i^2$.

4.1.2 The Canonical Link function is the identity link

Match up expression 28 against the canonical distribution 13. Since the numerator of the first term in 28 is $y_i \cdot \mu_i$, that means the Normal mean parameter μ_i is playing the role of θ_i. As a result,

$$\theta_i = \mu_i \tag{29}$$

The function $q(\mu_i)$ that I introduced above is trivially simple, $\theta_i = \mu_i = q(\mu_i)$. Whatever μ_i you put in, you get the same thing back out.

If you believed the result in section 3.2 , then you know that $g()$ is the same as $q()$. So the canonical link is the **identity link**, $g(\mu) = \mu$.

You arrive at the same conclusion through the route suggested by equation 21. Note that the exponential family function $c(\theta)$ corresponds to

$$c(\theta_i) = \frac{1}{2}\mu_i^2 \tag{30}$$

and

$$\frac{dc(\theta_i)}{d\theta_i} = c'(\theta_i) = \mu_i \tag{31}$$

You "put in θ_i" and you "get back" μ_i. The reverse must also be true. If you have the inverse of $c'(\theta_i)$, which was called $[c']^{-1}$ before, you know that

$$[c']^{-1}(\mu_i) = \theta_i \tag{32}$$

As Firth observed, the canonical link is such that $g(\mu_i) = [c']^{-1}(\theta_i)$. Since we already know from $\theta_i = \mu_i$, this means that $[c']^{-1}$ must be the identity function, because

$$\mu_i = \theta_i = [c']^{-1}(\theta_i) \tag{33}$$

4.1.3 We reproduced the same old OLS, WLS, and GLS

We can think of the linear predictor as fitting directly into the Normal mean. That is to say, the "natural parameter" θ is equal to the mean parameter μ, and the mean equals the linear predictor:

$$\theta_i = \mu_i = \eta_i = X_i b$$

In other words, the model we are fitting is $y \sim N(\eta_i, \sigma_i^2)$ or $y \sim N(X_i b, \sigma_i^2)$.
This is the Ordinary Least Squares model if you assume that

$$i.\ \sigma_i^2 = \sigma^2\ for all\ i$$

and

$$ii.\ Cov(y_i, y_j) = 0\ for all\ i \neq j$$

If you keep assumption *ii* but drop *i*, then this implies Weighted Least Squares.
If you drop both *i* and *ii*, then you have Generalized Least Squares.

4.2 Poisson.

Consider a Poisson dependent variable. It is tradition to refer to the single important parameter in the Poisson as λ, but since we know that λ is equal to the mean of the variable, let's make our lives simple by using μ for the parameter. For an individual *i*,

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{34}$$

The parameter μ_i is known to equal the expected value of the Poisson distribution. The variance is equal to μ_i as well. As such, it is a "one-parameter distribution."

If you were following along blindly with the Normal example, you would (wrongly) assume that you can use the identity link and just put in $\eta_i = X_i b$ in place of μ_i. Strangely enough, if you do that, you would not have a distribution that fits into the canonical exponential family.

4.2.1 Massage the Poisson into canonical form

Rewrite the Poisson as:

$$f(y_i|\mu_i) = e^{(-\mu_i + ln(\mu^{y_i}) - ln(y_i!))}$$

which is the same as

$$f(y_i|\mu) = exp[y_i ln(\mu_i) - \mu_i - ln(y_i!)] \tag{35}$$

There is no scale factor (dispersion parameter), so we don't have to compare against the general form of the exponential family, equation 13. Rather, we can line this up against the simpler form in equation 12

4.2.2 The canonical link is the log

In 35 we have the first term $y_i ln(\mu_i)$, which is not exactly parallel to the exponential family, which requires $y\theta$. In order to make this match up against the canonical exponential form, note:

$$\theta_i = ln(\mu_i) \tag{36}$$

So the mysterious function $q()$ is $ln(\mu_i)$. So the canonical link is

$$g(\mu_i) = ln(\mu_i) \tag{37}$$

You get the same answer if you follow the other route, as suggested by 21. Note that in 35, $c(\theta) = exp(\theta)$ and since

$$\mu_i = \frac{dc(\theta_i)}{d\theta_i} = exp(\theta_i) \tag{38}$$

The inverse function for $exp(\theta_i)$ is $ln()$,so $[c']^{-1} = ln$, and as a result

$$g(\mu_i) = [c']^{-1} = ln(\mu_i) \tag{39}$$

The Poisson is a "one parameter" distribution, so we don't have to worry about dispersion. And the standard approach to the modeling of Poisson data is to use the log link.

4.2.3 Before you knew about the GLM

In the standard "count data" (Poisson) regression model, as popularized in political science by Gary King in the late 1980s, we assume that the mean of y_i follows this Poisson distribution with an expected value of $exp[X_i b]$. That is to say, we assume

$$\mu_i = exp[X_i b]$$

Put $exp[X_i b]$ in place of μ_i in expression 40, and you have the Poisson count model.

$$f(y|X,b) = \frac{e^{-exp[X_i b]}[X_i b]^y}{y!} \tag{40}$$

When you view this model in isolation, apart from GLM, the justification for using $exp[X_i b]$ instead of $X_i b$ is usually given by "hand waving" arguments, contending that "we need to be sure the average is 0 or greater, so *exp* is a suitable transformation" or "we really do think the impact of the input variables is multiplicative and exponential."

The canonical link from the GLM provides another justification that does not require quite so much hand waving.

4.3 Binomial

A binomial model is used to predict the number of successes that we should expect out of a number of tries. The probability of success is fixed at p_i and the number of tries is n_i, and the probability of y_i successes given n_i tries, is

$$f(y_i|n_i,p_i) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) p_i^{y_i}(1-p_i)^{n_i-y_i} = \frac{n_i!}{y_i!(n_i-y_i)!} p_i^{y_i}(1-p_i)^{n_i-y_i} \tag{41}$$

In this case, the role of the mean, μ_i, is played by the letter p_i. We do that just for variety :)

4.3.1 Massage the Binomial into the canonical exponential form

I find it a bit surprising that this fits into the same exponential family, actually. But it does fit because you can do this trick. It is always true that

$$x = e^{ln(x)} = exp[ln(x)] \tag{42}$$

Ordinarily, we write the binomial model for the probability of y successes out of n trials

$$\left(\begin{array}{c} n \\ y \end{array} \right) p^y(1-p)^{n-y} = \frac{n!}{y!(n-y)!} p^y(1-p)^{n-y} \tag{43}$$

But that can be massaged into the exponential family

$$e^{ln\left(\left(\begin{array}{c} n \\ y \end{array} \right) p^y(1-p)^{n-y} \right)} = e^{ln\left(\begin{array}{c} n \\ y \end{array} \right) + ln(p^y) + ln((1-p)^{n-y})} \tag{44}$$

$$= e^{ln\left(\begin{array}{c} n \\ y \end{array} \right) + yln(p) + (n-y)ln(1-p)} = e^{ln\left(\begin{array}{c} n \\ y \end{array} \right) + yln(p) + nln(1-p) - yln(1-p)} \tag{45}$$

Note that because $ln(a) - ln(b) = ln(\frac{a}{b})$, we have:

$$e^{yln(\frac{p}{1-p}) + nln(1-p) + ln\left(\begin{array}{c} n \\ y \end{array} \right)}$$

this clearly fits within the simplest form of the exponential distribution.

$$exp[y_i ln(\frac{p_i}{1-p_i}) - (-n_i \cdot ln(1-p_i)) + ln\left(\begin{array}{c} n_i \\ y_i \end{array} \right)] \tag{46}$$

4.3.2 Suppose $n_i = 1$

The probability of success on a single trial is p_i. Hence, the expected value with one trial is p_i. The expected value of y_i when there are n_i trials is $\mu_i = n_i \cdot p_i$. The parameter n_i is a fixed constant for the $i'th$ observation.

When we do a logistic regression, we typically think of each observation as one test case. A survey respondent says "yes" or "no" and we code $y_i = 1$ or $y_i = 0$. If just one test of the random process is run per observation. That means y_i is either 0 or 1.

So, in the following, lets keep it simple by fixing $n_i = 1$. And, in that context,

$$\mu_i = p_i \tag{47}$$

4.3.3 The canonical link is the logit

Then eyeball equation 46 against the canonical form 12, you see that if you think the probability of a success is p_i, then you have to transform p_i so that:

$$\theta = ln(\frac{p_i}{1-p_i}) \tag{48}$$

Since $p_i = \mu_i$, then the mysterious function $q()$ is $ln(\frac{\mu_i}{1-\mu_i})$. That is the "canonical link" for the Binomial distribution. It is called the **logit link** and it maps from the mean of observations, $\mu_i = p_i$ back to the natural parameter θ_i.

4.3.4 Before you knew about the GLM, you already knew this

That's just the plain old "logistic regression" equation for a $(0,1)$ dependent variable. The logistic regression is usually introduced either as

$$P(y_i = 1|X_i, b) = \frac{1}{1+exp(-X_i b)} = \frac{exp(X_i b)}{1+exp(X_i b)} \tag{49}$$

Or, equivalently,

$$ln\left[\frac{p_i}{1-p_i}\right] = X_i b \tag{50}$$

It is easy to see that, if the canonical link is used, $\theta_i = \eta_i$.

4.4 Gamma

Gamma is a random variable that is continuous on $(0, +\infty)$. The probability of observing a score y_i depends on 2 parameters, the "shape" and the "scale". Call the "shape" α (same for all observations, so no subscript is used) and the "scale" β_i. The probability distribution for the Gamma is

$$y_i \sim f(\alpha, \beta_i) = \frac{1}{\beta_i^{\alpha}\Gamma(\alpha)} y_i^{(\alpha-1)} e^{-y_i/\beta_i} \tag{51}$$

The mean of this distribution is $\alpha\beta_i$. We use can input variables and parameters to predict a value of β_i, which in turn influences the distribution of observed outcomes.

The canonical link is $1/\mu_i$. I have a separate writeup about GammaGLM models.

4.5 Other distributions (may fill these in later).

Pareto $f(y|\theta) = \theta y^{-\theta-1}$

Exponential $f(y|\theta) = \theta e^{-y\theta}$ or $f(y|\theta) = \frac{1}{\theta}e^{-y/\theta}$ (depending on your taste)

Negative Binomial

$$f(y|\theta) = \begin{pmatrix} y+r-1 \\ r-1 \end{pmatrix} \theta^r (1-\theta)^y$$

r is a known value.

Extreme Value (Gumbel)

$$f(y|\theta) = \frac{1}{\theta} exp\left\{ \frac{(y-\theta)}{\phi} - exp\left[\frac{(y-\theta)}{\phi} \right] \right\}$$

5 Maximum likelihood. How do we estimate these beasts?

The parameter fitting process should maximize the probability that the model fits the data.

5.1 Likelihood basics

According to classical maximum likelihood theory, one should calculate the probability of observing the whole set of data, for observations i=1,...,N, given a particular parameter:

$$L(\theta|X_1,...,X_N) = f(X_1|\theta) \cdot f(X_2|\theta) \cdot ... \cdot f(X_N|\theta)$$

The log of the likelihood function is easier to work with, and maximizing that is logically equivalent to maximizing the likelihood.

Remember $ln(exp(x)) = x$?

With the more general form of exponential distribution, equation 13, the log likelihood for a single observation equals:

$$l_i(\theta_i, \phi_i|y_i) = log\left\{ exp\left[\frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i) \right] \right\}$$

$$= \frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i) \tag{52}$$

The dispersion parameter ϕ_i is often treated as a **nuisance parameter**. It is estimated separately from the other parameters, the ones that affect θ_i.

5.2 Bring the "regression coefficients" back in: θ_i depends on b

Recall the comments in section 2.2 which indicated that we can view the problem at hand as a sequence of transformations. The log likelihood contribution of a particular case lnL_i=l_i depends on θ_i, and θ_i depends on μ_i, and μ_i depends on η_i, and $\eta_i$depends on b. So you'd have to think of some gnarly thing like

$$l_i(b) = f(\theta_i(\mu_i(\eta_i(b))))$$

The "chain rule" for derivatives applies to a case where you have a "function of a function of a function...". In this case it implies, as noted in McCullagh and Nelder (p. 41), that the score equation for a coefficient b_k as

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial b_k} \tag{53}$$

The letter U is a customary (very widely used) label for the score equations. It is necessary to simultaneously solve p of these **score equations** (one for each of the p parameters being estimated):

$$\begin{aligned} \frac{\partial l}{\partial b_1} &= U_1 &&= 0 \\ \frac{\partial l}{\partial b_2} &= U_2 &&= 0 \\ &\vdots &&= 0 \\ \frac{\partial l}{\partial b_p} &= U_p &&= 0 \end{aligned}$$

That assumptions linking b to θ reduces the number of separate coefficients to be estimated quite dramatically. We started out with $\theta = (\theta_1, \theta_2, ..., \theta_N)$ natural location parameters, one for each case, but now there are just p coefficients in the vector b.

6 Background information necessary to simplify the score equations

6.1 Review of ML handout #2

When you maximize something by optimally choosing an estimate $\hat{\theta}$, take the first derivative with respect to θ and set that result equal to zero. Then solve for for θ. The **score function** is the first derivative of the log likelihood with respect to the parameter of interest.

If you find the values which maximize the likelihood, it means the score function has to equal zero, and there is some interesting insight in that (Gill, p.22).

From the general theory of maximum likelihood, as explained in the handout ML#2, it is true that:

$$MLFact\,\#1:\ E\left[\frac{\partial l_i}{\partial \theta_i}\right] = 0 \tag{54}$$

and

$$MLFact\,\#2:\ E\left[\frac{\partial^2 l_i}{\partial \theta_i^2}\right] = -Var\left[\frac{\partial l_i}{\partial \theta_i}\right] \tag{55}$$

6.2 GLM Facts 1 and 2

GLM Facts 1 & 2 are–almost invariably–stated as simple facts in textbooks, but often they are not explained in detail (see Gill, p. 23, McCullagh and Nelder 1989 p. 28; Firth 1991, p. 61 and Dobson 2000, p. 47-48).

6.2.1 GLM Fact #1:

For the exponential family as described in 13, the following is true.

$$GLM\,Fact\,\#1:\ E(y_i) = \mu_i = \frac{dc(\theta_i)}{d\theta_i} \tag{56}$$

In words, "the expected value of y_i is equal to the derivative of $c(\theta_i)$ with respect to θ_i." In Firth, the notation for $dc(\theta_i)/d\theta_i$ is $\dot{c}(\theta_i)$. In Dobson, it is $c'(\theta_i)$.

The importance of this result is that if we are given an exponential distribution, we can calculate its mean by differentiating the $c()$ function.

Proof.

Take the first derivative of l_i(equation 52) with respect to $\theta_i$and you find:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{1}{\phi_i}\left[y_i - \frac{dc(\theta_i)}{d\theta_i}\right] \tag{57}$$

This is the "score" of the i'th case. Apply the expected value operator to both sides, and taking into account ML Fact #1 stated above (54), the whole thing is equal to 0:

$$E\left[\frac{\partial l_i}{\partial \theta_i}\right] = \frac{1}{\phi_i}\left[E[y_i] - E\left[\frac{dc(\theta_i)}{d\theta_i}\right]\right] = 0 \tag{58}$$

which implies

$$E[y_i] = E\left[\frac{dc(\theta_i)}{d\theta_i}\right] \tag{59}$$

Two simplifications immediately follow. First, by definition, $E[y_i] = \mu$. Second, because $c(\theta_i)$ does not depend on y_i, then $dc/d\theta_i$ is the same for all values of y_i, and since the expected value of a constant is just the constant (recall $E(A) = A$), then

$$E\left[\frac{dc(\theta_i)}{d\theta_i}\right] = \frac{dc(\theta_i)}{d\theta_i} \tag{60}$$

As a result, expression 59 reduces to

$$\mu_i = \frac{dc(\theta_i)}{d\theta_i} \tag{61}$$

And the proof is finished!

6.2.2 GLM Fact #2: The Variance function appears!

The second ML fact (55), leads to a result about the variance of y as it depends on the mean. The variance of y is

$$GLM\,Fact\#2:\ Var(y_i) = \phi_i \cdot \frac{d^2c(\theta_i)}{d\theta_i^2} = \phi_i V(\mu_i) \tag{62}$$

The variance of y_i breaks into two parts, one of which is the dispersion parameter ϕ_i and the other is the so-called "variance function" $V(\mu_i)$. This can be found in McCullagh & Nelder (1989, p. 29; Gill, p. 29; or Firth p. 61).

Note $V(\mu_i)$ this does not mean the variance of μ_i. It means the "variance as it depends on μ_i". The variance of observed y_i combines the 2 sorts of variability, one of which is sensitive to the mean, one of which is not. which represents the impact of, the second derivative of $c(\theta_i)$, sometimes labeled $\ddot{c}(\theta_i)$ or $c''(\theta_i)$, times the scaling element ϕ_i.

Proof.

Apply the variance operator to 57.

$$Var\left[\frac{\partial l}{\partial \theta_i}\right] = \frac{1}{\phi_i^2} Var\left[y - \frac{dc}{d\theta_i}\right] \tag{63}$$

Since (as explained in the previous section) $dc/d\theta_i$ is a constant, then this reduces to

$$Var\left[\frac{\partial l}{\partial \theta_i}\right] = \frac{1}{\phi_i^2} Var\left[y_i\right] \tag{64}$$

Put that result aside for a moment. Take the partial derivative of 57 with respect to $\theta_i$and the result is

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = \frac{1}{\phi_i}\left[\frac{\partial y_i}{\partial \theta_i} - \frac{d^2c(\theta_i)}{d\theta_i^2}\right] \tag{65}$$

I wrote this out term-by-term in order to draw your attention to the fact that $\frac{\partial y_i}{\partial \theta_i} = 0$ because y_i is treated as a constant this point. So the expression reduces to:

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = -\frac{1}{\phi_i}\frac{d^2c(\theta_i)}{d\theta_i^2} \tag{66}$$

And, of course, it should be trivially obvious (since the expected value of a constant is the constant) that

$$E\left[\frac{\partial^2 l_i}{\partial \theta_i^2}\right] = -\frac{1}{\phi_i}\frac{d^2c(\theta_i)}{d\theta_i^2} \tag{67}$$

At this point, we use ML Fact #2, and replace the left hand side with $-Var[\frac{\partial l_i}{\partial \theta_i}]$, which results in

$$Var[\frac{\partial l_i}{\partial \theta_i}] = \frac{1}{\phi_i}\frac{d^2c(\theta_i)}{d\theta_i^2} \tag{68}$$

Next, put to use the result stated in expression (64). That allows the left hand side to be replaced:

$$\frac{1}{\phi_i^2} Var\left[y_i\right] = \frac{1}{\phi_i}\frac{d^2c(\theta_i)}{d\theta_i^2} \tag{69}$$

and the end result is:

$$Var[y_i] = \phi_i \frac{d^2 c(\theta_i)}{d\theta_i^2} \tag{70}$$

The Variance function, $V(\mu)$, is defined as

$$V(\mu_i) = \frac{d^2 c(\theta_i)}{d\theta_i^2} = c''(\theta_i) \tag{71}$$

The relabeling amounts to

$$Var(y_i) = \phi_i V(\mu_i)$$

6.2.3 Variance functions for various distributions

Firth observes that the variance function characterizes the members in this class of distributions, as in:

Normal $V(\mu) = 1$

Poisson $V(\mu) = \mu$

Binomial $V(\mu) = \mu(1-\mu)$

Gamma $V(\mu) = \mu^2$

7 Score equations for the model with a noncanonical link.

Our objective is to simplify this in order to calculate the k parameter estimates in $\hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k)$ that solve the score equations (one for each parameter):

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial b_k} = 0 \tag{72}$$

The "big result" that we want to derive is this:

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{1}{\phi_i V(\mu_i) g'(\mu_i)} [y_i - \mu_i] x_{ik} = 0 \tag{73}$$

That's a "big result" because because it looks "almost exactly like" the score equation from Weighted (or Generalized) Least Squares.

7.1 How to simplify expression 72.

1. Claim: $\partial l / \partial \theta = [y_i - \mu_i]/\phi_i$. This is easy to show. Start with the general version of the exponential family:

$$l(\theta_i | y_i) = \frac{y_i \theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i) \tag{74}$$

Differentiate:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{1}{\phi_i}\left[y_i - \frac{dc(\theta_i)}{d\theta_i}\right] \tag{75}$$

Use GLM Fact #1,$\mu_i = dc(\theta_i)/d\theta_i$.

$$\frac{\partial l_i}{\partial \theta_i} = \frac{1}{\phi_i}\left[y_i - \mu_i\right] \tag{76}$$

This is starting to look familiar, right? Its the difference between the observation and its expected (dare we say, predicted?) value. A residual, in other words.

2. Claim: $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu)}$

   Recall GLM Fact #1, $\mu_i = dc/d\theta_i$. Differentiate μ_i with respect to θ_i:

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{d^2 c(\theta_i)}{d\theta_i^2} \tag{77}$$

   In light of GLM Fact #2,

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{Var(y_i)}{\phi_i} \tag{78}$$

   And $Var(y_i) = \phi_i \cdot V(\mu)$, so 77 reduces to

$$\frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i) \tag{79}$$

   For continuous functions, it is true that

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} \tag{80}$$

   and so the conclusion is

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)} \tag{81}$$

3. Claim: $\partial \eta_i / \partial b_k = x_{ik}$

   This is a simple result because we are working with a linear model!

   The link function indicates

$$\eta_i = X_i b = b_1 x_{i1} + \ldots + b_p x_{ip}$$

$$\frac{\partial \eta_i}{\partial b_k} = x_{ik} \tag{82}$$

The Big Result: Put these claims together to simplify 72

There is one score equation for each parameter being estimated. Insert the results from the previous steps into 72. (See, eg, McCullagh & Nelder, p. 41, MM&V, p. 331, or Dobson, p. 63).

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{1}{\phi_i V(\mu_i)} [y_i - \mu_i] \frac{\partial \mu_i}{\partial \eta_i} x_{ik} = 0 \tag{83}$$

Note $\partial\mu_i/\partial\eta_i = 1/g'(\mu_o)$ because

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)} \tag{84}$$

As a result, we find:

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{1}{\phi_i V(\mu_i) g'(\mu_i)} [y_i - \mu_i] x_{ik} = 0 \tag{85}$$

7.2 The Big Result in matrix form

In matrix form, y and $\mu$are column vectors and X is the usual data matrix. So

$$X'W(y-\mu) = 0 \tag{86}$$

The matrix W is NxN square, but it has only diagonal elements $\frac{1}{\phi_i}\frac{1}{V(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}$. More explicitly, 86 would be:

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ & & & \\ x_{1p} & x_{2p} & & x_{Np} \end{bmatrix} \begin{bmatrix} \frac{1}{\phi_i V(\mu_1)}\frac{\partial \mu_i}{\partial \eta_i} & 0 & \cdots & 0 \\ 0 & \frac{1}{\phi_2 V(\mu_2)}\frac{\partial \mu_2}{\partial \eta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{\phi_N V(\mu_N)}\frac{\partial \mu_N}{\partial \eta_N} \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_N - \mu_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{87}$$

There are p rows here, each one representing one of the score equations.
The matrix equation in 86 implies

$$X'Wy = X'\mu \tag{88}$$

The problem remains to find estimate of b so that these equations are satisfied.

8 Score equations for the Canonical Link

Recall we want to solve this score equation.

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial b_k} = 0$$

The end result will be the following

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{1}{\phi_i} [y_i - \mu_i] x_{ik} = 0 \tag{89}$$

Usually it is assumed that the dispersion parameter is the same for all cases. That means ϕ is just a scaling factor that does not affect the calculation of optimal $\hat{b}$. So this simplifies as

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} [y_i - \mu_i] x_{ik} = 0 \tag{90}$$

I know of two ways to show that this result is valid. The first approach follows the same route that was used for the noncanonical case.

8.1 Proof strategy 1

1. Claim 1:

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \eta_i}{\partial b_k} \tag{91}$$

This is true because the canonical link implies that

$$\frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} = 1 \tag{92}$$

Because the canonical link is chosen so that $\theta_i = \eta_i$, then it must be that $\partial\theta/\partial\eta = 1$ and $\partial\eta/\partial\theta = 1$. If you just think of the derivatives in 92 as fractions and rearrange the denominators, you can arrive at the right conclusion (but I was cautioned by calculus professors at several times about doing things like that without exercising great caution).

It is probably better to remember my magical q, where $\theta_i = q(\mu_i)$, and note that $\mu_i = g^{-1}(\eta_i)$, so expression 92 can be written

$$\frac{\partial q(\mu_i)}{\partial \mu_i} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \tag{93}$$

In the canonical case, $\theta_i = q(\mu_i) = g(\mu_i)$, and also $\mu_i = g^{-1}(\eta_i) = q^{-1}(\eta)$. Recall the definition of the link function includes the restriction that it is a monotonic function. For monotonic functions, the derivative of the inverse is (here I cite my first year calculus book: Jerrold Marsden and Alan Weinstein, *Calculus,* Menlo Park, CA: Benjamin-Cummings, 1980, p. 224):

$$\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = \frac{\partial q^{-1}(\eta_i)}{\partial \eta_i} = \frac{1}{\frac{\partial q(\mu_i)}{\partial \eta_i}} \tag{94}$$

Insert that into expression 93 and you see that the assertion in 92 is correct.

2. Claim 2: Previously obtained results allow a completion of this exercise.

   The results of the proof for the noncanonical case can be used. Equations 76 and 82 indicate that we can replace $\partial l_i/\partial\theta_i$ with $\frac{1}{\phi_i}[y_i-\mu_i]$and $\partial\eta_i/\partial b_k$ with x_{ik}.

8.2 Proof Strategy 2.

Begin with the log likelihood of the exponential family.

$$l(\theta_i|y_i)=\frac{y_i\theta_i-c(\theta_i)}{\phi_i}+h(y_i,\phi_i) \tag{95}$$

Step 1. Recall $\theta_i=\eta_i=X_ib$:

$$l(\theta_i|y_i)=\frac{y_i\cdot X_ib-c(X_ib)}{\phi_i}+h(y_i,\phi_i) \tag{96}$$

We can replace θ_i by X_ib because the definition of the canonical link ($\theta_i=\eta_i$) and the model originally stated $\eta_i=X_ib$.

Step 2. Differentiate with respect to b_k to find the $k'th$ score equation

$$\frac{\partial l_i}{\partial b_k}=\frac{y_ix_i-c'(X_ib)x_{ik}}{\phi_i} \tag{97}$$

The chain rule is used.

$$\frac{dc(X_ib)}{db_k}=\frac{dc(X_ib)}{d\theta_k}\frac{dX_ib}{db_k}=c'(X_ib)x_{ik} \tag{98}$$

Recall GLM Fact #1, which stated that $c'(\theta_i)=\mu_i$. That implies

$$\frac{\partial l}{\partial b_k}=\frac{y_ix_{ik}-\mu_ix_{ik}}{\phi_i} \tag{99}$$

and after re-arranging

$$\frac{\partial l}{\partial b_k}=\frac{1}{\phi_i}[y_i-\mu_i]x_{ik} \tag{100}$$

There is one of the score equations for each of p the parameters. The challenge is to choose a vector of estimates $\hat{b}=(\hat{b}_1,\hat{b}_2,\ldots,\hat{b}_{k-1},\hat{b}_p)$ that simultaneously solve all equations of this form.

8.3 The Matrix form of the Score Equation for the Canonical Link

In matrix form, the k equations would look like this:

$$\begin{bmatrix} x_{11} & x_{21} & x_{31} & \cdots & x_{(N-1)1} & x_{N1} \\ x_{12} & x_{22} & & & & x_{N2} \\ x_{13} & x_{23} & & & & x_{N3} \\ \vdots & & & & & \vdots \\ x_{1p} & x_{2p} & & \cdots & x_{(N-1)p} & x_{Np} \end{bmatrix} \begin{bmatrix} \frac{1}{\phi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\phi_2} & 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{\phi_N} \end{bmatrix} \begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \\ y_3-\mu_3 \\ \\ \vdots \\ y_{(N-1)}-\mu_{(N-1)} \\ y_N-\mu_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{101}$$

MM&V note that we usually assume ϕ_i is the same for all cases. If it is a constant, it disappears from the problem.

$$\begin{bmatrix} x_{11} & x_{21} & x_{31} & \cdots & x_{(N-1)1} & x_{N1} \\ x_{12} & x_{22} & & & & x_{N2} \\ x_{13} & x_{23} & & & & x_{N3} \\ \vdots & & & & & \vdots \\ x_{1p} & x_{2p} & & \cdots & x_{(N-1)p} & x_{Np} \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ y_3 - \mu_3 \\ \\ \vdots \\ y_{(N-1)} - \mu_{(N-1)} \\ y_N - \mu_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{102}$$

Assuming $\phi_i = \phi$, the matrix form is as follows. y and μ are column vectors and X is the usual data matrix. So this amounts to:

$$\frac{1}{\phi} X'(y-\mu) = X'(y-\mu) = 0 \tag{103}$$

$$X'y = X'\mu \tag{104}$$

Although this appears to be not substantially easier to solve than the noncanonical version, it is in fact quite a bit simpler. In this expression, the only element that depends on b is μ. In the noncanonical version, we had both W and μ that were sensitive to b.

9 ML Estimation: Newton-based algorithms.

The problem outlined in 73 or 101 is not easy to solve for optimal estimates of b. The value of the mean, μ_i depends on the unknown coefficients, b, as well as the values of the X's, the y's, and possibly other parameters. And in the noncanonical system, there is the additional complication of the weight matrix W. There is no "closed form" solution to the problem.

The Bottom Line: we need to find the roots of the score functions.

Adjust the estimated values of the b's so that the score functions equal to 0.

The general approach for iterative calculation is as follows:

$$b^{t+1} = b^t + Adjustment$$

The Newton and Fisher scoring methods differ only slightly, in the choice of the *Adjustment*.

9.1 Recall Newton's method

Review my Approximations handout and Newton's method. One can approximate a function $f()$ by the sum of its value at a point and a linear projection (f' means derivative of f):

$$f(b^{t+1}) = f(b^t) + (b^{t+1} - b^t) f'(b^t) \tag{105}$$

This means that the value of f at b^{t+1} is approximated by the value of f at b^t plus an approximating increment.

The Newton algorithm for root finding (also called the Newton-Raphson algorithm) is a technique that uses Netwon's approximation insight to find the roots of an first order equation. If this is a score equation (first order condition), we want the left hand size to be zero, that means that we should adjust the parameter estimate so that

$$0 = f(b^t) + (b^{t+1} - b^t) f'(b^t) \tag{106}$$

or

$$b^{t+1} = b^t - \frac{f(b^t)}{f'(b^t)} \tag{107}$$

This gives us an **iterative procedure** for recalculating new estimates of b.

9.2 Newton-Raphson method with several parameters

The Newton approach to approximation is the basis for both the Newton-Raphson algorithm and the Fisher "method of scoring."

Examine formula 107. That equation holds whether b is a single scalar or a vector, with the obvious exception that, in a vector context, the denominator is thought of as the inverse of a matrix, $1/f'(b) = [f'(b)]^{-1}$. Note that $f(b)$ is a column vector of p derivatives, one for each variable in the vector b (it is called a 'gradient'). And the $f'(b)$ is the matrix of second derivatives. For a three-parameter problem, for example:

$$f(b_1, b_{2,}, b_3) = \begin{bmatrix} \frac{\partial l(b|y)}{\partial b_1} \\ \frac{\partial l(b|y)}{\partial b_2} \\ \frac{\partial l(b|y)}{\partial b_3} \end{bmatrix} \tag{108}$$

$$f'(b_1, b_{2,}, b_3) = \left[\frac{\partial^2 l}{\partial b \partial b'}\right] = \begin{bmatrix} \frac{\partial^2 l(b|y)}{\partial b_1^2} & \frac{\partial^2 l(b|y)}{\partial b_1 \partial b_2} & \frac{\partial^2 l(b|y)}{\partial b_1 \partial b_3} \\ \frac{\partial^2 l(b|y)}{\partial b_1 \partial b_2} & \frac{\partial^2 l(b|y)}{\partial b_2^2} & \frac{\partial^2 l(b|y)}{\partial b_2 \partial b_3} \\ \frac{\partial^2 l(b|y)}{\partial b_1 \partial b_3} & \frac{\partial^2 l(b|y)}{\partial b_2 \partial b_3} & \frac{\partial^2 l(b|y)}{\partial b_3^2} \end{bmatrix} \tag{109}$$

That matrix of second derivatives is also known as the Hessian. By custom, it is often called H.

The Newton-Raphson algorithm at step t updates the estimate of the parameters in this way:

$$b^{t+1} = b^t - \left[\frac{\partial^2 l(b^t|y)}{\partial b \partial b'}\right]^{-1} \left[\frac{\partial l(b^t|y)}{\partial b}\right] \tag{110}$$

If you want to save some ink, write $b^{t+1} = b^t - H^{-1} f(b)$ or $b^{t+1} = b^t - H^{-1} \partial l / \partial b$.

We have the negative of the Hessian matrix in this formula. From the basics of matrix calculus, if H is "negative semi-definite," it means we are in the vicinity of a local maximum in the likelihood function. The iteration process will cause the likelihood to increase with each recalculation. (Note, saying H is negative definite is the same as saying that $-H$ is not positive definite). In the vicinity of the maximum, all will be well, but when the likelihood is far from maximal, it may not be so.

The point of caution here is that the matrix H^{-1} is not always going to take the estimate of the new parameter in "the right direction". There is nothing that guarantees the matrix is negative semi-definite, assuring that the fit is "going down" the surface of the likelihood function.

9.3 Fisher Scoring

R.A. Fisher was a famous statistician who pioneered much of the maximum likelihood theory. In his method of scoring, instead of the negative Hessian, one should instead use the information matrix, which is the negative of the expected value of the Hessian matrix.

$$I(b) = -E\left[\frac{\partial^2 l(b^t|y)}{\partial b \partial b'}\right] \tag{111}$$

The matrix $E[\partial^2 l/\partial b \partial b']$ is always negative semi-definite, so it has a theoretical advantage over the Newton-Raphson approach. ($-E[\partial^2 l/\partial b \partial b']$ is positive semi-definite).

There are several different ways to calculate the information matrix. In Dobson's *An Introduction to Generalized Linear Models*, 2ed, the following strategy is used. The item in the j'th row, k'th column of the information matrix is known to be

$$I_{jk}(b) = E[U(b_j)U(b_k)] \tag{112}$$

I should track down the book where I read that this is one of three ways to calculate the expected value of the Hessian matrix in order to get the Information. Until I remember what that was, just proceed. Fill in the values of the score equations $U(b_j)$ and $U(b_k)$ from the previous results:

$$E[[\sum_{i=1}^{N} \frac{y_i - \mu_i}{\phi V(\mu_i) g'(\mu_i)} x_{ij}][\sum_{l=1}^{N} \frac{y_l - \mu_l}{\phi V(\mu_l) g'(\mu_l)} x_{lk}] \tag{113}$$

That gives $N \times N$ terms in a sum, but almost all of them are 0. $E[(y_i - \mu_i)(y_l - \mu_l)] = 0$ if $i \neq l$, since the cases are statistically independent. That collapses the total to N terms,

$$\sum_{i=1}^{N} \frac{E(y_i - \mu_i)^2}{[\phi V(\mu_i) g'(\mu_i)]^2} x_{ij} x_{ik} \tag{114}$$

The only randomly varying quantity is y_i, so the denominator is already reduced as far as it can go. Recall that the definition of the variance of y_i is equal to the numerator, $Var(y_i) = E(y_i - \mu_i)^2$ and $Var(y_i) = \phi V(\mu_i)$, so we can do some rearranging and end up with

$$\sum_{i=1}^{N} \frac{1}{\phi V(\mu)[g'(\mu_i)]^2} x_{ij} x_{ik} \tag{115}$$

which is the same as

$$\sum_{i=1}^{N} \frac{1}{\phi V(\mu)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ij} x_{ik} \tag{116}$$

The part that includes the X data information can be separated from the rest. Think of the not-X part as a weight.

Using matrix algebra, letting x_j represent the $j'th$ column from the data matrix X,

$$I_{jk}(b) = x_j' W x_k \tag{117}$$

The Fisher Information Matrix is simply compiled by filling in all of its elements in that way, and the whole information matrix is

$$I(b) = X'WX \tag{118}$$

The Fisher scoring equations, using the information matrix, are

$$b^{t+1} = b^t + \left[I(b^t)\right]^{-1} U(b^t) \tag{119}$$

Get rid of the inverse on the right hand side by multiplying through on the left by $I(b^t)$:

$$I(b^t)b^{t+1} = I(b^t)b^t + U(b^t) \tag{120}$$

Using the formulae that were developed for the information matrix and the score equation, the following is found (after about 6 steps, which Dobson wrote down clearly):

$$X'WXb^{t+1} = X'Wz \tag{121}$$

The reader should become very alert at this point because we have found something that should look very familiar. The previous equation is the VERY RECOGNIZABLE normal equation in regression analysis. It is the matrix equation for the estimation of a weighted regression model in which z is the dependent variable and X is a matrix of input variables. That means that a computer program written to do regression can be put to use in calculating the b estimates for a GLM. One simply has to obtain the "new" (should I say "constructed") variable z and run regressions. Over and over.

The variable z appears here after some careful re-grouping of the right hand side. It is something of a magical value, appearing as if by accident (but probably not an accident in the eyes of the experts). The column vector z has individual elements equal to the following (t represents calculations made at the t'th iteration).

$$z_i = x_i' b^t + (y_i - \mu_i^t)\left(\frac{\partial \eta_i^t}{\partial \mu_i}\right) \tag{122}$$

Since $\eta_i = x_i' b^t$, this is

$$z_i = \eta_i^t + (y_i - \mu_i^t)\left(\frac{\partial \eta_i^t}{\partial \mu_i}\right) \tag{123}$$

This is a Newtonian approximation formula. It gives an approximation of the linear predictor's value, starting at the point (μ_i^t, η_i^t) and moving "horizontally" by the distance $(y_i - \mu_i^t)$. The variable z_i essentially represents our best guess of what the unmeasured "linear predictor" is for a case in which the observed value of the dependent variable is y_i. The estimation uses μ_i^t, our current estimate of the mean corresponding to y_i, with the link function to make the approximation.

9.3.1 General blithering about Fisher Scoring and ML

I have encountered a mix of opinion about H and I and why some authors emphasize H and others emphasize I. In the general ML context, the Information matrix is more difficult to find because it requires us to calculate the expected value, which depends on a complicated distribution. That is William Greene's argument *(Econometric Analysis,* 5th edition, p. 939). He claims that most of the time the Hessian itself, rather than its expected value, is used in practice. In the GLM, however, we have special structure–the exponential family–which simplifies the problem.

In a general ML context, many excellent books argue in favor of using Fisher's information matrix rather than the Hessian matrix. This approach is called "Fisher scoring" most of the time, but the terminology is sometimes ambiguous. Eugene Demidenko's *Mixed Models: Theory and Applications* (New York: Wiley, 2004). Demidenko outlines several reasons to prefer the Fisher approach. In the present context, this is the most relevant (p. 86):

> The negative Hessian matrix, $-\partial^2 l/\partial\theta^2$, or *empirical* information matrix, may not be positive definite (more precisely, not nonnegative definite) especially when the current approximation is far from the MLE. When this happens, the NR algorithm slows down or even fails. On the contrary, the *expected* information matrix, used in the FS algorithm, is always positive definite...

Incidentally, it does not matter if we use the ordinary N-R algorithm or Fisher scoring with the canonical link in a GLM. The expected value of the Hessian equals the observed value when a canonical link is used. McCullagh & Nelder observe that$-H$ equals I.

The reader might want to review the ML theory. The Variance/Covariance matrix of the estimates of b is correctly given by the inverse of $I(b)$. As mentioned in the comparison of I and H, my observation is that, in the general maximum likelihood problem, $I(b)$ can't be calculated, so an approximation based on the observed second derivatives is used. It is frequently asserted that the approximations are almost as good.

I have found several sources which claim that the Newton-Raphson method will lead to slightly different estimates of the Variance/Covariance matrix of b than the Fisher scoring method.

10 Alternative description of the iterative approach: IWLS

The glm() procedure in R's stat library uses a routine called "Iterated Weighted Least Squares." As the previous section should indicate, the calculating algorithm in the N-R/Fisher framework will boil down to the iterated estimation of a weighted least squares problem. The interesting thing, in my opinion, is that McCullagh & Nelder prefer to present the IWLS as their solution strategy. Comprehending IWLS as the solution requires the reader to make some truly heroic mental leaps. Neverhtheless, the IWLS approach is a frequent starting place. After describing the IWLS algorithm, McCullagh & Nelder then showed that IWLS is equivalent to Fisher scoring. I don't mean simply equivalent in estimates of the b's, but in fact algorithmically equivalent. I did not understand this argument on first reading, but I do now. The two approaches are EXACTLY THE SAME. Not just similar. Mechanically identical.

I wonder if the following is true. I *believe* that it probably is. In 1972, they had good algorithms for calculating weighted least squares problems, but not such good access to maximum

likelihood software. So McCullagh and Nelder sought an estimating procedure for their GLM that would use existing tools. The IWLS approach is offered as a calculating algorithm, and then in the following section, McCullagh & Nelder showed that if one were to use Fisher scoring, one would arrive at the same iterative calculations.

Today, there are good general purpose optimization algorithms for maximum likelihood that could be used to calculate GLM estimates. One could use them instead of IWLS, probably. The algorithm would not be so obviously tailored to the substance of the problem, and a great deal of insight would be lost.

If you study IWLS, it makes your mental muscles stronger and it also helps you to see how all of the different kinds of regression "fit together".

10.1 Start by remembering OLS and GLS

Remember the estimators from linear regression in matrix form:

OLS	*WLS and GLS*
$\hat{y} = X\hat{b}$	$\hat{y} = X\hat{b}$
$minimize\,(y-\hat{y})'(y-\hat{y})$	$minimize\,(y-\hat{y})'W(y-\hat{y})$
$\hat{b} = (X'X)^{-1}X'y$	$\hat{b} = (X'WX)^{-1}X'Wy$
$Var(\hat{b}) = \sigma^2(X'X)^{-1}$	$Var(\hat{b}) = \sigma^2(X'WX)^{-1}$

10.2 Think of GLM as a least squares problem

Something like this would be nice as an objective function:

$$Sum\,of\,Squares \sum(\mu_i - \hat{\mu}_i)^2 \tag{124}$$

If we could choose $\hat{b}$ to make the predicted mean fit most closely against the true means, life would be sweet!

But recall in the GLM that we don't get to observe the mean, so we don't have μ_i to compare against $\hat{\mu}_i$. Rather we just have observations gathered from a distribution. In the case of a logit model, for example, we do not observe μ_i, but rather we observe 1 or 0. In the Poisson model, we observe count values, not λ_i.

Maybe we could think of the problem as a way of making the linear predictor fit most closely against the observations:

$$Sum\,of\,Squares \sum(\eta_i - \hat{\eta}_i)^2$$

We certainly can calculate $\hat{\eta}_i = \hat{b}_0 + \hat{b}_1 \cdot x_{11}$. But we can't observe η_i, the "true value" of the linear predictor, any more than we can observe the "true mean" μ_i.

And, for that matter, if we could observe either one, we could just use the link function g to translate between the two of them.

We can approximate the value of the unobserved linear predictor, however, by making an educated guess. Recall Newton's approximation scheme, where we approximate an unknown value by taking a known value and then projecting toward the unknown. If we–somehow magically–knew the mean value for a particular case, μ_i, then we would use the link function to figure out what the linear predictor should be:

$$\eta_i = g(\mu_i)$$

If we want to know the linear predictor at some neighboring value, say y_i, we could use Newton's approximation method and calculate an approximation of $g(y_i)$ as

$$\widetilde{g(y_i)} = g(\mu_i) + (y_i - \mu_i) \cdot g'(\mu_i) \tag{125}$$

The symbol $\widetilde{g(y_i)}$ means "approximate value of the linear predictor" and we could call it $\widetilde{\eta}_i$. However, possibly because publishers don't like authors to use lots of expensive symbols, we instead use the letter z_i to refer to that approximated value of the linear predictor.

$$z_i = g(\mu_i) + (y_i - \mu_i) \cdot g'(\mu_i) \tag{126}$$

If we use that estimate of the linear predictor, then we could think of the GLM estimation process as a process of calculating $\hat{\eta}_i = \hat{b}_0 + \hat{b}_1 x_1 + \ldots$ to minimize:

$$Sum\,of\,Squares \sum (z_i - \hat{\eta}_i)^2 \tag{127}$$

This should help you see the basic idea that the IWLS algorithm is exploiting.

The whole exercise here is premised on the idea that you know the mean, μ_i, and in real life, you don't. That's why an iterative procedure is needed. Make a guess for μ_i, then make an estimate for z_i, and repeat.

10.3 The IWLS algorithm.

1. Begin with μ_i^0, starting estimates of the mean of y_i.

2. Calculate a new variable z_i^0 (this is to be used as an "approximate" value of the response variableη_i).
$$z_i^0 = g(\mu_i^0) + (y_i - \mu_i^0) g'(\mu_i^0) \tag{128}$$
Because $\eta_i^0 = g(\mu_i^0)$, sometimes we write
$$z_i^0 = \eta_i^0 + (y_i - \mu_i^0) g'(\mu_i^0) \tag{129}$$

This is a Newton-style first-order approximation. It estimates the linear predictor's value that corresponds to the observed value of y_i– starting at the value of the $g(\mu_i^0)$ and adding the increment implied by moving the distance $(y_i - \mu_i^0)$ with the slope $g'(\mu_i)$.

If $g(\mu_i^0)$ is undefined, such as $ln(0)$, then some workaround is required.

3. Estimate b^0 by weighted least squares, minimizing the squared distances

$$\sum w_i^0 (z_i^0 - \eta_i)^2 = \sum w_i^0 (z_i^0 - X\widehat{b^0})^2 \tag{130}$$

Information on the weights is given below. The superscript w_i^0 is needed because the weights have to be calculated at each step, just as z^0 and b^0 must be re-calculated.

In matrix form, this is

$$b^0 = (X'W^0X)^{-1}X'W^0z^0 \tag{131}$$

4. Use b^0 to calculate $\eta^1 = X_i b^0$ and then calculate a new estimate of the mean as

$$\mu_i^1 = g^{-1}(\eta_i^1) \tag{132}$$

5. Repeat step 1, replacing μ^0 by μ^1.

10.4 Iterate until convergence

Repeat that process again and again, until the change from one iteration to the next is very small. A common tolerance criterion is 10^{-6}, as in

$$\frac{b^{t+1} - b^t}{b^t} < 10^{-6} \tag{133}$$

When the iteration stops, the parameter estimate of b is found.

10.5 The Variance of $\hat{b}$.

When the iteration stops, the parameter estimate of b is found. The $Var(b)$ from the last stage is used, so

$$Var(b) = (X'WX)^{-1} \tag{134}$$

10.6 The weights in step 3.

Now, about the weights in the regression. If you set the weights to equal this value

$$w_i = \frac{1}{\phi V(\mu_i)[g'(\mu_i)]^2} \tag{135}$$

Then the IWLS algorithm is identical to Fisher scoring.

10.7 Big insight from the First Order Condition

The first order conditions, the k score equations of the fitting process at each step, are

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^{N} \frac{1}{\phi_i V(\mu_i)[g'(\mu_i)]^2} [z_i - \eta_i] x_{ik} = 0 \tag{136}$$

In a matrix, the weights are seen as

$$W = \begin{bmatrix} \frac{1}{\phi_1 V(\mu_1)[g'(\mu_1)]^2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\phi_1 V(\mu_2)[g'(\mu_2)]^2} & 0 & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \frac{1}{\phi_{N-1} V(\mu_{N-1})[g'(\mu_{N-1})]^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\phi_N V(\mu_N)[g'(\mu_N)]^2} \end{bmatrix} \tag{137}$$

Next, 136 the first order condition is written with matrices as

$$X'W[z-\eta]=0 \tag{138}$$

Recall the definition of $\eta = Xb$, so

$$X'W[z-Xb]=0 \tag{139}$$

$$X'Wz=X'WXb \tag{140}$$

$$b=(X'WX)^{-1}X'Wz \tag{141}$$

Wow! That's just like the WLS formula! We are "regressing z on X".