ECON 616: Lecture Five: Introduction to Bayesian Inference

Ed Herbst

Last Updated: March 13, 2025

Modes of Interence

- Previously, we focussed on frequentist inference (repeated sampling prodecures)
- measures of accuracy and performance that we used to assess the statistical procedures were pre-experimental
- However, many statisticians and econometricians believed that post-experimental reasoning should be used to assess inference procedures
- wherein only the actual observation Y^T is relevant and not the other observations in the sample space that could have been observed

Example

Suppose Y_1 and Y_2 are independently and identically distributed and

$$P_\theta\{Y_i = \theta - 1\} = \frac{1}{2}, \quad P_\theta\{Y_i = \theta + 1\} = \frac{1}{2}$$

Consider the following confidence set

$$C(Y_1, Y_2) = \begin{cases} \frac{1}{2}(Y_1 + Y_2) & \text{if} \quad Y_1 \neq Y_2 \\ Y_1 - 1 & \text{if} \quad Y_1 = Y_2 \end{cases}$$

From a pre-experimental perspective $C(Y_1, Y_2)$ is a 75% confidence interval.

However, from a post-experimental perspective, we are a "100% confident" that $C(Y_1, Y_2)$ contains the"true" θ if $Y_1 \neq Y_2$, whereas we are only "50% percent" confident if $Y_1 = Y_2$.

Some Principles

Does it make sense to report a pre-experimental measure of accuracy, when it is known to be misleading after seeing the data?

Conditionality Principle: If an experiment is selected by some random mechanism independent of the unknown parameter θ, then only the experiment actually performed is relevant.

Most also agree with

Sufficiency Principle: Consider an experiment to determine the value of an unknown parameter θ and suppose that $\mathcal{S}(\cdot)$ is a sufficient statistic. If $\mathcal{S}(Y_1) = \mathcal{S}(Y_2)$ then Y_1 and Y_2 contain the same evidence with respect to θ.

Likelihood Principle

The combination of the quite reasonable Conditionality Principle and the Sufficiency Principle lead to the more controversial Likelihood Principle (see discussion in Robert (1994)).

Likelihood Principle: All the information about an unknown parameter θ obtainable from an experiment is contained in the likelihood function of θ given the data. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to another.

Frequentist maximum-likelihood estimation and inference typically violates the LP!

Bayesian methods do not

Bayesian Models

A Bayesian model consists of:

- parametric probability distribution for the data, which we will characterize by the density $p(Y^T|\theta)$
- prior distribution $p(\theta)$.

The density $p(Y^T|\theta)$ interpreted as a function of θ with fixed Y^T is the likelihood function.

The posterior distribution of the parameter θ, that is, the conditional distribution of θ given Y_T, can be obtained through Bayes theorem:

$$p(\theta|Y^T) = \frac{p(Y^T|\theta)p(\theta)}{\int p(Y^T|\theta)p(\theta)d\theta}$$

Bayesian Models continued

- can interpret this formula as an inversion of probabilities.
- think of the parameter θ as "cause" and the data Y^T as "effect"
- formula allows the calculation of the probability of a particular "cause" given the observed "effect" based on

the probability of the "effect" given the possible "causes"

Unlike in the frequentist framework, the parameter θ is regarded as a random variable.

This does, however, not imply that Bayesians consider parameters to be determined in a random experiment.

The calculus of probability is used to characterize the state of knowledge

Elephant in Room

Any inference in a Bayesian framework is to some extent sensitive to the choice of prior distribution $p(\theta)$.

The prior reflects the initial state of mind of an individual and is therefore "subjective"

Many econometricians believe that the result of a scientific inquiry should not depend on the subjective beliefs and very sceptical of Bayesian methods.

But all analysis involves some subjective choices!

Introduction to Bayesian Statistics

- denote the sample space by $\mathcal{Y}$ with elements Y^T.
- Probability distribution P will be defined on the product space $\Theta \otimes \mathcal{Y}$.
- The conditional distribution of θ given Y^T is denoted by P_{Y^T}
- P_θ denotes the conditional distribution of Y^T given θ

An Example

The parameter space is $\Theta = \{0, 1\}$,

the sample space is $\mathcal{Y} = \{0, 1, 2, 3, 4\}$.

	0	1	2	3	4
$P_{\theta=0}(Y)$	.75	.140	.04	.037	.033
$P_{\theta=1}(Y)$	.70	.251	.04	.005	.004

Suppose we consider $\theta = 0$ and $\theta = 1$ as equally likely a priori. Moreover, suppose that the observed value is $Y = 1$. The marginal probability of $Y = 1$ is

$$\begin{aligned} & P\{Y = 1 \mid \theta = 0\} P\{\theta = 0\} + P\{Y = 1 \mid \theta = 1\} P\{\theta = 1\} \\ = \; & 0.140 \cdot 0.5 + 0.251 \cdot 0.5 = 0.1955 \qquad (1) \end{aligned}$$

Example, Continued

The posterior probabilities for θ being zero or one are

$$
\begin{aligned}
P\{\theta = 0|Y = 1\} &= \frac{P\{Y = 1|\theta = 0\}P\{\theta = 0\}}{P\{Y = 1\}} = \frac{0.07}{0.1955} = 0.358 \\
P\{\theta = 1|Y = 1\} &= \frac{P\{Y = 1|\theta = 1\}P\{\theta = 1\}}{P\{Y = 1\}} = \frac{0.1255}{0.1955} = 0.642
\end{aligned}
$$

Thus, the observation $Y = 1$ provides evidence in favor of $\theta = 1$.

Example 2

Consider the linear regression model:

$$y_t = x_t'\theta + u_t, \quad u_t \sim iid\mathcal{N}(0,1), \tag{2}$$

which can be written in matrix form as $Y = X\theta + U$. We assume that $X'X/T \xrightarrow{p} Q_{XX}$ and $X'Y \xrightarrow{p} Q_{XY} = Q_{XX}\theta$. The dimension of θ is k. The likelihood function is of the form

$$p(Y|X,\theta) = (2\pi)^{-T/2} \exp\left\{ Y - X\theta)'(Y - X\theta)\right\}. \tag{3}$$

Suppose the prior distribution is of the form

$$\theta \sim \mathcal{N}\left(0_{k\times 1}, \tau^2 \mathcal{I}_{k\times k}\right) \tag{4}$$

with density

$$p(\theta) = (2\pi\tau^2)^{-k/2} \exp\left\{-\frac{1}{2\tau^2}\theta'\theta\right\} \tag{5}$$

For small values of τ the prior concentrates near zero, whereas for larger values of τ it is more diffuse.

Example 2, Continued

According to Bayes Theorem the posterior distribution of θ is proportional to the product of prior density and likelihood function

$$p(\theta|Y,X) \propto p(\theta)p(Y|X,\theta). \tag{6}$$

The right-hand-side is given by

$$p(\theta)p(Y|X,\theta) \propto (2\pi)^{-\frac{T+k}{2}}\tau^{-k}\exp\bigg\{-\frac{1}{2}[Y'Y - \theta'X'Y - Y'X\theta - \theta'X'X$$
$$- \tau^{-2}\theta'\theta]\bigg\}. \tag{7}$$

Example 2, Continued

The exponential term can be rewritten as follows

$$
\begin{aligned}
& Y'Y - \theta' X' Y - Y' X \theta - \theta' X' X \theta - \tau^{-2} \theta' \theta \\
&= Y'Y - \theta' X' Y - Y' X \theta + \theta' (X'X + \tau^{-2} \mathcal{I}) \theta \qquad (8) \\
&= \Big(\theta - (X'X + \tau^{-2}\mathcal{I})^{-1} X' Y \Big)' \Big(X'X + \tau^{-2}\mathcal{I} \Big) \\
&\quad \Big(\theta - (X'X + \tau^{-2}\mathcal{I})^{-1} X' Y \Big) \\
&\quad + Y'Y - Y'X(X'X + \tau^{-2}\mathcal{I})^{-1} X' Y.
\end{aligned}
$$

Thus, the exponential term is a quadratic function of θ.

Example 2, Continued

The exponential term is a quadratic function of θ. This information suffices to deduce that the posterior distribution of θ must be a multivariate normal distribution

$$\theta | Y, X \sim \mathcal{N}(\tilde{\theta}_T, \tilde{V}_T) \tag{9}$$

with mean and covariance

$$\begin{aligned} \tilde{\theta}_T &= (X'X + \tau^{-2}\mathcal{I})^{-1} X'Y & (10) \\ \tilde{V}_T &= (X'X + \tau^{-2}\mathcal{I})^{-1}. & (11) \end{aligned}$$

The maximum likelihood estimator for this problem is $\hat{\theta}_{mle} = (X'X)^{-1}X'Y$ and its asymptotic (frequentist) sampling variance is $T^{-1}Q_{XX}^{-1}$.

- Assumption that both likelihood function and prior are Gaussian made the derivation of the posterior simple.
- The pair of prior and likelihood is called conjugate
- leads to a posterior distribution that is from the same family

Takeaway

As $\tau \longrightarrow \infty$ the prior becomes more and more diffuse and the posterior distribution becomes more similar to the sampling distribution of $\hat{\theta}_{mle}|\theta$:

$$\theta|Y, X \overset{approx}{\sim} \mathcal{N}\left(\hat{\theta}_{mle}, (X'X)^{-1}\right). \tag{12}$$

If $\tau \longrightarrow 0$ the prior becomes dogmatic and the sample information is dominated by the prior information. The posterior converges to a point mass that concentrates at $\theta = 0$.

In large samples (fixed τ, $T \longrightarrow \infty$) the effect of the prior becomes negligibleand the sample information dominates

$$\theta|Y, X \overset{approx}{\sim} \mathcal{N}\left(\hat{\theta}_{mle}, T^{-1} Q_{XX}^{-1}\right). \quad \square \tag{13}$$

Estimation and Inference

- In principle, all the information with respect to θ is summarized in the posterior $p(\theta|Y)$ and we could simply report the posterior density to our audience.

- However, in many situations our audience prefers results in terms of point estimates and confidence intervals, rather than in terms of a probability density.

- we might be interested to answer questions of the form: do the data favor model $\mathcal{M}_1$ or $\mathcal{M}_2$?

Adopt a decision theoretic approach

Decision Theoretic Approach

decision rule $\delta(Y^T)$ that maps observations into decisions, and a loss function $L(\theta, \delta)$ according to which the decisions are evaluated.

$$\begin{aligned} \delta(Y^T) &: \mathcal{Y} \mapsto \mathcal{D} && (14) \\ L(\theta, \delta) &: \Theta \otimes \mathcal{D} \mapsto R^+ && (15) \end{aligned}$$

$\mathcal{D}$ denotes the decision space.

The goal is to find decisions that minimize the posterior expected loss $E_{Y^T}[L(\theta, \delta(Y^T))]$.

The expectation is taken conditional on the data x, and integrates out the parameter θ.

Point Estimation

the goal is to construct a point estimate $\delta(Y^T)$ of θ. It involves two steps:

- Find the posterior $p(\theta|Y^T)$.
- Determine the optimal decision $\delta(Y^T)$.

The optimal decision depends on the loss function $L(\theta, \delta(Y^T))$.

Example 1, Continued

Consider the zero-one loss function

$$L(\theta, \delta) = \left\{ \begin{array}{cc} 0 & \delta = \theta \\ 1 & \delta \neq \theta \end{array} \right\}. \tag{16}$$

The posterior expected loss is $E_Y[L(\theta, \delta)] = 1 - E_Y\{\theta = \delta\}$ The optimal decision rule is

$$\delta = \text{argmax}_{\theta' \in \Theta} \ P_Y\{\theta = \theta'\} \tag{17}$$

the point estimator under the zero-one loss is equal to the parameter value that has the highest posterior probability. We showed that

$$\begin{aligned} P\{\theta = 0|Y = 1\} &= 0.358 && (18) \\ P\{\theta = 1|Y = 1\} &= 0.642 && (19) \end{aligned}$$

Thus $\delta(Y = 1) = 1$.

Example 2, Continued

The quadratic loss function is of the form $L(\theta, \delta) = (\theta - \delta)^2$

The optimal decision rule is obtained by minimizing

$$\min_{\delta \in \mathcal{D}} E_{Y^T}[(\theta - \delta)^2] \tag{20}$$

It can be easily verified that the solution to the minimization problem is of the form $\delta(Y^T) = E_{Y^T}[\theta]$.

Thus, the posterior mean $\tilde{\theta}_T$ is the optimal point predictor under quadratic loss.

Asymptotically

Suppose data are generated from the model $y_t = x_t'\theta_0 + u_t$. Asymptotically the Bayes estimator converges to the "true" parameter θ_0

$$\begin{aligned}
\tilde{\theta}_T &= (X'X + \tau^{-2}\mathcal{I})^{-1}X'Y \qquad (21)\\
&= \theta_0 + \left(\frac{1}{T}X'X + \frac{1}{\tau^2 T}\mathcal{I}\right)^{-1}\left(\frac{1}{T}X'U\right)\\
&\xrightarrow{p} \theta_0
\end{aligned}$$

The disagreement between two Bayesians who have different priors will asymptotically vanish. $\square$

Testing Theory

Consider the hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_1 = \Theta / \Theta_0$.

Hypothesis testing can be interpreted as estimating the value of the indicator function $\{\theta \in \Theta_0\}$.

Consider the loss function

$$L(\theta, \delta) = \begin{cases} 0 & \delta = \{\theta \in \Theta_0\} & \text{correct decision} \\ a_0 & \delta = 0,\ \theta \in \Theta_0 & \text{Type 1 error} \\ a_1 & \delta = 1,\ \theta \in \Theta_1 & \text{Type 2 error} \end{cases} \tag{22}$$

Note that the parameters a_1 and a_2 are part of the econometricians preferences.

Optimal Decision Rule

$$\delta(Y^T) = \begin{cases} 1 & P_{Y^T}\{\theta \in \Theta_0\} \geq a_1/(a_0 + a_1) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

The expected loss is

$$E_{Y^T} L(\theta, \delta) = \{\delta = 0\} a_0 P_{Y^T}\{\theta \in \Theta_0\} + \{\delta = 1\} a_1 [1 - P_{Y^T}\{\theta \in \Theta_0\}]$$

Thus, one should accept the hypothesis $\theta \in \Theta_0$ (choose $\delta = 1$) if

$$a_1 P_{Y^T}\{\theta \in \Theta_1\} = a_1[1 - P_{Y^T}\{\theta \in \Theta_0\}] \leq a_0 P_{Y^T}\{\theta \in \Theta_0\} \tag{24}$$

Bayes Factors

Bayes Factors: ratio of posterior probabilities and prior probabilities in favor of that hypothesis:

$$B(Y^T) = \frac{\text{Posterior Odds}}{\text{Prior Odds}} = \frac{P_{Y^T}\{\theta \in \Theta_0\}/P_{Y^T}\{\theta \in \Theta_1\}}{P\{\theta \in \Theta_0\}/P\{\theta \in \Theta_1\}} \quad (25)$$

Example 1, Continued

Suppose the observed value of Y is 2. Note that

$$P_{\theta=0}\{Y \geq 2\} = 0.110 \tag{26}$$
$$P_{\theta=1}\{Y \geq 2\} = 0.049 \tag{27}$$

The frequentist interpretation of this result would be that there is significant evidence against $H_0 : \theta = 1$ at the 5 percent level.

Frequentist rejections are based on unlikely events that did not occur!!

The Bayesian answers in terms of posterior odds is

$$\frac{P_{Y=2}\{\theta = 0\}}{P_{Y=2}\{\theta = 1\}} = 1 \tag{28}$$

and in terms of the Bayes Factor $B(Y) = 1$. $Y = 2$ does not favor one versus the other model.

Example 2, Continued

Suppose we only have one regressor $k = 1$.

Consider the hypothesis $H_0 : \theta < 0$ versus $H_1 : \theta \geq 0$. Then,

$$P_{Y^T}\{\theta < 0\} = P\left\{\frac{\theta - \tilde{\theta}_T}{\sqrt{\tilde{V}_T}} < -\frac{\tilde{\theta}_T}{\sqrt{\tilde{V}_T}}\right\} = \Phi\left(-\tilde{\theta}_T/\sqrt{\tilde{V}_T}\right) \quad (29)$$

where $\Phi(\cdot)$ denotes the cdf of a $\mathcal{N}(0,1)$. Suppose that $a_0 = a_1 = 1$

H_0 is accepted if

$$\Phi\left(-\tilde{\theta}_T/\sqrt{\tilde{V}_T}\right) \geq 1/2 \quad \text{or} \quad \tilde{\theta}_T < 0 \quad (30)$$

Example 2, Continued

Suppose that $y_t = x_t\theta_0 + u_t$. Note that

$$\frac{\tilde{\theta}_T}{\sqrt{\tilde{V}_T}} = \sqrt{(\frac{1}{\tau^2} + \sum x_t^2)^{-1}} \sum x_t y_t \tag{31}$$

$$= \sqrt{T}\theta_0 \frac{\frac{1}{T}\sum x_t^2}{\sqrt{\frac{1}{T}\sum x_t^2 + \frac{1}{\tau^2 T}}} + \frac{\frac{1}{\sqrt{T}}\sum x_t u_t}{\sqrt{\frac{1}{T}\sum x_t^2 + \frac{1}{\tau^2 T}}} \tag{32}$$

$\tilde{\theta}_T/\sqrt{\tilde{V}_T}$ diverges to $+\infty$ if $\theta_0 > 0$ and $P_{Y^T}\{\theta < 0\}$ converges to zero.

Vice versa, if $\theta_0 < 0$ then $\tilde{\theta}_T/\sqrt{\tilde{V}_T}$ diverges to $-\infty$ and $P_{Y^T}\{\theta < 0\}$ converges to one.

Thus for almost all values of θ_0 (except $\theta_0 = 0$) the Bayesian test will provide the correct answer asymptotically.

Point Hypotheses

Suppose in the context of Example~2 we would like to test $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$.

Since $P\{\theta = 0\} = 0$ it follows that $P_{Y^T}\{\theta = 0\} = 0$ and the null hypothesis is never accepted!

This observations raises the question: are point hypotheses realistic?

Only, if one is willing to place positive probability λ on the event that the null hypothesis is true.

A modification of the prior

Consider the modified prior

$$p^*(\theta) = \lambda \Delta[\{\theta = 0\}] + (1-\lambda)p(\theta)$$

where $\Delta[\{\theta = 0\}]$ is a point mass or dirac function.

The marginal density of Y^T can be derived as follows

$$\begin{aligned}
\int p(Y^T|\theta)p^*(\theta)d\theta &= \lambda \int p(Y^T|\theta)\Delta[\{\theta = 0\}]d\theta \\
&\quad +(1-\lambda)\int p(Y^T|\theta)p(\theta)d\theta \\
&= \lambda \int p(Y^T|0)\Delta[\{\theta = 0\}]d\theta \\
&\quad +(1-\lambda)\int p(Y^T|\theta)p(\theta)d\theta \\
&= \lambda p(Y^T|0) + (1-\lambda)\int p(Y^T|\theta)p(\theta)d\theta
\end{aligned}$$

Evidence for $\theta = 0$

The posterior probability of $\theta = 0$ is given by {

$$
\begin{aligned}
P_{Y^T}\{\theta = 0\} &= \lim_{\epsilon \longrightarrow \mathbf{0}} P_{Y^T}\{0 \leq \theta \leq \epsilon\} && (33) \\
&= \lim_{\epsilon \longrightarrow \mathbf{0}} \frac{\lambda \int_{\mathbf{0}}^{\epsilon} p(Y^T|\theta)\Delta[\{\theta = 0\}]d\theta + (1-\lambda)\int_{\mathbf{0}}^{\epsilon} p(Y^T|\theta)p(\theta)d\theta}{\lambda p(Y^T|0) + (1-\lambda)\int p(Y^T|\theta)p(\theta)d\theta} \\
&= \frac{\lambda p(Y^T|0)}{\lambda p(Y^T|0) + (1-\lambda)\int p(Y^T|\theta)p(\theta)d\theta}. && (34)
\end{aligned}
$$

}

Example 2, Continued

Assume that $\lambda = 1/2$. In order to obtain the posterior probability that $\theta = 0$ we have to evaluate

$$p(Y|X, \theta = 0) = (2\pi)^{-T/2} \exp\left\{-\frac{1}{2} Y'Y\right\} \qquad (35)$$

and calculate the marginal data density

$$p(Y|X) = \int p(Y|X, \theta) p(\theta) d\theta. \qquad (36)$$

Typically, this is a pain! However, since everything is normal here, we can show:

$$\begin{aligned} p(Y|X) &= (2\pi)^{-T/2} \tau^{-k} |X'X + \tau^{-2}|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'Y - Y'X(X'X + \tau^{-2}\mathcal{I})^{-1}X'Y]\right\}. \end{aligned}$$

Posterior Odds

the posterior odds ratio in favor of the null hypothesis is given by

$$\frac{P_{Y^T}\{\theta = 0\}}{P_{Y^T}\{\theta \neq 0\}} = \tau^k |X'X + \tau^{-2}|^{1/2}$$
$$\times \exp\left\{-\frac{1}{2}[Y'X(X'X + \tau^{-2}\mathcal{I})^{-1}X'Y]\right\} \tag{37}$$

Taking logs and standardizing the sums by T^{-1} yields

$$\ln\left[\frac{P_{Y^T}\{\theta = 0\}}{P_{Y^T}\{\theta \neq 0\}}\right] = -\frac{T}{2}\left(\frac{1}{T}\sum x_t y_t\right)'\left(\frac{1}{T}\sum x_t x_t' + \frac{1}{\tau^2 T}\right)^{-1}$$

$$\times\left(\tfrac{1}{T}\sum x_t y_t\right) + \tfrac{k}{2}\ln T + \tfrac{1}{2}\ln\left|\tfrac{1}{T}\sum x_t x_t' + \tfrac{1}{\tau^2 T}\right| + k\ln\tau$$

Assessing Posterior Odds

Assume that Data Were Generated from $y_t = x_t'\theta_0 + u_t$.

$$
\begin{aligned}
& Y'X(X'X+\tau^{-2})^{-1}X'Y \\
&= \theta_0'X'X(X'X+\tau^{-2})^{-1}X'X\theta_0 + U'X(X'X+\tau^{-2})^{-1}X'U \\
&\quad + U'X(X'X+\tau^{-2})^{-1}X'X\theta_0 + \theta_0'X(X'X+\tau^{-2})^{-1}X'U \\
&= T\theta_0'\left(\frac{1}{T}\sum x_t x_t'\right)^{-1}\theta_0 + \sqrt{T}2\left(\frac{1}{\sqrt{T}}\sum x_t u_t\right)'\theta_0 \\
&\quad + \left(\frac{1}{\sqrt{T}}\sum x_t u_t\right)'\left(\frac{1}{T}\sum x_t x_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum x_t u_t\right) + O_p(1).
\end{aligned}
$$

Asymptotics

If the null hypothesis is satisfied $\theta_0 = 0$ then

$$\ln\left[\frac{P_{Y^T}\{\theta = 0\}}{P_{Y^T}\{\theta \neq 0\}}\right] = \frac{k}{2}\ln T + small \longrightarrow +\infty. \tag{38}$$

That is, the posterior odds in favor of the null hypothesis converge to infinity and the posterior probability of $\theta = 0$ converges to one.

On the other hand, if the alternative hypothesis is true $\theta_0 \neq 0$ then

$$\ln\left[\frac{P_{Y^T}\{\theta = 0\}}{P_{Y^T}\{\theta \neq 0\}}\right] = -\frac{T}{2}\theta_0'\left(\frac{1}{T}\sum x_t x_t'\right)^{-1}\theta_0 + small \longrightarrow -\infty.$$

and the posterior odds converge to zero, which implies that the posterior probability of the null hypothesis being true converges to zero.

Summing up

Bayesian test is consistent in the following sense.

- If the null hypothesis is "true" then the posterior probability of H_0 converges in probability to one as $T \longrightarrow \infty$.
- If the null hypothesis is false then the posterior probability of H_0 tends to zero

Thus, asymptotically the Bayesian test procedure has no "Type 1" error.

Understanding this

consider the marginal data density $p(Y|X)$ in Example~2. The terms that asymptotically dominate are

$$\begin{aligned}
\ln p(Y|X) &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}(Y'Y - Y'X(X'X)^{-1}X'Y) - \frac{k}{2}\ln T + sm \\
&= \ln p(Y|X, \hat{\theta}_{mle}) - \frac{k}{2}\ln T + small \\
&= \text{maximized likelihood function} - \text{penalty.} \qquad (
\end{aligned}$$

The marginal data density has the form of a penalized likelihood function.

The maximized likelihood function captures the goodness-of-fit of the regression model in which θ is freely estimated.

The second term penalizes the dimensionality to avoid overfitting the data.

Confidence Sets

The frequentist definition is that $C_{Y^T} \subseteq \Theta$ is an α confidence region if

$$P_\theta\{\theta \in C_{Y^T}\} \geq 1 - \alpha \quad \forall \theta \in \Theta \tag{41}$$

A Bayesian confidence set is defined as follows. $C_{Y^T} \subseteq \Theta$ is α credible if

$$P_{Y^T}\{\theta \in C_{Y^T}\} \geq 1 - \alpha \tag{42}$$

A highest posterior density region (HPD) is of the form

$$C_{Y^T} = \{\theta : p(\theta | Y^T) \geq k_\alpha\} \tag{43}$$

where k_α is the largest bound such that

$$P_{Y^T}\{\theta \in C_{Y^T}\} \geq 1 - \alpha$$

The HPD regions have the smallest size among all α credible regions of the parameter space Θ.

Example 2, Continued

The Bayesian highest posterior density region with coverage $1 - \alpha$ for θ_j is of the form

$$C_{Y^T} = \left[\tilde{\theta}_{T,j} - z_{crit}[\tilde{V}_T]_{jj}^{1/2} \leq \theta_j \leq \tilde{\theta}_{T,j} + z_{crit}[\tilde{V}_T]_{jj}^{1/2}\right]$$

where $[\tilde{V}_T]_{jj}$ is the j'th diagonal element of $\tilde{V}_T$, and z_{crit} is the $\alpha/2$ critical value of a $\mathcal{N}(0, 1)$.

In the Gaussian linear regression model the Bayesian interval is very similar to the classical confidence interval, but its statistical interpretation is quite different. $\square$