# Maximum Likelihood Estimation

JEFF GILL
**Departments of Government and Mathematics/Statistics**
**Center for Data Science**
***American University***

## Overview

- Previously we have seen "closed form" estimators for quantities of interest, such as $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

- Moving to nonlinear models for categorical and limited support outcomes requires a more flexible process.

- Maximum Likelihood Estimation (Fisher 1922, 1925) is a classic method that finds the value of the estimator "most likely to have generated the observed data, assuming the model specification is correct."

- There is both an abstract idea to absorb and a mechanical process to master.

## Model Background

- Suppose we care about some outcome of interest $\mathbf{Y}$, and determine that it has distribution $f()$.
- The stochastic component is:
$$\mathbf{Y} \sim f(\mu, \tau).$$
- The systematic component is:
$$\mu = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$
- This setup is very general and covers all of the nonlinear regression models we will cover.
- You have seen the linear model in a similar form before:
$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\epsilon}_i = N(0, \sigma^2).$$

# Model Background

- But now we are going to think of it in this more general way, for example:

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \mathbf{X}_i\boldsymbol{\beta}.$$

- An even more general way specifies a link function:

$$g(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$
$$\mathbf{Y}_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) + \boldsymbol{\epsilon}_i$$
$$\mathbf{Y}_i = g^{-1}(\mu_i) + \boldsymbol{\epsilon}_i$$

- We typically write this in expected value terms:

$$\mathbb{E}[Y|\mathbf{X}, \boldsymbol{\beta}] = \boldsymbol{\mu}$$

## The Likelihood Function

- Assume that:

  $$x_1, x_1, \ldots, x_n \sim \text{ iid } f(x|\theta),$$

  where $\theta$ is a parameter that is critical to the data generation process (DGP).

- Since these values are independent, the joint distribution of the observed data is just the product of their individual PDF/PMFs:

  $$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- But once we observe the the data $\mathbf{x}$ is fixed.

- It is $\theta$ that is unknown, so rewrite the joint distribution function according to:

  $$f(\mathbf{x}|\theta) = L(\theta|\mathbf{x}).$$

- Note that this is a purely *notational* change, nothing is different mathematically.

## The Likelihood Function

- Fisher (1922) justifies this because at this point we know $\mathbf{x}$.

$$f(\mathbf{x}|\theta) \;\longrightarrow\; L(\theta|\mathbf{x}).$$

- A semi-Bayesian justification works as follows, we want to perform:

$$p(\mathbf{x}|\theta) = \frac{p(\mathbf{x})}{p(\theta)} p(\theta|\mathbf{x}).$$

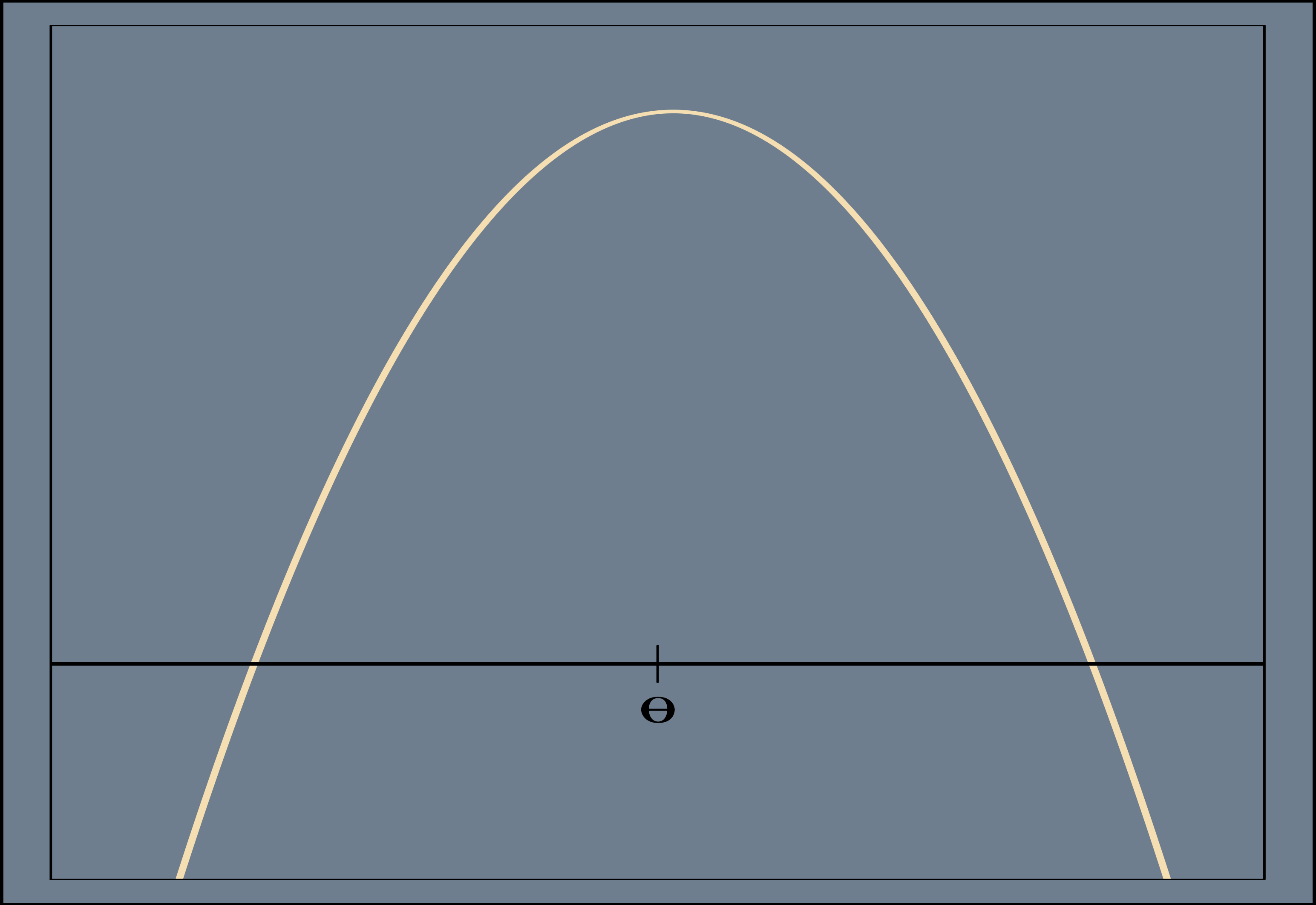
  but $p(\mathbf{x}) = 1$ since the data has already occurred, and if we put a finite uniform prior on $\theta$ over its finite allowable range (support), then $p(\theta) = 1$.

- Therefore:

$$p(\mathbf{x}|\theta) = \frac{1}{1} p(\theta|\mathbf{x}) = p(\theta|\mathbf{x}).$$

- The only caveat here is the finiteness of the support of $\theta$.

# Generic Likelihood Function Illustration



# Poisson MLE

- Start with the Poisson PMF for $x_i$:

$$p(X = x_i) = f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!},$$

which requires the assumptions: non-concurrence of arrivals, the number of arrivals is proportion to the time of study, this rate is constant over the time, and there is no serial correlation of arrivals.

- The likelihood function is created from the joint distribution:

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-\theta}\theta^{x_1}}{x_1!}\frac{e^{-\theta}\theta^{x_2}}{x_2!}\cdots\frac{e^{-\theta}\theta^{x_n}}{x_n!} = e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}.$$

- Suppose we have the data: $\mathbf{x} = \{5, 1, 1, 1, 0, 0, 3, 2, 3, 4\}$, then the likelihood function is:

$$L(\theta|\mathbf{x}) = \frac{e^{-10\theta}\theta^{20}}{207360},$$

which is the probability of observing *this* exact sample.

# Poisson MLE

- It is often easier to deal the logarithm of the MLE:

$$\log L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x}) = \log\left(e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}\right) = -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right).$$

- For our small example this is:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

- Importantly, for the family of functions that we will use the likelihood function and the log-likelihood function have the same mode (maximum of the function) for $\theta$.

- They are both guaranteed to be concave to the x-axis.

# Obtaining the Poisson MLE

- Freshman calculus: where is the maximum of the function? At the point when first derivative of the function equals zero.
- So take the first derivative, set it equal to zero, and solve.
- $\frac{d}{d\theta}\ell(\theta|\mathbf{x}) \equiv 0$ is called the likelihood equation.
- For the example:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

Taking the derivative, and setting equal to zero:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -10 + 20\theta^{-1} \equiv 0,$$

so that $20\theta^{-1} = 10$, and therefore $\hat{\theta} = 2$ (note the hat).

# Obtaining the Poisson MLE

- More generally:

$$\ell(\theta|\mathbf{x}) = -n\theta + \sum_{i=1}^{n} \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i \equiv 0$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\mathbf{x}}$$

- It is *not* true that the MLE is always the data mean.

## General Steps

- This process is import to us:
  1. Identify the PMF or PDF.
  2. Create the likelihood function from the joint distribution of the observed data.
  3. Change to the log for convenience.
  4. Take the first derivative with respect to the parameter of interest.
  5. Set equal to zero.
  6. Solve for the MLE.

## Poisson Example in R

```
# POISSON LIKELIHOOD AND LOG-LIKELIHOOD FUNCTION
llhfunc<-function(X,p,do.log=TRUE) {
        d <- rep(X,length(p))
        q.vec <- rep(length(y.vals),length(p)); p.vec <- rep(p,q.vec)
        print(q.vec)
        d.mat <- matrix(dpois(d,p.vec,log=do.log),ncol=length(p))
        print(d.mat)
        if (do.log==TRUE) apply(d.mat,2,sum)
        else apply(d.mat,2,prod)
}
```

## Poisson Example in R

```
# HERE'S A TEST FUNCTION
y.vals<-c(1,3,1,5,2,6,8,11,0,0)
llhfunc(y.vals,c(4,30))
[1] 10 10
         [,1]    [,2]
 [1,] -2.6137 -26.599
 [2,] -1.6329 -21.588
 [3,] -2.6137 -26.599
 [4,] -1.8560 -17.782
 [5,] -1.9206 -23.891
 [6,] -2.2615 -16.172
 [7,] -3.5142 -13.395
 [8,] -6.2531 -10.089
 [9,] -4.0000 -30.000
[10,] -4.0000 -30.000
[1]  -30.666 -216.114
```

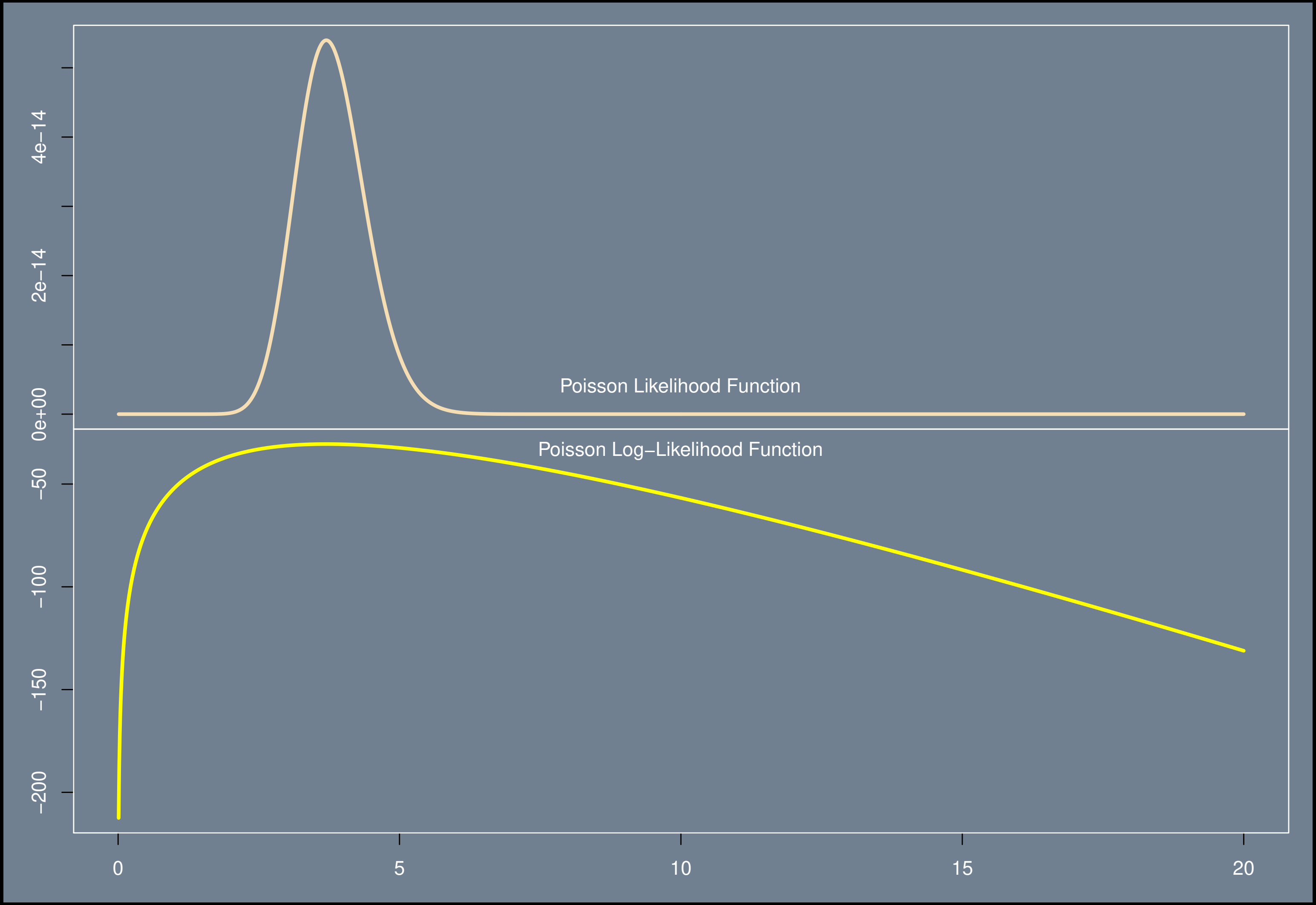
## Poisson Example in R

```
# USE THE R CORE FUNCTION FOR OPTIMIZING, par=STARTING VALUES,
# control=list(fnscale=-1) INDICATES A MAXIMIZATION, bfgs=QUASI-NEWTON ALGORITHM
mle <- optim(par=1,fn=llhfunc,X=y.vals,control=list(fnscale=-1),method="BFGS")

# MAKE A PRETTY GRAPH OF THE LOG AND NON-LOG VERSIONS
ruler <- seq(from=.01, to=20, by= .01)
poison.ll <- llhfunc(y.vals,ruler)
poison.l <- llhfunc(y.vals,ruler,do.log=FALSE)

par(oma=c(3,3,1,1),mar=c(0,0,0,0),mfrow=c(2,1))
plot(ruler,poison.l,col="wheat",type="l",xaxt="n",lwd=3)
text(mean(ruler),mean(poison.l),"Poisson Likelihood Function")
plot(ruler,poison.ll,col="yellow",type="l",lwd=3)
text(mean(ruler),mean(poison.ll)/2,"Poisson Log-Likelihood Function")
```

## Maximum Likelihood Estimation [15]



## Measuring the Uncertainty of the MLE

- The first derivative measures slope and the second derivative measures “curvature” of the function at a given point.
- The more peaked the function is at the MLE, the more “certain” the data are about this estimator.
- The square root of the negative inverse of the expected value of the second derivative is the SE of the MLE.
- In multivariate terms for vector $\boldsymbol{\theta}$, we take the negative inverse of the expected *Hessian*.
- Poisson example:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \frac{d}{d\theta}\left(\frac{d}{d\theta}\ell(\theta|\mathbf{x})\right) = -\theta^{-2}\sum_{i=1}^{n} x_i$$

- The expected value (estimate) of $\theta$ is the MLE, so:

$$SE(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^{n} x_i} = \frac{\bar{\mathbf{x}}^2}{n\bar{\mathbf{x}}} = \frac{\bar{\mathbf{x}}}{n}.$$

## Multivariable MLE

- Now $\boldsymbol{\theta}$ is a vector of coefficients to be estimated (eg. regression).

- The Score Function is:

$$\dot{\ell}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\mathbf{x})$$

  which we use to get the MLE $\hat{\boldsymbol{\theta}}$.

- The Hessian Matrix is:

$$\mathbf{H} = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

  which we use to get the SE of the MLE.

- The information matrix is:

$$\mathbf{I} = -\mathbb{E}(f)\left[ \left. \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\hat{\boldsymbol{\theta}}} \right] \equiv \mathbb{E}(bbbh)\left[ \left. \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}'} \right|_{\hat{\boldsymbol{\theta}}} \right]$$

  where the equivalence of these forms is called the *information equality*.

- The variance-covariance of $\hat{\boldsymbol{\theta}}$ is produced by:

$$\boldsymbol{\Sigma} = \mathbf{I}^{-1}$$

## Normal MLE

- Normal/Gaussian PDF:

$$f(x|\mu,\sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

- Likelihood function for $\mu$ and $\sigma^2$ from joint PDF for the data:

$$L(\mu,\sigma|\mathbf{x}) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right] = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

- Log-likelihood:

$$\ell(\mu,\sigma|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

# Normal MLE

- MLE for $\mu$ ($\sigma^2$ a nuisance parameter):

$$\frac{d}{d\mu}\ell(\mu,\sigma|\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)(2)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \equiv 0$$

$$0 = \sum_{i=1}^{n} x_i - n\mu$$

- So:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\mathbf{x}}.$$

- Again, it is not true that the MLE is always the data mean; the normal is an especially elegant form.

## Normal MLE

- The standard error for $\hat{\mu}$ is:

$$\begin{aligned}\frac{d^2}{d\mu^2}\ell(\mu,\sigma|\mathbf{x}) &= \frac{d}{d\mu}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right)\\ &= \frac{1}{\sigma^2}\frac{d}{d\mu}\left(\sum_{i=1}^{n}x_i-n\mu\right)\\ &= -\frac{n}{\sigma^2} = \mathbf{H}\end{aligned}$$

$$SE(\mu) = \mathbb{E}[-\mathbf{H}^{-1}] = \frac{\sigma^2}{n}.$$

# Normal MLE

- MLE for $\sigma^2$ ($\mu$ a nuisance parameter):

$$\frac{d}{d(\sigma^2)}\ell(\mu,\sigma|\mathbf{x}) = \frac{d}{d(\sigma^2)}\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 \equiv 0$$

$$0 = -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\hat{\sigma}^2 = \boxed{\frac{1}{n}}\sum_{i=1}^{n}(x_i-\mu)^2$$

- This is biased in finite samples by: $\frac{n}{n-1}$

  because of $\boxed{\frac{1}{n}}$.

- Note the difference between the SE of $\hat{\mu}$ and the MLE of $\sigma^2$.

# Normal MLE

- The standard error for $\sigma^2$ is:

$$\frac{d^2}{d(\sigma^2)^2}\ell(\mu,\sigma|\mathbf{x}) = \frac{d}{d(\sigma^2)}\left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$\mathbf{H} = \frac{n}{2}\left(\sigma^2\right)^{-2} - \left(\sigma^2\right)^{-3}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \left(\sigma^2\right)^{-2}\left[\frac{n}{2} - \left(\sigma^2\right)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

$$-\mathbf{H}^{-1} = -\frac{-\sigma^4}{\frac{n}{2} - \left(\sigma^2\right)^{-1}\underbrace{\sum_{i=1}^{n}(x_i-\mu)^2}_{\mathbb{E}[\;]=n\sigma^2}}$$

$$\mathbb{E}_x[-\mathbf{H}^{-1}] = \frac{2\sigma^4}{-n+2(\sigma^2)^{-1}n\sigma^2} = \frac{2\sigma^4}{-n+2n} = \frac{2\sigma^4}{n}$$

# Properties of the MLE (Birnbaum 1962)

- Consistency:

$$\text{plim}\hat{\theta} = \theta.$$

- Asymptotic Normality:

$$\hat{\theta} \underset{a}{\sim} N\left(\theta, I(\theta)^{-1}\right) \quad \text{where } I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right].$$

- Asymptotic Efficiency: no other estimator has lower variance, the variance of the MLE meets the Crámer-Rao Lower Bound.

- Invariance To Reparameterization:

$$\gamma = c(\theta) \implies \hat{\gamma} = c(\hat{\theta}).$$

# Exponential PDF MLE

- Assume: $x_i, i = 1, \ldots, n$ iid with $f(x|\theta) = \frac{1}{\theta}\exp[-x/\theta]$.
- This gives:

$$L(\theta|\mathbf{x}) = \theta^{-n}\exp\left[-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right] \qquad \ell(\theta|\mathbf{x}) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

- Taking the derivative, setting equal to zero, and solving, gives:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \equiv 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

- The curvature for this estimate is:

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \frac{d}{d\theta}\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i\right) = n\theta^{-2} - 2\theta^{-3}\sum_{i=1}^{n} x_i = n\theta^{-2}\left[1 - 2\theta^{-1}\bar{x}\right] = \mathbf{H}.$$

- So the variance of the MLE is given by:

$$-\mathbf{H}^{-1} = -\frac{\theta^2}{n(1 - 2\theta^{-1}\bar{x})} = \frac{\theta^2/n}{2\theta^{-1}\bar{x} - 1} \quad \Longrightarrow \quad \mathbb{E}\left[-\mathbf{H}^{-1}\right] = \frac{\bar{x}^2/n}{2\bar{x}^{-1}\bar{x} - 1} = \frac{\bar{x}^2}{n}.$$

# Maximum Likelihood Estimation for Linear Regression

- Assume: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N(0, \Sigma), \quad \Sigma = \sigma^2\mathbf{I}$.
- Normally we would estimate according to $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- Start with the likelihood function for iid $\mathbf{e}$:

$$L(\mathbf{e}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{e}'\mathbf{e}\right].$$

- Now plug in the estimate of $\mathbf{e}$, $e_i = y_i - \mathbf{X}_i\boldsymbol{\beta}$ (where $\boldsymbol{\beta}$ is yet to be estimated):

$$L(\boldsymbol{\beta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

$$\ell(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

## Maximum Likelihood Estimation for Linear Regression

- Now take the derivative with regard to $\boldsymbol{\beta}$, and set equal to zero:

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) &= \frac{\partial}{\partial\boldsymbol{\beta}}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \\
&= -\frac{1}{2\sigma^2}(-\mathbf{X})'(2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \equiv 0 \\
0 &= \mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
0 &= \mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

## Maximum Likelihood Estimation for Linear Regression

- Now take the derivative with regard to $\boldsymbol{\beta}$, and set equal to zero:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \\
&= -\frac{1}{2\sigma^2}(-\mathbf{X})'(2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \equiv 0 \\
0 &= \mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
0 &= \mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

- Here, $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$ is called the *normal equation*.

## Maximum Likelihood Estimation for Linear Regression

- Now treat $\mu$ as the nuisance parameter:

$$\begin{aligned}
\frac{d}{d\sigma^2}\ell(\sigma^2) &= \frac{d}{d\sigma^2}\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \\
&= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\mathbf{e}'\mathbf{e} \equiv 0 \\
0 &\equiv -n\sigma^2 + \mathbf{e}'\mathbf{e} \\
\hat{\sigma}^2 &= \boxed{\frac{1}{n}}\mathbf{e}'\mathbf{e}.
\end{aligned}$$

- Note that this is biased by:

$$\frac{n}{n-1}$$

because of $\boxed{\frac{1}{n}}$.

## MLEs Are Not Guaranteed To Exist

- Let $X_1, X_2, \ldots, X_n$ be iid from $B(1, p)$, where $p \in (0, 1)$.
- If we get the sample $(0, 0, \ldots, 0)$, then the MLE is obviously $\bar{X} = 0$.
- But this is not an admissible value, so the MLE does not exist.

# Likelihood Problem #1

- What if the likelihood function is flat around the mode?

- This is an indeterminate (and fortunately rare) occurrence.

- We say that the model is “non-identified” because the likelihood function cannot discriminate between alternative MLE values.

- Usually this comes from a model specification that has ambiguities.

## Likelihood Problem #2

- What if the likelihood function has more than one mode?

- Then it is difficult to choose one, even if we had perfect knowledge about the shape of the function.

- This model is identified provided that there is some criteria for picking a mode.

- Usually this comes from complex model specifications, like nonparametrics.

## Root and Mode finding with Newton-Raphson

- Newton's method exploits the properties of a Taylor series expansion around some given point.
- General form (to be derived):

$$x^{(1)} \cong x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})},$$

- The Taylor series expansion gives the relationship between the value of a mathematical function at point, $x_0$, and the function value at another point, $x_1$, given (with continuous derivatives over the relevant support) as:

$$\begin{aligned} f(x_1) = f(x_0) + (x_1 - x_0)f'(x_0) &+ \frac{1}{2!}(x_1 - x_0)^2 f''(x_0) \\ &+ \frac{1}{3!}(x_1 - x_0)^3 f'''(x_0) + \ldots, \end{aligned}$$

  where $f'$ is the first derivative with respect to $x$, $f''$ is the second derivative with respect to $x$, and so on.

## Root and Mode finding with Newton-Raphson

- Note that it is required that $f()$ have continuous derivatives over the relevant support. Infinite precision is achieved with the infinite extending of the series into higher order derivatives and higher order polynomials (of course the factorial component in the denominator means that these are rapidly decreasing increments).

- This process is both unobtainable and unnecessary, and only the first two terms are required as a step in an iterative process.

- The point of interest is $x_1$ such that $f(x_1) = 0$. This value is a root of the function, $f()$ in that it provides a solution to the polynomial expressed by the function.

# Root and Mode finding with Newton-Raphson

- It is also the point where the function crosses the x-axis in a graph of $x$ versus $f(x)$. This point could be found in one step with an infinite Taylor series:

$$\begin{aligned} 0 = f(x_0) + (x_1 - x_0)f'(x_0) + \frac{1}{2!}(x_1 - x_0)^2 f''(x_0) + \ldots \\ + \frac{1}{\infty!}(x_1 - x_0)^{\infty} f^{(\infty)}(x_0) + \ldots. \end{aligned}$$

- While this is impossible, we can use the first two terms to get closer to the desired point:

$$0 \cong f(x_0) + (x_1 - x_0)f'(x_0).$$

- Now rearrange to produce at the $(j+1)^{\text{th}}$ step:

$$x^{(j+1)} \cong x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})},$$

  so that progressively improved estimates are produced until $f(x^{(j+1)})$ is sufficiently close to zero.

- It has been shown that this method converges quadratically to a solution provided that the selected starting point is reasonably close to the solution, although the results can be very bad if this condition is not met.

## Newton-Raphson for Statistical Problems

- The Newton-Raphson algorithm, when applied to mode finding in an MLE statistical setting, substitutes $\beta^{(j+1)}$ for $x^{(j+1)}$ and $\beta^{(j)}$ for $x^{(j)}$ (where the $\beta$ values are iterative estimates of the parameter vector) and $f()$ is the score function.

- For a likelihood function: $L(\boldsymbol{\beta}|\mathbf{X})$, the score function is the first derivative with respect to the parameters of interest:

$$\dot{\ell}(\boldsymbol{\beta}|\mathbf{X}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\mathbf{X}).$$

  Setting $\dot{\ell}(\boldsymbol{\beta}|\mathbf{X})$ equal to zero and solving gives the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$.

- The goal is to estimate a $k$-dimensional $\hat{\boldsymbol{\beta}}$ estimate, given data and a model. The applicable multivariate likelihood updating equation is now provided by:

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} - \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x})\left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x})\right)^{-1}.$$

- Which can be rewritten as:

$$\left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x})\right)(\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x})$$

## Newton-Raphson for Statistical Problems

- When $\boldsymbol{\beta}^{(j)}$ is the maximum likelihood coefficient vector, the quantity in the denominator is the *Hessian*.

- At each step of these Newton-Raphson steps, a system of equations determined by the multivariate normal equations must be solved:

$$\underbrace{\mathbf{A}}_{\text{angle}} \underbrace{(\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})}_{\text{direction: } \delta\boldsymbol{\beta}} = \underbrace{\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x})}_{\text{size of direction: } \mathbf{u}},$$

which builds “linear structure in the parameter vector,” and leads to estimates from the system of linear equations: $\delta\boldsymbol{\beta} = \mathbf{A}^{-1}\mathbf{u}$.

- It is computationally convenient to solve on each iteration by least squares, so that the problem of mode finding reduces to a repeated weighted least squares application in which the inverse of the diagonal values of the second derivative matrix in the denominator are the appropriate weights (this is a diagonal matrix by the iid assumption).

## Newton-Raphson for Statistical Problems

- The Newton-Raphson algorithm when applied to ML mode finding treats the score function as $f()$ to produce *iterative* estimates from the Taylor series:

$$\beta^{(j+1)} = \beta^{(j)} - \frac{\frac{\partial}{\partial\beta}\ell(\beta^{(j)}|\mathbf{y})}{\frac{\partial^2}{\partial\beta\partial\beta}\ell(\beta^{(j)}|\mathbf{y})}.$$

- Generalize this by allowing multiple coefficients, changing the parameter $\beta$ to the vector $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} - \left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})\right)^{-1}\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y}).$$

- Where the Hessian is:

$$\mathbf{H} = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y}).$$

## Newton-Raphson for Statistical Problems

- For exponential family distributions and natural link functions, the observed and expected Hessian matrix are identical (Fahrmeir and Tutz, 1994, p.39; Lehmann and Casella, 1998, pp.124-8).

- So it is common to replace this calculation with forms that are equivalent for the exponential family, such as the Fisher information:

$$\mathbf{A}_{Fisher} = -\mathbb{E}\left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})\right)$$

 (Fisher 1925), or the square of the score function:

$$\mathbf{A}_{BHHH} = \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})'\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})\right].$$

 sometimes called the *BHHH* method (Berndt, Hall, Hall, and & Hausman 1974).

- From now on the slides will just say $\mathbf{A}$, under the assumption it's either of these forms.

## Newton-Raphson for Statistical Problems

- At each step of the Newton-Raphson algorithm there is a system of multivariate normal equations:

$$\underbrace{\mathbf{A}}_{\text{angle}} \underbrace{(\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})}_{\text{direction: } \delta\boldsymbol{\beta}} = \underbrace{\frac{\partial}{\partial \boldsymbol{\beta}^{(j)}} \ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})}_{\text{size of direction: } \mathbf{u}},$$

  builds “linear structure in the parameter vector.”

- This creates a linear system of equations according to:

$$\underset{(p\times 1)}{\delta\boldsymbol{\beta}} = \underset{(p\times p)}{\mathbf{A}^{-1}} \underset{(p\times 1)}{\mathbf{u}}$$

  where we obtain new estimates by iterating over:

$$\begin{aligned} \boldsymbol{\beta}^* &= \boldsymbol{\beta} + \delta\boldsymbol{\beta} \\ &= \boldsymbol{\beta} + \mathbf{A}^{-1}\mathbf{u} \end{aligned}$$

# Newton-Raphson for Statistical Problems

- Given this system of equations, it is computationally convenient to solve on each iteration by *weighted least squares.*
- Therefore the problem of mode finding reduces to a repeated weighted least squares application in which the inverse of the diagonal values of $\mathbf{A}$ are the appropriate weights.
- So we now review weighted least squares. . .

# Review of Weighted Least Squares

- A standard technique for compensating for non-constant error variance in LMs is to insert a diagonal matrix of weights, $\mathbf{\Omega}$, into the calculation of $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ such that the heteroscedasticity is mitigated.

- The $\mathbf{\Omega}$ matrix is created by taking the error variance of the $i^{\text{th}}$ case (estimated or known), $v_i$, and assigning the inverse to the $i^{\text{th}}$ diagonal: $\mathbf{\Omega}_{ii} = \frac{1}{v_i}$. The idea is that large error variances are reduced by multiplication of the reciprocal.

- Starting with $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, observe that there is heteroscedasticity in the error term so: $\boldsymbol{\epsilon}_i = \epsilon v_i$, where the shared (minimum) variance is $\epsilon$ (i.e. non-indexed), and differences are reflected in the $v_i$ term.

## Review of Weighted Least Squares

- Really simple example: a heteroscedastic error vector: $\mathbf{E} = [1, 2, 3, 4]$. Then $\epsilon = 1$, and the $\mathbf{v}$ vector is $[1, 2, 3, 4]$. So by the logic above, the $\boldsymbol{\mu}$ matrix for this example is:

$$\boldsymbol{\Omega} = \begin{bmatrix} \frac{1}{v_1} & 0 & 0 & 0 \\ 0 & \frac{1}{v_2} & 0 & 0 \\ 0 & 0 & \frac{1}{v_3} & 0 \\ 0 & 0 & 0 & \frac{1}{v_4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

- Premultiply each term by the square root of the $\boldsymbol{\Omega}$ matrix (a Cholesky factorization given that $\mathbf{A}$ is a positive definite, but greatly simplified here since $\boldsymbol{\Omega}$ is diagonal): $\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{Y} = \boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\epsilon}$.

- So if the heteroscedasticity in the error term is expressed as the diagonals of a matrix: $\boldsymbol{\epsilon} \sim (0, \sigma^2\mathbf{V})$, then this gives: $\boldsymbol{\epsilon} \sim (0, \boldsymbol{\Omega}\sigma^2\mathbf{V}) = (0, \sigma^2)$, and the heteroscedasticity is “removed.”

- Now instead of minimizing $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, we minimize $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, and the weighted least squares estimator is found by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{Y}$.

# Review of Weighted Least Squares, GLS in R

```
diastolic.pressure.df <-
  read.table("http://people.hmdc.harvard.edu/~jgill/bloodpressure.data",header=FALSE))
dimnames(diastolic.pressure.df)[[2]] <- c("age","pressure")
summary(diastolic.pressure.df)
                  age           pressure
        Min.   :20.00     Min.   : 63.00
        1st Qu.:30.25     1st Qu.: 71.00
        Median :40.00     Median : 77.00
        Mean   :39.57     Mean   : 79.11
        3rd Qu.:49.00     3rd Qu.: 85.75
        Max.   :59.00     Max.   :109.00

attach(diastolic.pressure.df)
unweighted.lm <- lm(pressure~age)
```

# Review of Weighted Least Squares, GLS in R

```
summary(unweighted.lm)
        Residuals:
      Min         1Q    Median         3Q        Max
-16.47859  -5.78765  -0.07844   5.61173  19.78132

Coefficients:
                        Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)             56.15693      3.99367     14.061    < 2e-16
age                      0.58003      0.09695      5.983   2.05e-07
---
Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-Squared: 0.4077,     Adjusted R-squared: 0.3963
F-statistic: 35.79 on 1 and 52 DF,  p-value: 2.05e-07
```

## Review of Weighted Least Squares, GLS in R

```
# REGRESS ABSOLUTE VALUE RESIDUALS ON PREDICTOR -> SD FUNCTION
resid.fit <- lm(abs(unweighted.lm$residuals)~age)
# OBTAIN FITTED VALUES FOR THE WEIGHTS
weights.fit <- 1/(resid.fit$fitted.values)^2
# USE THESE WEIGHTS FOR A GLS REGRESSION
weighted.lm <- lm(pressure~age,weights=weights.fit)
summary(weighted.lm)
        Residuals:
        Min      1Q  Median      3Q     Max
        -2.0230 -0.9939 -0.0327  0.9250  2.2008
Coefficients:
                          Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)               55.56577      2.52092    22.042    < 2e-16
age                       0.59634       0.07924     7.526   7.19e-10
---
Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-Squared: 0.5214,     Adjusted R-squared: 0.5122
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

## Iterative Weighted Least Squares

- Suppose that the individual variances used to make the reciprocal diagonal values for $\boldsymbol{\Omega}$ are unknown and cannot be easily estimated.

- It is known that they are a function of the mean of the outcome variable: $v_i = f(\mathbb{E}[Y_i])$.

- So if the expected value of the outcome variable, $\mathbb{E}[Y_i] = \mu$, and the form of the relation function, $f()$, are known then this is a straightforward estimation procedure.

- Unfortunately, it is not always possible to know the exact form of the relationship between the mean function and the variance structure.

## Iterative Weighted Least Squares

- A solution to this problem is to iteratively estimate the weights, improving the variance estimate on each cycle from the mean function:

$$\text{diag}(\mathbf{\Omega}) = \min_{\boldsymbol{\beta}} \left[ (\mathbf{A}^{-1}\mathbf{u} - \delta\boldsymbol{\beta})'(\mathbf{A}^{-1}\mathbf{u} - \delta\boldsymbol{\beta}) \right]$$

- Recall that:

$$\mathbf{A} = -\mathbb{E}\left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{y})\right) \qquad \mathbf{u} = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^{(j)}|\mathbf{x}) \qquad \delta\boldsymbol{\beta} = (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})$$

- Since $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$, then the coefficient estimate, $\hat{\boldsymbol{\beta}}$, provides a mean estimate and vice versa, and for exponential family forms this process leads to the MLE.

- Furthermore, the matrix produced by: $\hat{\sigma}^2(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}$ converges in probability to the variance matrix of $\hat{\boldsymbol{\beta}}$ as desired (here $\hat{\sigma}^2 = \sum \text{diag}(\mathbf{\Omega})$).

## Derivation of IWLS

- Define the current (or starting) point of the linear predictor by:

$$\underset{(n\times 1)}{\hat{\mathbf{e}}_0} = \underset{(n\times p)(p\times 1)}{\mathbf{X}'\boldsymbol{\beta}_0}$$

with fitted value $\hat{\boldsymbol{\mu}}_0$ from $g^{-1}(\hat{\mathbf{e}}_0)$.

- Form the "adjusted dependent variable" according to:

$$\underset{(n\times 1)}{z_0} = \underset{(n\times 1)}{\hat{\mathbf{e}}_0} + \underset{\text{diag}(n\times n)}{\left(\left.\frac{\partial\eta}{\partial\mu}\right|_{\hat{\boldsymbol{\mu}}_0}\right)} \underset{(n\times 1)}{(\mathbf{y}-\hat{\boldsymbol{\mu}}_0)}$$

which is a linearized form of the link function applied to the data. As an example of this derivative function, the Poisson form looks like:

$$\eta = \log(\mu) \implies \frac{\partial\eta}{\partial\mu} = \frac{1}{\mu}$$

# Derivation of IWLS

- Form the *quadratic weight matrix*, which is the variance of $z$:

$$\underset{(n\times n)}{w_0^{-1}} = \left(\frac{\partial\eta}{\partial\mu}\bigg|_{\hat{\boldsymbol{\mu}}_0}\right)^2 v(\mu)|_{\hat{\boldsymbol{\mu}}_0}$$

where $v(\mu)$ is the variance function: $\frac{\partial}{\partial\theta}b'(\theta) = b''(\theta)$ from McCullagh & Nelder's standard setup.

- This process is necessarily iterative because both $z$ and $w$ depend on the current fitted value, $\boldsymbol{\mu}_0$.

- Recall the following:

$$b'(\theta) = \mu \qquad\qquad b''(\theta) = v(\mu) = \frac{\partial}{\partial\theta}\mu$$

$$\sum_j \boldsymbol{\beta}_j\mathbf{x}_j = \eta \implies \frac{\partial\eta}{\partial\beta_j} = \mathbf{x}_j \qquad\qquad \ell(y_j,\theta) = \frac{y_j\theta - b(\theta)}{\phi} + c(y_j,\phi)$$

with the simplifications: $a(\phi) = \phi$, and no grouping/clustering.

# Derivation of IWLS

- General Scheme:

1. Construct $z$, $w$. Regress $z$ on the covariates with weights to get a new interim estimate:

$$\underset{(p\times 1)}{\hat{\boldsymbol{\beta}}_1} = \left[ \underset{(p\times n)}{\mathbf{X}'} \underset{(n\times n)}{w_0} \underset{(n\times p)}{\mathbf{X}} \right]^{-1} \underset{(p\times n)}{\mathbf{X}'} \underset{(n\times n)}{w_0} \underset{(n\times 1)}{z_0}$$

2. Use the coefficient vector estimate to update the linear predictor:

$$\hat{\mathbf{e}}_1 = \mathbf{X}'\hat{\boldsymbol{\beta}}_1$$

3. Iterate:

$$\begin{aligned} z_1, w_1 &\Longrightarrow \hat{\boldsymbol{\beta}}_2, \hat{\mathbf{e}}_2 & \qquad \hat{\boldsymbol{\beta}}_2, \hat{\mathbf{e}}_2 &\Longrightarrow z_2, w_2 \\ z_2, w_2 &\Longrightarrow \hat{\boldsymbol{\beta}}_3, \hat{\mathbf{e}}_3 & \qquad \hat{\boldsymbol{\beta}}_3, \hat{\mathbf{e}}_3 &\Longrightarrow z_3, w_3 \\ z_3, w_3 &\Longrightarrow \hat{\boldsymbol{\beta}}_4, \hat{\mathbf{e}}_4 & \qquad \hat{\boldsymbol{\beta}}_4, \hat{\mathbf{e}}_4 &\Longrightarrow z_4, w_4 \end{aligned}$$

(and so on...).

## Justification

- First obtain the first derivative WRT $\boldsymbol{\beta}_j$ coefficient starting with the log likelihood in canonical form for a single data point, apply the chain rule:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_j} = \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \boldsymbol{\beta}_j}$$

- Now simplify using the following properties/results:

$$\begin{aligned}\frac{\partial \mu}{\partial \theta} &= \frac{\partial}{\partial \theta} b'(\theta) = v(\mu) \\ \frac{\partial \eta}{\partial \boldsymbol{\beta}_j} &= \mathbf{x}_j \\ \frac{\partial \mu}{\partial \eta} &= w_0 \frac{\partial \eta}{\partial \mu} v(\mu) \\ \frac{\partial \ell}{\partial \theta} &= \frac{y-\mu}{\phi},\end{aligned}$$

which produces:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_j} = \frac{w_0}{\phi}(y-\mu)\frac{\partial \eta}{\partial \mu}\mathbf{x}_j = \mathbf{U}_j$$

# Justification

- This can also be expressed in vector form for the full coefficient vector:

$$\underset{(p\times 1)}{\mathbf{u}} = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\mathbf{y}, \boldsymbol{\beta})$$

- Use Fisher scoring with gradient vector and the negative expected value of the Hessian, for example for coefficients $r$ and $s$:

$$\mathbf{A}_{rs} = -\mathbb{E}\left( \frac{\partial^2}{\partial \boldsymbol{\beta}_r \partial \boldsymbol{\beta}'_s} \ell(\mathbf{y}, \boldsymbol{\beta}) \right).$$

This is the value in the $r^{th}$ row and $s^{th}$ column.

## Justification

- What does $\mathbf{A}$ look like?

$$\mathbf{A}_{rs} = -E\frac{\partial \mathbf{u}_r}{\partial \boldsymbol{\beta}_s} = -E\frac{\partial}{\partial \boldsymbol{\beta}_s}\left[\sum_{i=1}^{n} w_i(y-\mu)\frac{\partial \eta}{\partial \mu}\mathbf{x}_r\right]$$

$$= -E\sum_{i=1}^{n}\left[(y-\mu)\frac{\partial}{\partial \boldsymbol{\beta}_s}\left(w_i\frac{\partial \eta}{\partial \mu}\mathbf{x}_r\right) + \left(w_i\frac{\partial \eta}{\partial \mu}\mathbf{x}_r\right)\frac{\partial}{\partial \boldsymbol{\beta}_s}(y-\mu)\right]$$

$$= -\sum_{i=1}^{n}\left[0+\left(w_i\frac{\partial \eta}{\partial \mu}\mathbf{x}_r\right)\frac{\partial}{\partial \boldsymbol{\beta}_s}(-\mu)\right] = \sum_{i=1}^{n} w_i\mathbf{x}_r\frac{\partial \eta}{\partial \mu}\frac{\partial \mu}{\partial \boldsymbol{\beta}_s} = \sum_{i=1}^{n} w_i\mathbf{x}_r\frac{\partial \eta}{\partial \boldsymbol{\beta}_s} = \sum_{i=1}^{n} w_i\mathbf{x}_r\mathbf{x}_s$$

- So the full $\mathbf{A}$ matrix is a weighted sums of products matrix of the covariates, in matrix form:

$$\mathbf{A} = \mathbf{X}'w\mathbf{X}$$

and if we express $\mathbf{U}$ in matrix form as:

$$\mathbf{U} = \mathbf{X}'w\mathbf{y}^*$$

(using $\mathbf{U} = A\delta\boldsymbol{\beta}$ and $\mathbf{y}^* = \mathbf{X}\delta\boldsymbol{\beta}$) then the weighted LM form is apparent.