SS9055B: Generalized Linear Models

Section 4: Introduction Generalized Linear Models

Objectives

This section of the course will introduce you to the class of generalized linear models and likelihood inference for these models. By the end of this section you should be able to:

- explain the importance of generalized linear models in modern statistics,
- show that a model fits into the class of generalized linear models by identifying the random component, systematic component, and link function,
- derive likelihood equations,
- apply the iteratively (re-)weighted least squares algorithm to compute the MLEs, and
- obtain the approximate covariance matrix for the MLEs.

Introduction

Enough with the introductory material! We will now begin our discussion of generalized linear models (GLM) in earnest. The three specific extensions of the linear regression model that we will work with during the course are the logistic regression model for binomial data, the Poisson log-linear model for count data, and the multinomial model for categorical data. Maximum likelihood theory and applications for these models date to the mid-20th century and they were considered separately for a long time. One of the major achievements of statistics in the last century was the development of the theory of the exponential family of distributions which lead to the realization that these models – and many other linear models – can be fit into the same framework. This allows us to study general methods of estimation and inference that can be applied to a wide range of models without needing to develop tools separately.

To motivate the structure of a GLM, consider what we teach introductory statistics students about linear regression models that don't satisfy the assumptions of linearity, normality, and

constant variance (heteroscedasticity). The standard multiple linear regression model can be written as:

$$Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon_i, \quad i = 1, \ldots, N.$$

Usually we assume that the errors, ϵ_i, are independent draws from an identical normal distribution with mean zero and constant variance, σ^2.

Suppose that we fit this model and we find that the residuals form a curved pattern. What do we tell students to do? One thing we tell them to do is to transform the response variable. This is one way to achieve linearity.

Suppose that we fit this model and we find that the residuals do not have constant variance. Most often, they have a funnel shape so that the residuals are larger for larger values of the predictors. What do we tell students to do? One thing we tell them is to transform the response variable. This is called stabilizing the variance.

Suppose that we fit this model and we find that the residuals are not normally distributed. Perhaps the distribution of the residuals is positively skewed so that values below the regression line lie closer to the mean. What do we tell students to do? One thing we tell them to do is to transform the response variable to achieve normality.

What should we tell students if different transformations are required to achieve linearity, normality, and to stabilize the variance?

The problem here is that we are trying to solve problems with three aspects of the model – linearity of the mean, normality of the residuals, and constant variance – with the same tool. The framework of GLMs allow us to separate linearity of the mean and the distribution of the errors by separately defining the random component of the model (the distribution of the errors), the systematic component (the linear predictor), and a link function which relates the two.

Generalized Linear Model Framework

Models within the GLM framework have a certain structure defined by three components: the random component, the systematic component, and the link function.

Random Component

The first component of a GLM defines the distribution of the of the data given the mean. Standard linear regression models assume that the data are normally distributed about the mean. The methods developed for generalized linear modeling allow the data to come more generally from a larger class of distributions which form the exponential families. Specifically, we are going to consider distributions in a subset of the exponential family called the exponential dispersion famillt. Generally, the random variable Y indexed by θ is a member of the

exponential dispersion family if we can write the density of Y (either the pdf for continuous random variables or the pmf for discrete random variables) as:

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

The parameter θ is called the natural or canonical parameter and may not be equal to the parameter that you are used to seeing for different distributions (e.g., p for the binomial or λ for the Poisson). Instead, θ may be defined by some one-to-one transformation of the usual parameter. The function $b(\theta)$ is called the cumulant function and ϕ is the dispersion parameter. One important property of the exponential dispersion family is that the mean and variance of Y can be obtained directly from the cumulant function and dispersion function:

1. $\mu = E(Y) = b'(\theta)$
2. $\text{Var}(Y) = b''(\theta)a(\phi)$.

The proof of this property will be an exercise on the next assignment.

For the most part, we will actually limit ourselves to a simplified version of the exponential family where $a(\phi) = \phi/\omega$ so that

$$f(y|\theta,\phi) = \exp\left(\omega\left[\frac{y\theta - b(\theta)}{\phi}\right] + c(y,\phi)\right)$$

where ω is a known weight and ϕ is a constant to be estimated. An example of this is weighted linear regression in which each Y_i is the average of ω_i *iid* random variables and so the variance of Y_i is σ^2/ω_i for some known value of (like the number of observations if Y_i is an average). In this case:

1. $\mu = E(Y) = b'(\theta)$
2. $\text{Var}(Y) = \frac{\phi b''(\theta)}{\omega}$.

Our assumption in the GLM framework will be that the observations $Y_1, \ldots, Y_N$ come from the same member of the exponential family so that the functions $b(\cdot)$ and $c(\cdot,\cdot)$ and the parameter ϕ in common. However, the weights ω_i and natural parameters θ_i may vary by observation. In particular, we will assume that the weights are known and model the natural parameters as functions of covariates and regression coefficients, as defined by the final two components.

Systematic Component

The second component of the a GLM describes how the predictor variables combine to affect the mean of the response. Consider our simple linear regression model again:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, N$$

where the residuals are independent with constant variance σ^2. We could rewrite this as:

$$E(Y_i) = \eta_i$$
$$\text{Var}(Y_i) = \sigma^2$$

where:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

This term is called the linear predictor. In a GLM, the mean as a function of the linear predictor.

Link Function

The final component of a GLM describes the relationship between the random and systematic components of the model. Specifically, it determines the relationship between the mean of the data distribution and the linear predictor. Given the random component, Y, and the linear predictor, η, the link function is a one-to-one function $g(\cdot)$ such that:

$$g(\mu) = \eta \text{ and } \mu = g^{-1}(\eta).$$

The simple linear regression model employs the identity link function,

$$g(\mu) = \mu,$$

so that the mean is equal to the linear predictor, $\mu = \eta$. However, this need not be the case. Note that the expression for the mean above allows us to extend this relationship to the canonical parameter. Specifically, if $\mu = g^{-1}(\eta)$ and $\mu = b'(\theta)$ then $b'(\theta) = g^{-1}(\eta)$. The canonical link function for a model is the function $g(\cdot)$ such that $g(\mu) = \theta$. These link functions have special properties which we will encounter later, but the choice of the link function should be based on your knowledge of the system being studied.

Full Form

Putting these three components together, we can write the full form of the GLM by saying that $Y_1, \ldots, Y_n$ are independent random variables such that the pdf/pmf of Y_i is:

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

where the mean of Y_i, $\mu_i = b'(\theta_i)$, is defined by the linear predictor and link function as

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

where $x_{i1}, \ldots, x_{ip}$ are the values of p covariates or predictors associated with the i^{th} observation. We can think of this as a linear model for the transformed mean of the distribution of Y_i. Note

that the form of the exponential dispersion family somewhat hides the role of the regression coefficients, $\boldsymbol{\beta}$, because this parameter does not appear directly in the density. It is important to keep in mind that the natural parameter, θ_i, is a function of $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ defined through $b'(\theta)$ and $g(\mu)$.

Maximum Likelihood Inference for GLMs

The primary advantage of the GLM framework is that it unifies methods of inference for a range of different models. We will consider maximum likelihood inference in depth, but this also applies to Bayesian inference.

Likelihood Equations

Suppose that we have a model which falls within the GLM framework. We have assumed that the responses from our n observations are independent conditional on the observed predictor variables. So the likelihood function is:

$$L(\boldsymbol{\beta}, \phi|\boldsymbol{y}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\beta}, \phi)$$

and the log-likelihood is

$$l(\boldsymbol{\beta}, \phi|\boldsymbol{y}) = \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\beta}, \phi)$$

where

$$\log f(y_i|\boldsymbol{\beta}, \phi) = \omega_i \left[\frac{y_i\theta_i - b(\theta_i)}{\phi}\right] + c(y_i, \phi)$$

since each Y_i belongs to the exponential dispersion family with $b(\cdot)$, $c(\cdot, \cdot)$ and ϕ in common.

The likelihood functions we will work with will all be regular, smooth functions and so we can find the maximum by differentiating to identify critical points. The results equations that we need to solve to compute the MLE of $\boldsymbol{\beta}$,

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = 0, \quad j = 0, \ldots, p$$

are called the likelihood equations. The individual equations are derived by applying the chain rule:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{1}$$

Considering each component in turn we get:

i) $\frac{\partial l_i}{\partial \theta_i} = \omega_i \frac{y_i - b'(\theta_i)}{\phi} = \omega_i \frac{y_i - \mu_i}{\phi}$,

ii) $\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \left(\frac{\partial}{\partial \theta_i} b'(\theta_i)\right)^{-1} = (b''(\theta_i))^{-1} = \frac{\phi}{\omega_i \text{Var}(Y_i)}$,

iii) $\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}\right) = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1}$ which can't be simplified until the link function is specified, and

iv) $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$ where we set $x_{i0} = 1$ for all $i = 1, \ldots, n$.

Substituting these elements back into equation (1) yields

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

so that the likelihood equations are

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, \ldots, p.$$

The unique solutions to these equations are the MLEs. Note again that the vector of regression coefficeints, $\boldsymbol{\beta}$, is hidden within the mean for each observation, μ_i. Different link functions produce different values of $\partial \mu_i / \partial \eta_i$ and hence produce different estimates of $\boldsymbol{\beta}$ for the same data.

Note that the likelihood equations can now be written down very easily once the GLM is completely defined – i.e., once we have selected a distribution for the data and a link function. All we need to know are the mean and variance of each observation in terms of the parameters $\boldsymbol{\beta}$ and ϕ, and the first derivative of the link function.

Asymptotic Covariance Matrix

Since we are dealing with distributions in the exponential family the regularity conditions for asymptotic normality of the MLEs are going to be satisfied and do not need to be checked (we will not do the math to show this). This means that we can approximate the variance-covariance matrix of the parameters by the inverse information matrix:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = I^{(n)}(\boldsymbol{\beta})^{-1}.$$

We cannot apply the simplification that the information for the sample of size n is n times the information because the response variables are not identically distributed. Their means change depending on the values of the covariates. However, the j, k entry of the information matrix can still be computed fairly easily:

$$-E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) = -E\left(\frac{\partial}{\partial \beta_k} \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right) = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

In matrix form

$$I(\boldsymbol{\beta}) = X'W(\boldsymbol{\beta})X$$

where $W(\boldsymbol{\beta})$ is the $N \times N$ diagonal matrix with i^{th} entry

$$w_i = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

We can then estimate the asymptotic variance of $\hat{\boldsymbol{\beta}}$ by:

$$\widehat{\text{cov}(\hat{\boldsymbol{\beta}})} = (X'\hat{W}X)^{-1}$$

where $\hat{W} = W(\hat{\boldsymbol{\beta}})$. This expression should look familiar to you because it is the variance matrix for the coefficients in a weighted regression model except that we are estimating the weights based on estimates of the coefficients instead of assuming that they are known. Note that the weights in this equation, w_i (double-u i), are distinct from the weights in the exponential dispersion family, ω_i (omega i), both in value and in notation.

Computing the MLEs

In section we derived the likelihood equations for a general model within the GLM framework. However, these equations cannot be solved analytically in most cases. Instead, we need to apply numerical methods to compute the MLEs. This could be done by applying any canned optimization routine, like those available through `optim()` function in `R`. However, there is a very nice generic algorithm for solving the likelihood equations based on the Newton-Raphson algorithm that is both quick and stable and, as you will see, produces the asymptotic covariance matrix directly as part of the optimization algorithm.

Newton-Raphson Algorithm

The the most common algorithms for optimization (i.e., maximizing or miniziming a function) are the Newton-Raphson (NR) algorithm[1] and its extensions. We'll start by considering optimization in 1-dimension and then derive the formulas for multiple dimensions.

Suppose that we wish to optimize a function $f(x)$, $x \in \Re$, which is at least twice differnetiable. The NR algorithm works by approximating $f(x)$ by a quadratic function about some point x_0,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2}.$$

[1] It's tempting to call this the NRA, but that acronym has already been taken so we'll stick with NR algorithm.

To identify the critical points we need to solve $f'(x) = 0$. Differentiating both sides wrt x yields

$$f'(x) \approx f'(x_0) + f''(x_0)(x - x_0)$$

and setting this equal to 0 and solving for x we find that

$$x = x_0 - f'(x_0)/f''(x_0).$$

The NR algorithm works by using this equation to update our guesses of the critical points until we reach convergence. Given a current guess, $x^{(t)}$, we set

$$x^{(t+1)} = x^{(t)} - f'(x^{(t)})/f''(x^{(t)})$$

and iterate until the difference between $x^{(t)}$ and $x^{(t+1)}$ is sufficiently small.

Suppose now that $f(\boldsymbol{x})$ is a multivariate function. Then we approximate $f(\boldsymbol{x})$ by

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \frac{df(\boldsymbol{x}_0)}{d\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)' \frac{d^2 f(x_0)}{d\boldsymbol{x} d\boldsymbol{x}'}(\boldsymbol{x} - \boldsymbol{x}_0)$$

and the NR algorithm update becomes

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\frac{d^2 f(x^{(t)})}{d\boldsymbol{x}\ d\boldsymbol{x}'} \right)^{-1} \frac{df(\boldsymbol{x}^{(t)})}{d\boldsymbol{x}}.$$

Applying this to the problem of finding the MLEs, suppose that we wish to maximize $l(\boldsymbol{\beta})$. The NR iteration is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\frac{d^2 l(\boldsymbol{\beta}^{(t)})}{d\boldsymbol{\beta}\ d\boldsymbol{\beta}'} \right)^{-1} \frac{dl(\boldsymbol{\beta}^{(t)})}{d\boldsymbol{\beta}}.$$

Sometimes you will see this written as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\frac{d^2 l(\boldsymbol{\beta}^{(t)})}{d\boldsymbol{\beta}\ d\boldsymbol{\beta}'} \right)^{-1} u(\boldsymbol{\beta}^{(t)})$$

where $u(\boldsymbol{\beta}) = dl(\boldsymbol{\beta})/d\boldsymbol{\beta}$ is called the score function. This is really just the derivative of the log-likelihood, but it is important enough in its own right to deserve its own name.

Fisher Scoring

One modified version of the NR algorithm that is particularly applicable to GLMs is the Fisher scoring algorithm. The NR algorithm depends on the second derivative of the log-likelihood evaluated at the observed data. Explicitly, we need to compute the matrix

$$-\left(\frac{d^2}{d\boldsymbol{\beta}\ d\boldsymbol{\beta}'} l(\boldsymbol{\beta}^{(t)}, \boldsymbol{y}) \right) \tag{2}$$

which has j, k entry

$$- \frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\boldsymbol{\beta}, \boldsymbol{y}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$$

where $\boldsymbol{y}$ represents the observed data. This matrix looks almost like the information matrix, but the two are not equal. The information matrix has j, k entry

$$-E\left(\frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\boldsymbol{\beta}, \boldsymbol{Y})\right).$$

The matrix in equation (2) is found by replacing the random variable $\boldsymbol{Y}$ with the observed value $\boldsymbol{y}$ and removing the expected value (since there are no longer any random quantities). For this reason, the matrix is called the observed information matrix.

Fisher scoring replaces the observed information matrix with the expected information (or just the information)

$$I(\boldsymbol{\beta}^{(t)}) = -E\left(\frac{d^2}{d\boldsymbol{\beta}\ d\boldsymbol{\beta}'} l(\boldsymbol{\beta}^{(t)}, \boldsymbol{Y})\right).$$

An iteration in the Fisher Scoring algorithm is then defined as:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I^{-1}(\boldsymbol{\beta}^{(t)}) u(\boldsymbol{\beta}^{(t)})$$

where $I^{-1}(\boldsymbol{\beta}^{(t)})$ represents the inverse of the information matrix given the current values of the parameters, $\boldsymbol{\beta}^{(t)}$. Equivalently $\boldsymbol{\beta}^{(t+1)}$ is the solution to:

$$I(\boldsymbol{\beta}^{(t)}) \boldsymbol{\beta}^{(t+1)} = I(\boldsymbol{\beta}^{(t)}) \boldsymbol{\beta}^{(t)} + u(\boldsymbol{\beta}^{(t)}).$$

Both algorithms work by providing successive quadratic approximations to the likelihood. The only difference is the value of the second order term. One advantage of the Fisher Scoring algorithm is that it immediately produces the approximate covariance matrix for the parameters as a by-product. Consider that at convergence $\boldsymbol{\beta}^{(t)} \approx \boldsymbol{\beta}^{(t+1)}$ and we set $\hat{\boldsymbol{\beta}} = \beta^{(t+1)}$. Then:

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}}) = I^{-1}(\boldsymbol{\beta}^{(t+1)}) \approx I^{-1}(\boldsymbol{\beta}^{(t)})$$

which is the expected information computed on the final iteration. The gain may be small on modern computers, but this means that we have to do no extra work to compute variance estimates for the parameters.

ML as Reweighted Least Squares

To gain some further insight into the algorithm for fitting GLMs we can compare the Fisher Scoring algorithm with the least squares method of fitting linear regression models. Suppose that we have a simple linear regression model,

$$z_i = x_i' \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, N,$$

or in matrix form

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

If $\epsilon \sim N(0, \sigma^2 I_N)$ then the MLEs,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z},$$

are also the simple least squares estimates. More generally,if $\epsilon \sim N(0, \sigma^2 V)$ for known matrix $\boldsymbol{V}$ then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{z}.$$

Equivalently, $\hat{\boldsymbol{\beta}}$ is the solution of the normal equations,

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{z}. \tag{3}$$

These are referred to as weighted least squares estimates if $\boldsymbol{V}$ is diagonal and generalized least squares estimates if $\boldsymbol{V}$ is not diagonal.

To compare this with the Fisher Scoring algorithm we need to rewrite the score equation in a more useful manner. Recall that

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \eta_i}{\partial \mu_i}.$$

In matrix form we get

$$u(\boldsymbol{\beta}) = \frac{dl}{d\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

where $\boldsymbol{W}$ is the matrix of weights we computed previously, $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n)$ where:

$$w_i = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

and $\boldsymbol{D}$ is the diagonal matrix with entries $\frac{\partial \mu_i}{\partial \eta_i}$. Further, recall that:

$$I(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}.$$

Substituting these expressions into the equations for a single iteration of the Fisher Scoring algorithm we get

$$\boldsymbol{X}'\boldsymbol{W}_t\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}'\boldsymbol{W}_t\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{X}'\boldsymbol{W}_t\boldsymbol{D}_t^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

or

$$\boldsymbol{X}'\boldsymbol{W}_t\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}'\boldsymbol{W}_t z(\boldsymbol{\beta}^{(t)}) \tag{4}$$

where

$$z(\boldsymbol{\beta}^{(t)}) = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{D}_t^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(t)}).$$

The quantities W_t and D_t represent the values of the matrices W and D computed at the current values of the parameters, $\boldsymbol{\beta}^{(t)}$.

What does this show? Comparing equations (3) and (4) we can see that they have exactly the same form. This shows that each iteration of the Fisher Scoring algorithm is equivalent to conducting weighted least squares using a transformed version of the original response variables, $z(\boldsymbol{\beta}^{(t)})$, sometimes called pseudo-data which depends on the current values of the parameters. On each iteration of the algorithms we update the transformed data and compute the weight matrix given our current estimates of the parameters, and then we compute least squares estimates as if we had normal data. For this reason, the algorithm is known as iteratively reweighted least squares.

As an alternative interpretation we can consider that least squares estimation is equivalent to fitting a linear regression model with normal errors. In essence, the Fisher scoring algorithm is equivalent to fitting linear regression models to a transformation of the original data where the transformation uses the current values of the parameters to achieve the best possible normal approximation. The reason we restrict to the exponential family of distributions is that, in essence, these are the distributions which are well approximated by the normal.