R404: Advanced Estimation Techniques

Topic: Bayesian Methods in Econometrics

Andrey Vassilev

2016/2017

Lecture Contents

1. Review: the philosophy of the Bayesian approach

2. Extending the Bayesian approach

3. The linear regression model in a Bayesian framework

Review: the philosophy of the Bayesian approach

A classical estimation example

- Consider a sample of iid observations $x_1, \ldots, x_n$ coming from a random variable $\xi \sim N(\mu, \sigma^2)$.
- We are interested in obtaining an estimate of the mean.
- A standard approach would be to use the method of maximum likelihood (MML).
- We construct the likelihood function (recall the convention to write it as if conditioning on the data):

$$
\begin{aligned}
L(\mu, \sigma^2 | x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}}.
\end{aligned}
\tag{1}
$$

A classical estimation example

- According to the MML, we maximize L w.r.t. μ.
- It is well-known (or, if your recollections are hazy, you can derive it) that the solution is given by the estimator $\hat{\mu} = \sum_{i=1}^{n} x_i / n$, i.e. the sample mean.
- By definition, the statistic $\hat{\mu}$ is a random variable.
- Consequently, if we keep repeating the experiment and regenerating the n observations, we'll obtain new samples $\tilde{x}_1, \dots, \tilde{x}_n, \tilde{\tilde{x}}_1, \dots, \tilde{\tilde{x}}_n$ etc. and therefore new values of $\hat{\mu}$.
- For each of those samples the corresponding value $\hat{\mu}$ will be our estimate of the unknown parameter μ.

A classical estimation example

- When we speak of the statistical properties of $\hat{\mu}$ like unbiasedness or consistency, we are implicitly referring to an ability to repeat the experiment many times or to extend the sample size n within an experiment.
- In this context, any probabilistic reasoning about $\hat{\mu}$ is based on a *classical* notion of probability as the theoretical limit of the ratio of occurrences of an event to the total number of trials (i.e. the relative frequency).
- It can be argued that this notion of probability is "objective" – it derives from an experiment and reflects mechanisms external to an observer.
- At the same time it is operational only when repeatability is ensured.
  - A football player is allowed to shoot one penalty but misses. Does it matter that he is the best scorer in his team?

Subjective probability

- The idea of probability is often used in contexts where the classical interpretation as the limit of the relative frequency is not applicable.
- Consider a person making the following statement:

  The probability that there is life on Mars is 1/1000000.
- What is this person trying to say?
  - If we could repeatedly try to find life on Mars, this would occur on average once in a million attempts?
  - If we could recreate the universe over and over, the planet Mars would materialize and there would be life on it once in a million runs?
- **The term "probability" in this context is used as a measure of the *subjective* degree of certainty in the correctness of a statement.**

Subjective probability

- "Probability" in the above sense can be interpreted as a way to define a fair bet.
- Take the statement

  The probability that team A will win the game against team B is 2/5.
- This can be construed to mean that the person making the statement considers it fair if he has to pay 2 dollars to enter a bet paying back 5 dollars if team A wins.
- The is related to the concept of *odds*. If the probability of an event is $\frac{m}{m+n}$, then the associated odds would be $m : n$ (m to n).
- Thus, I'm willing to bet m dollars to get a profit of n dollars (recover the initial m and get additional n dollars) and I think neither side of the bet is getting an unfair advantage.
- Clearly, subjective assessments of probability can differ in this context.

Axiomatic foundations of subjective probability

- The above considerations may seem informal but they turn out to be consistent with a formal definition of probability.
- Given a measurable space $(\Omega, \mathcal{F})$ and a relation $\preceq$ over the elements of the σ-algebra $\mathcal{F}$ having particular properties, it can be shown that the relation $\preceq$ induces a probability measure P on the space $(\Omega, \mathcal{F})$.
- The "particular properties" are basically a formal way of expressing the idea that one event is "more likely" or "more plausible" than another.
- The relation is called *relative likelihood*.
- The corresponding probability measure is known as *subjective probability*.

Two interpretations of probability

- To summarise, the notion of probability can be used in (at least) two contexts:
    - To measure how likely a particular outcome of an “experiment” is. This encompasses all sorts of situations whose outcomes can be considered objective: coin tosses, recording data with some degree of imprecision (measurement error) etc.
    - To measure the personal degree of certainty or conviction about something. This is by definition subjective and assessments about one and the same event can (will!) differ between different people.
- With some simplification, objective probability can be thought of as applicable to the modelling of data generation processes with a stochastic element, while subjective probability is applicable to situations where we need to quantify our personal certainty (or ignorance) about something.

Subjective probability in a statistical context

- As a consequence of the above, a probability distribution can be used to measure our degree of uncertainty about the precise value of a numerical variable.
- This is applicable to unknown parameters that are to be estimated from data.
- In our example of estimating the mean μ of the normally distributed random variable, taking a subjective probability point of view, it would be perfectly acceptable to treat μ as a random variable with certain properties.
- This would merely be a way to formalize our degree of knowledge about μ and is therefore applicable even in situations when we know that μ is in fact a (unknown) constant.

The Bayesian approach

- Since both the objective and the subjective notions of probability ultimately lead to the same mathematical object, we can also combine them, provided that this combination has a meaningful interpretation.

The Bayesian approach

- Since both the objective and the subjective notions of probability ultimately lead to the same mathematical object, we can also combine them, provided that this combination has a meaningful interpretation.
- We can treat the available data as generated by an appropriate stochastic mechanism.
- We can also treat model parameters as random if we agree to work with a subjectivist interpretation of probability.
- This allows us to combine the sample $\mathbf{x} = (x_1, \ldots, x_n)$ and the parameters in a joint probability distribution (or density, in appropriate contexts).
- In our example, the joint density would involve $\mathbf{x}$ and μ: $f(\mathbf{x}, \mu)$. (It would also depend on σ^2 but we omit that for simplicity.)

The Bayesian approach

- This approach allows us to formalize parameter uncertainty but cannot magically help with getting more data or re-running the data generation process.
- Since in many practical situations the sample $\mathbf{x}$ is fixed and getting more data is difficult or impossible, the object of interest is actually the conditional density of the parameter(s) given available data, i.e. $f(\mu|\mathbf{x})$.
- The density $f(\mu|\mathbf{x})$ can be obtained by means of Bayes' formula for probability density functions, hence the term Bayesian approach:

$$f(\mu|\mathbf{x}) = \frac{f(\mathbf{x}, \mu)}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\mu) f(\mu)}{f(\mathbf{x})}. \tag{2}$$

The Bayesian approach

- The density $f(\mathbf{x}|\mu)$ is the *likelihood* function for the model. It reflects the data generation mechanism and thus captures the "objective" component of the density $f(\mu|\mathbf{x})$.
- The density $f(\mu)$ is called the *prior* density. It is formed on the basis of information outside the model (prior to observing the data) and can incorporate a variety of factors like:
  - theoretical considerations
  - results from previous studies and experiments
  - subjective judgement on what is reasonable
  - the modeller's uncertainty about the parameters
- While the prior corresponds to the "subjective" component of the density $f(\mu|\mathbf{x})$, the above factors show that it need not be *ad hoc* or arbitrary but can rigorously incorporate additionally available information.

The Bayesian approach

- The density $f(\mu|\mathbf{x})$ is called the *posterior* density. It combines information coming from the data via the likelihood with additional information coming from the prior.

- For a given sample the last component of formula (2) – the density $f(\mathbf{x})$ – is a constant and does not affect the computations substantively but acts only as a scaling factor. For that reason, formula (2) is often written as

$$f(\mu|\mathbf{x}) \propto f(\mathbf{x}|\mu) f(\mu), \tag{3}$$

  where $\propto$ denotes proportionality.

- Sometimes the expression on the right-hand side of $\propto$ in (3) is called the *kernel* of the posterior density.

- The above construction obviously does not depend on the dimension of μ and will remain valid in a multidimensional setting.

Interpretations of Bayes’ formula

- We can interpret (3) as a mechanism to update an initial body of information in light of new empirical evidence (data).
- Alternatively, (3) can be interpreted as a mechanism to inform empirical analysis by introducing already available information.
- In any case this presents the issue of how information is encoded in a prior distribution.
- At a minimum, this is done by choosing a distribution with appropriate support and calibrating its parameters to ensure, for example, a specific mean and variance.

Some advantages and limitations of the Bayesian approach

- Bayesian methods can alleviate small-sample problems where the data do not contain enough information to use a classical approach.
- Bayesian methods can be used to impose theoretically motivated constraints on model parameters.
- There is the risk that a strong prior will predetermine the outcome of an analysis. This means we can essentially force any result we want.
- Bayesian methods can be computationally demanding.

How does it work?

Revisiting the μ estimation example in a Bayesian framework

Let's go back to the example of estimating μ from a sample of iid observations from a $N(\mu, \sigma^2)$ random variable with σ^2 known.

The likelihood (1) contains the expression

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}\left((x_i-\hat{\mu})-(\mu-\hat{\mu})\right)^2 \\
&= \sum_{i=1}^{n}(x_i-\hat{\mu})^2 - \sum_{i=1}^{n}2(x_i-\hat{\mu})(\mu-\hat{\mu}) + n(\mu-\hat{\mu})^2 \\
&= \sum_{i=1}^{n}(x_i-\hat{\mu})^2 - 2(\mu-\hat{\mu})\underbrace{\sum_{i=1}^{n}(x_i-\hat{\mu})}_{0} + n(\mu-\hat{\mu})^2 \\
&= \sum_{i=1}^{n}(x_i-\hat{\mu})^2 + n(\mu-\hat{\mu})^2 = \nu s^2 + n(\mu-\hat{\mu})^2,
\end{aligned}
\tag{4}
$$

where $\nu = n-1$ and $s^2 = \nu^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$.

How does it work?

Revisiting the μ estimation example in a Bayesian framework

Therefore, the likelihood (1) can be written as

$$L(\mu, \sigma^2 | x_1, \dots, x_n) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \left(\nu s^2 + n(\mu - \hat{\mu})^2 \right) \right). \quad (5)$$

Suppose our prior information on μ can be summarised by the density

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_a^2} \exp\left(-\frac{1}{2\sigma_a^2} (\mu - \mu_a)^2 \right), \quad (6)$$

where μ_a is the prior mean and σ_a^2 is the prior variance of the person conducting the analysis.

How does it work?

Revisiting the μ estimation example in a Bayesian framework

Combining (5) and (6) by means of (3), we obtain

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto \exp\left(-\frac{1}{2}\left[\frac{(\mu-\mu_a)^2}{\sigma_a^2}+\frac{n}{\sigma^2}(\mu-\hat{\mu})^2\right]\right) \\ &\propto \exp\left(-\left(\frac{\sigma_a^2+\sigma^2/n}{2\sigma_a^2\sigma^2/n}\right)\left(\mu-\frac{\hat{\mu}\sigma_a^2+\mu_a\frac{\sigma^2}{n}}{\sigma_a^2+\sigma^2/n}\right)^2\right). \end{aligned} \tag{7}$$

Thus, the posterior distribution of μ is normal with mean

$$\mathbb{E}[\mu]=\frac{\hat{\mu}\sigma_a^2+\mu_a\sigma^2/n}{\sigma_a^2+\sigma^2/n}=\frac{\hat{\mu}(\sigma^2/n)^{-1}+\mu_a(\sigma_a^2)^{-1}}{(\sigma^2/n)^{-1}+(\sigma_a^2)^{-1}}$$

and variance

$$\mathbb{D}[\mu]=\frac{\sigma_a^2\sigma^2/n}{\sigma_a^2+\sigma^2/n}=\frac{1}{(\sigma^2/n)^{-1}+(\sigma_a^2)^{-1}}.$$

How does it work?

Revisiting the μ estimation example in a Bayesian framework

The preceding result can be clarified if we introduce the notation $h_0 = (\sigma^2/n)^{-1}$ and $h_a = (\sigma_a^2)^{-1}$. These are called *precision parameters*.

Then, we can equivalently write

$$\mathbb{E}[\mu] = \frac{h_0\hat{\mu} + h_a\mu_a}{h_0 + h_a},$$

$$\mathbb{D}[\mu] = \frac{1}{h_0 + h_a}.$$

In other words, the posterior mean turns out to be a weighted average of the sample mean and the prior mean.

Extending the Bayesian approach

Bayesian updating with new data

- We can use Bayes' theorem to update our information sequentially as new data arrive.
- Let the initial sample be $\mathbf{x_1}$ and the prior density be $f(\mu)$, leading to a posterior $f(\mu|\mathbf{x_1}) \propto f(\mu)f(\mathbf{x_1}|\mu)$.
- Suppose we obtain an additional sample $\mathbf{x_2}$.
- Then the posterior $f(\mu|\mathbf{x_1})$ can be treated as a prior with respect to the new sample and, using Bayes' theorem, we get

$$f(\mu|\mathbf{x_1}, \mathbf{x_2}) \propto f(\mu|\mathbf{x_1})f(\mathbf{x_2}|\mu), \tag{8}$$

where $f(\mu|\mathbf{x_1}, \mathbf{x_2})$ is the posterior obtained with the two samples merged.

Bayesian updating with new data

- Formula (8) can also be written as

$$f(\mu|\mathbf{x_1},\mathbf{x_2}) \propto f(\mu)f(\mathbf{x_1}|\mu)f(\mathbf{x_2}|\mu). \tag{9}$$

- Since the likelihood function for the merged samples $\mathbf{x_1}$ and $\mathbf{x_2}$ is $f(\mathbf{x_1}|\mu)f(\mathbf{x_2}|\mu)$, the new posterior will be the same regardless of whether we obtain the samples sequentially or we have the full sample $(\mathbf{x_1},\mathbf{x_2})$ from the start.

- Clearly, this approach generalizes to the case of more than one sample.

Marginal and conditional posterior densities

- As noted, the Bayesian approach works the same way when we are interested in a vector of parameters $\boldsymbol{\theta}$. In this context we denote the joint posterior density by $f(\boldsymbol{\theta}|\mathbf{x})$.

- In some cases we are interested only in a subset of the parameters $\boldsymbol{\theta}$, i.e. given $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})'$, we would like to separate out the posterior information for $\boldsymbol{\theta_1}$ only.

- In other words, we are interested in the marginal posterior density of the vector $\boldsymbol{\theta_1}$.

- It can be obtained as

$$f(\boldsymbol{\theta_1}|\mathbf{x}) = \int_{R_{\boldsymbol{\theta_2}}} f(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}|\mathbf{x})\, d\boldsymbol{\theta_2} = \int_{R_{\boldsymbol{\theta_2}}} f(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \mathbf{x}) f(\boldsymbol{\theta_2}|\mathbf{x})\, d\boldsymbol{\theta_2}, \qquad (10)$$

  where $R_{\boldsymbol{\theta_2}}$ is the domain of $\boldsymbol{\theta_2}$ and $f(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \mathbf{x})$ is the conditional posterior density of $\boldsymbol{\theta_1}$ for given $\boldsymbol{\theta_2}$ and $\mathbf{x}$.

Marginal and conditional posterior densities

- The expression following the second equality sign in (10),

$$\int_{R_{\boldsymbol{\theta_2}}} f(\boldsymbol{\theta_1}|\boldsymbol{\theta_2},\mathbf{x})f(\boldsymbol{\theta_2}|\mathbf{x})\,d\boldsymbol{\theta_2},$$

  shows that the marginal posterior density $f(\boldsymbol{\theta_1}|\mathbf{x})$ can be interpreted as the result of averaging the conditional posterior density $f(\boldsymbol{\theta_1}|\boldsymbol{\theta_2},\mathbf{x})$ by using the marginal posterior density $f(\boldsymbol{\theta_2}|\mathbf{x})$ as the weight function.

- The integration operation in (10) serves to eliminate the information on the parameters we are **not** interested in, leaving only the posterior information on the relevant parameters.

Point estimates in a Bayesian framework

- The posterior summarises all the available information (sample and non-sample) about the parameters of interest.

- The downside is that the posterior is ultimately a probability distribution, while we may need simpler characterizations of the parameters.

- One example is a situation where we want to produce specific numerical values for the unknown parameters, i.e. we want *point estimates*.

- Once we have the posterior for the parameter vector $\boldsymbol{\theta}$, the Bayesian way using it to produce a point estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ is in a decision-theoretic framework.

Point estimates in a Bayesian framework

- Decision theory requires us to have *loss function* $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ that measures how "harmful" deviations of the estimates $\hat{\boldsymbol{\theta}}$ from the true value $\boldsymbol{\theta}$.
- Since in our case $\boldsymbol{\theta}$ is a random variable, the loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ will also be random, even though the sample $\mathbf{x}$ is fixed.
- Thus, we need to work with the expected loss function.
- Since the posterior summarizes the information on $\boldsymbol{\theta}$, it is natural to construct the expectation with respect to $f(\boldsymbol{\theta}|\mathbf{x})$.
- We are then in a position to compute the expected loss associated with an estimate $\hat{\boldsymbol{\theta}}$.

Point estimates in a Bayesian framework

- Then, the optimal point estimate would be the one that minimizes the expectation of the loss function:

$$\hat{\boldsymbol{\theta}}^{*} = \min_{\hat{\boldsymbol{\theta}}} \mathbb{E}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] = \min_{\hat{\boldsymbol{\theta}}} \int_{R_{\boldsymbol{\theta}}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{11}$$

- The expected loss function is called *risk* or *risk function*.

- We are implicitly assuming that both the expectation and the minimum exist.

Point estimates in a Bayesian framework

- As an example, take the quadratic loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'C(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, where C is a fixed symmetric positive definite matrix.

- Then the posterior expectation of $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is

$$\begin{aligned} \mathbb{E}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] =& \mathbb{E}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'C(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})] \\ =& \mathbb{E}[((\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}]) - (\hat{\boldsymbol{\theta}} - \mathbb{E}[\boldsymbol{\theta}]))'C((\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}]) - (\hat{\boldsymbol{\theta}} - \mathbb{E}[\boldsymbol{\theta}]))] \\ =& \mathbb{E}[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}])'C(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}])] + (\hat{\boldsymbol{\theta}} - \mathbb{E}[\boldsymbol{\theta}])'C(\hat{\boldsymbol{\theta}} - \mathbb{E}[\boldsymbol{\theta}]), \end{aligned} \tag{12}$$

  where the second term in the last equality is not stochastic and can be taken out of the expectation. (Note that in going from the second to the third line terms of the form $\mathbb{E}[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}])'C(\hat{\boldsymbol{\theta}} - \mathbb{E}[\boldsymbol{\theta}])]$ will disappear since $\mathbb{E}[\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta}]] = \mathbf{0}$.)

- At the same time, it is precisely the last term of (12) that depends on $\hat{\boldsymbol{\theta}}$ and determines the minimum.

Point estimates in a Bayesian framework

- Clearly for a positive definite C the minimum is attained at $\hat{\boldsymbol{\theta}}^* = \mathbb{E}[\boldsymbol{\theta}]$.
- Thus, for the quadratic loss function the optimal point estimate is given by the mean[1] of the respective posterior distribution.

[1] Assuming it exists in the first place.

Point estimates in a Bayesian framework

- Clearly for a positive definite C the minimum is attained at $\hat{\boldsymbol{\theta}^*} = \mathbb{E}[\boldsymbol{\theta}]$.
- Thus, for the quadratic loss function the optimal point estimate is given by the mean[1] of the respective posterior distribution.
- As further examples, for a univariate parameter θ and an absolute deviation loss function $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ (plus technical assumptions), the optimal point estimate can be shown to be the median.
- A zero-one loss function of the form

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \begin{cases} 0, \ \hat{\boldsymbol{\theta}} = \boldsymbol{\theta} \\ 1, \ \hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta} \end{cases},$$

will yield a mode of the posterior as the optimal point estimate (again under technical assumptions).

[1] Assuming it exists in the first place.

Bayesian credible regions

- The Bayesian counterpart of a confidence interval is called a *credible region*.
- If we have the posterior density $f(\boldsymbol{\theta}|\mathbf{x})$, we can compute the probability that the vector of parameters $\boldsymbol{\theta}$ belongs to a given subset $\bar{R}$ of the space of parameters:

$$P(\boldsymbol{\theta} \in \bar{R}|\mathbf{x}) = \int_{\bar{R}} f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{13}$$

Bayesian credible regions

- The Bayesian counterpart of a confidence interval is called a *credible region*.
- If we have the posterior density $f(\boldsymbol{\theta}|\mathbf{x})$, we can compute the probability that the vector of parameters $\boldsymbol{\theta}$ belongs to a given subset $\bar{R}$ of the space of parameters:

$$P(\boldsymbol{\theta} \in \bar{R}|\mathbf{x}) = \int_{\bar{R}} f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{13}$$

- We can also approach the problem from the opposite direction: for a fixed probability $P(\boldsymbol{\theta} \in \bar{R}|\mathbf{x})$, find a region $\bar{R}$ such that (13) holds.
- In general such a region will not be unique.

Bayesian credible regions

- If the posterior is unimodal, in some cases a unique credible region can be obtained by imposing an additional requirement known as *highest posterior density*.

- This requirement is intuitively described along the lines “the posterior density values over that region should not be smaller than those for any other region with the same probability”...

- ...but a more precise characterization would be related to the fact that they minimize the volume enclosed in the parameter space among all regions with the same probability.

- For example, a unimodal symmetric density in the case of one parameter would produce a credible region that is an interval centred on the mode and containing the largest values of the posterior density (or, equivalently, being the smallest in length).

Bayesian credible regions

- It should be remembered that, despite the superficial similarity, Bayesian credible regions are conceptually different from classical confidence intervals.
- A classical confidence interval is considered random and probabilistic statements relate to the probability that this random interval will cover the fixed (but unknown) value of the parameter of interest.
- In contrast, a Bayesian credible region is considered deterministic and probabilistic statements relate to the probability with which the random parameter of interest will fall in that pre-specified region.

Marginal distribution of the observations

- Sometimes we are interested in the marginal density of the observations $f(\mathbf{x})$.
- These can be derived as follows:

$$f(\mathbf{x}) = \int_{R_{\boldsymbol{\theta}}} f(\boldsymbol{\theta}, \mathbf{x})\, d\boldsymbol{\theta} = \int_{R_{\boldsymbol{\theta}}} f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{14}$$

- The expression after the second equality sign in (14) means that the marginal density can be viewed as averaging the likelihood function by using the prior density as a weighting function.

Predictive probability density functions

- One important issue arising in a forecasting context is that of the distribution of still unobserved events $\tilde{\mathbf{x}}$.
- To answer that, we consider the joint density of the unobserved data $\tilde{\mathbf{x}}$ and the parameters $\boldsymbol{\theta}$, given already observed data $\mathbf{x}$:

$$f(\tilde{\mathbf{x}}, \boldsymbol{\theta}|\mathbf{x}) = f(\tilde{\mathbf{x}}|\boldsymbol{\theta}, \mathbf{x}) f(\boldsymbol{\theta}|\mathbf{x}). \tag{15}$$

- The predictive density of $\tilde{\mathbf{x}}$ conditional on $\mathbf{x}$ can be obtained after integrating (15) w.r.t. $\boldsymbol{\theta}$:

$$f(\tilde{\mathbf{x}}|\mathbf{x}) = \int_{R_{\boldsymbol{\theta}}} f(\tilde{\mathbf{x}}, \boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = \int_{R_{\boldsymbol{\theta}}} f(\tilde{\mathbf{x}}|\boldsymbol{\theta}, \mathbf{x}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{16}$$

- The last expression in (16) means that the predictive density can be interpreted as averaging the conditional predictive density with the posterior used as a weighting function.

Point prediction

- The predictive density $f(\tilde{\mathbf{x}}|\mathbf{x})$ can be used to produce point forecasts just like the posterior is used to produce point estimates.
- The principle is the same as for point estimates of parameters.
- To produce a point prediction $\hat{\mathbf{x}}$ for the yet unobserved data $\tilde{\mathbf{x}}$, we need a loss function $L(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$.
- We then construct the risk function and minimize it w.r.t. $\hat{\mathbf{x}}$:

  $$\min_{\hat{\mathbf{x}}} \int_{R_{\tilde{\mathbf{x}}}} L(\tilde{\mathbf{x}}, \hat{\mathbf{x}}) f(\tilde{\mathbf{x}}|\mathbf{x}) d\tilde{\mathbf{x}}, \tag{17}$$

  where $R_{\tilde{\mathbf{x}}}$ is the set of possible values of $\tilde{\mathbf{x}}$.
- The solution to (17), if it exists, is the optimal point prediction for $\tilde{\mathbf{x}}$.
- We can establish counterparts to point estimate results on the mean, median etc. corresponding to specific loss functions.

Prediction regions

- Given a predictive density $f(\tilde{\mathbf{x}}|\mathbf{x})$, we can compute the probability that future observations $\tilde{\mathbf{x}}$ belong to a given subset $\bar{R}$ of their domain:

$$P(\tilde{\mathbf{x}} \in \bar{R}|\mathbf{x}) = \int_{\bar{R}} f(\tilde{\mathbf{x}}|\mathbf{x})\, d\tilde{\mathbf{x}}. \tag{18}$$

- Similarly to credible regions, we can also study the inverse problem, where we seek a region $\bar{R}$ satisfying (18) – a *prediction region* – for a given probability $P(\tilde{\mathbf{x}} \in \bar{R}|\mathbf{x})$.

- Uniqueness generally cannot be guaranteed in this case but we can develop the concept of a “highest predictive density” region along the lines of highest posterior density regions.

Bayesian hypothesis testing

- In a Bayesian context, hypothesis testing is a comparison of equally standing alternatives. This stands in contrast to the classical treatment, where the null hypothesis is the object of special interest.
- In general, we study statements H_0 and H_1 relating to different models M_0 and M_1, given data $\mathbf{x}$.
- For a model $M_i,\ i = 0, 1,$ the respective likelihood function is $f_i(\mathbf{x}|\boldsymbol{\theta_i})$, while the prior is $f_i(\boldsymbol{\theta_i})$.
- The subscripts i in the likelihoods and the priors indicate that it is possible to compare hypotheses for different data generating mechanisms and different priors.
- It is still possible to analyse simpler cases, e.g. when the parameters $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, corresponding to hypotheses H_0 and H_1, take values in different domains of the same parametric space.

Bayesian hypothesis testing

- If the true model is model M_i, the marginal density of the data (*marginal likelihood function*) is

$$f(\mathbf{x}|M_i) = \int_{R_{\boldsymbol{\theta_i}}} f_i(\mathbf{x}|\boldsymbol{\theta_i}) f_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}. \tag{19}$$

- Then, a *Bayes factor* is defined as the ratio of the marginal likelihood functions for the two models:

$$B_{01}(\mathbf{x}) := \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}. \tag{20}$$

- A Bayes factor greater than 1 can be interpreted as evidence that M_0 (or, respectively, hypothesis H_0) is more plausible compared to model M_1 (hypothesis H_1).

Bayesian hypothesis testing

- It is also possible to have prior probabilities on the models themselves and take these probabilities into account when testing hypotheses.
- If $P(M_i)$ is the prior probability of model M_i, then we can calculate the posterior probability of this model given the data $\mathbf{x}$ using Bayes' formula:

$$P(M_i|\mathbf{x}) = \frac{P(M_i)f(\mathbf{x}|M_i)}{\sum_{j=0}^{1} P(M_j)f(\mathbf{x}|M_j)}. \tag{21}$$

- The *posterior odds* of model M_0 compared to model M_1 is defined as the ratio of the posterior probabilities of the two models:

$$K_{01}(\mathbf{x}) := \frac{P(M_0|\mathbf{x})}{P(M_1|\mathbf{x})} = \frac{P(M_0)}{P(M_1)} \times \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)} = \frac{P(M_0)}{P(M_1)} B_{01}(\mathbf{x}). \tag{22}$$

Bayesian hypothesis testing

- Equation (22) illustrates why the quantity $B_{01}(\mathbf{x})$ is called a Bayes factor: it is the quantity that multiplies the *prior odds* $P(M_0)/P(M_1)$ to adjust it with information coming from the data and thus obtain the posterior odds. In other words, the Bayes factor tells us whether the data increase or decrease the plausibility of one model compared to another.
- The above considerations were framed in terms of model comparison but they can equally be applied to hypothesis testing.
- Thus, if we have the posterior odds $K_{01}(\mathbf{x})$ needed to compare hypothesis H_0 to H_1, we can use it to decide which hypothesis is more plausible.
- In the absence of prior information in favour of one hypothesis, the hypotheses can be viewed as equally likely and the prior odds can be taken to be 1. Then the posterior odds is equal to the Bayes factor and $B_{01}(\mathbf{x})$ suffices to choose between the hypotheses tested.

Bayesian hypothesis testing

- If we focus on accepting or rejecting hypothesis H_0 in favour of H_1, then using only the posterior odds – for instance, by applying rules such as "Accept H_0 if $K_{01}(\mathbf{x}) \geq 1$, else reject H_0." – is a relatively blunt decision-making tool.
- In such cases, it is recommended to work in a decision-theoretic framework by introducing an explicit loss function $L(H_i, \hat{H}_j)$, where H_i is the true situation and $\hat{H}_j$ is the hypothesis accepted by the researcher.
- We can then compute the risk associated with accepting $\hat{H}_j$ and choose the risk-minimizing hypothesis $\hat{H}_j$.

Additional information on Bayesian methods

Informative and non-informative priors

- Up to this point we implicitly assumed that the priors were chosen to reflect additional available information. That is, they were taken to be *informative*.
- We can, however, imagine situations in which we lack non-experimental information, requiring a *non-informative* prior.
- It is intuitive to treat the parameter values as equally probable in such cases.
- This approach seems natural but gives rise to technical difficulties.
- For example, attempting to use the uniform distribution as a prior for a continuous parameter either implicitly introduces information (if we define it over a finite interval, thus limiting the support of the distribution), or else leads to a divergent integral.

Additional information on Bayesian methods

Informative and non-informative priors

- If a non-informative “density” (under the above definition) leads to a divergent integral, it obviously cannot be a proper probability density function.
- Nevertheless, we can agree to consider such priors if the resulting posteriors are proper, i.e. if they integrate to a finite quantity.
- Such situations can be construed as taking only sample information into account.
- In many cases this leads to results that coincide with the classical results for large enough samples.

Additional information on Bayesian methods

Simulation techniques: Markov chain Monte Carlo

- In a general formulation like ours, the posterior densities can take a huge variety of forms.
- In most cases closed-form solutions cannot be obtained and we need to resort to numerical methods.
- If the parameter space is high-dimensional, generating a random sample from the posterior or integrating it leads to serious computational difficulties.
- In the past, this problem was circumvented by using the so-called *conjugate priors*, i.e. priors which lead to the posterior belonging to the same class of distributions. In addition, conjugate priors that were easy to study were favoured.

Additional information on Bayesian methods

Simulation techniques: Markov chain Monte Carlo

- With the growth of computing power over the past few decades, Bayesian methods have progressively gravitated towards a powerful class of numerical simulation methods known as Markov chain Monte Carlo (MCMC).
- The main idea of MCMC is to draw a large enough sample from a Markov chain that asymptotically has the same distribution as the posterior.
- The properties of that sample are then studied to learn about the posterior.
- Most contemporary empirical studies in a Bayesian framework employ some form of MCMC or similar methods.

The linear regression model in a Bayesian framework

Formulation

We study the familiar regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}, \tag{23}$$

where there are T observations and K parameters. We take the matrix X to be non-stochastic (the same approach works for a stochastic X if it is independent of the error $\mathbf{e}$), and assume that $\operatorname{rank}(X) = K$ and $\mathbf{e} \in N(\mathbf{0}, \sigma^2 I_T)$.

The likelihood function for (23) under our normality assumption is

$$\ell(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})}{2\sigma^2}\right]. \tag{24}$$

Formulation

In accordance with Bayesian principles we take $\boldsymbol{\beta}$ and σ^2 as random and having a prior density $f(\boldsymbol{\beta}, \sigma)$ with the kernel given below:

$$f(\boldsymbol{\beta}, \sigma) \propto \sigma^{-m} \exp\left[-\frac{1}{2\sigma^2}[\eta + (\boldsymbol{\beta} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})]\right], \tag{25}$$

where $m > K + 1, \eta > 0$ and $\boldsymbol{\Psi}$ is symmetric and positive definite.

This is an example of a conjugate prior.

Interpretation of the prior

The choice of (25) as our prior can be understood if we consider the decomposition of $f(\boldsymbol{\beta}, \sigma)$ as

$$f(\boldsymbol{\beta}, \sigma) = f(\boldsymbol{\beta}|\sigma) f(\sigma), \tag{26}$$

where

$$f(\boldsymbol{\beta}|\sigma) = \frac{1}{(2\pi)^{K/2}\sigma^K[\det(\boldsymbol{\Psi})]^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})\right] \tag{27}$$

and

$$f(\sigma) \propto \sigma^{-(m-K)} \exp\left[-\frac{\eta}{2\sigma^2}\right], \quad \sigma > 0. \tag{28}$$

In other words, conditional on σ, the vector $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Psi}$. The distribution of σ that corresponds to the kernel (28) can be obtained as a nonlinear transformation of a gamma-distributed random variable.

The posterior

Given the prior (25) and the likelihood (24) the posterior of the model can be shown to be

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma|\mathbf{y}) \propto & \frac{1}{\sigma^{T+m}} \exp\left[-\frac{1}{2\sigma^2}\left[(T-K)\hat{\sigma}^2 + (\boldsymbol{\beta}-\mathbf{b})'X'X(\boldsymbol{\beta}-\mathbf{b})\right]\right] \\ & \times \exp\left[-\frac{1}{2\sigma^2}\left[\eta + (\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right]\right], \end{aligned} \tag{29}$$

where $\hat{\sigma}^2 = RSS/(T-K)$ and $\mathbf{b}$ are the OLS estimates for the variance and the coefficient vector.
The kernel (29) can be written equivalently as

$$f(\boldsymbol{\beta}, \sigma|\mathbf{y}) \propto \frac{1}{\sigma^{T+m}} \exp\left[-\frac{1}{2\sigma^2}[(\boldsymbol{\beta}-\boldsymbol{\beta}_*)'(\boldsymbol{\Psi}^{-1}+X'X)(\boldsymbol{\beta}-\boldsymbol{\beta}_*) + \xi]\right], \tag{30}$$

where

$$\begin{aligned} \boldsymbol{\beta}_* &= (\boldsymbol{\Psi}^{-1}+X'X)^{-1}(\boldsymbol{\Psi}^{-1}\boldsymbol{\mu} + X'X\mathbf{b}), \\ \xi &= \eta + (T-K)\hat{\sigma}^2 + \boldsymbol{\mu}'\boldsymbol{\Psi}^{-1}\boldsymbol{\mu} + \mathbf{b}'X'X\mathbf{b} - \boldsymbol{\beta}_*'(\boldsymbol{\Psi}^{-1}+X'X)\boldsymbol{\beta}_*. \end{aligned}$$

The posterior

The marginal posterior density for β can be obtained in a standard manner by integrating (30) w.r.t. σ. It has the kernel

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}) &\propto \left[\xi + (\boldsymbol{\beta} - \boldsymbol{\beta}_*)'(\boldsymbol{\Psi}^{-1} + X'X)(\boldsymbol{\beta} - \boldsymbol{\beta}_*)\right]^{-(T+m-1)/2} \\ &\propto \left[v + (\boldsymbol{\beta} - \boldsymbol{\beta}_*)'\frac{v}{\xi}(\boldsymbol{\Psi}^{-1} + X'X)(\boldsymbol{\beta} - \boldsymbol{\beta}_*)\right]^{-(v+K)/2}, \end{aligned} \tag{31}$$

where $v = T + m - K - 1$.

This is the kernel of a multivariate t-distribution.

The posterior

A random vector distributed with kernel (31) will have mean $\boldsymbol{\beta}_*$ for $T+m-K>2$ and covariance matrix

$$\left(\frac{\xi}{T+m-K-3}\right)(\boldsymbol{\Psi}^{-1}+X'X)^{-1} \;\; T+m-K>3.$$

If we denote $\mathbf{h}=\left[\frac{v}{\xi}(\boldsymbol{\Psi}^{-1}+X'X)\right]$, then the distribution of the j-th element of $\boldsymbol{\beta}$, β_j, can be obtained by taking into account that $(\beta_j-\beta_{*j})\,/\,[h^{-1}(j,j)]^{1/2}$ has univariate t-distribution with v degrees of freedom. The notation in the denominator stands for the square root of the j-th diagonal element of the matrix $\mathbf{h}^{-1}$.

The posterior

It can be shown that the multivariate t-distribution forms a class that is closed with respect to linear transformations (up to a change in the parameters of the distribution).
We can also use the fact that

$$\frac{T+m-K-1}{K\xi}(\mathbf{B}-\beta_*)'(\Psi^{-1}+X'X)(\mathbf{B}-\beta_*) \;\sim\; F(K,v),$$

where $\mathbf{B}$ denotes a random vector with density kernel (31).
Similarly, if C is a $J \times K$ matrix, then

$$\frac{T+m-K-1}{J\xi}[C(\mathbf{B}-\beta_*)]'[C(\Psi^{-1}+X'X)^{-1}C']^{-1}[C(\mathbf{B}-\beta_*)] \;\sim\; F(J,v).$$

This can be used to study the properties of β, including to construct Bayesian credible regions and test hypotheses.

Readings

Readings:

William H. Greene. *Econometric Analysis*, 7th edition. Chapter 16.

Additional readings:

Gary Koop. *Bayesian Econometrics*. 2003. Wiley.

Allen Downey. *Think Bayes: Bayesian Statistics in Python*.