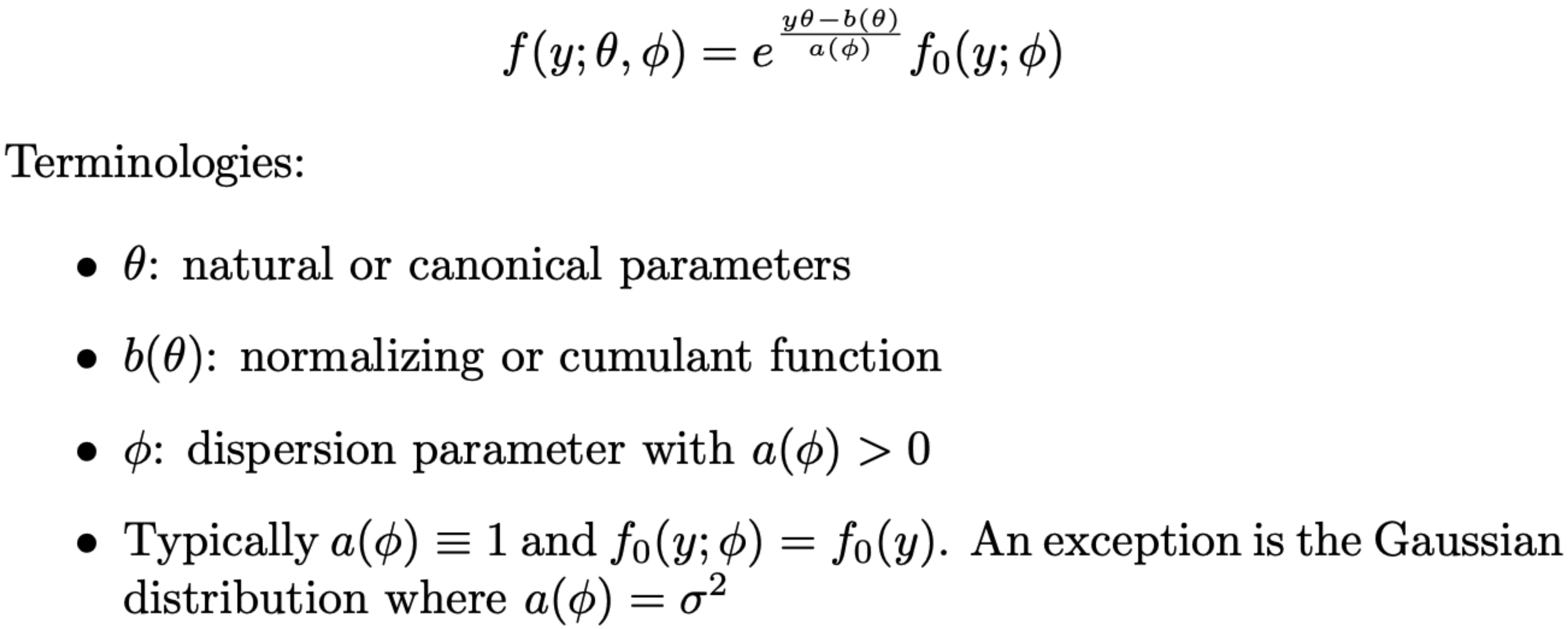
Lecture 2
Exponential dispersion family and GLM

Today’s topics:

- The exponential dispersion family
- Exponential family distribution for GLM
- Likelihood score equations for parameter estimation
- Reading: Agresti Chapters 4.1-4.2, Faraway Chapter 8.1-8.2

The exponential dispersion family

- A random variable Y follows an exponential dispersion family distribution and has the density $f(y;\theta,\phi)$ of the form



$$f(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)}} f_0(y;\phi)$$

Terminologies:

- θ: natural or canonical parameters
- $b(\theta)$: normalizing or cumulant function
- ϕ: dispersion parameter with $a(\phi) > 0$
- Typically $a(\phi) \equiv 1$ and $f_0(y;\phi) = f_0(y)$. An exception is the Gaussian distribution where $a(\phi) = \sigma^2$

- “density” here includes the possibility of discrete atoms.
- Above definition is not the most general form of the exponential family distribution

Some well-known examples

- Normal distribution for continuous data

$$f(y;\mu,\sigma) = e^{\frac{y\mu-\mu^2/2}{\sigma^2}}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}\right]$$

Compare with the general form of exponential dispersion family

- $\theta = \mu, b(\theta) = \theta^2/2, a(\phi) = \sigma^2$
- $f_0(y;\phi) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}$ Gaussian density of $N(0,\sigma^2)$

- Mean: $\mu = \theta = b'(\theta)$
- Variance: $\sigma^2 = b''(\theta)a(\phi)$

Some well-known examples

- Bernoulli distribution for binary data

$$f(y;p) = p^y(1-p)^{1-y} = e^{y\log\frac{p}{1-p} + \log(1-p)}$$
$$= e^{y\theta - \log[1+e^\theta]}$$

  - $\theta = \log(\frac{p}{1-p})$, $b(\theta) = \log[1 + e^\theta]$, $a(\phi) = 1$
  - $f_0(y;\phi) = 1$

  - Mean: $\mu = p = \frac{e^\theta}{1+e^\theta} = b'(\theta)$
  - Variance: $\sigma^2 = p(1-p) = \frac{e^\theta}{(1+e^\theta)^2} = b''(\theta)a(\phi)$

Some well-known examples

- Binomial distribution for counts data

$$
\begin{aligned}
f(y;p,n) = \binom{n}{y} p^y (1-p)^{n-y} &= e^{y \log \frac{p}{1-p} + n \log(1-p)} \binom{n}{y} \\
&= e^{y\theta - n \log[1+e^\theta]} \binom{n}{y}
\end{aligned}
$$

  - $\theta = \log(\frac{p}{1-p})$, $b(\theta) = n\log[1+ \ e^\theta]$, $a(\phi) = 1$
  - $f_0(y;\phi) = \binom{n}{y}$

  - Mean: $\mu = np = n\frac{e^\theta}{1+e^\theta} = b'(\theta)$
  - Variance: $\sigma^2 = np(1-p) = \frac{ne^\theta}{(1+e^\theta)^2} = b''(\theta)a(\phi)$

Some well-known examples

- Poisson distribution for counts data

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = e^{y\log\lambda-\lambda}\frac{1}{y!} = e^{y\theta-e^\theta}\frac{1}{y!}$$

  - $\theta = \log(\lambda), b(\theta) = e^\theta, a(\phi) = 1$
  - $f_0(y;\phi) = \frac{1}{y!}$

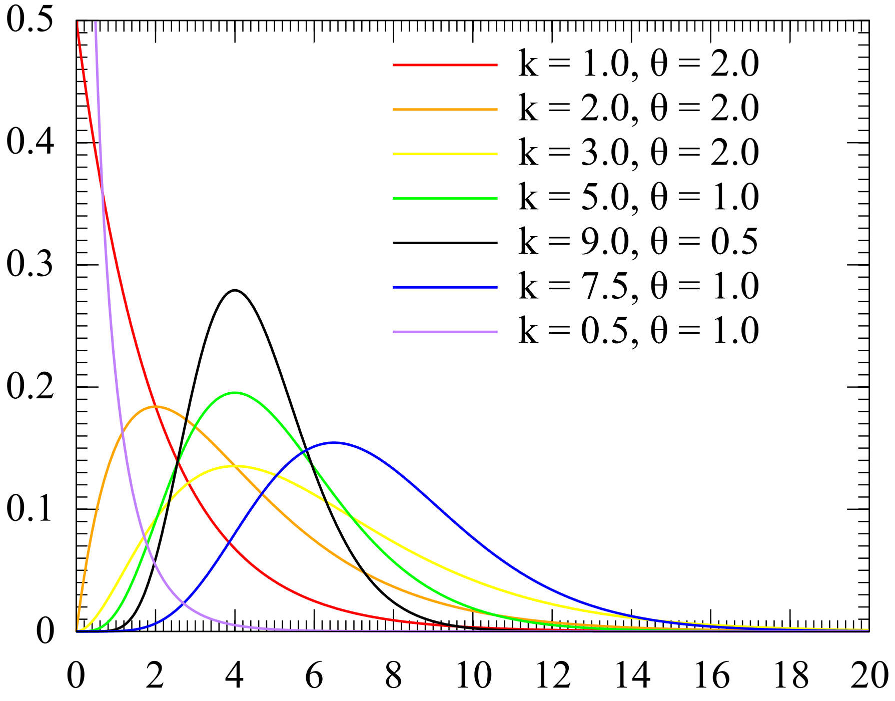
  - Mean: $\mu = \lambda = e^\theta = b'(\theta)$
  - Variance: $\sigma^2 = \lambda = e^\theta = b''(\theta)a(\phi)$

Some additional examples

- Gamma distribution for positive real-valued data

$$f(y;k,\theta) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} e^{-y/\theta}$$

$$= e^{\frac{-\frac{1}{k\theta} y + \log\left(\frac{1}{k\theta}\right)}{1/k}} \frac{y^{k-1} k^k}{\Gamma(k)}$$



  - Canonical parameter $\tilde{\theta} = -\frac{1}{k\theta}, b(\tilde{\theta}) = \log(-\tilde{\theta})$
  - $a(\phi) = 1/k$
  - $f_0(y;\phi) = \frac{y^{k-1}k^k}{\Gamma(k)}$

  - Mean: $\mu = k\theta = -\frac{1}{\tilde{\theta}} = b'(\tilde{\theta})$
  - Variance: $\sigma^2 = \ k\theta^2 = \frac{\mu^2}{k} = \frac{a(\phi)}{\tilde{\theta}^2} = b''(\tilde{\theta})a(\phi)$

Moment relationships

- The exponential family has some special properties that can make our calculation easier
  - Calculate mean and variance of Y

$$\mu = \mathbb{E}(y) = b'(\theta)$$

$$V_\theta = \mathrm{Var}(y) = b''(\theta)a(\phi)$$

  - Why? As $\int f(y;\theta,\phi)dy = 1$, we have

$$e^{b(\theta)/a(\phi)} = \int e^{y\theta/a(\phi)} f_0(y;\phi)dy$$

    - Take derivatives with respect to θ

Moment relationships

- The exponential family has some special properties that can make our calculation easier
    - Calculate mean and variance of Y

$$\mu = \mathbb{E}(Y) = b'(\theta)$$

$$V_\theta = \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

- The above relationship also indicates that

$$\frac{\partial \mu}{\partial \theta} = \frac{\mathrm{Var}(Y)}{a(\phi)} > 0$$

- Mapping from θ to μ is one to one increasing

Exponential family distribution for GLM

- Assume that each observation y_i follows an exponential family with the canonical parameter θ_i and a shared dispersion parameter ϕ

- $\mu_i = \mathbb{E}(y_i)$ is a function of X_i defined by a pre-specified link function

$$g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$$

  - Because of one-to-one mapping, θ_i is also a function of X_i

As a special link function for exponential families, we define

- Canonical link function:
  Define the transformation function $g(\cdot)$ so that:

$$g(\mu_i) = \theta_i = X_i^T \beta$$

Canonical link function examples

- Gaussian: $\theta_i = \mu_i = X_i^T \beta$

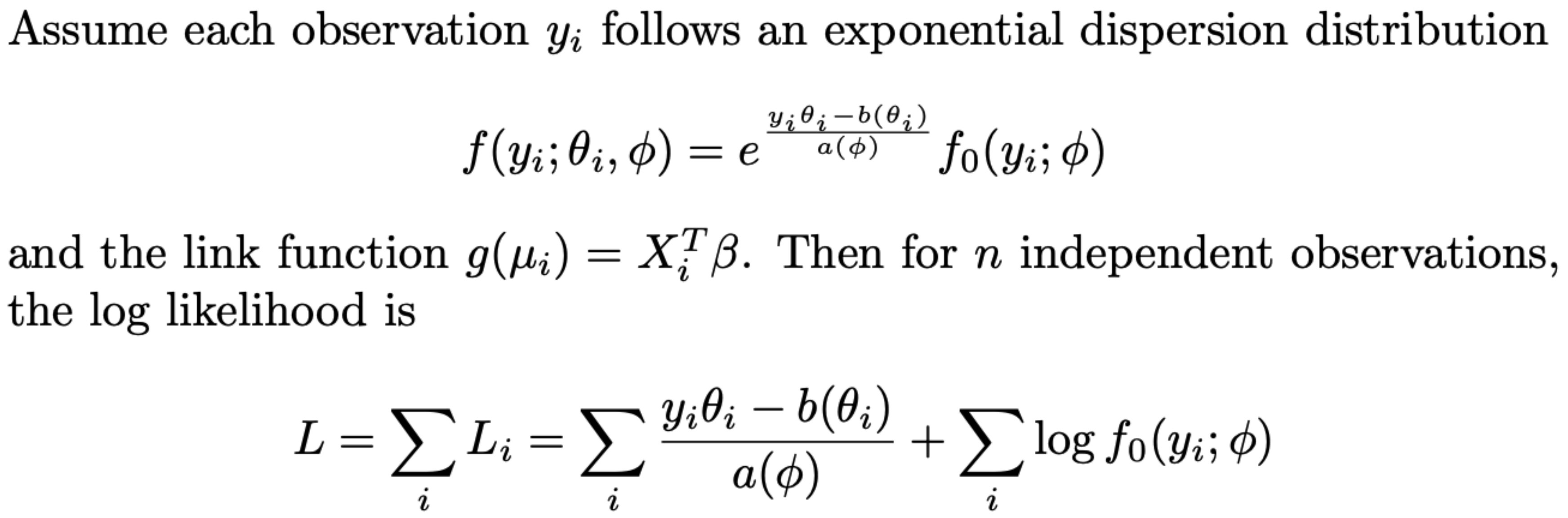
- Binomial and Bernoulli distribution: $\theta_i = \log(\frac{p_i}{1-p_i}) = X_i^T \beta$
  - Called the logit function

- Poisson distribution: $\theta_i = \log(\mu_i) = X_i^T \beta$

- Why do we use the canonical link?
  - The canonical parameter θ always have an unrestrictive support
  - Computational convenience (see later)
  - Easy interpretation

Likelihood score equations

- We now use the maximum likelihood method to solve for the GLM and estimate β



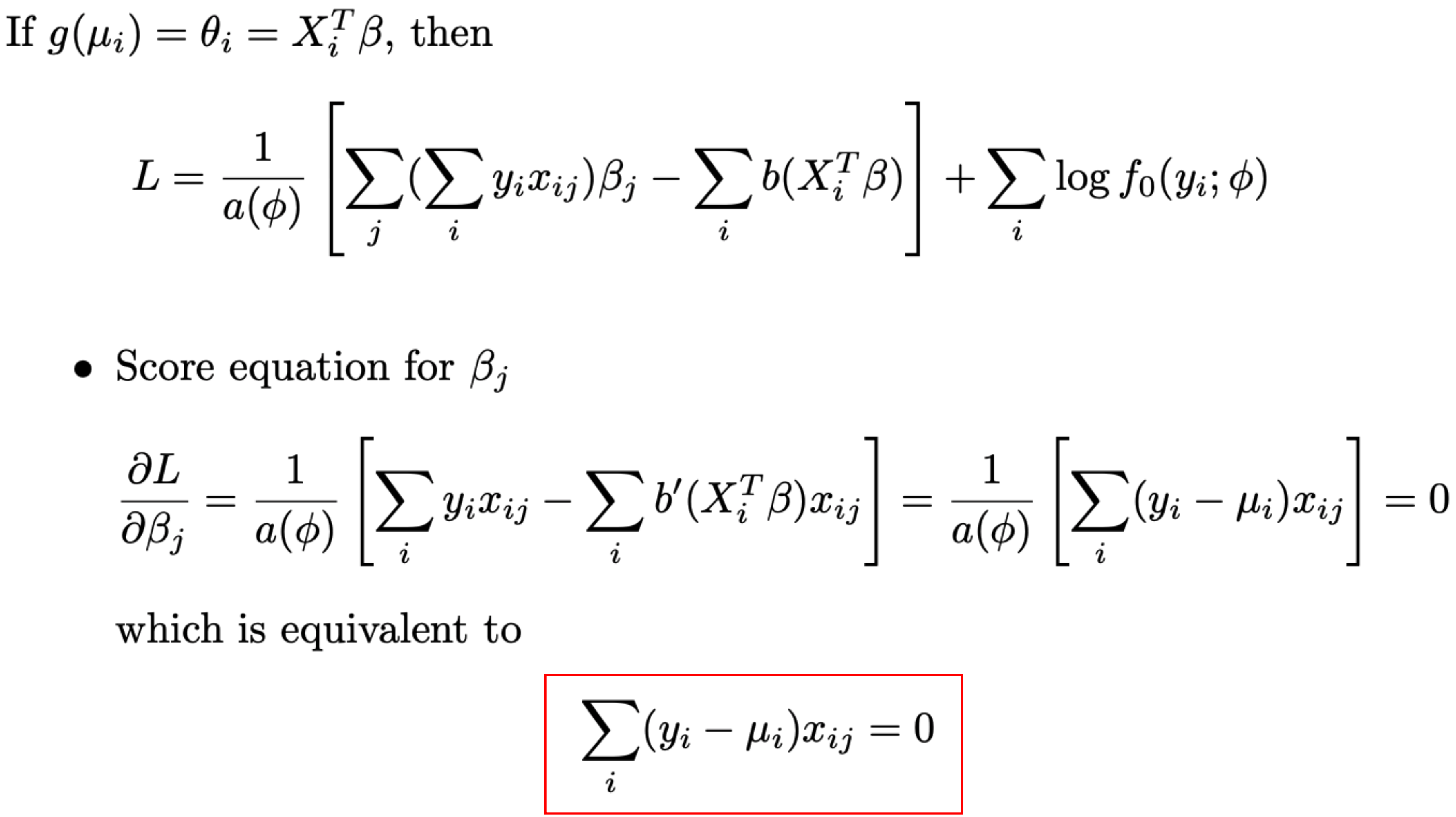
Assume each observation y_i follows an exponential dispersion distribution

$$f(y_i;\theta_i,\phi) = e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}} f_0(y_i;\phi)$$

and the link function $g(\mu_i) = X_i^T\beta$. Then for n independent observations, the log likelihood is

$$L = \sum_i L_i = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i \log f_0(y_i;\phi)$$

Likelihood score equation for the canonical link



If $g(\mu_i) = \theta_i = X_i^T \beta$, then

$$L = \frac{1}{a(\phi)} \left[\sum_j (\sum_i y_i x_{ij}) \beta_j - \sum_i b(X_i^T \beta) \right] + \sum_i \log f_0(y_i; \phi)$$

- Score equation for β_j

$$\frac{\partial L}{\partial \beta_j} = \frac{1}{a(\phi)} \left[\sum_i y_i x_{ij} - \sum_i b'(X_i^T \beta) x_{ij} \right] = \frac{1}{a(\phi)} \left[\sum_i (y_i - \mu_i) x_{ij} \right] = 0$$

which is equivalent to

$$\boxed{\sum_i (y_i - \mu_i) x_{ij} = 0}$$

Likelihood score equation for the canonical link

- Examples

Gaussian model:

$$\sum_i (y_i - X_i^T \beta) x_{ij} = 0$$

Poisson model:

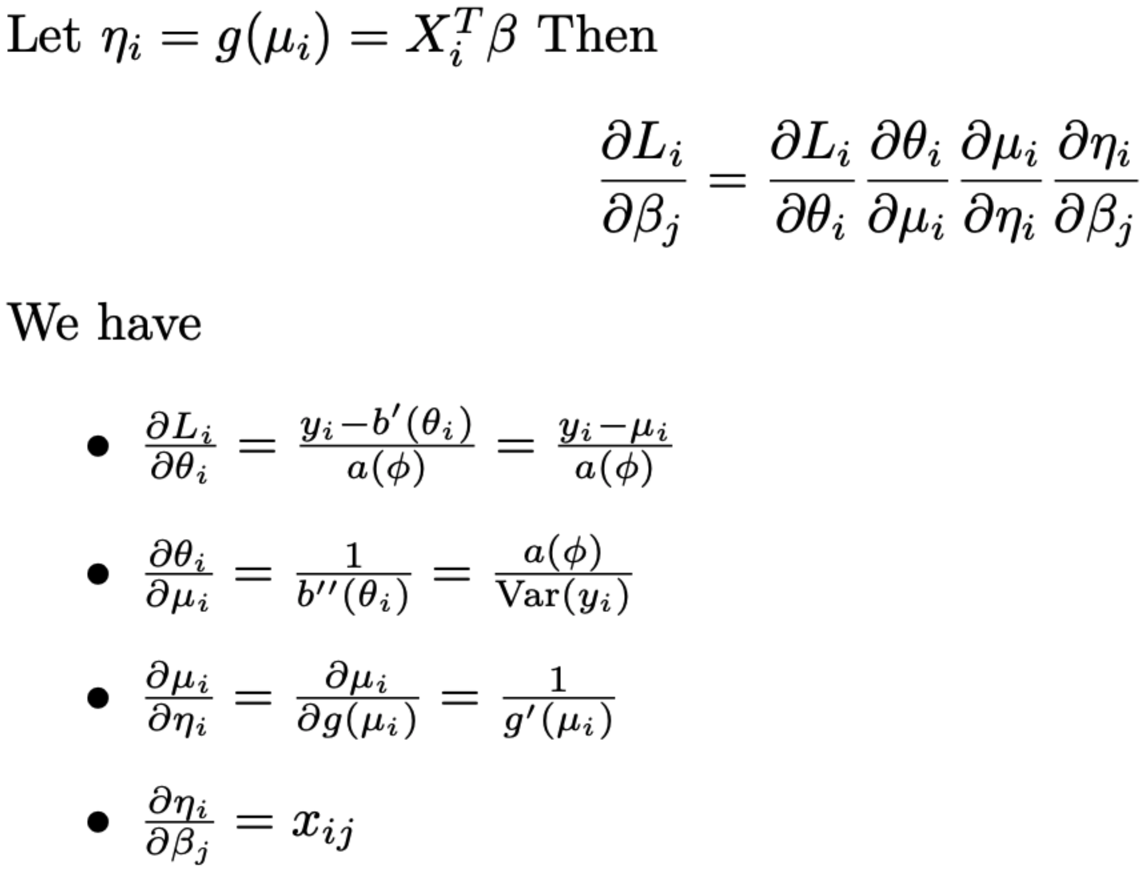
$$\sum_i (y_i - e^{X_i^T \beta}) x_{ij} = 0$$

- L is a concave function of $\beta = (\beta_1, \cdots, \beta_p)$

$$\frac{\partial}{\partial \beta} \left[\sum_i (y_i - \mu_i) X_i \right] = -\sum_i \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta} X_i^T = -\sum_i \frac{\mathrm{Var}(y_i)}{a(\phi)} X_i X_i^T \prec 0$$

  - Easy optimization to find the solution (will discuss computation later)

Likelihood score equation for a general link



Let $\eta_i = g(\mu_i) = X_i^T \beta$ Then

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

We have

- $\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{\mathrm{Var}(y_i)}$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial g(\mu_i)} = \frac{1}{g'(\mu_i)}$
- $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Likelihood score equation for a general link

- The score equations can be written as

$$\boxed{\frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(y_i)} \frac{1}{g'(\mu_i)} = 0}$$

  - μ_i and Var(y_i) are both functions of $\beta = (\beta_1, \cdots, \beta_p)$
  - The score equations only depend on the mean and variance of y_i
  - Matrix form of the score equation:

$$\dot{L}(\beta) = X^T D V^{-1}(y - \mu) = 0$$

where $V = \mathrm{diag}(\mathrm{Var}(y_1), \cdots, \mathrm{Var}(y_n))$ and $D = \mathrm{diag}(g'(\mu_1), \cdots, g'(\mu_n))^{-1}$,
$y = (y_1, \cdots, y_n)$ and $\mu = (\mu_1, \cdots, \mu_n)$.

  - L is not necessarily a concave function of β

Likelihood score equation for a general link

Special cases

- If the link function is the canonical link, then $D = \frac{1}{a(\phi)} V$, thus the score equation becomes

  $$\frac{1}{a(\phi)} X^T (y - \mu) = 0$$

  the same as we derived earlier

- If we assume that $g(\mu_i) = \mu_i = X_i^T \beta$, then the estimating (score) equation becomes

  $$\sum_i \frac{(y_i - X_i^T \beta) X_i}{\mathrm{Var}(y_i)} = 0$$

  which looks like weighted least square (difference: weights can depend on β)