Some background on choice e.g.

Consider a general family of distributions:

subfamily of exponential family that includes Binomial/Bernoulli and Poisson

vary w/ covariates

$$\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi).$$

$$f(y_i;\theta_i,\phi) = \exp\left\{ \underbrace{\frac{y_i\theta_i}{a_i(\phi)} + c(y_i,\theta_i)}_{(*)} - \underbrace{\frac{b(\theta_i)}{a_i(\phi)}}_{(**)} \right\}$$

recall exponential family w/ parameter $\underline{\theta} = (\theta_1,\dots,\theta_s)^T$ is of the form

$$f(y;\underline{\theta}) = h(y)\exp\left\{ \underbrace{\sum_{j=1}^{s} g_j(\underline{\theta})\, T_j(y_i)}_{(*)} - \underbrace{B(\underline{\theta})}_{(**)} \right\}$$

Assumes: $T_1(y_i) = y_i$ & $g_1(\underline{\theta}) = \dfrac{\theta_i}{a_i(\phi)}$ (subfamily of exponential family).

Similar to single parameter exponential family except for dispersion term $a_i(\phi)$.

Example (Normal model): $EY_i = \mu_i$

$$f(y_i;\mu_i,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$\log f(y_i;\mu_i,\sigma) = \log\frac{1}{\sqrt{2\pi}\,\sigma} - \frac{(y_i-\mu_i)^2}{2\sigma^2}$$

$$= -\log(\sqrt{2\pi}\,\sigma) - \frac{y_i^2 - 2\mu_i y_i + \mu_i^2}{2\sigma^2}$$

$$= \frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \log(\sqrt{2\pi}\,\sigma) - \frac{y_i^2}{2\sigma^2}$$

So,

$a_i(\phi) = \sigma^2$

$\theta_i = \mu_i$

$b(\theta_i) = \dfrac{\mu_i^2}{2} = \dfrac{\theta_i^2}{2}$

$c(y_i,\phi) = -\log(\sqrt{2\pi}\,\sigma) - \dfrac{y_i^2}{2\sigma^2}$ (only depends on σ, not μ_i)

We can learn something about this distribution by considering it's mean and variance. Because we don't have an explicit form of the density, we rely on two facts:

See HW3.

1. $\mathrm{E}\left[\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right] = 0.$

"derivative of log density wrt parameter of interest θ_i has expection 0"

2. $\mathrm{E}\left[\frac{\partial^2 \log f(Y_i;\theta_i,\phi)}{\partial\theta_i^2}\right] + \mathrm{E}\left[\left(\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right)^2\right] = 0.$

These will also come up later when we talk about information matrix.

For $\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)$,

Using 1:

$$\frac{\partial}{\partial\theta_i}\log f(y_i;\theta_i,\phi) = \frac{1}{a_i(\phi)}(y_i - b'(\theta_i))$$

$$E\left[\frac{1}{a_i(\phi)}(Y_i - b'(\theta_i))\right] = 0 \Rightarrow b'(\theta_i) = E[Y_i] \Rightarrow$$ information about the mean is contained in $b'(\theta_i)$

E.g. Normal model

$$b(\theta_i) = \frac{\theta_i^2}{2} \Rightarrow b'(\theta_i) = \frac{2\theta_i}{2} = \theta_i \underset{\text{from the family form}}{=} \mu_i \underset{\text{defined}}{=} E[Y_i] \checkmark$$

Using 2:

$$\frac{\partial^2}{\partial\theta_i^2}\log f(y_i;\theta_i,\phi) = -\frac{b''(\theta_i)}{a_i(\phi)} \Rightarrow E\left[\frac{\partial^2}{\partial\theta_i^2}\log f(Y_i;\theta_i,\phi)\right] = -\frac{b''(\theta_i)}{a(\phi)}$$

$$E\left[\left(\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right)^2\right] = E\left[\left(\frac{1}{a_i(\phi)}(Y_i - b'(\theta_i))\right)^2\right] = \frac{1}{a_i(\phi)^2}E\left[(Y_i-\mu_i)^2\right]$$

$$\Rightarrow -\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{a_i(\phi)^2}\mathrm{Var}\,Y_i = 0 \Rightarrow \mathrm{Var}[Y_i] = a_i(\phi)\,b''(\theta_i)$$

Thoughts:

- Variance depends on i
- $\mathrm{Var}[Y_i]$ positive $\Rightarrow b''(\theta_i)$ positive $\forall$ values of θ_i
  - $\Rightarrow b(\theta_i)$ strictly convex
  - $b'(\theta_i)$ monotone increasing so b'^{-1} exists.

Example (Bernoulli model):

$$f(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\log f(y_i; p_i) = y_i \log p_i + (1-y_i)\log(1-p_i)$$

$$= y_i \boxed{\log \frac{p_i}{1-p_i}} + \boxed{\log(1-p_i)}$$

Comparing to general form:

$$\frac{y_i \boxed{\theta_i} - \boxed{b(\theta_i)}}{a_i(\phi)} + c(y_i, \phi)$$

$$\Rightarrow \theta_i = \log\left(\frac{p_i}{1-p_i}\right) \Rightarrow p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$$

$$b(\theta_i) = -\log(1-p_i)$$

$$= -\log\left(1 - \frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)$$

$$= -\log\left(\frac{1}{1+\exp(\theta_i)}\right)$$

$$= \log(1+\exp(\theta_i))$$

Notice: $b'(\theta_i) = \frac{1}{1+\exp(\theta_i)} \exp(\theta_i) = p_i$ ✓

$$a_i(\phi) = 1.$$

$$a_i(\phi)\, b''(\theta_i) = -\left(-(1+\exp(\theta_i))^{-2}\right)\exp(\theta_i)\exp(\theta_i) + (1+\exp(\theta_i))^{-1}\exp(\theta_i)$$

$$= \frac{(1+\exp(\theta_i))\exp(\theta_i) - \exp(\theta_i)\exp(\theta_i)}{(1+\exp(\theta_i))^2} = \frac{\exp(\theta_i)}{(1+\exp(\theta_i))^2}$$

$$= \frac{\exp(\theta_i)}{(1+\exp(\theta_i))} \cdot \frac{1}{1+\exp(\theta_i)}$$

$$= p_i(1-p_i) = \operatorname{Var} Y_i$$

Finally, back to modelling. Our **goal** is to build a relationship between the mean of Y_i and covariates $\boldsymbol{x}_i$.

choose $x_i^T \beta = g(\mu_i) = g(E[Y_i])$.

what to choose for g?

$\mu_i = g^{-1}(x_i^T \beta)$

we know $\mu_i = b'(\theta_i)$

we know this exists!

$\Rightarrow b'(\theta_i) = g^{-1}(x_i^T \beta)$ OR $\theta_i = b'^{-1}(g^{-1}(x_i^T \beta))$

If we choose $\underbrace{g^{-1} = b'}$, then this will clean up nicely! i.e. $\theta_i = \underline{x}_i^T \underline{\beta}$.

"canonical" or "natural" link function.

$\Rightarrow$ log-likelihood is

$$\ell(\underline{\beta}, \phi \mid \{Y_i, \underline{x}_i\}_{i=1}^n) = \sum_{i=1}^n \left\{ \frac{Y_i \underline{x}_i^T \underline{\beta} - b(\underline{x}_i^T \underline{\beta})}{a_i(\phi)} + c(Y_i, \phi) \right\}$$

Example (Bernoulli model, cont'd):

$b(\theta_i) = \log(1 + \exp(\theta_i))$

$b'(\theta_i) = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = \frac{\exp(\underline{x}_i^T \underline{\beta})}{1+\exp(\underline{x}_i^T \underline{\beta})}$ canonical link

$\Big\downarrow$ mean

$p_i = \frac{\exp(\underline{x}_i^T \underline{\beta})}{1+\exp(x_i^T \beta)}$

(same form as before)

1.4 Marginal and Conditional Likelihoods

Consider a model which has $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ are the parameters of interest and $\boldsymbol{\theta}_2$ are nuisance parameters.

not what we are interested in performing inference for

When dimension of $\underline{\theta}_2$ is large MLE's of $\underline{\theta}_1$ can be biased for small samples and inconsistent in large samples.

One way to improve estimation for $\boldsymbol{\theta}_1$ is to find a one-to-one transformation of the data $\boldsymbol{Y}$ to $(\boldsymbol{V}, \boldsymbol{W})$ such that either

$$f_Y(\underline{y}; \underline{\theta}_1, \underline{\theta}_2) = f_{W,V}(\underline{w}, \underline{v}; \underline{\theta}_1, \underline{\theta}_2)$$

$$= f_{W,V}(\underline{w} \mid \underline{v}; \underline{\theta}_1, \underline{\theta}_2)\, \underline{f_V(\underline{v}; \underline{\theta}_1)}$$

"marginal"

alternative likelihood

OR $$f_Y(\underline{y}; \underline{\theta}_1, \underline{\theta}_2) = \underline{f_{W,V}(\underline{w} \mid \underline{v}; \underline{\theta}_1)}\, f_V(\underline{v}; \underline{\theta}_1, \underline{\theta}_2)$$

"conditional"

either way looking to split density into a piece that doesn't depend on $\underline{\theta}_2$ (nuisance parameters)

The key feature is that one component of each contains only the parameter of interest.

$\underline{\theta}_1$

Example (Neyman-Scott problem): Let $Y_{ij}, i = 1, \dots, n, j = 1, 2$ be intependent normal random variables with possible different means μ_i but the same variance σ^2.

$$\underset{}{Y_{ij}}, \underbrace{i = 1, \dots, n}_{n \text{ groups}}, \underbrace{j = 1, 2}_{\substack{\text{ONLY 2} \\ \text{obs per group}}} \overset{iid}{\sim} N(\underset{\substack{\uparrow \\ \text{group mean}}}{\mu_i}, \underset{\text{common variance.}}{\sigma^2})$$

$$\underline{\theta} = (\overbrace{\mu_1, \dots, \mu_n}^{\text{nuisance params}}, \sigma^2)^T$$

n+1 parameters.

Our goal is to estimate σ^2. Should we be able to?

Yes: lots of groups!

No: only 2 obs per group.

Usual asymptotic assumptions as n grows,

Here as n grows, # groups grows $\Rightarrow$ # parameters grows.

Following the usual arguments,

$$\hat{\mu}_{i,\text{MLE}} = \frac{Y_{i1} + Y_{i2}}{2}$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^2 (Y_{ij} - \hat{\mu}_{i,\text{MLE}})^2$$

$$= \frac{1}{2n} \sum_{i=1}^n \left\{ \left(Y_{i1} - \frac{Y_{i1} + Y_{i2}}{2} \right)^2 + \left(Y_{i2} - \frac{Y_{i1} + Y_{i2}}{2} \right)^2 \right\}$$

$$= \frac{1}{2n} \sum_{i=1}^n \left\{ \underline{\left(\frac{Y_{i1} - Y_{i2}}{2} \right)^2} + \underline{\left(\frac{Y_{i2} - Y_{i1}}{2} \right)^2} \right\}$$

terms are equivalent

$$= \frac{1}{n} \sum_{i=1}^n \frac{1}{4} (Y_{i1} - Y_{i2})^2$$

$$E[\hat{\sigma}^2_{\text{MLE}}] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{4}(Y_{i_1}-Y_{i_2})^2\right]$$

$$\text{(iid)} \quad = \frac{1}{4} E\left[(Y_{i_1}-Y_{i_2})^2\right]$$

$$= \frac{1}{4} E\left[\left((Y_{i_1}-\mu_i)-(Y_{i_2}-\mu_i)\right)^2\right]$$

$$= \frac{1}{4} E\left[(Y_{i_1}-\mu_i)^2 - 2(Y_{i_1}-\mu_i)(Y_{i_2}-\mu_i) + (Y_{i_2}-\mu_i)^2\right]$$

$$= \frac{1}{4}\left[\sigma^2 - 0 + \sigma^2\right]$$

$$= \frac{\sigma^2}{2} \neq \sigma^2 \,!$$

So as $n \to \infty$, $\hat{\sigma}^2_{\text{MLE}} \xrightarrow[p]{} \frac{\sigma^2}{2}$ by WLLN

This seems bad.

Happens because # of nuisance parameters grows w/ n (# groups).

A reworking of the data seems more promising. Let,

$$V_i = \frac{Y_{i1} - Y_{i2}}{\sqrt{2}} \qquad \text{and} \qquad W_i = \frac{Y_{i1} + Y_{i2}}{\sqrt{2}}$$