# Lecture 1: Overview of Bayesian Econometrics

Hedibert Freitas Lopes

The University of Chicago Booth School of Business
5807 South Woodlawn Avenue, Chicago, IL 60637
http://faculty.chicagobooth.edu/hedibert.lopes

hlopes@ChicagoBooth.edu

# Outline

# Example $i$. Sequential learning

- John claims some discomfort and goes to the doctor.
- The doctor believes John may have the disease A.
- $\theta = 1$: John has disease A; $\theta = 0$: he does not.
- The doctor claims, based on his expertise (H), that

$$P(\theta = 1|H) = 0.70$$

- Examination $X$ is related to $\theta$ as follows

$$\begin{cases} P(X = 1|\theta = 0) = 0.40, & \text{positive test given no disease} \\ P(X = 1|\theta = 1) = 0.95, & \text{positive test given disease} \end{cases}$$

# Observe $X = 1$

Exam's result: $X = 1$

$$\begin{aligned}
P(\theta = 1|X = 1) &\propto P(X = 1|\theta = 1)P(\theta = 1) \\
&\propto (0.95)(0.70) = 0.665 \\
P(\theta = 0|X = 1) &\propto P(X = 1|\theta = 0)P(\theta = 0) \\
&\propto (0.40)(0.30) = 0.120
\end{aligned}$$

Consequently

$$\begin{aligned}
P(\theta = 0|X = 1) &= 0.120/0.785 = 0.1528662 \text{ and} \\
P(\theta = 1|X = 1) &= 0.665/0.785 = 0.8471338
\end{aligned}$$

The information $X = 1$ increases, for the doctor, the probability that John has the disease $A$ from 70% to 84.71%.

# Posterior predictive

John undertakes the test $Y$, which relates to $\theta$ as follows[1]

$$P(Y=1|\theta=1)=0.99 \quad \text{and} \quad P(Y=1|\theta=0)=0.04$$

Then, the predictive of $Y=0$ given $X=1$ is given by

$$\begin{aligned}
P(Y=0|X=1) &= P(Y=0|X=1,\theta=0)P(\theta=0|X=1)\\
&+ P(Y=0|X=1,\theta=1)P(\theta=1|X=1)\\
&= P(Y=0|\theta=0)P(\theta=0|X=1)\\
&+ P(Y=0|\theta=1)P(\theta=1|X=1)\\
&= (0.96)(0.1528662)+(0.01)(0.8471338)\\
&= 15.52\%
\end{aligned}$$

**Key condition:** $X$ and $Y$ are conditionally independent given $\theta$.

[1] Recall that $P(X=1|\theta=1)=0.95$ and $P(X=1|\theta=0)=0.40$.

# Model criticism

Suppose the observed result was $Y = 0$. This is a reasonably unexpected result as the doctor only gave it roughly 15% chance.

He should at least consider rethinking the model based on this result. In particular, he might want to ask himself

1. Did 0.7 adequately reflect his $P(\theta = 1|H)$?
2. Is test $X$ really so unreliable?
3. Is the sample distribution of $X$ correct?
4. Is the test $Y$ so powerful?
5. Have the tests been carried out properly?

# Observe $Y = 0$

Let $H_2 = \{X = 1, Y = 0\}$. Then, Bayes theorem leads to

$$\begin{aligned}
P(\theta = 1|H_2) &\propto P(Y = 0|\theta = 1)P(\theta = 1|X = 1) \\
&\propto (0.01)(0.8471338) = 0.008471338 \\
P(\theta = 0|H_2) &\propto P(Y = 0|\theta = 0)P(\theta = 0|X = 1) \\
&\propto (0.96)(0.1528662) = 0.1467516
\end{aligned}$$

Therefore,

$$P(\theta = 1|X = 1, Y = 0) = \frac{P(Y = 0, \theta = 1|X = 1)}{P(Y = 0|X = 1)} = 0.0545753$$

$$P(\theta = 1|H_i) = \begin{cases} 0.7000 & , H_0\text{: before X and Y} \\ 0.8446 & , H_1\text{: after X=1 and before Y} \\ 0.0546 & , H_2\text{: after X=1 and Y=0} \end{cases}$$

# Example *ii*. Normal-normal

Consider a simple measurement error model

$$X = \theta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

where

$$\theta \sim N(\theta_0, \tau_0^2).$$

The quantities $(\sigma^2, \theta_0, \tau_0^2)$ are known.

The posterior distribution of $\theta$ (after $X = x$ is observed) is

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

More precisely,

$$
\begin{aligned}
p(\theta|x) &\propto \exp\{-0.5(\theta^2 - 2\theta x)/\sigma^2\}\exp\{-0.5(\theta^2 - 2\theta\theta_0)/\tau_0^2\} \\
&\times \exp\{-0.5(\theta^2(1/\sigma^2 + 1/\tau_0^2) + 2\theta(x/\sigma^2 + \theta_0/\tau_0^2)\} \\
&= \exp\{-0.5(\theta^2/\tau_1^2 + 2\theta\tau_1^2(x/\sigma^2 + \theta_0/\tau_0^2)/\tau_1^2\} \\
&= \exp\{-0.5(\theta^2 + 2\theta\theta_1)/\tau_1^2\}.
\end{aligned}
$$

Therefore, $\theta|x$ is normally distributed with

$$E(\theta|x) = \tau_1^2(x/\sigma^2 + \theta_0/\tau_0^2)$$

and

$$V(\theta|x) = \left(1/\sigma^2 + 1/\tau_0^2\right)^{-1}.$$

Notice that

$$E(\theta|x) = \omega\theta_0 + (1-\omega)x$$

where

$$\omega = \frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

measures the relative information contained in the prior distribution with respect to the total information (prior plus

# Illustration

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$.

Model: $(X|\theta) \sim N(\theta, (40)^2)$.

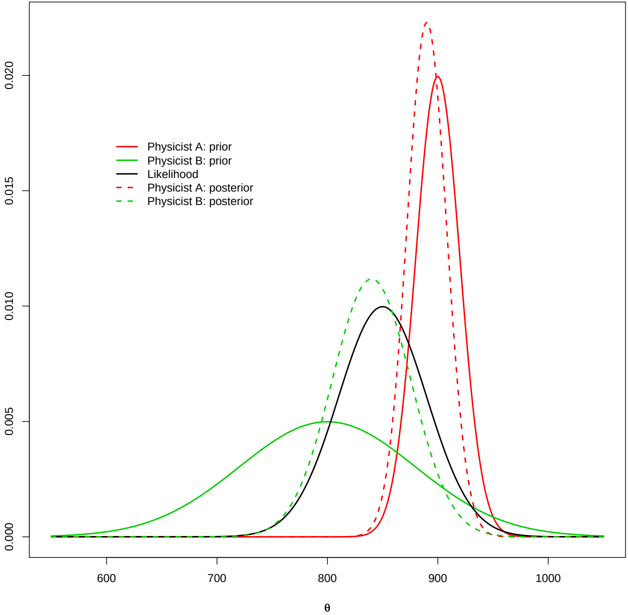
Observation: $X = 850$

$$\begin{aligned}(\theta|X = 850, H_A) &\sim N(890, (17.9)^2)\\(\theta|X = 850, H_B) &\sim N(840, (35.7)^2)\end{aligned}$$

Information (precision)

Physicist A: from 0.002500 to 0.003120 (an increase of 25%)

Physicist B: from 0.000156 to 0.000781 (an increase of 400%)


Physicist A: prior
Physicist B: prior
Likelihood
Physicist A: posterior
Physicist B: posterior
0.000
0.005
0.010
0.015
0.020
600
700
800
900
1000
θ

# Turning the Bayesian crank

We usually decompose

$$p(\theta, x|H)$$

into

$$p(\theta|H) \quad \text{and} \quad p(x|\theta, H)$$

The prior predictive distribution

$$p(x|H) = \int_\Theta p(x|\theta, H)p(\theta|H)\,d\theta = E_\theta[p(x|\theta, H)]$$

if of key importance in Bayesian model assessment.

# Posterior distribution

The posterior distribution of $\theta$ is obtained, after $x$ is observed, by Bayes' Theorem:

$$
\begin{aligned}
p(\theta|x,H) &= \frac{p(\theta,x|H)}{p(x|H)} \\
&= \frac{p(x|\theta,H)p(\theta|H)}{p(x|H)} \\
&\propto p(x|\theta,H)p(\theta|H).
\end{aligned}
$$

# Posterior predictive distribution

Let $y$ be a new set of observations conditionally independent of $x$ given $\theta$, ie.

$$p(x, y|\theta) = p(x|\theta, H)p(y|\theta, H).$$

Then,

$$\begin{aligned}
p(y|x, H) &= \int_\Theta p(y, \theta|x, H)d\theta \\
&= \int_\Theta p(y|\theta, x, H)p(\theta|x, H)d\theta \\
&= \int_\Theta p(y|\theta, H)p(\theta|x, H)d\theta \\
&= E_{\theta|x}\left[p(y|\theta, H)\right]
\end{aligned}$$

In general, but not always (time series, for example) $x$ and $y$ are independent given $\theta$.

It might be more useful to concentrate on prediction rather than on estimation since the former is *verifiable*.

$x$ and $y$ can be (and usually are) observed; $\theta$ can not!

# Sequential Bayes theorem

Experimental result: $x_1 \sim p_1(x_1|\theta)$

$$p(\theta|x_1) \propto l_1(\theta;x_1)p(\theta)$$

Experimental result: $x_2 \sim p_2(x_2|\theta)$

$$\begin{aligned} p(\theta|x_2,x_1) &\propto l_2(\theta;x_2)p(\theta|x_1) \\ &\propto l_2(\theta;x_2)l_1(\theta;x_1)p(\theta) \end{aligned}$$

Experimental results: $x_i \sim p_i(x_i|\theta)$, for $i = 3,\ldots,n$

$$\begin{aligned} p(\theta|x_n,\ldots,x_1) &\propto l_n(\theta;x_n)p(\theta|x_{n-1},\ldots,x_1) \\ &\propto \left[\prod_{i=1}^{n} l_i(\theta;x_i)\right] p(\theta) \end{aligned}$$

# Model probability

Suppose that the competing models can be enumerated and are represented by the set

$$M = \{M_1, M_2, \ldots\}$$

and that the *true model* is in $M$ (Bernardo and Smith, 1994).

The posterior model probability of model $M_j$ is given by

$$Pr(M_j|y) = \frac{f(y|M_j)Pr(M_j)}{f(y)}$$

# Ingredients

Prior predictive density of model $M_j$

$$f(y|M_j) = \int f(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j$$

Prior model probability of model $M_j$

$$Pr(M_j)$$

Overall prior predictive

$$f(y) = \sum_{M_j \in M} f(y|M_j)Pr(M_j)$$

# Posterior odds

The posterior odds of model $M_j$ relative to $M_k$ is given by

$$\underbrace{\frac{Pr(M_j|y)}{Pr(M_k|y)}}_{\text{posterior odds}} = \underbrace{\frac{Pr(M_j)}{Pr(M_k)}}_{\text{prior odds}} \times \underbrace{\frac{f(y|M_j)}{f(y|M_k)}}_{\text{Bayes factor}} .$$

The Bayes factor can be viewed as the weighted likelihood ratio of $M_j$ to $M_k$.

The main difficulty is the computation of the marginal likelihood or normalizing constant $f(y|M_j)$.

Therefore, the posterior model probability for model $j$ can be obtained from

$$\frac{1}{Pr(M_j|y)} = \sum_{M_k \in M} B_{kj} \frac{Pr(M_k)}{Pr(M_j)}.$$

# Bayes factor

Jeffreys (1961) recommends the use of the following rule of thumb to decide between models $j$ and $k$:

| $\log_{10} B_{jk}$ | $B_{jk}$ | Evidence against $k$ |
|---|---|---|
| 0.0 to 0.5 | 1.0 to 3.2 | Not worth more than a bare mention |
| 0.5 to 1.0 | 3.2 to 10 | Substantial |
| 1.0 to 2.0 | 10 to 100 | Strong |
| $> 2$ | $> 100$ | Decisive |

Kass and Raftery (1995) argue that "it can be useful to consider twice the natural logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics". Their slight modification is:

| $2 \log_e B_{jk}$ | $B_{jk}$ | Evidence against $k$ |
|---|---|---|
| 0.0 to 2.0 | 1.0 to 3.0 | Not worth more than a bare mention |
| 2.0 to 6.0 | 3.0 to 20 | Substantial |
| 6.0 to 10.0 | 20 to 150 | Strong |
| $> 10$ | $> 150$ | Decisive |

# Marginal likelihood

A basic ingredient for model assessment is given by the predictive density

$$f(y|M) = \int f(y|\theta, M)p(\theta|M)d\theta \ ,$$

which is the normalizing constant of the posterior distribution.

The predictive density can now be viewed as the likelihood of model $M$.

It is sometimes referred to as predictive likelihood, because it is obtained after marginalization of model parameters.

The predictive density can be written as the expectation of the likelihood with respect to the prior:

$$f(y) = E_p[f(y|\theta)].$$

# Example iii. Multiple linear regression

The standard Bayesian approach to multiple linear regression is

$$y_i = x_i'\beta + \epsilon_i$$

for $i = 1, \ldots, n$, $x_i$ a $q$-dimensional vector of regressors and residuals $\epsilon_i$ iid $N(0, \sigma^2)$.

In matrix notation,

$$(y|X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I_n)$$

where $y = (y_1, \ldots, y_n)$, $X = (x_1, \ldots, x_n)'$ is the $(n \times q)$, design matrix and $q = p + 1$.

# Example iii. Maximum likelihood estimation

It is well known that

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\sigma}^2 &= \frac{S_e}{n-q} = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-q}\end{aligned}$$

are the OLS estimates of $\beta$ and $\sigma^2$, respectively.

The conditional and unconditional sampling distributions of $\hat{\beta}$ are

$$\begin{aligned}(\hat{\beta}|\sigma^2, y, X) &\sim N(\beta, \sigma^2(X'X)^{-1}) \\ (\hat{\beta}|y, X) &\sim t_{n-q}(\beta, S_e(X'X)^{-1})\end{aligned}$$

respectively, with

$$(\hat{\sigma}^2|\sigma^2) \sim IG\left((n-q)/2, ((n-q)\sigma^2/2\right).$$

# Example iii. Conjugate prior

The prior distribution of $(\beta, \sigma^2)$ is $NIG(b_0, B_0, n_0, S_0)$, i.e.

$$\begin{aligned} \beta|\sigma^2 &\sim N(b_0, \sigma^2 B_0) \\ \sigma^2 &\sim IG(n_0/2, n_0 S_0/2) \end{aligned}$$

for known hyperparameters $b_0, B_0, n_0$ and $S_0$.

For clarification, when $\sigma^2 \sim IG(a, b)$, if follows that

$$p(\sigma^2) \propto (\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

with

$$E(\sigma^2) = \frac{b}{a-1} \quad \text{and} \quad V(\sigma^2) = \frac{b^2}{(a-1)^2(a-2)}$$

# Example iii. Conditionals

It is easy to show that

$$(\beta|\sigma^2, y, X) \sim N(b_1, \sigma^2 B_1)$$

where

$$\begin{aligned} B_1^{-1} &= B_0^{-1} + X'X \\ B_1^{-1} b_1 &= B_0^{-1} b_0 + X'y. \end{aligned}$$

It is also easy to show that

$$(\sigma^2|\beta, y, X) \sim IG(n_1/2, n_1 S_{11}(\beta)/2)$$

where

$$\begin{aligned} n_1 &= n_0 + n \\ n_1 S_{11}(\beta) &= n_0 S_0 + (y - X\beta)'(y - X\beta). \end{aligned}$$

# Example iii. Marginals

It can be shown that

$$(\sigma^2|y,X) \quad \sim \quad IG(n_1/2, n_1S_1/2)$$

where

$$n_1S_1 = n_0S_0 + (y - Xb_1)'y + (b_0 - b_1)'B_0^{-1}b_0.$$

Consequently,

$$(\beta|y,X) \sim t_{n_1}(b_1, S_1B_1).$$

# Example iii. MLE versus Bayes

Distributions of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$

$$
\begin{aligned}
(\hat{\sigma}^2|\sigma^2, y, X) &\sim IG\left((n-q)/2, ((n-q)\sigma^2/2\right) \\
(\hat{\beta}|\beta, y, X) &\sim t_{n-q}(\beta, S_e(X'X)^{-1}).
\end{aligned}
$$

Marginal posterior distributions of $\beta$ and $\sigma^2$

$$
\begin{aligned}
(\sigma^2|y, X) &\sim IG(n_1/2, n_1 S_1/2) \\
(\beta|y, X) &\sim t_{n_1}(b_1, S_1 B_1).
\end{aligned}
$$

Vague prior: When $B_0^{-1} = 0$, $n_0 = -q$ and $S_0 = 0$

$$
\begin{aligned}
b_1 &= \hat{\beta} \\
B_1 &= (X'X)^{-1} \\
n_1 &= n - q \\
n_1 S_1 &= (y - X\hat{\beta})'y = (y - X\hat{\beta})'(y - X\hat{\beta}) = (n-q)\hat{\sigma}^2 \\
S_1 B_1 &= \hat{\sigma}^2 (X'X)^{-1}.
\end{aligned}
$$

# Example iii. Predictive

The predictive density can be obtained by

$$p(y|X) = \int p(y|X, \beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)d\beta d\sigma^2$$

or (via Bayes' theorem) by

$$p(y|X) = \frac{p(y|X, \beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)}{p(\beta|\sigma^2, y, X)p(\sigma^2|y, X)}$$

which is valid for all $(\beta, \sigma^2)$.

Closed form solution for the multiple normal linear regression:

$$(y|X) \sim t_{n_0}(Xb_0, S_0(I_n + XB_0X')).$$

# Real data exercise

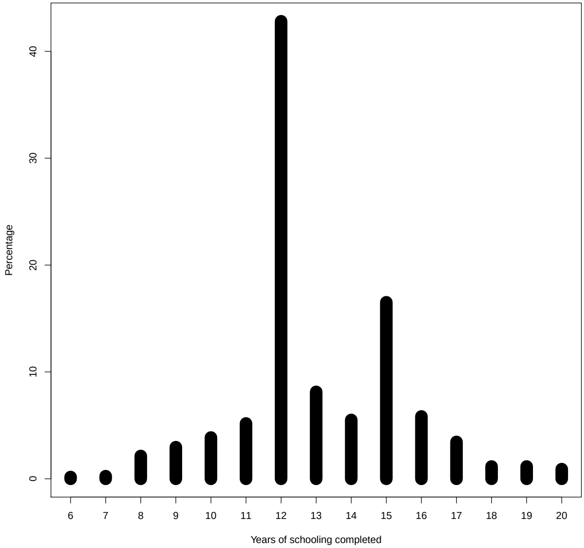
To better understand the differential role of the prior in estimation and model comparison, consider the following simple linear regression application, illustrated using a sample of $n = 1,217$ observations from the National Longitudinal Survey of Youth (NLSY):

- $\mathcal{M}_0 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$
- $\mathcal{M}_1 : y_i = \beta_0 \qquad\qquad + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$

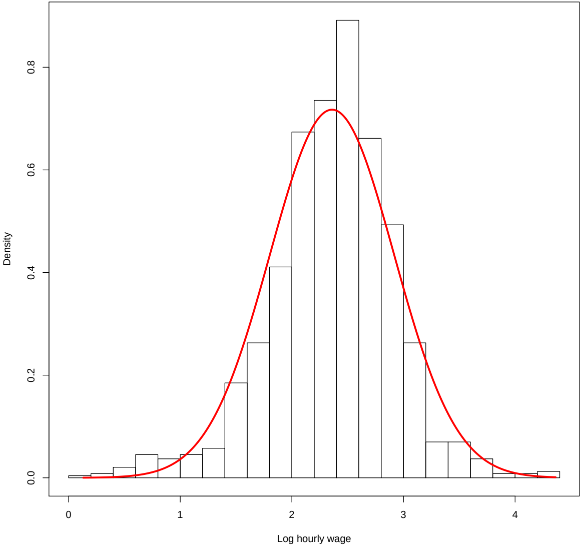
$y_i$: log hourly wage received by individual $i$.

$x_i$: education in years of schooling completed by individual $i$.

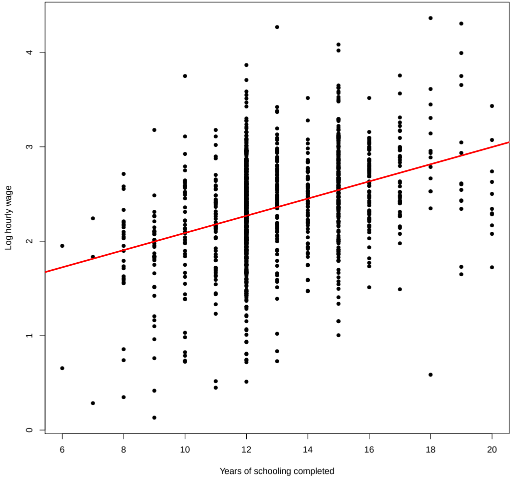
# Years of schooling completed



# Log hourly wage



# MLE regression




$\hat{\beta} = (1.17766, 0.09101)'$ and $\hat{\sigma}^2 = 0.2668455$.

Recall the conjugate prior for $(\beta, \sigma^2)$ is

$$\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0) \quad \text{and} \quad \sigma^2 \sim IG(n_0/2, n_0 S_0/2).$$

Let us assume that $b_0 = 0$, $n_0 = 6$ and $S_0 = 0.1333$.

Let us consider two different prior variance for $\beta$:

- Prior I: $B_0 = 1.0 \times 10^1 I_2$,
- Prior II: $B_0 = 1.0 \times 10^{100} I_2$.

| Posterior summary for the complete model $\mathcal{M}_0$ | | | | |
|---|---|---|---|---|
| Parameter | Prior I | | Prior II | |
| | Post. Mean | Post Std. | Post Mean | Post Std. |
| $\beta_0$ | 1.17439 | (0.08626) | 1.17439 | (0.08637) |
| $\beta_1$ | 0.09125 | (0.00654) | 0.09125 | (0.00655) |
| $\sigma^2$ | 0.26587 | (0.01073) | 0.26587 | (0.01073) |
| **Posterior summary for the restricted model $\mathcal{M}_1$** | | | | |
| $\beta_0$ | 2.35963 | (0.01592) | 2.35963 | (0.01591) |
| $\sigma^2$ | 0.30815 | (0.01244) | 0.30815 | (0.01244) |
| **Log Bayes factor of $\mathcal{M}_0$ versus $\mathcal{M}_1$** | | | | |
| $\log B_{01}$ | 84.7294 | | -29.8886 | |

$$\begin{aligned}\log B_{01}(\text{Prior I}) &= \log p(y|X,\mathcal{M}_0,\text{Prior I}) - \log p(y|X,\mathcal{M}_1,\text{Prior I})\\ &= -2458.713 - (-2543.442) = 84.7294\end{aligned}$$

$$\begin{aligned}\log B_{01}(\text{Prior II}) &= \log p(y|X,\mathcal{M}_0,\text{Prior II}) - \log p(y|X,\mathcal{M}_1,\text{Prior II})\\ &= -2686.406 - (-2656.517) = -29.8886\end{aligned}$$

# Example iv. Stochastic volatility

One of the most used models in financial econometrics is the diffusive stochastic volatility model, where log-returns are normally distributed

$$y_t|\theta_t, H \sim N(0; e^{\theta_t})$$

with heteroscedasticity modeled as

$$\theta_t|\theta_{t-1}, \gamma, H \sim N(\alpha + \beta\theta_{t-1}, \sigma^2)$$

for $t = 1, \ldots, T$ and $\gamma = (\alpha, \beta, \sigma^2)$, known for now.

The model is completed with

$$\theta_0|\gamma, H \sim N(m_0, C_0)$$

for known hyperparameters $(m_0, C_0)$.

# Example iv. Posterior distribution

For $\theta = (\theta_1, \ldots, \theta_T)'$, it follows that

$$
\begin{aligned}
p(\theta|y,H) \quad \propto & \quad \prod_{t=1}^{T} e^{-\theta_t/2} \exp\left\{-\frac{1}{2} y_t^2 e^{-\theta_t}\right\} \\
\times & \quad \prod_{t=1}^{T} \exp\left\{-\frac{1}{2\sigma^2}(\theta_t - \alpha - \beta\theta_{t-1})^2\right\} \\
\times & \quad \exp\left\{-\frac{1}{2C_0}(\theta_0 - m_0)^2\right\}
\end{aligned}
$$

Unfortunately, closed form solutions are rare!

- How to compute $E(\theta_{43}|y,H)$ or $V(\theta_{11}|y,H)$?
- How to obtain a 95% credible region for $(\theta_{35}, \theta_{36}|y,H)$?
- How to sample from $p(\theta|y,H)$?
- How to compute $p(y|H)$ or $p(y_{T+1}, \ldots, y_{T+k}|y,H)$?