Exponential Families

version: 2020-02-04 · 07:41:59

Exponential Families

- **Definition:** If $\boldsymbol{Y} \sim f(\boldsymbol{y} \,|\, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$, then $\boldsymbol{Y}$ is said to belong to an exponential family if

$$f(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \exp\left\{ \boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta}) \right\} \cdot h(\boldsymbol{y}),$$

  where

  - $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^d$ are the *natural parameters*.

    ($\boldsymbol{\eta}$ must have the same dimension as $\boldsymbol{\theta}$ for upcoming results to hold.)

  - $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{y})$ are the *sufficient statistics*.
  - $\Psi(\boldsymbol{\eta})$ is called the log-partition function, or sometimes the cumulant-generating function.

- **Natural Parametrization:** Since each value of $\boldsymbol{\theta}$ defines a different PDF, $\boldsymbol{\eta}(\boldsymbol{\theta})$ *must* be a bijection. Therefore, we might as well parametrize the exponential family by $\boldsymbol{\eta}$, in which case $f(\boldsymbol{y} \,|\, \boldsymbol{\eta})$ is said to be in its *canonical form*.

Examples

Binomial Distribution

$$Y \sim \text{Binomial}(n, \rho) \implies$$

$$\begin{aligned} p(y \mid \rho) &= \binom{n}{y} \rho^y (1-\rho)^{n-y} \\ &= \exp\left\{ y \cdot \underbrace{\log\left(\frac{\rho}{1-\rho}\right)}_{\eta} - \underbrace{[-n \log(1-\rho)]}_{\Psi(\eta)} \right\} \cdot \binom{n}{y} \end{aligned}$$

Examples

Multivariate Normal Distribution

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies$$

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) - \frac{1}{2} \log|\boldsymbol{\Sigma}| \right\} \cdot \underbrace{h(\boldsymbol{y})}_{(2\pi)^{d/2}}$$

$$= \exp\left\{ -\frac{1}{2}\Big[\underbrace{\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\boldsymbol{y}')}_{\operatorname{vec}(\boldsymbol{\Sigma}^{-1})' \operatorname{vec}(\boldsymbol{y}\boldsymbol{y}')} - 2\boldsymbol{y}'[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}] + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \log|\boldsymbol{\Sigma}| \Big] \right\} h(\boldsymbol{y})$$

$$\implies$$

$$\boldsymbol{T} = (-\tfrac{1}{2}\boldsymbol{y}\boldsymbol{y}', \boldsymbol{y}), \qquad \boldsymbol{\eta} = (\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), \qquad \Psi(\boldsymbol{\eta}) = -\frac{1}{2}(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \log|\boldsymbol{\Sigma}|).$$

(Some redundancy since $\boldsymbol{y}\boldsymbol{y}'$ and $\boldsymbol{\Sigma}^{-1}$ are symmetric matrices, but formulas get complicated)

Examples

- **Model:** $\boldsymbol{Y} \sim f(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y}), \quad \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{y}).$
- **Exponential families:**

  Poisson, Gamma (and Exponential), Multinomial (and Binomial), Negative-Binomial (and Geometric), Dirichlet (and Beta), Wishart (and Chi-Square).

Examples

- **Model:** $\boldsymbol{Y} \sim f(\boldsymbol{y} \,|\, \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y}), \quad \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{y})$.
- **Exponential families:**

  Poisson, Gamma (and Exponential), Multinomial (and Binomial), Negative-Binomial (and Geometric), Dirichlet (and Beta), Wishart (and Chi-Square).
- **Not Exponential families:**

  Student-t (and Cauchy), Weibull, Unif$(0, \theta)$.

Moments of Sufficient Statistics

- **Exponential Family:** $\boldsymbol{Y} \sim f(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y}), \quad \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{y})$.
- **Expectation of $\boldsymbol{T}$**:

$$
\begin{aligned}
\text{(since RHS is a PDF)} \qquad 1 &= \int \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
\text{(take } \tfrac{\partial}{\partial \boldsymbol{\eta}} \text{ on both sides)} \qquad \boldsymbol{0} &= \frac{\partial}{\partial \boldsymbol{\eta}} \int \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \boldsymbol{\eta}} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&= \int [\boldsymbol{T} - \nabla\Psi(\boldsymbol{\eta})] f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y}
\end{aligned}
$$

$$
\underbrace{\int \boldsymbol{T} \cdot f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y}}_{=E[\boldsymbol{T} \mid \boldsymbol{\eta}]} = \nabla\Psi(\boldsymbol{\eta}) \underbrace{\int f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y}}_{=1}
$$

$\implies E[\boldsymbol{T} \mid \boldsymbol{\eta}] = \nabla\Psi(\boldsymbol{\eta})$.

Moments of Sufficient Statistics

- **Exponential Family:** $\boldsymbol{Y} \sim f(\boldsymbol{y} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y}), \quad \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{y})$.
- **Variance of $\boldsymbol{T}$**:

$$
\begin{aligned}
1 &= \int \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
\boldsymbol{0} &= \frac{\partial}{\partial \boldsymbol{\eta}} \int \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&= \int [\boldsymbol{T} - \nabla\Psi(\boldsymbol{\eta})] f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y} \\
\text{(take } \tfrac{\partial}{\partial \boldsymbol{\eta}} \text{ on both sides again)} \qquad &= \int \frac{\partial}{\partial \boldsymbol{\eta}} [\boldsymbol{T} - \nabla\Psi(\boldsymbol{\eta})] f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y} \\
\nabla^2\Psi(\boldsymbol{\eta}) \int f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y} &= \int [\boldsymbol{T} - \underbrace{\nabla\Psi(\boldsymbol{\eta})}_{E[\boldsymbol{T} \mid \boldsymbol{\eta}]}][\boldsymbol{T} - \nabla\Psi(\boldsymbol{\eta})]' f(\boldsymbol{y} \mid \boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y}
\end{aligned}
$$

$\implies \operatorname{var}[\boldsymbol{T} \mid \boldsymbol{\eta}] = \nabla^2\Psi(\boldsymbol{\eta}) \implies \nabla^2\Psi(\boldsymbol{\eta})$ is positive definite.

Inference

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, where $\bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{Y}_i)$.

Inference

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, where $\bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{Y}_i)$.
- **Score function:** $\nabla\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}} - \nabla\Psi(\boldsymbol{\eta})]$

  $\implies$ MLE satisfies $\nabla\Psi(\hat{\boldsymbol{\eta}}) = \bar{\boldsymbol{T}}$.

Inference

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, where $\bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{Y}_i)$.
- **Score function:** $\nabla\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}} - \nabla\Psi(\boldsymbol{\eta})]$

  $\implies$ MLE satisfies $\nabla\Psi(\hat{\boldsymbol{\eta}}) = \bar{\boldsymbol{T}}$.
- **Expected Fisher Information:**

$$\mathcal{I}(\boldsymbol{\eta}) = E[-\nabla^2\ell(\boldsymbol{\eta} \mid \boldsymbol{Y})] = n\,E[\nabla^2\Psi(\boldsymbol{\eta})] = n\nabla^2\Psi(\boldsymbol{\eta}).$$

  $\implies$ Asymptotic theory $\hat{\boldsymbol{\eta}} \approx \mathcal{N}(\boldsymbol{\eta}_0, \mathcal{I}(\boldsymbol{\eta}_0)^{-1})$ is more effectively applied in practice since Observed Fisher Information is $\hat{\mathcal{I}} = \mathcal{I}(\hat{\boldsymbol{\eta}}) = n\nabla^2\Psi(\boldsymbol{\eta})$.

  (usually expectation can't be calculated analytically)

Inference

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, where $\bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{T}(\boldsymbol{Y}_i)$.
- **Score function:** $\nabla\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}} - \nabla\Psi(\boldsymbol{\eta})]$

  $\implies$ MLE satisfies $\nabla\Psi(\hat{\boldsymbol{\eta}}) = \bar{\boldsymbol{T}}$.

- **Expected Fisher Information:**

$$\mathcal{I}(\boldsymbol{\eta}) = E[-\nabla^2\ell(\boldsymbol{\eta} \mid \boldsymbol{Y})] = n\, E[\nabla^2\Psi(\boldsymbol{\eta})] = n\nabla^2\Psi(\boldsymbol{\eta}).$$

  $\implies$ Asymptotic theory $\hat{\boldsymbol{\eta}} \approx \mathcal{N}(\boldsymbol{\eta}_0, \mathcal{I}(\boldsymbol{\eta}_0)^{-1})$ is more effectively applied in practice since Observed Fisher Information is $\hat{\mathcal{I}} = \mathcal{I}(\hat{\boldsymbol{\eta}}) = n\nabla^2\Psi(\boldsymbol{\eta})$.

  (usually expectation can't be calculated analytically)

- **Question:** How to compute MLE $\hat{\boldsymbol{\eta}}$?

Newton-Raphson Method

- **Problem:** Find a minimum of $f : \mathbb{R}^d \to \mathbb{R}$.
- **Quadratic case:** $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}'\boldsymbol{x} + c$, with $\boldsymbol{A}_{dxd}$ is positive definite.

  (Using Cholesky $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$, show that $\boldsymbol{A}^{-1}$ exists and is +ve definite)

  - *Multivariate complete-the-square:*

$$
\begin{aligned}
f(\boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - 2\underbrace{\boldsymbol{b}'\boldsymbol{A}^{-1}}_{\boldsymbol{\mu}'}\boldsymbol{A}\boldsymbol{x} + c \\
&= \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}_{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{\mu}'\boldsymbol{x} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}} - \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + c,
\end{aligned}
$$

$\implies$ Unique minimum of $f(\boldsymbol{x})$ is $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.

Newton-Raphson Method

- **Problem:** Find a minimum of $f : \mathbb{R}^d \to \mathbb{R}$.
- **Non-Quadratic case:** Iterative method.
  - *Initial guess*: $\boldsymbol{x}_0$
  - *Iterations:* At step $n+1$, find 2nd order Taylor expansion of $f(\boldsymbol{x})$ around $\boldsymbol{x} = \boldsymbol{x}_n$:

$$
\begin{aligned}
f(\boldsymbol{x}) &\approx f(\boldsymbol{x}_n) + \underbrace{\boldsymbol{g}_n'}_{\nabla f(\boldsymbol{x}_n)'}(\boldsymbol{x} - \boldsymbol{x}_n) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_n)' \underbrace{\boldsymbol{H}_n}_{\nabla^2 f(\boldsymbol{x}_n)}(\boldsymbol{x} - \boldsymbol{x}_n) \\
&= \frac{1}{2}\Big[\boldsymbol{x}'\boldsymbol{H}_n\boldsymbol{x} - 2(\boldsymbol{H}_n\boldsymbol{x}_n - \boldsymbol{g}_n)'\boldsymbol{x}\Big] + \text{const} \\
&= \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{H}_n(\boldsymbol{x} - \boldsymbol{\mu}) + \text{const}, \qquad \boldsymbol{\mu} = \boldsymbol{H}_n^{-1}(-\boldsymbol{g}_n + \boldsymbol{H}_n\boldsymbol{x}_n) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\;\; = \boldsymbol{x}_n - \boldsymbol{H}_n^{-1}\boldsymbol{g}_n.
\end{aligned}
$$

$\implies$ Let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \boldsymbol{H}_n^{-1}\boldsymbol{g}_n = \boldsymbol{x}_n - [\nabla^2 f(\boldsymbol{x}_n)]^{-1}\nabla f(\boldsymbol{x}_n)$. where typically $\frac{1}{10} \le C \le 1$ (compromise between relative and absolute error).

Newton-Raphson Method

- **Problem:** Find a minimum of $f : \mathbb{R}^d \to \mathbb{R}$.
- **Non-Quadratic case:** Iterative method.
  - *Initial guess*: $\boldsymbol{x}_0$
  - *Iterations:* $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - [\nabla^2 f(\boldsymbol{x}_n)]^{-1} \nabla f(\boldsymbol{x}_n)$.
  - *Stopping Condition:* Algorithm terminates when $N_{\max}$ steps have been reached (perhaps without convergence), or when

$$\max_{1 \le i \le d} \frac{|x_{n,i} - x_{n-1,i}|}{C + |x_{n,i} + x_{n-1,i}|} < \varepsilon,$$

where typically $\frac{1}{10} \le C \le 1$ (compromise between relative and absolute error).

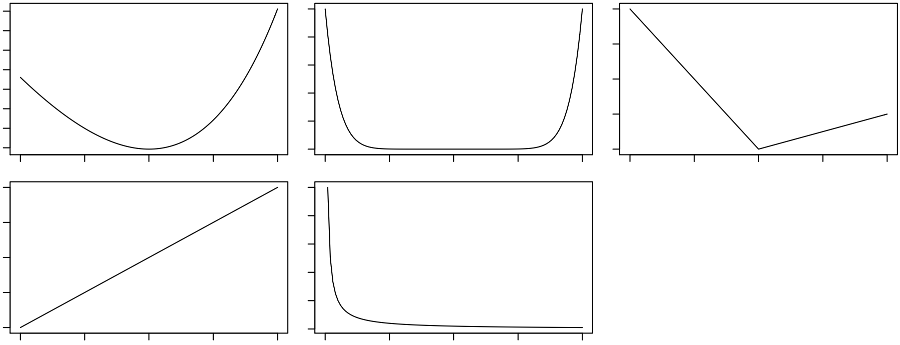
Convex Functions

Newton-Raphson algorithm fails in all sorts of situations, but works relatively well when $f(\boldsymbol{x})$ is a convex function:

$$f(\rho \cdot \boldsymbol{x}_1 + (1-\rho) \cdot \boldsymbol{x}_2) \;\le\; \rho \cdot f(\boldsymbol{x}_1) + (1-\rho) \cdot f(\boldsymbol{x}_2),$$

$\forall\, x_1, x_2$ and $\rho \in (0, 1)$.

$f(\boldsymbol{x})$ is strictly convex if "$\le$" is replaced by "$<$". Examples of convex functions:



Convex Functions

- **Definition:** $f(\rho \cdot \boldsymbol{x}_1 + (1-\rho) \cdot \boldsymbol{x}_2) \;\le\; \rho \cdot f(\boldsymbol{x}_1) + (1-\rho) \cdot f(\boldsymbol{x}_2)$,

  $\forall\, \boldsymbol{x}_1, \boldsymbol{x}_2$ and $\rho \in (0,1)$. Strictly convex if "$\le$" is replaced by "$<$".

- **Properties:**

  1. If $\nabla^2 f(\boldsymbol{x})$ is positive definite then $f(\boldsymbol{x})$ is strictly convex.

  2. Sum of convex functions is convex.

  3. f, g convex and $\nabla g(\boldsymbol{x}) \ge 0 \implies h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ convex.

  4. $f(\boldsymbol{x})$ (strictly) convex $\implies f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ (strictly) convex.

  5. If $f(\boldsymbol{x})$ is convex and $\boldsymbol{x}_0$ is a local minimum of f, then x_0 is a global minimum.

  6. If $f(\boldsymbol{x})$ is strictly convex, then it has either a unique global minimum or no minimum at all.

Convex Functions

Application to Exponential Families

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\} h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, $\quad \bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{Y}_i)$.
- **Expected Fisher-Information:** If $\boldsymbol{\eta}$ is the true parameter value, then

$$\mathcal{I}(\boldsymbol{\eta}) = -\nabla^2 \ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n\nabla^2 \Psi(\boldsymbol{\eta}) = \text{var}(\boldsymbol{T} \mid \boldsymbol{\eta})^{-1}.$$

Therefore:

Convex Functions

Application to Exponential Families

- **Data:** $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \exp\{\boldsymbol{T}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\}h(\boldsymbol{y})$.
- **Loglikelihood:** $\ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n[\bar{\boldsymbol{T}}'\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})]$, $\quad \bar{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{T}(\boldsymbol{Y}_i)$.
- **Expected Fisher-Information:** If $\boldsymbol{\eta}$ is the true parameter value, then

$$\mathcal{I}(\boldsymbol{\eta}) = -\nabla^2 \ell(\boldsymbol{\eta} \mid \boldsymbol{Y}) = n\nabla^2\Psi(\boldsymbol{\eta}) = \text{var}(\boldsymbol{T} \mid \boldsymbol{\eta})^{-1}.$$

Therefore:

- $-\ell(\boldsymbol{\eta} \mid \boldsymbol{Y})$ is a strictly convex function.
- If the MLE $\hat{\boldsymbol{\eta}}$ exists, then it is unique.
- Newton-Raphson is well-suited to find $\hat{\boldsymbol{\eta}}$. The NR updates are given by

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n + [\nabla^2\Psi(\boldsymbol{\eta}_n)]^{-1}[\bar{\boldsymbol{T}} - \nabla\Psi(\boldsymbol{\eta}_n)].$$

Application

Generalized Linear Models

- **Model:**

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \exp\{T_i \eta_i - \Psi(\eta_i)\} h(y_i), \qquad \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

- **Loglikelihood:**

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} T_i \boldsymbol{x}_i' \boldsymbol{\beta} - \Psi(\boldsymbol{x}_i' \boldsymbol{\beta})$$

- **Hessian:**

$$\tfrac{\partial^2}{\partial \boldsymbol{\beta}^2} \ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = -\boldsymbol{X}' \left[\Psi^{(2)}(\boldsymbol{X}\boldsymbol{\beta}) \right] \boldsymbol{X}, \qquad \text{where}$$

$$\Psi^{(2)}(\eta) = \tfrac{\mathrm{d}^2}{\mathrm{d}\eta^2} \Psi(\eta), \qquad \Psi^{(2)}(\boldsymbol{X}\boldsymbol{\beta}) = \operatorname{diag}\left(\Psi^{(2)}(\boldsymbol{x}_1' \boldsymbol{\beta}), \ldots, \Psi^{(2)}(\boldsymbol{x}_n' \boldsymbol{\beta}) \right).$$

$\implies$ $-\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X})$ is strictly convex since $\boldsymbol{X}' \left[\Psi^{(2)}(\boldsymbol{X}\boldsymbol{\beta}) \right] \boldsymbol{X} = \operatorname{var}(\boldsymbol{X}'\boldsymbol{z})$, where $\operatorname{var}(\boldsymbol{z}) = \Psi^{(2)}(\boldsymbol{x}_i' \boldsymbol{\beta})$.

GLM: Common Cases

1. Poisson Regression (for count data)

- **Model:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_i), \quad \lambda_i = \exp(\boldsymbol{x}_i' \beta).$

  $\implies E[y \mid \boldsymbol{x}] = \exp(\boldsymbol{x}' \beta).$

- **Log-Likelihood:**

$$\ell(\beta \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} y_i \cdot \boldsymbol{x}_i' \beta - \exp(\boldsymbol{x}_i' \beta)$$

- **R command:**

```
M <- glm(y ~ x1 + x2, family = "poisson")
```

GLM: Common Cases

2. Binomial Regression (for success/failure data)

- **Model:** $y_i \mid \boldsymbol{x}_i, N_i \overset{\text{ind}}{\sim} \text{Binomial}(N_i, \rho_i)$,

$$\rho_i = \frac{1}{1 + \exp(-\boldsymbol{x}_i'\boldsymbol{\beta})} \qquad \iff \qquad \boldsymbol{x}_i'\boldsymbol{\beta} = \log\left(\frac{\rho_i}{1-\rho_i}\right) = \text{logit}(\rho_i).$$

- **Log-Likelihood:**

$$\begin{aligned}\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) &= \sum_{i=1}^n y_i \log\left(\frac{\rho_i}{1-\rho_i}\right) + N_i \log(1-\rho_i) \\ &= \sum_{i=1}^n y_i \boldsymbol{x}_i'\boldsymbol{\beta} - N_i \log\left\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}\end{aligned}$$

- **Logistic Regression:** Special name for the common case where $N_i \equiv 1$.

Logistic Regression
Example

- **Model:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i), \qquad \rho_i = [1 + \exp(-\boldsymbol{x}_i'\beta)]^{-1}.$
- **Titanic Data:** 4-way contingency table of the $n = 2201$ passengers on the Titanic in the following categories:
  - $\texttt{Class} \in \{\texttt{1st}, \texttt{2nd}, \texttt{3rd}, \texttt{Crew}\}$.
  - $\texttt{Sex} \in \{\texttt{Male}, \texttt{Female}\}$.
  - $\texttt{Age} \in \{\texttt{Child}, \texttt{Adult}\}$.
  - $\texttt{Survived} \in \{\texttt{No}, \texttt{Yes}\}$.

Application of GLM/NR

Heteroscedastic Linear Regression

- **Usual Linear Regression:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$.

  Model has *homoscedastic errors*: $\text{var}(y \mid \boldsymbol{x}) \equiv \sigma^2$ is constant.

- **Heteroscedastic Linear Regression:**

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_i^2), \qquad \sigma_i = \sigma(\boldsymbol{x}_i),$$

  such that $\text{var}(y \mid \boldsymbol{x}) = \sigma^2(\boldsymbol{x})$ is not constant (depends on $\boldsymbol{x}$).

Application of GLM/NR

Heteroscedastic Linear Regression

- **Model:** (ignore mean term for now)

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2), \qquad \sigma_i^2 = \exp(\boldsymbol{x}_i' \beta).$$

- **Log-Likelihood:**

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2} \sum_{i=1}^{n} \frac{y_i^2}{\exp(\boldsymbol{x}_i' \beta)} + \boldsymbol{x}_i' \beta.$$

- **Convexity:**

  Let $g(\eta) = a \cdot \exp(\eta) + \eta$, for $\eta \in \mathbb{R}$, $a > 0$.

  $\implies \frac{\mathrm{d}^2}{\mathrm{d}\eta^2} g(\eta) = a \cdot \exp(\eta) > 0 \implies g(\eta)$ is convex.

  $\implies -\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} g(\boldsymbol{x}_i' \beta)$ is also convex.

Heteroscedastic Linear Regression

- **Simplified Model:** $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathcal{N}\big(0, \exp(\boldsymbol{x}_i'\beta)\big) \implies$

$$y_i^2 \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \underbrace{\text{Gamma}\left(\tfrac{1}{2}, 2\mu_i\right)}_{\mu_i \cdot \chi^2_{(1)}}, \qquad \mu_i = \exp(\boldsymbol{x}_i'\beta).$$

  - *Gamma parametrization:*

$$z \sim \text{Gamma}(\alpha, \lambda) \qquad \implies \qquad \begin{aligned} E[Y] &= \alpha\lambda \\ \text{var}(Y) &= \alpha\lambda^2. \end{aligned}$$

- **Gamma Regression:**

$$z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Gamma}(1/\tau, \tau\mu_i), \qquad \mu_i = g^{-1}(\boldsymbol{x}_i'\beta)$$

$$\implies \qquad E[z \mid \boldsymbol{x}] = g^{-1}(\boldsymbol{x}'\beta), \qquad \text{var}(z \mid \boldsymbol{x}) = \tau \cdot E[z \mid \boldsymbol{x}]^2$$

  - $g(\mu)$: Link function.
  - τ: Dispersion parameter.

Gamma Regression

- **Model:** $z_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Gamma}(1/\tau, \tau\mu_i), \qquad \mu_i = g^{-1}(\boldsymbol{x}_i'\beta).$
- **Log-Likelihood:**

$$\ell(\beta, \tau \mid \boldsymbol{z}, \boldsymbol{X}) = \sum_{i=1}^n \left[\frac{\log g^{-1}(\boldsymbol{x}_i'\beta) - z_i/g^{-1}(\boldsymbol{x}_i'\beta)}{\tau}\right] - n \log\Gamma(1/\tau) + \sum_{i=1}^n \frac{\log(z_i)}{\tau}$$

- **Properties:**
  - $\ell(\beta, \tau \mid \boldsymbol{z}, \boldsymbol{X})$ convex if $\mu(\boldsymbol{x}) = \exp(\boldsymbol{x}'\beta)$.
  - $\hat\beta = \arg\max_\beta \ell(\beta, \tau \mid \boldsymbol{z}, \boldsymbol{X})$ doesn't depend on τ.
  - Two independent convex problems:

    (i) find $\hat\beta$, then (ii) find $\hat\tau = \arg\max_\tau \ell(\hat\beta, \tau \mid z, \boldsymbol{X})$.
- **R Command:** `glm(z ~ X, family = Gamma("log"))`

Heteroscedastic Linear Regression

- **Full Model:**

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{w}_i \overset{\text{ind}}{\sim} \mathcal{N}\big(\boldsymbol{x}_i'\beta, \exp(\boldsymbol{w}_i'\gamma)\big).$$

Can think of $\boldsymbol{x}$ and $\boldsymbol{w}$ as subsets of a single set of covariates $\boldsymbol{\mathcal{X}}$, e.g.,

$$\boldsymbol{x} = (\text{Age}, \text{Height}, \text{Weight}), \qquad \boldsymbol{w} = (\log(\text{Age}), \text{Height}/\text{Weight}).$$

- **Maximum Likelihood Estimation:**
  - **Initial Value:** $\beta_0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad \gamma_0 = \boldsymbol{0}.$
  - **Iterative fitting:** Given (β_n, γ_n),
    - $\beta_{n+1} = (\boldsymbol{X}'\Lambda_n\boldsymbol{X})^{-1}\boldsymbol{X}'\Lambda_n\boldsymbol{y}, \qquad \Lambda_n = \begin{bmatrix} \exp(-\boldsymbol{w}_1'\gamma) & & \\ & \ddots & \\ & & \exp(-\boldsymbol{w}_n'\gamma) \end{bmatrix}.$

      This is just MLE of β for $y_i \overset{\text{ind}}{\sim} \mathcal{N}\big(\boldsymbol{x}_i'\beta, \exp(\boldsymbol{w}_i'\gamma_n)\big)$.
    - $\gamma_{n+1} = \texttt{coef(glm(}\boldsymbol{u}_{n+1}{}^2 \sim \texttt{W, family = Gamma("log")))}, \quad \boldsymbol{u}_{n+1} = \boldsymbol{y} - \boldsymbol{X}\beta_{n+1}.$

      This is just MLE of γ for $u_{i,n+1}^2 \overset{\text{ind}}{\sim} \text{Gamma}\big(1, \exp(\boldsymbol{w}_i'\gamma)\big)$.

Heteroscedastic Linear Regression

Example

- **Model:** $y_i \mid \boldsymbol{x}_i, \boldsymbol{w}_i \overset{\text{ind}}{\sim} \mathcal{N}\big(\boldsymbol{x}_i'\beta, \exp(\boldsymbol{w}_i'\gamma)\big)$.
- **SENIC Dataset:** Study on the Efficiency of Nosocomial Infection Control (SENIC). $n = 113$ US hospitals with following measurements:
  - `length`: Average length of stay of patients in days.
  - `age`: Average age of patients.
  - `inf`: Probability of acquiring infection in hospital.
  - `cult`: Culturing ratio, i.e. $100 \times \frac{\text{cultures performed}}{\text{\# of patients with no infection}}$.
  - `xray`: Chest X-ray ratio (defined as above).
  - `beds`: Number of beds.
  - `school`: Medical school affiliation ($1 =$ no, $2 =$ yes).
  - `region`: US geographic region ($1 =$ NC, $2 =$ NE, $3 =$ S, $4 =$ W).
  - `pat`: Number of patients.
  - `nurs`: Number of nurses.
  - `serv`: Available facilities (at given hospital).

More Resources

- Useful R functions for `lm`, `glm` and other regression models (e.g., in package `survival`): `coef`, `vcov`, `confint`, `predict`, `fitted`, `residuals`, `summary`, `effects`, `formula`.
- Article by Carl Morris (1982) on Exponential Families with so-called "quadratic variance functions" (easy to read and considered a great breakthrough in statistical theory).
- Simplified version by Morris & Lock (2009) with a nice figure relating the different EF distributions.
- **hlm**: Efficient implementation of the heteroskedastic linear regression model (HLM).