Prof. H. Bensmail

Stat-664: Adv Stat Infer

Spring Semester 2006

# Chapter 3:

# Bayesian inferences

February 1, 2006

**Bayes theorem**: **Discrete case**

$H_i$: hypothesis

$E_i$: effect

the conditional probability that the hypothesis $H_i$ holds given the effect $E_j$ is

$$p(H_i|E_j) = \frac{p(H = H_i, E = E_j)}{P(E_j)}$$

$$= \frac{p(E_j|H_i)p(H_i)}{p(E_j)}$$

$p(E_j)$ = marginal or total probability of observing the effect of $E_j$

$$p(E_j) = \sum_{i=1}^{m} p(E_j|H_i)p(H_i)$$

$$= E_H[p(E_j|H_i)]$$

so $p(H_i|E_j) \propto p(E_j|H_i)p(H_i)$

suppose additional evidence $E_i^{'}$ accrues, then

$$p(H_i|E_j^{'}, E_j) \propto p(E_j^{'}|E_j, H_i)p(H_i|E_j)$$

**Example**:

suppose a rare disease whose frequency is 1 out of 5000.

A person is tested, and the false positive is $0.05$ and the false negative is $0.01$

let's $\theta$ be a random variable taking value 1 if the person is diseased and 0 if not.

$Y$ is a random variable taking value 1 when the test is positive, and value 0 when the test is negative

False positive = $P(Y = 1|\theta = 0)$

False negative= $P(Y = 0|\theta = 1)$

$$p(\theta = 1) = 0.0002; \text{ disease}$$

$$p(Y = 1|\theta = 0) = 0.05$$

$$p(Y = 0|\theta = 1) = 0.01$$

$$p(Y = 0|\theta = 0) = 0.95$$

$$p(Y = 1|\theta = 1) = 0.99$$

**Question**: what is the probability that the person which has a positive test is diseased?

Bayes theorem:

$$p(H_i|E_j) = \frac{p(E_j|H_i)p(H_i)}{p(E_j)} = \frac{p(E_j|H_i)p(H_i)}{E_H(E_j|H_i)}$$

$$= \frac{p(E_j|H_i)p(H_i)}{\sum_{i=1}^{m} p(E_j|H_i)p(H_i)}$$

$p(\theta = 1|Y = 1) =$

$$\frac{p(\theta = 1)p(Y = 1|\theta = 1)}{p(\theta = 1)p(Y = 1|\theta = 1) + p(\theta = 0)p(Y = 1|\theta = 0)}$$

$$= 0.0039$$

**Bayes theorem**: **continuous case**

- $\theta, y$ are continuous valued
- $M$ is a model and $(\theta, y|M)$ exists.

  example: $y$ is the evidence, $\theta$ is hypothesis (missing data, etc), $M$ is model: gaussian or $t\ldots$

$$\begin{aligned}
\underbrace{h(\theta,y)}_{\text{joint}} &= \underbrace{g(\theta)}_{\text{marginal of } \theta} \quad f(y|\theta) \\
&= \underbrace{m(y)}_{\text{marginal of } y} \quad p(\theta|y)
\end{aligned}$$

$$\begin{aligned}
g(\theta) &= \int h(\theta,y)dy = \int p(\theta|y)m(y)dy \\
&= E_y[p(\theta|y)] \\
m(y) &= \int h(\theta,y)dy = \int f(y|\theta)g(\theta)d\theta \\
&= E_\theta[f(y|\theta)] \\
p(\theta|y) &= \frac{g(\theta)f(y|\theta)}{m(y)} \\
&\propto g(\theta)f(y|\theta)
\end{aligned}$$

$$\begin{aligned}
\mathbf{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \mathbf{g}(\boldsymbol{\theta})\mathbf{L}(\boldsymbol{\theta}|\mathbf{y}) \\
= \frac{g(\theta)L(\theta|y)}{\int g(\theta)L(\theta|y)d\theta}
\end{aligned}$$

if we have a uniform over $\theta$, then

$$\begin{aligned}
\mathbf{p}(\boldsymbol{\theta}|\mathbf{y}) &\propto \frac{L(\theta|y)}{\int L(\theta|y)d\theta} \\
&\propto \mathbf{L}(\boldsymbol{\theta}|\mathbf{y})
\end{aligned}$$

**Example**:

$$y = \mu + a + e$$

- $\mu$ : known
- $a$ : additive effect (unknown)
- $e$ :deviation

$$a \sim N(0, v_a)$$
$$e \sim N(0, v_e)$$
$$(y|\mu, a, v_a, v_e) \sim N(\mu + a, v_e)$$

if the problem is inferring $a$ from $y$

$$p(a|y, \mu, v_a, v_e) \sim N(h^2(y-\mu); v_a(1-h^2))$$

where

$$h^2 = \frac{v_a}{v_a + v_e}$$

Hint1: we can use:

$$E(y_1|y_2) = m_1 + V_{12}V_{22}^{-1}(y_2 - m_2)$$
$$Var(y_1|y_2) = V_{11} - V_{12}V_{22}^{-1}V_{21}$$

Hint2: use $\theta = a,$ and calculate

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

- prior mode of $a$ is $0,$ the posterior mode $h^2(y-\mu)$
- $p(a > 0) = 0.5$, but $p(a > 0|y)$?
- calculate $p(y)$?

**<u>Posterior distribution</u>**:

$\theta = (\theta_1^t, \theta_2^t)$,

example in a linear case, $\theta_1$ is location and $\theta_2$ is dispersion

$$p(\theta_1, \theta_2|y) \propto L(\theta_1, \theta_2|y)g(\theta_1, \theta_2)$$

we can write

$$\begin{aligned} g(\theta_1, \theta_2) &= g(\theta_1|\theta_2)g(\theta_2) \\ &= g(\theta_2|\theta_1)g(\theta_1) \end{aligned}$$

then the marginal posteriors are:

$$\begin{aligned} p(\theta_1|y) &= \int p(\theta_1, \theta_2|y)d\theta_2 \\ p(\theta_2|y) &= \int p(\theta_1, \theta_2|y)d\theta_1 \end{aligned}$$

if $\theta_1$ is the primary parameter, we also have:

$$\begin{aligned} p(\theta_1|y) &= \int p(\theta_1, \theta_2|y)d\theta_2 \\ &= \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2 \\ &= E_{(\theta_2|y)}[p(\theta_1|\theta_2, y)] \end{aligned}$$

$(\theta_1|\theta_2, y)$ is the uncertainty of inferences about $\theta_1$ when nuisance parameter $\theta_2$ is known.

$$\begin{aligned} p(\theta_1|\theta_2, y) &= \frac{p(\theta_1, \theta_2|y)}{p(\theta_2|y)} \\ &\propto \mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) \\ &\propto L(\theta_1, \theta_2|y)p(\theta_1, \theta_2) \\ &\propto L(\theta_1, \theta_2|y)p(\theta_1|\theta_2) \\ &\propto \mathbf{L}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\mathbf{p}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \end{aligned}$$

$L(\theta_1|\theta_2, y)$: is likelihood of $\theta_1$ with $\theta_2$ treated as known.

so the conditional posterior distribution can be identified by inspection of the joint posterior distribution and retains only the parts that vary with the parameters of interests.

**Example**:

suppose $y = (y_1,..,y_n)$ are independently drawn from

$$y \sim N(\mu, \sigma^2)$$

$$p(\mu|a,b) = \frac{1}{b-a}$$

$$p(\sigma^2|c,d) = \frac{1}{d-c}$$

$$\begin{aligned} p(\mu,\sigma^2|y,\theta) &\propto L(\mu,\sigma^2|y,\theta) \times g(\mu,\sigma^2) \\ &= \prod_{i=1}^{n} (\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ &\quad \times \frac{1}{(b-a)(d-c)} \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{\sum(y_i-\bar{y})^2 + n(\mu-\bar{y})^2}{2\sigma^2}\right\} \end{aligned}$$

so

$$\begin{aligned} p(\sigma^2|\mathbf{y},a,b,c,d) &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{\sum(y_i-\bar{y})^2}{2\sigma^2}\right\} \\ &\quad \times \int_a^b \frac{\sqrt{2\pi\sigma^2/n}}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{n(\mu-\bar{y})^2}{2\sigma^2}} d\mu \\ &= (\sigma^2)^{-(n-1)/2} \exp\left\{-\frac{\sum(y_i-\bar{y})^2}{2\sigma^2}\right\} \\ &\quad \times \left[\Phi\left(\frac{b-\bar{y}}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\bar{y}}{\sigma/\sqrt{n}}\right)\right] \end{aligned}$$

$$\begin{aligned} p(\mu|y,\theta) &= \int p(\mu,\sigma^2|y,\theta) d\sigma^2 \\ &\propto \int_c^d (\sigma^2)^{-\left(\frac{n-2}{2}+1\right)} e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2} d\sigma^2 \end{aligned}$$

where $S_\mu = \frac{1}{2}\sum(y_i-\mu)^2$

if $c = 0$ and $d = \infty$, we use inverse gamma

$$p(\mu|y,\theta) \propto \frac{\Gamma(\frac{n-2}{2})}{S_\mu^{\frac{n-2}{2}}} \propto S_\mu^{-\frac{n-2}{2}}$$

$$\begin{aligned} p(\mu|y,\theta) &\propto \left[1 + \frac{(\mu-\bar{y})^2}{(n-3)\frac{\hat{S}^2}{n}}\right]^{-\frac{n-3+1}{2}} \\ &\propto t-dist\left(n-3, \bar{y}, \frac{\hat{S}^2(n-3)}{n(n-5)}\right) \text{ if } c = 0, d = \infty \end{aligned}$$

where $\hat{S}^2 = \frac{1}{n-3}\sum(y_i-\bar{y})^2$

**Example**:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$i = 1,\ldots,n$$
$$\varepsilon_i \sim t(0,\sigma^2,\upsilon)$$
$$y_i = \beta_0 + \beta_1 x_i + \frac{e_i}{\sqrt{\omega_i}}$$
$$e_i \sim N(0,\sigma^2)$$
$$\omega_i \sim G_a(\frac{\upsilon}{2},\frac{\upsilon}{2})$$

$\theta = (\beta_0,\beta_1,\sigma^2,\omega_1,\ldots,\omega_n)$
priors:

$$p(\theta|\upsilon) = p(\beta_0)p(\beta_1)p(\sigma^2)\prod_{i=1}^{n} p(\omega_i|\upsilon)$$

Then

$$\begin{aligned} p(\theta|y,\upsilon) &\propto L(\theta|y,\upsilon)p(\theta) \\ &\propto \prod_{i=1}^{n}\left(\frac{\sigma^2}{\omega_i}\right)^{-1/2}\exp\left\{\frac{\omega_i}{2}\,\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}\right\} \\ &\quad\times\prod_{i=1}^{n}(\omega_i^{\upsilon/2-1}e^{-\frac{\upsilon\omega_i}{2}})\; p(\beta_0)p(\beta_1)p(\sigma^2) \\ &\propto \prod_{i=1}^{n}\left(\frac{\sigma^2}{\omega_i}\right)^{-1/2}\omega_i^{\upsilon/2-1} \\ &\quad\times\exp\left\{\frac{\omega_i}{2}\,\frac{(y_i-\beta_0-\beta_1 x_i)^2+\upsilon\sigma^2}{\sigma^2}\right\} \qquad (*) \\ &\quad\times p(\beta_0)p(\beta_1)p(\sigma^2) \end{aligned}$$

where

$$p(\omega_i|\upsilon) \propto \omega_i^{\frac{\upsilon}{2}-1}e^{-\frac{\upsilon\omega_i}{2}}$$

**Conditional posterior distribution of $\omega_i$**

We have:

$p(\beta_0, \beta_1, \sigma^2, \omega_1, \ldots, \omega_n | y, \upsilon)$ = given in $(*)$

Independency of $\omega_i$ means that:

$p(\beta_0, \beta_1, \sigma^2, \omega_i | y, \upsilon) =$

$$\propto \left(\frac{\sigma^2}{\omega_i}\right)^{-1/2} \omega_i^{\frac{\upsilon}{2}-1}$$

$$\times \exp\left\{\frac{\omega_i}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2 + \upsilon\sigma^2}{\sigma^2}\right\} \qquad (*)$$

$$\times\, p(\beta_0) p(\beta_1) p(\sigma^2)$$

Using formula $p(\theta_1|\theta_2, y) = \frac{p(\theta_1,\theta_2|y)}{p(\theta_2|y)}$

$p(\omega_i | \beta_0, \beta_1, \sigma^2, y, \upsilon) = \frac{p(\beta_0,\beta_1,\sigma^2,\omega_i|y,\upsilon)}{p(\beta_0,\beta_1,\sigma^2|y,\upsilon)}$

$$\propto \left(\frac{\sigma^2}{\omega_i}\right)^{-1/2} \omega_i^{\upsilon/2-1}$$

$$\times \exp\left\{-\frac{\omega_i}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2 + \upsilon\sigma^2}{\sigma^2}\right\}$$

so

$$p(\omega_i | \beta_0, \beta_1, \sigma^2, y, \upsilon) \propto \omega_i^{\frac{\upsilon+1}{2}-1} \exp\left(-\frac{\omega_i S_i}{2}\right)$$

where

$$S_i = \frac{(y_i - \beta_0 - \beta_1 x_i)^2 + \upsilon\sigma^2}{\sigma^2}$$

which means that

$$(\omega_i | \beta_0, \beta_1, \sigma^2, y, \upsilon) \sim G_a\left(\frac{\upsilon + 1}{2}, \frac{\upsilon\sigma^2 + (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

**Conditional posterior distribution of $\beta_0, \beta_1$**

$$p(\beta_0,\beta_1,\sigma^2,\omega|y,\upsilon) \propto \prod_{i=1}^{n}(\sigma^2)^{-1/2}\omega_i^{\frac{\upsilon+1}{2}-1}\exp\left(-\frac{\omega_i S_i}{2}\right) \times p(\beta_0)p(\beta_1)p(\sigma^2)$$

$$p(\beta_0,\beta_1|\sigma^2,\omega,y,\upsilon) = \frac{p(\beta_0,\beta_1,\sigma^2,\omega|y,\upsilon)}{p(\sigma^2,\omega|y,\upsilon)}$$

is a joint posterior distribution with constant $\sigma^2, \omega$ constant:
so

$$p(\beta_0,\beta_1|\sigma^2,\omega,y,\upsilon) \propto \prod_{i=1}^{n}\exp\left\{\frac{\omega_i}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2\right\} \times p(\beta_0)p(\beta_1)$$

if

$$\beta_0 \sim N(\alpha_0, \sigma^2_{\beta_0})$$

and

$$\beta_1 \sim N(\alpha_1, \sigma^2_{\beta_1})$$

then:

$$p(\beta_0,\beta_1|\sigma^2,\omega,y,\upsilon) \propto \exp[-\frac{1}{2\sigma^2}\{\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2 + \frac{\sigma^2}{\sigma^2_{\beta_0}}(\beta_0-\alpha_0)^2 + \frac{\sigma^2}{\sigma^2_{\beta_1}}(\beta_1-\alpha_1)^2\}]$$

If we reformulate:

$$\beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix};\ W = diag(\omega_i);\ X = [\mathbf{1}\ \mathbf{x}]'_{n\times 2}$$

$$\mathbf{1} = \{1\}, \mathbf{x} = (x_1,..,x_n)'$$

then we can write the above as:

$$\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2 = (\mathbf{y}-X\beta)^t W(\mathbf{y}-X\beta)$$

Let $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = (X^t WX)^{-1}X^t W\mathbf{y}$

then we can write:

$$(\mathbf{y}-X\beta)^t W(\mathbf{y}-X\beta) = (\mathbf{y}-X\hat{\beta}-X\beta+X\hat{\beta})^t W(\mathbf{y}-X\hat{\beta}-X\beta+X\hat{\beta})$$
$$= (\mathbf{y}-X\hat{\beta})^t W(\mathbf{y}-X\hat{\beta}) + (\beta-\hat{\beta})^t X^t WX(\beta-\hat{\beta})$$

then

$$
\begin{aligned}
& p(\beta_0, \beta_1 | y, \omega, \sigma^2, \upsilon, \alpha_0, \alpha_1, \beta_0, \beta_1) \\
\propto\ & \exp\left\{-\frac{1}{2\sigma^2}((\mathbf{y} - X\hat{\beta})^t W(\mathbf{y} - X\hat{\beta})\right\} \\
& + (\beta - \hat{\beta})^t X^t WX(\beta - \hat{\beta}) + \frac{\sigma^2(\beta_0 - \alpha_0)^2}{\sigma^2_{\beta_0}} + \frac{\sigma^2(\beta_1 - \alpha_1)^2}{\sigma^2_{\beta_1}})\} \\
\propto\ & \exp\left\{-\frac{1}{2\sigma^2}((\beta - \hat{\beta})^t X^t WX(\beta - \hat{\beta}) + \frac{\sigma^2(\beta_0 - \alpha_0)^2}{\sigma^2_{\beta_0}} + \frac{\sigma^2(\beta_1 - \alpha_1)^2}{\sigma^2_{\beta_1}})\right\} \\
\propto\ & \exp\left\{-\frac{1}{2\sigma^2}((\beta - \hat{\beta})^t X^t WX(\beta - \hat{\beta}) + (\beta - \boldsymbol{\alpha})^t \Lambda(\beta - \boldsymbol{\alpha}))\right\}
\end{aligned}
$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}$,

$$
\Lambda = \begin{pmatrix} \sigma^2/\sigma^2_{\beta_0} & 0 \\ 0 & \sigma^2/\sigma^2_{\beta_1} \end{pmatrix} = \begin{pmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{pmatrix}
$$

Use the following relationship:

(1)Univariate case:
$M(z-m)^2 + B(z-b)^2 = (M+B)(z-c)^2 + \frac{MB}{M+B}(m-b)^2$
where: $c = (M+B)^{-1}(Mm+Bb)$
(2)Multivariate case:
$(z-m)^t M(z-m) + (z-b)^t B(z-b)$
$= (z-c)^t(M+B)(z-c) + (m-b)^t M(M+B)^{-1}B(m-b)$
where:$c = (M+B)^{-1}(Mm+Bb)$

we had

$$p(\beta_0, \beta_1 | \mathbf{y}, \omega, \sigma^2, \upsilon, \alpha_0, \alpha_1, \beta_0, \beta_1)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}((\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^t X^t WX(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\boldsymbol{\alpha})^t \Lambda(\boldsymbol{\beta}-\boldsymbol{\alpha}))\right\}$$

which become by using (2)

$$\propto \exp\left\{\begin{array}{c} -\frac{1}{2\sigma^2}((\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^t(X^t WX+\Lambda)(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}) + \\ (\hat{\boldsymbol{\beta}}-\boldsymbol{\alpha})^t X^t WX(X^t WX+\Lambda)^{-1}\Lambda(\hat{\boldsymbol{\beta}}-\boldsymbol{\alpha})) \end{array}\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^t(X^t WX+\Lambda)(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right\}$$

where $\bar{\beta} = (X^t WX+\Lambda)^{-1}(X^t WX\hat{\beta}+\Lambda\alpha)$

$$p(\beta_0, \beta_1 | \mathbf{y}, \omega, \sigma^2, \upsilon, \alpha_0, \alpha_1, \beta_0, \beta_1) \sim N(\bar{\beta}, var?)$$

where

$$Var = (X^t WX+\Lambda)^{-1}\sigma^2$$

**Conditional distribution of $\beta_1 | \beta_0, rest$**

Here, we think of $\beta_0$ as if it is known.
Regression becomes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \rightarrow$$

$$r_i = y_i - \beta_0 = \beta_1 x_i + \varepsilon_i$$

$$p(\beta_0, \beta_1 | \mathbf{y}, \omega, \sigma^2, \upsilon, \alpha_0, \alpha_1)$$

$$\propto \exp\left\{\begin{array}{c} -\frac{1}{2\sigma^2}\left(\sum_{i=1}^n \omega_i (r_i - \beta_1 x_i)\right)^2 + \\ \frac{\sigma^2}{\sigma^2_{\beta_1}}(\beta_1-\alpha_1)^2 + \frac{\sigma^2}{\sigma^2_{\beta_0}}(\beta_0-\alpha_0)^2) \end{array}\right\}$$

Take $\beta_0$ as constant:

$$p(\beta_1 | \beta_0, \mathbf{y}, \omega, \sigma^2, \upsilon, \alpha_0, \alpha_1)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_i(r_i-\beta_1x_i)^2+\frac{\sigma^2}{\sigma_{\beta_1}^2}(\beta_1-\alpha_1)^2\right)$$
$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_ix_i^2(\beta_1-r_i/x_i)^2+\lambda_1(\beta_1-\alpha_1)^2\right)$$
$$\propto \exp\left(\begin{array}{l} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\omega_ix_i^2+\lambda_1)(\beta_1-\bar{\beta}_{1i})^2 \\ +(\omega_ix_i^2\lambda_1)(\omega_ix_i^2+\lambda_1)^{-1}(\frac{r_i}{x_i}-\alpha_1)^2 \end{array}\right)$$
$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\omega_ix_i^2+\lambda_1)(\beta_1-\bar{\beta}_{1.i})^2\right)$$

where $\bar{\beta}_{1.i} = (\omega_ix_i^2+\lambda_1)^{-1}(\omega_ix_ir_i+\lambda_1\alpha_1)$
using the following formula:
$M(z-m)^2+B(z-b)^2 = (M+B)(z-c)^2+\frac{MB}{M+B}(m-b)^2$
where: $c = (M+B)^{-1}(Mm+Bb)$

similarly

$$(\beta_1|\beta_0,\mathbf{y},\omega,\sigma^2,\upsilon,\alpha_0,\alpha_1) \sim N(\bar{\beta}_{1,0},var?)$$

$\bar{\beta}_{1,0} = (X^tWX+\lambda_1)^{-1}(X^tWr+\lambda_1\alpha_1) = \frac{\sum\omega_ix_i(y_i-\beta_0)+\lambda_1\alpha_1}{\sum\omega_ix_i^2+\lambda_1}$
$var = (X^tWX+\lambda_1)^{-1}\sigma^2 = \frac{\sigma^2}{\sum\omega_ix_i^2+\lambda_1}$

## <u>**Conditional posterior of $\beta_0$ given $\beta_1$**</u>

We use

$$t_i = y_i-\beta_1x_i = \beta_0+\varepsilon_i$$

similarly:

$$p(\beta_0|\beta_1,\mathbf{y},\omega,\sigma^2,\upsilon,\alpha_0,\alpha_1)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta_0-\bar{\beta}_{0,1})^t(\mathbf{1}^tW\mathbf{1}+\lambda_0)(\beta_0-\bar{\beta}_{0,1})\right\}$$

where $\bar{\beta}_{0,1} = (\mathbf{1}^tW\mathbf{1}+\lambda_0)^{-1}(\mathbf{1}^tW\mathbf{1}+\lambda_0\alpha_0)$
so

$$\beta_0|\beta_1,\mathbf{y},\omega,\sigma^2,\upsilon,\alpha_0,\alpha_1 \sim N(\bar{\beta}_{0,1},var)$$
$$var = \frac{\sigma^2}{\sum_{i=1}^{n}\omega_i+\lambda_0}$$

## <u>**Conditional posterior distribution of $\sigma^2$/rest**:</u>

$p(\sigma^2|\beta_0,\beta_1,\mathbf{y},\omega,\upsilon)$?

$$
\begin{aligned}
& p(\sigma^2,\beta_0,\beta_1,\omega|\mathbf{y},\upsilon) \\
\propto\ & \prod_{i=1}^{n}(\sigma^2)^{-1/2}\omega_i^{\frac{\upsilon+1}{2}-1}\exp\left(-\frac{\omega_i S_i}{2}\right) \\
& \times p(\beta_0)p(\beta_1)p(\sigma^2)
\end{aligned}
$$

$$
S_i = \frac{(y_i-\beta_0-\beta_1 x_i)^2+\upsilon\sigma^2}{\sigma^2}
$$

Retaining only the term in $\sigma^2$, so we have

$$
\begin{aligned}
& p(\sigma^2|\beta_0,\beta_1,\omega,\mathbf{y},\upsilon) \\
\propto\ & (\sigma^2)^{-n/2}\times\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2\right]\right\}p(\sigma^2)
\end{aligned}
$$

if

$$
\sigma^2 \sim \text{lognormal}(0,\omega) \propto \exp\left[-\frac{(\log\sigma^2)^2}{2\omega}\right]\frac{1}{\sigma^2}
$$

$$
\begin{aligned}
=\ & (\sigma^2)^{-(n+2)/2} \\
& \times\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2+\sigma^2\frac{(\log\sigma^2)^2}{\omega}\right]\right\}
\end{aligned}
$$

***which is not a recognizable form***.
if, we use instead a scaled inverse $\chi^2$
so $\sigma^2 \sim I\chi^2(q,r)$

means

$$
p(\sigma^2|q,r) \propto (\sigma^2)^{-\left(\frac{q}{2}+1\right)}\exp\left(\frac{qr}{2\sigma^2}\right)
$$

then

$$
\begin{aligned}
& p(\sigma^2|\beta_0,\beta_1,\omega,\mathbf{y},\upsilon) \\
\propto\ & (\sigma^2)^{-n/2}\times\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2\right]\right\}p(\sigma^2) \\
\propto\ & (\sigma^2)^{-\left(\frac{n+q}{2}+1\right)}\exp\left(-\frac{q^*r^*}{2\sigma^2}\right)
\end{aligned}
$$

where

$$
q^* = n+q \qquad r^* = \frac{ns^2+qr}{n+q} \qquad s^2 = \frac{\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1 x_i)^2}{n}
$$

which is a scaled inverse Chi-square process with parameters $q^*$ and $r^*$

**Conditional posterior distribution of $\beta_0, \beta_1 / \mathbf{y}, \omega, hyper$**

using the scaled chi-square distribution for $\sigma^2$
$p(\sigma^2) \sim I\chi^2 \propto (\sigma^2)^{-\left(\frac{q}{2}+1\right)} \exp\left(-\frac{qr}{2\sigma^2}\right)$

we had:

$$p(\sigma^2, \beta_0, \beta_1, \omega | \mathbf{y}, \upsilon) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \omega_i^{\frac{\upsilon+1}{2}-1} \exp\left(-\frac{\omega_i S_i}{2}\right) \times p(\beta_0) p(\beta_1) p(\sigma^2)$$

then taking $\omega$ as fixed:

$$p(\sigma^2, \beta_0, \beta_1 | \omega, \mathbf{y}, \upsilon) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \exp\left(-\frac{\omega_i S_i}{2}\right) \times p(\beta_0) p(\beta_1) p(\sigma^2)$$

$$p(\beta_0, \beta_1 | \omega, \mathbf{y}, \upsilon, q, r) = \int p(\beta_0, \beta_1, \sigma^2 | \omega, \mathbf{y}, \upsilon, q, r) d\sigma^2$$

$$\propto p(\beta_0) p(\beta_1) \int (\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\left\{-\frac{\omega_i}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}\right\} \times (\sigma^2)^{-\left(\frac{q}{2}+1\right)} \exp\left(-\frac{qr}{2\sigma^2}\right) d\sigma^2$$

$$\propto p(\beta_0) p(\beta_1) \int (\sigma^2)^{-\left(\frac{n+q}{2}+1\right)} \exp\left\{-\frac{(\mathbf{y} - X\beta)^t W (\mathbf{y} - X\beta) + qr}{2\sigma^2}\right\}$$

using the following
$[\sum \omega_i (y_i - \beta_0 - \beta_1 x_i)^2 = (\mathbf{y} - X\boldsymbol{\beta})^t W (\mathbf{y} - X\boldsymbol{\beta})]$

then we have:

$$p(\beta_0, \beta_1 | \omega, \mathbf{y}, \upsilon, q, r) \propto p(\beta_0) p(\beta_1) [(\mathbf{y} - X\boldsymbol{\beta})^t W (\mathbf{y} - X\boldsymbol{\beta}) + qr]^{-\left(\frac{n+q}{2}\right)}$$

using the relationship:

$$(\mathbf{y} - X\boldsymbol{\beta})^t W (\mathbf{y} - X\boldsymbol{\beta}) = (\mathbf{y} - X\hat{\boldsymbol{\beta}})^t W (\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) X^t W X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

we have

$$p(\beta_0, \beta_1 | \omega, \mathbf{y}, \upsilon, q, r) \propto p(\beta_0) p(\beta_1) [(\mathbf{y} - X\boldsymbol{\beta})^t W (\mathbf{y} - X\boldsymbol{\beta}) + qr]^{-\left(\frac{n+q}{2}\right)}$$

$$\propto p(\beta_0) p(\beta_1) \left[ (\mathbf{y} - X\hat{\boldsymbol{\beta}})^t W (\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) X^t W X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + qr \right]^{-\left(\frac{n+q}{2}\right)}$$

$$\propto p(\beta_0) p(\beta_1) \left[ + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) X^t W X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(\mathbf{y} - X\hat{\boldsymbol{\beta}})^t W (\mathbf{y} - X\hat{\boldsymbol{\beta}}) + qr} \right]^{-\left(\frac{n+q}{2}\right)}$$

$$\propto p(\beta_0)p(\beta_1)\left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^t X^t W X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{k^2(n-2+q)}\right]^{-\left(\frac{n-2+q+2}{2}\right)}$$

where

$$k^2 = \frac{(\mathbf{y}-X\hat{\boldsymbol{\beta}})^t W(\mathbf{y}-X\hat{\boldsymbol{\beta}})+qr}{n-2+q}$$

which will depend on the prior of $\beta_0$ and $\beta_1$

It will be a truncated t-distribution if priors of $\beta_0$ and $\beta_1$ are uniform.

**<u>Joint Posterior density of $\beta_0, \beta_1$ and $\sigma^2$.</u>**

$$p(\beta_0,\beta_1,\sigma^2|y,\upsilon) \propto \int p((\beta_0,\beta_1,\sigma^2,\omega|y,\upsilon)d\omega$$

$$\propto \prod_{i=1}^n \int_0^{+\infty}\left(\left(\frac{\sigma^2}{\omega_i}\right)^{-1/2}\times\omega_i^{\upsilon/2-1}\exp\left\{-\frac{\omega_i}{2}\left(\frac{(y_i-\beta_0-\beta_1x_i)^2+\upsilon\sigma^2}{\sigma^2}\right)\right\}\right)d\omega_i$$
$$\times p(\beta_0)p(\beta_1)p(\sigma^2)$$
$$\propto p(\beta_0)p(\beta_1)p(\sigma^2)\prod_{i=1}^n(\sigma^2)^{-\frac{n}{2}}\int_0^{+\infty}\omega_i^{\left(\frac{\upsilon+1}{2}-1\right)}\exp\{-\frac{\omega_i s_i}{2}\}d\omega_i$$

the formula inside the integral is a gamma density function with parameter

$$Ga\left(\frac{\upsilon+1}{2},\frac{s_i}{2}\right), \text{ where}$$

$$s_i = \frac{(y_i-\beta_0-\beta_1x_i)^2+\upsilon\sigma^2}{\sigma^2}$$

so

$$p(\beta_0,\beta_1,\sigma^2|y,\upsilon) \propto p(\beta_0)p(\beta_1)p(\sigma^2)$$
$$\times(\sigma^2)^{-n/2}\prod_{i=1}^n\Gamma\left(\frac{\upsilon+1}{2}\right)\left(\frac{s_i}{2}\right)^{-\frac{\upsilon+1}{2}}$$
$$\propto p(\beta_0)p(\beta_1)p(\sigma^2)(\sigma^2)^{-n/2}\prod_{i=1}^n\left[\frac{(y_i-\beta_0-\beta_1x_i)^2+\upsilon\sigma^2}{\sigma^2}\right]^{-\frac{\upsilon+1}{2}}$$
$$\propto p(\beta_0)p(\beta_1)p(\sigma^2)\prod_{i=1}^n(\sigma^2)^{-\frac{1}{2}}\left[1+\frac{(y_i-\beta_0-\beta_1x_i)^2}{\upsilon\sigma^2}\right]^{-\frac{\upsilon+1}{2}}$$

## 5.5 Bayesian updating:

Bayesian learning in a continuous case:

Suppose that data accrue sequentially as $y_1,..,y_K$ and that the problem is to infer a parameter vector $\theta$.

$$p(\theta|y_1,..,y_K) = \frac{g(\theta)p(y_1,\ldots,y_K|\theta)}{m(y_1,\ldots,y_K)}$$
$$= \frac{g(\theta)p(y_1|\theta)p(y_2|y_1,\theta)\ldots.p(y_K|y_1,\ldots,y_{K-1},\theta)}{m(y_1)m(y_2|y_1)\ldots.m(y_K|y_1,\ldots,y_{K-1})}$$
$$\propto g(\theta)p(y_1|\theta)p(y_2|y_1,\theta)\ldots.p(y_K|y_1,\ldots,y_{K-1},\theta)$$

<u>Bayesian learning</u>

$$
\begin{aligned}
p(\theta|y_1,..,y_K) &\propto \underbrace{g(\theta)p(y_1|\theta)}_{p(\theta|y_1)}\ p(y_2|y_1,\theta)\ldots.p(y_K|y_1,\ldots,y_{K-1},\theta) \\
&\propto p(\theta|y_1)\times p(y_2|y_1,\theta)\ldots.p(y_K|y_1,\ldots,y_{K-1},\theta) \\
&\propto \underbrace{\underbrace{p(\theta|y_1,y_2)}p(y_3|y_1,y_2,\theta)}\ldots\ldots p(y_K|y_1,\ldots,y_{K-1},\theta) \\
&\propto p(\theta|y_1,y_2,y_3)\ldots\ldots p(y_K|y_1,\ldots,y_{K-1},\theta) \\
&\ \vdots \\
&\propto \underbrace{p(\theta|y_1,..,y_{K-1})}_{\text{acts as a prior}}\ p(y_K|y_1,\ldots,y_{K-1},\theta)
\end{aligned}
$$

*This means that a bayesian analysis carried out the end of the process will lead to the same inference about $\theta$ as one carried out sequentially.*

If $y_1,\ldots,y_K$ are independent, then

$$
p(\theta|y_1,..,y_K) \propto g(\theta)\prod_{i=1}^{n} p(y_i|\theta)
$$

Application:

Suppose at a stage (1), measurements $y_1$ are taken on $n_1$different observations and at a stage (2) measuerements $y_2$ are taken on $n_2$different observations.

Suppose that the objective is to infer the parameter $a_1$ and $a_2$ such that

$$
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \begin{pmatrix} 1_1\mu_1 \\ 1_2\mu_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}
$$

where $\mu_1$ and $\mu_2$ are known location parameters common to the data collected in stage 1 and stage 2. and that

$$
\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} |\sigma_e^2 \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_e^2 \begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} \right)
$$

where $\sigma_e^2$ is known and $I_i$ is the identity matrix $(n_i \times n_i)$

a- Calculate $p(y_1,y_2|\mu_1,\mu_2,a_1,a_2,\sigma_e^2)$?

b- Calculate $p(a_1,a_2|y_1,y_2,\mu_1,\mu_2,,\sigma_e^2,\sigma_a^2)$? such that

$$
\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} |\sigma_a^2 \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_a^2 \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right)
$$

c-using your result from (b), calculate $p(a_1|y_1,\mu_1,,\sigma_e^2,\sigma_a^2)$ and $p(a_2|y_1,\mu_1,,\sigma_e^2,\sigma_a^2)$

d- calculate $p(a_2|y_1,y_2,\mu_1,\mu_2,\sigma_e^2,\sigma_a^2)$

## 5.6 Posterior quantile/modes/Mean/cov...

Marginal posterior distribution $p(\theta|y)$ gives an exact, complete description of the state of knowledge about an unknown, after observing the data. It can be parameter, hypothesis, model etc

When posterior distribution is derived we can seek information about:

- Highliting zones of relatively high density or probability
- modes (unimodality or multimodality)
- area where the true value of the parameter may be located

**Posterior probabilities**

The probability that parameter $\theta$ falls in some region $R$ of $\Theta$ is

$$p(\theta \in R|y) = \int_R p(\theta|y)d\theta$$

if $\theta = (\theta_1, \theta_2)$

$\theta_1$ : parameter of interest, $\theta_2$ : parameter of nuisance

$$\begin{aligned}
p(\theta_1 \in R_1|y) &= \int_{R_1} p(\theta_1|y)d\theta_1 \\
&= \int_{R_1}\int_{\Theta_2} p(\theta_2,\theta_1|y)d\theta_1 d\theta_2 \\
&= \int_{R_1}\int_{\Theta_2} p(\theta_1|\theta_2,y)p(\theta_2|y)d\theta_2 d\theta_1 \\
&= \int_{\Theta_2} \underbrace{\left[\int_{R_1} p(\theta_1|\theta_2,y)d\theta_1\right]}_{p(\theta_1 \in R_1|\theta_2,y)} p(\theta_2|y)d\theta_2 \\
&= E_{[\theta_2|y]}(p(\theta_1 \in R_1|\theta_2,y))
\end{aligned}$$

So the marginal probablity $p(\theta_1 \in R_1|y)$ is a weighted average of $p(\theta_1 \in R_1|\theta_2,y)$ at each value of $\theta_2$ with the weight function, the marginal $p(\theta_2|y)$.

- will be used by MCMC:
- useful when $\theta_2$ is not feasible to integrat over

suppose that $\theta_2^{(1)},\theta_2^{(2)},..\theta_2^{(m)}$ are drawn from $p(\theta_2|y)$, then

$\hat{p}(\theta_1 \in R_1|y) = \frac{1}{m}\sum_{i=1}^m p(\theta_1 \in R_1|\theta_2^{(i)},y)$

- alternative consistent estimator of $\int_{R_1} p(\theta_1|y)d\theta_1$ is

$$\hat{p}(\theta_1 \in R_1|y) = \frac{1}{m}\sum_{i=1}^m I(\theta_1^{(i)} \in R_1)$$

where $\theta_2^{(1)},\theta_2^{(2)},..\theta_2^{(m)}$ are Monte carlo draw from $[\theta_1|y]$

**Posterior quantile**

$$(\theta|y) \text{ is defined in } [\theta_L, \theta_U]$$

*Definition: α-quantile of* $(\theta|y)$ ***is the value*** $"q"$, *:*

$$p(\theta \le q|y) = \alpha$$

Example:

A commonly used quantile to characterize location of a posterior distribution is the posterior median. If $q = m = median,$ then

$$p(\theta \le q|y) = p(\theta > q|y) = 0.5$$

or

$$\int_{-\infty}^{q} p(\theta|y)d\theta = \int_{q}^{+\infty} p(\theta|y)d\theta$$

Quantiles are used for high-credibility sets which are sets of values $\theta$ that contains the true value at high probability.

A credibility set of size $(1-\alpha)$ is given by all value of $\theta \in [q_1, q_2]$ such that

$$1-\alpha = p(q_1 \le \theta \le q_2|y)$$

$$p(\theta \le q_1|y) = \frac{\alpha}{2}$$

$$p(\theta > q_2|y) = \frac{\alpha}{2}$$

So the credibility set of $\theta$ is $[q_1, q_2]$

if $\alpha = 5\%$ then $q_1 = 2.5\%$ quantile of $\theta|y$ and $q_2 = 97.5\%$ quantile of $\theta|y$
$[q_1, q_2]$ is called a $(1-\alpha)$ highest posterior density or *HPD*.

There are infinity of credibility set.
Particular case: ***Highest posterior density interval or HPD***
is cerdibility set where values inside have higher posterior density than values outside it.

Remark: the median of a posterior distribution can be used as a ”point estimate” of $\theta$.

Illustration

Suppose we want to estimate

$$\theta = (\underbrace{\theta_1}_{\text{scalar}}, \underbrace{\theta_2^t}_{\text{nuisance}})^t$$

Lemma: if $\hat{\theta}_1$ is the optimum (estimate) of $\theta_1$, then $p(\theta_1 \leq \hat{\theta}_1|y) = \frac{1}{2}$

Hint:

- Choose a loss function $L(\hat{\theta}_1, \theta_1|y) = a|\hat{\theta}_1 - \theta_1|$
- Calculate $E\left[L(\hat{\theta}_1, \theta_1|y)\right] = \int_{-\infty}^{+\infty}\int_{R_{\theta_2}} a|\hat{\theta}_1 - \theta_1|\, p(\theta_1, \theta_2|y)d\theta_1 d\theta_2$

$$E\left[L(\hat{\theta}_1, \theta_1|y)\right] = \int_{-\infty}^{+\infty} a|\hat{\theta}_1 - \theta_1|p(\theta_1|y)d\theta_1$$

we have

$$E\left[L(\hat{\theta}_1, \theta_1|y)\right] = a\left[\int_{-\infty}^{\hat{\theta}_1}(\hat{\theta}_1 - \theta_1)p(\theta_1|y)d\theta_1 + \int_{\hat{\theta}_1}^{+\infty}(\hat{\theta}_1 - \theta_1)p(\theta_1|y)d\theta_1\right]$$

$$f \propto \hat{\theta}_1 p(\theta_1 \leq \hat{\theta}_1|y) - \int_{-\infty}^{\hat{\theta}_1}\theta_1 p(\theta_1|y)d\theta_1 +$$

$$\int_{\hat{\theta}_1}^{+\infty}\theta_1 p(\theta_1|y)d\theta_1 - \hat{\theta}_1\left[1 - p(\theta_1 \leq \hat{\theta}_1|y)\right]$$

we are looking for $\hat{\theta}_1$ which minimizes above, so

$$\frac{\partial f}{\partial \hat{\theta}_1} = 0, \Rightarrow$$

$$\frac{\partial f}{\partial \hat{\theta}_1} = 2p(\hat{\theta}_1 \leq \hat{\theta}_1|y) - 1 = 0$$

$$2p(\theta_1 \leq \hat{\theta}_1|y) - 1 = 0$$

$$p(\theta_1 \leq \hat{\theta}_1|y) = 0.5$$

satified when $\hat{\theta}_1$ is the median.

**Posterior mode**:

$\tilde{\theta}$ is the value of the parameter $\theta$ having the highest density.

$$\begin{aligned}\tilde{\theta} &= \arg \max_{\Theta} \; p(\theta|y) \\ &= \arg \max_{\Theta} \; \left[ c\, L(\theta|y) g(\theta) \right] \\ &= \arg \max_{\Theta} \; \left[ \log \, L(\theta|y) + \log \, g(\theta) \right]\end{aligned}$$

The mode can also be used as a ”point estimate” of $\theta$. It is considered as a location of the posterior parameter.

Hint: following O’hagan (1994)

- Choose a loss function $L(\hat{\theta}_1, \theta_1|y) = \begin{matrix} 0 \text{ if } |\hat{\theta}_1 - \theta_1| \leq b \\ 1 \text{ if } |\hat{\theta}_1 - \theta_1| > b \end{matrix}$

- Calculate

$$\begin{aligned}E\left[ L(\hat{\theta}_1, \theta_1|y) \right] &= 0 \times p(|\hat{\theta}_1 - \theta_1| \leq b|y) + 1 \times p(|\hat{\theta}_1 - \theta_1| > b|y) \\ &= p(|\hat{\theta}_1 - \theta_1| > b|y) \\ &= p(\hat{\theta}_1 - \theta_1 > b|y) + p(\theta_1 - \hat{\theta}_1 > b|y) \\ &= \underbrace{\int_{-\infty}^{\hat{\theta}_1 - b} p(\theta_1|y) d\theta_1 + \int_{\hat{\theta}_1 + b}^{+\infty} p(\theta_1|y) d\theta_1}_{f}\end{aligned}$$

$$\frac{\partial f}{\partial \hat{\theta}_1} = 0, \Rightarrow p(\hat{\theta}_1 - b|y) = p(\hat{\theta}_1 + b|y)$$

If $\hat{\theta}_1$ is a mode, from $(*)$ $\begin{matrix} p(\hat{\theta}_1|y) \geq p(\hat{\theta}_1 - b|y) \\ p(\hat{\theta}_1|y) \geq p(\hat{\theta}_1 + b|y) \end{matrix}$ $\Rightarrow$ necessary condition

If $b \to 0$, $\begin{matrix} p(\hat{\theta}_1|y) \geq p(\hat{\theta}_1 - b|y) \\ p(\hat{\theta}_1|y) \geq p(\hat{\theta}_1 + b|y) \end{matrix}$ , is correct if $\hat{\theta}_1$ is the mode $\Rightarrow$ sufficient

**Example**: Joint and Marginal modes

Suppose that

$$n_1 \text{of } y_1 \sim N(\mu_1, \sigma^2),\ \text{Treatment 1,}$$

$$n_2 \text{ of } y_2 \sim N(\mu_2, \sigma^2),\ \text{Treatment 2}$$

$$n_3 \text{ of } y_3 \sim N(\mu_3, \sigma^2),\ \text{Treatment 3}$$

*we want to see if there is a difference between different treaments?*

$$p(\mu_1 - \mu_2|y) \text{ and } p(\mu_1 - \mu_3|y)?$$

Parameter $\theta_1 = \mu_1 - \mu_2,$ or $\mu_1 - \mu_3$
Nuisance parameter is $\theta_2 = \sigma^2$

Likelihood=

$$p(y_{11},..,y_{1n_1},y_{21}..,y_{2n_2},y_{31},..y_{3n_3},|\mu_1,\mu_2,\mu_3,\sigma^2)$$

$$\propto \prod_{i=1}^{3}\prod_{j=1}^{n_i}(\sigma^2)^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\mu_i)^2\right\}$$

$$\propto (\sigma^2)^{-\frac{n_1+n_2+n_3}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{3}\left[\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + n_i(\bar{y}_i-\mu_i)^2\right]\right\}$$

then

$$\hat{\sigma}^2 = \frac{\sum_i \sum_j (y_{ij}-\bar{y}_i)^2}{n_1+n_2+n_3}$$

and we have
$$\bar{y}_i \sim N(\mu_i, \tfrac{\sigma^2}{n_i})$$
$$\hat{\sigma}^2 \sim \chi^2_{(n_1+n_2+n_3-3)} \times \tfrac{\sigma^2}{n_1+n_2+n_3}$$

if we choose a prior

$$\sigma^2 \sim \chi^2(\upsilon, \tau^2)$$
$$\mu_i \sim U(\mu_{\min}, \mu_{\max})$$
$$\theta = (\mu_1, \mu_2, \mu_3, \sigma^2)$$
$$p(\mu_1, \mu_2, \mu_3, \sigma^2|y) \propto p(y|\theta)p(\mu_1, \mu_2, \mu_3)p(\sigma^2|\upsilon, \tau)$$
$$p(y|\theta) \propto (\sigma^2)^{-\frac{n_1+n_2+n_3}{2}} \exp\left\{ \begin{array}{c} -\frac{1}{2\sigma^2} \sum_i [\sum_j (y_{ij} - \mu_i)^2 \\ +n_i(\mu_i - \bar{y}_i)] \end{array} \right\}$$

since

$$p(\mu_1, \mu_2, \mu_3) \propto Cst$$
$$p(\sigma^2|\upsilon, \tau) \propto (\sigma^2)^{-\frac{\upsilon}{2}} e^{\upsilon\tau^2/2}$$

then

$$p(\mu_1, \mu_2, \mu_3, \sigma^2|y)$$
$$\propto f = (\sigma^2)^{-\frac{n+v+2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(n\hat{\sigma}^2 + \upsilon\tau^2 + \sum_i n_i(\mu_i - \bar{y}_i)^2\right)\right\}$$

So the mode of the joint posterior distribution is obtained by deriving the logarithm of ”f” with respect to all parameters.

$$\frac{\partial f}{\partial \mu_i} = 0 \Rightarrow \quad \approx{\mu}_i = \bar{y}_i, \; i = 1,2,3$$
$$\frac{\partial f}{\partial \sigma^2} = 0 \Rightarrow \overset{\approx}{\sigma}^2 = \frac{n\hat{\sigma}^2 + \upsilon\tau^2}{n + \upsilon + 2}$$

this gives us the mode of the joint distribution (posterior) of $(\mu_1, \mu_2, \mu_3, \sigma^2)$.
The marginal posterior distribution of $\sigma^2$ is obtained by:

$$
\begin{aligned}
p(\sigma^2|y) &= \int_{\mu_{\min}}^{\mu_{\max}} f\, d\mu_1 d\mu_2 d\mu_3 \\
&\propto (\sigma^2)^{-\frac{n_1+n_2+n_3-3+\upsilon+2}{2}} \exp\left\{-\frac{1}{2\sigma^2}(n\hat{\sigma}^2 + \upsilon\tau^2)\right\} \\
&\quad \times \prod_{i=1}^{3}\left(\Phi(\frac{\mu_{\max} - \bar{y}_i}{\sigma^2/n_i}) - \Phi(\frac{\mu_{\min} - \bar{y}_i}{\sigma^2/n_i})\right)
\end{aligned}
$$

when we use an improper (vague prior)

$$\mu_{\min} = \mu_{\max} \longrightarrow \pm\infty$$

so

$$
\begin{aligned}
p(\sigma^2|y) &\propto (\sigma^2)^{-\frac{n_1+n_2+n_3-3+\upsilon+2}{2}} \exp\left\{-\frac{1}{2\sigma^2}(n\hat{\sigma}^2 + \upsilon\tau^2)\right\} \\
&\sim \chi^2\left(\upsilon^* = n + \upsilon - 3, (\tau^*)^2 = \frac{n\hat{\sigma}^2 + \upsilon\tau^2}{\upsilon^*}\right)
\end{aligned}
$$

the mode of $(\sigma^2|y)$ is $\tilde{\sigma}^2 = \frac{\upsilon^* (\tau^*)^2}{\upsilon^*+2}$

**Remak**:

- The mode of the joint posterior distribution of $\sigma^2$ is $\overset{\approx}{\sigma}{}^2 = \frac{n\hat{\sigma}^2+\upsilon\tau^2}{n+\upsilon+2}$
- The mode of the marginal posterior distribution of $\sigma^2$ is

$$\tilde{\sigma}^2 = \frac{(n+\upsilon-3)(n\hat{\sigma}^2+\upsilon\tau^2)/(n+\upsilon-3)^2}{(n+\upsilon-1)}$$

They are different

Using the previous posterior distribution we have:

$$\begin{aligned}
& p(\mu_1,\mu_2,\mu_3|y) \\
& \propto \int p(\mu_1,\mu_2,\mu_3,\sigma^2|y)d\sigma^2 \text{ (has an } IG_a\text{)} \\
p(\mu_1,\mu_2,\mu_3|y) & \propto \left(n\hat{\sigma}^2+\upsilon\tau^2+\sum_i n_i(\mu_i-\bar{y}_i)^2\right)^{-\frac{n+\upsilon}{2}} \\
& \propto \left(1+\frac{\sum_i n_i(\mu_i-\bar{y}_i)^2}{n\hat{\sigma}^2+\upsilon\tau^2}\right)^{-\frac{n-3+\upsilon+3}{2}} \\
& \propto \left(1+\frac{(\mu-\bar{y})^t N(\mu-\bar{y})}{(n-3+\upsilon)c^2}\right)^{-\frac{n-3+\upsilon+3}{2}} \\
& \sim t(\bar{y},(n-3+\upsilon),var)
\end{aligned}$$

where:

- $N = diag(n_1,n_2,n_3)$
- $c^2 = \frac{n\hat{\sigma}^2+\upsilon\tau^2}{n-3+\upsilon}$
- $var = c^2N^{-1}\frac{n-3+\upsilon}{n-5+\upsilon}$

This is a Unimodale distribution with mode

$$\mu_i = \bar{y}_i$$

which is the same mode as the one for the joint posterior distribution
so the marginal distribution of $\mu_i$ is

$$\mu_i \sim t(\bar{y}_i, n - 3 + \upsilon, var)$$

$$var = c^2 \frac{n - 3 + \upsilon}{n_i(n - 5 + \upsilon)}$$

and then

$$(\mu_i - \mu_j) \sim t(\bar{y}_i - \bar{y}_i, n - 3 + \upsilon, var)$$

$$var = c^2 \frac{n - 3 + \upsilon}{(n - 5 + \upsilon)} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)$$

**Posterior Mean (vector)**

Mean

$$E(\theta_1|y) = E_{\theta_2|y}[E(\theta_1|\theta_2, y)]$$

Estimator:

$$\hat{E}(\theta_1|y) = \frac{1}{m}\sum_{i=1}^{m}\left[E(\theta_1|\theta_2^{(i)}, y)\right]$$

$$or$$

$$\tilde{\tilde{E}}(\theta_1|y) = \frac{1}{m}\sum_{i=1}^{m}\theta_1^{(i)}$$

where

- $\theta_2^{(i)}$ can be draws from $(\theta_2|y)$
- $\theta_1^{(i)}$ can be draws from $(\theta_1|y)$

**Posterior covariance (matrix)**

Defined as

$$
\begin{aligned}
cov((\theta_1, \theta_2|y) &= \int\int(\theta_1 - \bar{\theta}_1)(\theta_2 - \bar{\theta}_2)^t p(\theta_1, \theta_2|y) d\theta_1 d\theta_2 \\
&= \int\int(\theta_1 \theta_2^t) p(\theta_1, \theta_2|y) d\theta_1 d\theta_2 - \bar{\theta}_1 \bar{\theta}_2
\end{aligned}
$$

If $\theta_3$ is a third parameter (nuisance parameter), then

$$
\begin{aligned}
cov((\theta_1, \theta_2|y) = & \; E_{(\theta_3|y)}[Cov(\theta_1, \theta_2|\theta_3, y)] \\
& + Cov_{(\theta_3|y)}[E(\theta_1|\theta_3, y), E(\theta_2|\theta_3, y)]
\end{aligned}
$$