Lecture 2 - Technical Aspects of GLM estimation

- Topics Covered
  - First and Second Moment for the canonical exponential Family
  - Maximum Likelihood
  - Newton-Raphson
  - Fisher Information
  - Inference in GLMs

The exponential family: First Moment

- GLMs with the canonical exponential family can be estimated using the same technique and the same function with R (with slight adjustments to the syntax)

- Part of the reason is that they also have a similar form of the mean and variance of their distributions.

- To see this, start with one of the basic properties of all distribution functions:

$$\int f(y; \theta)\, dy = 1$$

- Differentiating with respect to θ

$$\int \frac{df(y; \theta)}{d\theta} dy = 0$$

- Any changes to the distribution through θ must cancel each other out over the support of y.

The exponential family: First Moment (cont)

- Recall that

$$f(y;\theta) = \exp\left(yb(\theta) + c(\theta) + d(y)\right)$$

- Differentiating with respect to θ

$$\begin{aligned} \frac{df(y;\theta)}{d\theta} &= \left(yb'(\theta) + c'(\theta)\right)\exp\left(yb(\theta) + c(\theta) + d(y)\right) \\ &= \left(yb'(\theta) + c'(\theta)\right)f(y;\theta) \end{aligned}$$

- Plugging into $\int \frac{df(y;\theta)}{d\theta} dy = 0$, we have:

$$\int \left(yb'(\theta) + c'(\theta)\right)f(y;\theta)\,dy = 0$$

- Breaking the integral into two parts:

$$b'(\theta)\int yf(y;\theta)\,dy + c'(\theta)\int f(y;\theta)\,dy = 0$$

- How do I simplify these components?

The exponential family: First Moment (cont)

- One definition and one property that are useful:

$$\mathrm{E}(y) = \int yf(y;\theta)\,dy \quad , \quad \int f(y;\theta)\,dy = 1$$

- Thus,

$$\begin{aligned} b'(\theta)\underbrace{\int yf(y;\theta)\,dy}_{=\mathrm{E}(y)} + c'(\theta)\underbrace{\int f(y;\theta)\,dy}_{=1} &= 0 \\ b'(\theta)\mathrm{E}(y) + c'(\theta) &= 0 \\ \Rightarrow \qquad \mathrm{E}(y) &= -\frac{c'(\theta)}{b'(\theta)} \end{aligned}$$

- Both $b(\theta)$ and $c(\theta)$ affect the mean of the y.
  - $c(\theta)$ is often called the "scale" function/parameter
  - $b(\theta)$ is often called the "shape" function, since it interacts with y.
- These can be most clearly seen when taking the log of the PDF:

$$\log(f(y;\theta)) = yb(\theta) + c(\theta) + d(y)$$

The exponential family: Second Moment

- To solve for variance, differentiate $\int \frac{df(y;\theta)}{d\theta} dy = 0$ with respect to θ

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = 0$$

- Recalling that:

$$\frac{df(y;\theta)}{d\theta} = \left(yb'(\theta) + c'(\theta)\right) f(y;\theta)$$

- We take a second derivative to get:

$$\begin{aligned}
\frac{d^2 f(y;\theta)}{d\theta^2} &= \left(yb''(\theta) + c''(\theta)\right) f(y;\theta) + \left(yb'(\theta) + c'(\theta)\right)^2 f(y;\theta) \\
&= \left(yb''(\theta) + c''(\theta)\right) f(y;\theta) + b'(\theta)^2 \left(y + \frac{c'(\theta)}{b'(\theta)}\right)^2 f(y;\theta) \\
&= \left(yb''(\theta) + c''(\theta)\right) f(y;\theta) + b'(\theta)^2 \left(y - \mathrm{E}(y)\right)^2 f(y;\theta)
\end{aligned}$$

- To complete the derivation, substitute into $\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = 0$

The exponential family: Second Moment (cont)

- Precisely,

$$\int \left(yb''(\theta) + c''(\theta)\right) f(y;\theta) + b'(\theta)^2 (y - \mathrm{E}(y))^2 f(y;\theta)\, dy = 0$$

- Using the same operations as before, first distribute the integral:

$$b''(\theta) \int y f(y;\theta)\, dy + c''(\theta) \int f(y;\theta)\, dy + b'(\theta)^2 \int (y - \mathrm{E}(y))^2 f(y;\theta)\, dy = 0$$

- Then impose the definition of expectations and variance:

$$b''(\theta)\mathrm{E}(y) + c''(\theta) + b'(\theta)^2 \mathrm{Var}(Y) = 0$$

- Finally, solving for variance:

$$\mathrm{Var}(Y) = -\frac{b''(\theta)\mathrm{E}(y) + c''(\theta)}{b'(\theta)^2}$$

The exponential family: Summary

- Thus, for the canonical exponential family of distributions,

$$f(y;\theta) = \exp(yb(\theta) + c(\theta) + d(y)),$$

  the mean and variance of the variables are precisely characterized by the functions $b(\theta)$ and $c(\theta)$

$$\begin{aligned} \mathrm{E}(y) &= -\frac{c'(\theta)}{b'(\theta)} \\ \mathrm{Var}(Y) &= -\frac{b''(\theta)\,\mathrm{E}(y) + c''(\theta)}{b'(\theta)^2} \end{aligned}$$

- Thus, the parameters we estimate are linked to the mean and variance through these equations.

Maximum Likelihood Estimation

- All of these properties are helpful for estimating relationships that are assumed to follow the canonical exponential family.

- As you might recall from 216, the likelihood function is written as:

$$L = \prod_{i=1}^{N} f(y_i; \theta)$$

- The Log-likelihood function, $l = \log(L)$, is

$$l = \sum_{i=1}^{N} \log(f(y_i; \theta))$$

- Within the exponential family,

$$l = \sum_{i=1}^{N} [a(y_i)b(\theta) + c(\theta) + d(y_i)]$$

- Remember that θ links to some underlying mean parameter of the model, μ, which is the mean of y, which itself links to the covariates by the link function

- When choosing optimal θ, only $b(\theta)$ and $c(\theta)$ and outcomes y_i matter.

Maximum Likelihood Estimation

- The derivative of the log-likelihood function with respect to some parameter θ is called the "score", U.

$$
\begin{aligned}
U \equiv \frac{dl}{d\theta} &= \sum_{i=1}^{N} \frac{d}{d\theta} \log f(y_i;\theta) \\
&= \sum_{i=1}^{N} \frac{\frac{d}{d\theta} f(y_i;\theta)}{f(y_i;\theta)}
\end{aligned}
$$

- The expected value of U is zero. To see this, note that

$$
\begin{aligned}
\mathrm{E}[U] &= \sum_{i=1}^{N} \mathrm{E}\left[\frac{\frac{d}{d\theta} f(y_i;\theta)}{f(y_i;\theta)} \right] \\
&= \sum_{i=1}^{N} \int \frac{\frac{d}{d\theta} f(y;\theta)}{f(y;\theta)} f(y;\theta)\, dy \\
&= \sum_{i=1}^{N} \int \frac{d}{d\theta} f(y;\theta)\, dy \\
&= \sum_{i=1}^{N} \frac{d}{d\theta} \underbrace{\int f(y;\theta)\, dy}_{=1} = 0
\end{aligned}
$$

Maximum Likelihood for Exponential Family

- To make this simple to start, let us assume that:

$$g(\mu) = \beta$$

- Under this assumption, we are essentially choosing one value of θ that is the same for every person, since the mean of y is assumed to be invariant to other covariates

- After estimating θ, then we can link to μ using the assumed distribution, and then β using the link function..

- Taking the derivative of l with respect to θ

$$U = \frac{dl}{d\theta} = \sum_{i=1}^{N} \frac{dl_i}{d\theta} = 0$$

- For univariate functions, this can be done by hand in some cases

- Though in practice, this is done using standard computational techniques, such as Newton-Raphson.

Univariate Numerical Optimization by Newton-Raphson

- The idea behind Newton-Raphson is pretty simple. Suppose you have a function $U(\theta)$, and you want to find the roots of the function.

$$U(\theta) = 0$$

- For Newton-Raphson, we iterate over different values for θ, trying to find a solution. θ^m is defined as the "mth" iteration (not to the power of m).

- Suppose that were are at a value θ^{m-1}, and would like to approximate the function $U(\theta)$ at θ^m. By a first-order Taylor series approximation:

$$U(\theta^m) = U(\theta^{m-1}) + \frac{dU(\theta)}{d\theta}\left(\theta^m - \theta^{m-1}\right)$$

- Substituting $U(\theta^m) = 0$, and solving for θ^m, we have

$$\begin{aligned}
0 &= U(\theta^{m-1}) + \frac{dU(\theta)}{d\theta}\left(\theta^m - \theta^{m-1}\right) \\
0 &= \frac{U(\theta^{m-1})}{\frac{dU(\theta)}{d\theta}} + \left(\theta^m - \theta^{m-1}\right) \\
\Rightarrow \quad \theta^m &= \theta^{m-1} - \frac{U(\theta^{m-1})}{\frac{dU(\theta^{m-1})}{d\theta}}
\end{aligned}$$

- The Newton-Raphson algorithm is based on this equation

Univariate Numerical Optimization by Newton-Raphson

- Newton-Raphson algorithm

  1. Begin with an initial guess, θ^0
  2. Solve for
  $$\theta^1 = \theta^0 - \frac{U(\theta^0)}{\frac{dU(\theta^0)}{d\theta}}$$
  3. If $\left|\theta^1 - \theta^0\right| < \epsilon$, then stop.
  4. If $\left|\theta^1 - \theta^0\right| > \epsilon$, then use θ^1 as initial guess and repeat from step 1.

- This always works when nicely behavior functions (continuous, differentiable) have a unique, global maximum.

- Other techniques are used when you cannot guarantee a unique global maximum. They all seem to have funny names (simulated annealing, particle swarm, etc..)

- Broyden's method is a variant of Newton-Raphson that approximates $\frac{dU(\theta^0)}{d\theta}$ using past changes in the function. Useful, but very slow. If you can take derivatives, you can speed up the process.

Newton-Raphson Example

- Here is a simple version of Newton-Raphson. We wish to find the value at which the following function is zero:

$$f(x) = (x-1)^2$$

- Obviously, we know the answer is $x = 1$. But, let's work through this iteratively.
- For newton-raphson, we need an initial guess. Let's say $x^0 = 0$
- Next, we need the derivative of the function.

$$\frac{df(x)}{dx} = 2x - 2$$

- Now, we iterate!

$$\begin{aligned} x^1 &= x^0 - \frac{f(x^0)}{\frac{df(x^0)}{dx}} \\ &= 0 - \frac{f(0)}{\frac{df(0)}{dx}} \\ x^1 &= 0 - \frac{1}{-2} = \frac{1}{2} \end{aligned}$$

Newton-Raphson Example

- Again!!

$$
\begin{aligned}
x^2 &= x^1 - \frac{f(x^1)}{\frac{df(x^1)}{dx}} \\
&= \frac{1}{2} - \frac{f(\frac{1}{2})}{\frac{df(\frac{1}{2})}{dx}} \\
x^2 &= \frac{1}{2} - \frac{\frac{1}{4}}{-1} = \frac{3}{4}
\end{aligned}
$$

- Check the value of $f(x)$

$$
f(\frac{3}{4}) = (\frac{3}{4} - 1)^2 = \frac{1}{16} \neq 0
$$

- Difference in x's: $|\frac{3}{4} - \frac{1}{2}| = \frac{1}{4}$

Newton-Raphson Example

- Again!!

$$
\begin{aligned}
x^3 &= x^2 - \frac{f(x^2)}{\frac{df(x^2)}{dx}} \\
&= \frac{3}{4} - \frac{f(\frac{3}{4})}{\frac{df(\frac{3}{4})}{dx}} \\
&= \frac{3}{4} - \frac{\frac{1}{16}}{-\frac{1}{2}} \\
x^2 &= \frac{3}{4} + \frac{1}{8} = 7/8
\end{aligned}
$$

- Check the value of $f(x)$

$$f(\frac{7}{8}) = (\frac{7}{8} - 1)^2 = \frac{1}{64}$$

- We are closer to 0 for the outcome.
- Difference in x's: $|\frac{3}{4} - \frac{7}{8}| = \frac{1}{8}$

Newton-Raphson Example

- Again!!

$$
\begin{aligned}
x^4 &= x^3 - \frac{f(x^3)}{\frac{df(x^3)}{dx}} \\
&= \frac{7}{8} - \frac{f(\frac{7}{8})}{\frac{df(\frac{7}{8})}{dx}} \\
&= \frac{7}{8} - \frac{\frac{1}{64}}{-\frac{1}{4}} \\
x^4 &= \frac{7}{8} + \frac{1}{16} = \frac{15}{16}
\end{aligned}
$$

- Check the value of $f(x)$

$$
f(\frac{15}{16}) = (\frac{15}{16} - 1)^2 = \left(\frac{1}{256}\right)
$$

- We are closer to 0 for the outcome.
- Difference in x's: $|\frac{15}{16} - \frac{14}{16}| = \frac{1}{16}$
- We'll stop here, but you keep going until the difference in x's is small enough.

Multivariate Newton Raphson

- Newton Raphson can be extended to a setting with multiple variables over which we maximize a function.

- Suppose that there are p variables, indexed β_j, $j = 1...p$, over which we are maximizing a function f

- For this case,

$$\frac{df}{d\beta_j} \equiv U_j(\beta) = 0$$

  must equal zero for all j, where β represents the $px1$ vector of β_j's

- A multi-variate first-order taylor-series expansion is written as:

$$\mathbf{U}^m = \mathbf{U}^{m-1} + \mathbf{J}^{m-1}\left(\beta^m - \beta^{m-1}\right)$$

  where:

  - $\mathbf{J}^{m-1}$ is the Jacobian matrix of U at iteration $m-1$
  - $\mathbf{U}^m$ is the $px1$ vector of scoring values at iteration m.

Multivariate Newton Raphson (cont.)

- As a reminder, the Jacobian is a *pxp* matrix with $\frac{dU_j}{d\beta_k}$ is the j^{th} row and k^{th} column.
- The element in the j^{th} row and k^{th} column of $\mathbf{J}$ is written as J_{jk}
- Trying to hit $\mathbf{U}^m = \mathbf{0}$ (all scores equal to zero) using the first-order approximation, we get:

$$\mathbf{0} = \mathbf{U}^{m-1} + \mathbf{J}^{m-1}\left(\beta^m - \beta^{m-1}\right)$$

- Rearranging:

$$\beta^m = \beta^{m-1} - \left(\mathbf{J}^{m-1}\right)^{-1}\mathbf{U}^{m-1}$$

- Again, we iterate until a solution.

Multivariate Maximum Likelihood for Exponential Family

- We now extend our earlier model to allow for a vector of covariates (which may include constants)

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

- Recall that μ_i links to to the mean of the distribution by θ_i
- Taking the derivative of l with respect to some parameter β_j

$$U_j = \frac{dl}{d\beta_j} = \sum_{i=1}^{N} \frac{dl_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\beta_j}$$

- $\frac{dl_i}{d\theta_i}$ is once again written as:

$$\begin{aligned}
\frac{dl_i}{d\theta_i} &= \frac{d}{d\theta_i}(y_i b(\theta_i) + c(\theta_i) + d(y_i)) \\
&= y_i b'(\theta_i) + c'(\theta_i) \\
&= b'(\theta_i)\left(y_i + \frac{c'(\theta_i)}{b'(\theta_i)}\right) = b'(\theta_i)(y_i - \mu_i)
\end{aligned}$$

- The last step is since $\mu_i = \mathrm{E}(Y_i) = -\frac{c'(\theta)}{b'(\theta)}$

Multivariate Maximum Likelihood for Exponential Family

- $\frac{d\theta_i}{d\mu_i}$ is the inverse of $\frac{d\mu_i}{d\theta_i}$:

$$\begin{aligned}\frac{d\mu_i}{d\theta_i} &= -\frac{c''(\theta_i)b'(\theta_i) - c'(\theta_i)b''(\theta_i)}{b'(\theta_i)^2} \\ &= -b'(\theta_i)\frac{c''(\theta_i) - c'(\theta_i)\frac{b''(\theta_i)}{b'(\theta_i)}}{b'(\theta_i)^2} = b'(\theta_i)\mathrm{Var}(Y_i)\end{aligned}$$

- Thus,

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{b'(\theta)\mathrm{Var}(Y_i)}$$

- Finally, since $g(\mu_i) = \mathbf{x}_i^T\beta$, we have:

$$\begin{aligned}\frac{dg(\mu_i)}{d\mu_i}\frac{d\mu_i}{d\beta_j} &= x_{ij} \\ \Rightarrow \qquad \frac{d\mu_i}{d\beta_j} &= \frac{x_{ij}}{\frac{dg(\mu_i)}{d\mu_i}}\end{aligned}$$

- Overall, we have that the derivative of the likelihood function (the "score") is:

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} \frac{x_{ij}}{\frac{dg(\mu)}{d\mu}} = 0$$

- To find the maximum likelihood estimates, U_j must be zero for all j.

Examples of Scoring Functions: Gaussian

- Gaussian regression with the identity link:
  - Identity link: $g(\mu_i) = \mu_i = x_i^T \beta$
  - Gaussian Distribution: $\text{Var}(Y_i) = \sigma$
- Thus, the score can be written as:

$$
\begin{aligned}
U_j &= \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} \frac{x_{ij}}{\frac{dg(\mu)}{d\mu}} = 0 \\
&= \sum_{i=1}^{N} \frac{\left(y_i - x_i^T \beta\right)}{\sigma} \frac{x_{ij}}{1} = 0 \\
&= \sum_{i=1}^{N} \left(y_i - x_i^T \beta\right) x_{ij} = 0
\end{aligned}
$$

- What does this remind you of?

Examples of Scoring Functions: Poisson

- Recall the Poisson distribution:

$$f(y;\theta) = \frac{\theta^y \exp[-\theta]}{y!}$$

- Poisson has a very cool property:
  - $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \theta_i$

- Assuming the identity link: $g(\mu_i) = \mu_i = x_i^T\beta = \theta_i$

- Thus, the score can be written as:

$$\begin{aligned} U_j &= \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} \frac{x_{ij}}{\frac{dg(\mu_i)}{d\mu_i}} = 0 \\ &= \sum_{i=1}^{N} \frac{\left(y_i - x_i^T\beta\right)x_{ij}}{x_i^T\beta} = 0 \end{aligned}$$

- We will use this a bit later when continuing the Poisson example

Multivariate Maximum Likelihood for Exponential Family

- The last piece for multivariate estimation of GLM models is the *information matrix*, **J**, which is made up of the elements J_{jk}
  - **J** is also called the "Fisher Information Matrix", named after Ronald Fisher.
  - Accuracy or (information given by X) around the maximum likelihood solution is defined by the curvature of the likelihood function at these points. This is why we call it information.
- The element J_{jk} is simply the covariance between score functions

$$J_{jk} = \mathrm{E}\left[U_j U_k\right]$$

- Importantly, for GLM models, J_{jk} is also the Jacobian matrix of the scoring functions (or, the Hessian matrix for the log-likelihood function)
- Thus, the information matrix is used in optimization, as well in variance-covariance estimation.

Information Matrix

- Using the formula for U_j, $\mathrm{E}\left[U_j U_k\right]$ can be written as:

$$\mathrm{E}\left[U_j U_k\right] = \mathrm{E}\left(\sum_{i=1}^{N} \frac{(y_i - \mu_i)}{\mathrm{Var}\,(Y_i)} \frac{x_{ij}}{\frac{dg(\mu_i)}{d\mu_i}} \sum_{l=1}^{N} \frac{(y_l - \mu_l)}{\mathrm{Var}\,(Y_l)} \frac{x_{ik}}{\frac{dg(\mu_l)}{d\mu_l}}\right)$$

- Expanding the summation into the square and cross-products

$$\mathrm{E}\left[U_j U_k\right] = \mathrm{E}\left(\sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\mathrm{Var}\,(Y_i)^2} \frac{x_{ij} x_{ik}}{\left(\frac{dg(\mu_i)}{d\mu_i}\right)^2}\right) + \mathrm{E}\left(\sum_{i=1}^{N}\sum_{l\neq i}^{N} \frac{(y_i - \mu_i)}{\mathrm{Var}\,(Y_i)} \frac{x_{ij}}{\frac{dg(\mu_i)}{d\mu_i}} \frac{(y_l - \mu_l)}{\mathrm{Var}\,(Y_l)} \frac{x_{lk}}{\frac{dg(\mu_l)}{d\mu_l}}\right)$$

- Since the expectation is only applied to random data (y's)

$$\begin{aligned} \mathrm{E}\left[U_j U_k\right] &= \left(\sum_{i=1}^{N} \frac{\mathrm{E}\,(y_i - \mu_i)^2}{\mathrm{Var}\,(Y_i)^2} \frac{x_{ij} x_{ik}}{\left(\frac{dg(\mu_i)}{d\mu_i}\right)^2}\right) \\ &+ \left(\sum_{i=1}^{N}\sum_{l\neq i}^{N} \frac{1}{\mathrm{Var}\,(Y_i)} \frac{x_{ij}}{\frac{dg(\mu_i)}{d\mu_i}} \frac{1}{\mathrm{Var}\,(Y_l)} \frac{x_{lk}}{\frac{dg(\mu_l)}{d\mu_l}} \mathrm{E}\left[(y_i - \mu_i)(y_l - \mu_l)\right]\right) \end{aligned}$$

- If observations are independent $\mathrm{E}\left[(y_i - \mu_i)(y_l - \mu_l)\right] = 0$ for all $i \neq l$. Finally,

$$J_{jk} = \mathrm{E}\left[U_j U_k\right] = \sum_{i=1}^{N} \frac{1}{\mathrm{Var}\,(Y_i)} \frac{x_{ij} x_{ik}}{\left(\frac{dg(\mu_i)}{d\mu_i}\right)^2}$$

Examples of Information Matrix

- We wish to simplify the following elements of the matrix **J**

$$J_{jk} = \mathrm{E}\left[U_j U_k\right] = \sum_{i=1}^{N} \frac{1}{\mathrm{Var}\left(Y_i\right)} \frac{x_{ij}x_{ik}}{\left(\frac{dg(\mu_i)}{d\mu_i}\right)^2}$$

- For **Gaussian**, assuming an identity link, we get:

$$J_{jk} = \mathrm{E}\left[U_j U_k\right] = \frac{1}{\sigma} \sum_{i=1}^{N} x_{ij}x_{ik}$$

- For **Poisson**, assuming an identity link, $\mathrm{Var}\left(Y_i\right) = x_i^T \beta$, we get:

$$J_{jk} = \mathrm{E}\left[U_j U_k\right] = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{x_i^T \beta}$$

- Let's now write out the entire procedure for Poisson and $\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2}$, where $x_{i1} = 1$ for all i (ie. a constant)
    - That is, $\mu_i = \beta_1 + \beta_2 x_{i2}$

Examples of Information Matrix

- Since $x_{i1} = 1$ for all i, J_{11} is written as:

$$J_{11} = \mathrm{E}[U_1 U_1] = \sum_{i=1}^{N} \frac{1}{\beta_1 + \beta_2 x_{i2}}$$

- J_{12} is written as:

$$J_{12} = \mathrm{E}[U_1 U_2] = \sum_{i=1}^{N} \frac{x_{i2}}{\beta_1 + \beta_2 x_{i2}}$$

- J_{21} is written as:

$$J_{21} = \mathrm{E}[U_2 U_1] = \sum_{i=1}^{N} \frac{x_{i2}}{\beta_1 + \beta_2 x_{i2}}$$

- J_{22} is written as:

$$J_{22} = \mathrm{E}[U_2 U_2] = \sum_{i=1}^{N} \frac{x_{i2}^2}{\beta_1 + \beta_2 x_{i2}}$$

- On your own, you should write this for the Gaussian distribution under the same link $\mu_i = \beta_1 + \beta_2 x_{i2}$.

Examples of Information Matrix

- Thus, we can write the matrix $\mathbf{J}$

$$\mathbf{J} = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \frac{1}{\beta_1+\beta_2 x_{i2}} & \sum_{i=1}^{N} \frac{x_{i2}}{\beta_1+\beta_2 x_{i2}} \\ \sum_{i=1}^{N} \frac{x_{i2}}{\beta_1+\beta_2 x_{i2}} & \sum_{i=1}^{N} \frac{x_{i2}^2}{\beta_1+\beta_2 x_{i2}} \end{pmatrix}$$

- Recalling that the score is written as:

$$U_j = \sum_{i=1}^{N} \frac{\left(y_i - x_i^T \beta\right) x_{ij}}{x_i^T \beta} = 0$$

- A matrix $\mathbf{U}$ of scoring functions can be written as:

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \frac{y_i-\beta_1-\beta_2 x_{i2}}{\beta_1+\beta_2 x_{i2}} \\ \sum_{i=1}^{N} \frac{(y_i-\beta_1-\beta_2 x_{i2}) x_{i2}}{\beta_1+\beta_2 x_{i2}} \end{pmatrix}$$

- So, by Newton Raphson, we find our solution by iterating the following:

$$\begin{pmatrix} \beta_1^{new} \\ \beta_2^{new} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \mathbf{J}^{-1}\mathbf{U}$$

- R uses "Iteratively Re-weighted Least Squares ", which is identical to this (though approached differently)

Predictions in GLM Models

- Predictions are central to applied applications
  - Predict clicking behavior on ads
  - Prediction intervals for stock prices
- A vast majority of R commands use "predict()" to generate a vector of predictions
- Example using Logit

```
glm_logit<-glm(nilf~age+educ,d,family=binomial(link="logit"))
glm_predict_1<-predict(glm_logit)
summary(glm_predict_1)
length(glm_predict_1)
nrow(d)
```

- What do you notice about the predictions?

Predictions in GLM Models

- There are two issues
  - The vector of predictions is, by default, the same length as the vector of feasible output
  - The predictions are on the scale of the link function, not the response
- Two solutions (respectively):
  - Define "newdata" as the original dataset, in this case "d".
  - Use option type="response".
- Example using Logit

```
glm_predict_2<-predict(glm_logit,newdata=d,type="response")
summary(glm_predict_2)
length(glm_predict_2)
nrow(d)
d$nilf_predict<-as.numeric(glm_predict_2)
```

- You can also extract standard errors of the predictions

```
glm_predict_3<-predict(glm_logit,newdata=d,type="response", se=TRUE)
```

- Command is similar for "lm" but without option for type.

Inference in GLM Models

- For inference regarding one parameter, use t-test as you would with OLS
  - Central limit theorem works for GLMs
  - The variance-covariance matrix of β's is $\mathbf{J}^{-1}$
- For joint-tests:
  - Use F-test and F-distribution for normal regression
  - Use "Likelihood Ratio" test and Chi-square distribution for all others
- Likelihood Ratios are a simple comparison of the "maximal model", i.e. the best we could do given the data, and the actual model:

$$D = 2\left(l(\beta_{max};y) - l(\widehat{\beta};y)\right)$$

- D is also called "deviance", and a summary of which is provided in regression results.
- $l(\beta_{max};y)$ is constructed by basically using y_i for μ_i in the likelihood function, and then calculating likelihood.

Derivation of Deviance

- Deviance is defined as follows

$$D = 2\left(l(\widehat{\beta}_{max};y) - l(\widehat{\beta};y)\right)$$

- The questions:
  - Where does the 2 come from?
  - How do we use this for inference?

- Write a second-order taylor series expansion of the likelihood function around some estimate $\widehat{\beta}$:

$$l(\beta;y) = l(\widehat{\beta};y) + \left(\beta - \widehat{\beta}\right)\mathbf{U}(\widehat{\beta}) - \frac{1}{2}\left(\beta - \widehat{\beta}\right)^T \mathbf{J}\left(\widehat{\beta}\right)\left(\beta - \widehat{\beta}\right)$$

- What is the value of $\mathbf{U}(\widehat{\beta})$ if $\widehat{\beta}$ is the solution to maximum likelihood?

- $\mathbf{U}(\widehat{\beta}) = 0$

Deriving Deviance

- Thus, we have:

$$l(\beta;y) = l(\widehat{\beta};y) - \frac{1}{2}\left(\beta - \widehat{\beta}\right)^T \mathbf{J}\left(\widehat{\beta}\right)\left(\beta - \widehat{\beta}\right)$$

- Rearranging

$$2\left(l(\widehat{\beta};y) - l(\beta;y)\right) = \left(\widehat{\beta} - \beta\right)^T \mathbf{J}\left(\widehat{\beta}\right)\left(\widehat{\beta} - \beta\right) \sim \chi^2(p)$$

- This is where the two comes from. To related deviance to this, recall that

$$\begin{aligned} D &= 2\left(l(\widehat{\beta}_{max};y) - l(\widehat{\beta};y)\right) \\ &= 2\left(l(\widehat{\beta}_{max};y) - l(\beta_{max};y)\right) - 2\left(l(\widehat{\beta};y) - l(\beta;y)\right) + 2\left(l(\beta_{max};y) - l(\beta;y)\right) \\ &\sim \qquad \chi^2(m) \qquad - \qquad \chi^2(p) \qquad + \qquad K \end{aligned}$$

- If K is small, then we have:

$$D \sim \chi^2(m-p)$$

Likelihood Ratio Test

- The likelihood ratio tests does exactly as the name suggests - compares the likelihood of two different models.
- Suppose that $\widehat{\beta}$ are the estimates from the full unrestricted model, and $\widehat{\beta}_A$ is an alternate set of parameter estimates that impose restrictions on the model.
- Deviance for unrestricted model:

$$D = 2\left(l(\widehat{\beta}_{max};y) - l(\widehat{\beta};y)\right)$$

- Deviance for restricted model:

$$D_A = 2\left(l(\widehat{\beta}_{max};y) - l(\widehat{\beta}_A;y)\right)$$

- Subtract D from D_A:

$$\Delta D = D_A - D = 2\left(l(\widehat{\beta};y) - l(\widehat{\beta}_A;y)\right)$$

- Then compare this value to $\chi^2(r,p)$, which is the value from a chi-squared distribution, where:
    - r is the number of restrictions.
    - p is the preferred probability of false rejection (note that programs, including R, may require the confidence level as opposed to probability of false rejection).

LR Test in R

- There are a few ways to execute the LR test in R.
- Can calculate the likelihood ratio directly.
- Using our previous Poisson example for hours worked, let's test for the joint effect of all education dummy categories.

```
poissonreg<-glm(hourslw~age+educ,subd,family=poisson(link="log"))
summary(poissonreg)
poissonreg2<-glm(hourslw~age,subd,family=poisson(link="log"))
summary(poissonreg2)
LR<-(poissonreg2$deviance-poissonreg$deviance)
```

- Then, we compare the LR to the Chi-square distribution

```
chi_crit<-qchisq(.95, df=4)
ifelse(LR>chi_crit,"Reject the restrictions", "Fail to reject the restrictions")
```

- Or, you can construct the P-value for false rejection

```
pchisq(LR, 4, lower.tail = FALSE)
```

LR Test in R

- There are a few ways to execute the LR test in R.
- The best is using the "lrtest" command from the "lmtest" library in R.
- Using our previous Poisson example for hours worked, let's test for the joint effect of all education dummy categories.

```
library(lmtest)
poissonreg<-glm(hourslw~age+educ,subd,family=poisson(link="log"))
summary(poissonreg)
lrtest(poissonreg,"educ")
```

- The results indicate the two models being tested, the log-likelihood for each, and the p-value from the LR test.
- Small p-values indicate that one can reject the joint restrictions.