The Bayesian approach to poverty measurement

Lecture 1: Introduction

Michel Lubrano

November 2022

Abstract

This lecture reviews the recent Bayesian literature on poverty measurement. After introducing Bayesian statistics, we show how Bayesian model criticism could help to revise the international poverty line. Using mixtures of lognormals to model income, we derive the posterior distribution for the FGT, Watts and Sen poverty indices, then for TIP curves (with an illustration on child poverty in Germany) and finally for Growth Incidence Curves. The relation of restricted stochastic dominance with TIP and GIC dominance is detailed with an example on UK data. Using panel data, we show how to decompose poverty into total, chronic and transient poverty, comparing child and adult poverty in East Germany when redistribution is introduced. When a panel is not available, a Gibbs sampler is used to build a pseudo panel. We illustrate poverty dynamics by examining the consequences of the Wall on poverty entry and poverty persistence in occupied West Bank.

Contents

1 Introduction

12 hours, divided in 6 sessions of two hours each, in the morning from 10:00 to 12:00, Room 15 at the IBD.

1. Wednesday November 9th
2. Wednesday November 16th
3. Wednesday November 23rd
4. Wednesday November 30th
5. Wednesday December 7th
6. Monday December 12th, 14h30-16h30, room 21, first floor at IBD

I will try to put my slides and the quoted papers on my web page:

https://perso.amse-aixmarseille.fr/lubrano/

My room is 1.34, first floor at IBD.

This lecture is based on a paper entitled *The Bayesian approach to poverty measurement* written together with Zhou XUN from Nanjing University, China. The paper is to be published in a Research Handbook entitled *Measuring Poverty and Deprivation* edited by Jacques Silber and published at Edward Elgar Publishing in 2023.

2 Detailed outline

1. Lecture 1: General introduction and a first look at Bayesian statistics
2. Lecture 2: Revising the IPL using Bayesian inference
3. Lecture 3: Modelling the income distribution using mixtures
4. Lecture 4: Poverty indices and poverty curves
5. Lecture 5: Restricted stochastic dominance
6. Lecture 6: Poverty dynamics.

3 Introduction to the topic

For long, standard errors were not reported for poverty or inequality indices, and this on two grounds. Data sets based on surveys included more than five thousands observations, so it was thought that the standard errors would have been very small. A second objection was the difficulty of computation (see for instance Davidson 2009 for the Gini index or Biewen and Jenkins 2006 for generalised entropy indices and complex sampling). These arguments are no longer tenable. We might well be interested in sub-groups, operating thus on reduced sample sizes. The Bayesian approach brings in feasible answers for small sample sizes and its simulation techniques make simple the computation of standard errors.

More precisely, a Bayesian approach to poverty measurement relies most of the time on a parametric modelling of the income distribution. Poverty indices, the TIP curve of Jenkins and Lambert (1997), the growth incidence curve of Ravallion and Chen (2003) are transformations of the parameters of this parametric income distribution. The purpose of Bayesian inference will be to provide draws of the posterior density of these quantities, using simulation methods. The same approach will be used to explore restricted stochastic dominance and poverty dynamics. The interested reader can find an introduction to Bayesian inference in Lindley (1971), and to the required simulation methods in Bauwens et al. (1999).

Another word of introduction. Jacques Drèze in his presidential address to the Econometric Society (Drèze 1972) made an interesting introduction to the interest that an economist should have in Bayesian econometrics. When Econometrics is viewed as *a scientific approach to quantitative empirical economics*, it leads to study the decisions that economic agents are taking under uncertainty. Following Savage, a decision problem involves three basic concepts:

1. the states of nature,
2. the acts of the decision maker,
3. the consequences.

The decision which is taken is the one that has the maximum expected utility according to the moral expectation theorem. The later requires among other things to define a probability measure on the states of nature. Following Drèze (1972), *Recent advances in statistical and econometric methodology*

enable us to derive from empirical observations a probability measure on the relevant events for many decision problems. This means having the necessary tools to revise the prior knowledge we have about the ordering of the states of nature by means of Bayes theorem.

4 Bayesian statistics

Bayesian inference is organized around **Bayes theorem**, a theorem which can be taught at different levels. Basically this theorem allows you to get information about causes of a phenomenon:

Bayes theorem = probability of causes

4.1 Bayes theorem

Let us consider two events A and B and let us suppose that we know the following probabilities:

$$\Pr(A), \qquad \Pr(B), \qquad \Pr(B|A).$$

The first two probabilities are marginal probabilities and the last is the conditional probability of B knowing the realization of A. What we are looking for is the conditional probability of A, knowing that B was realised. In other terms, does the realization of B teach us something about the probability that A was realised. A first formulation of Bayes theorem is as follows:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

A proof of this theorem can be found by noting that we can write $\Pr(A \cap B)$ in two different ways:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A).$$

The final formula of the theorem is obtained by dividing each term by $\Pr(B)$.

This theorem does not rely on the interpretation of probability in term of frequencies. If we flip a coin, we assume that the coin is well balanced and consequently we have the prior opinion that heads and tails have the same likeliness to appear. We do not need to flip the coin an infinite number of times and count the number of heads and tails. So in the above writing $\Pr(A)$ is the **prior probability** of A. It means that this probability was elicited or

built before any information on B. $\Pr(A|B)$ is the conditional probability of A, knowing B. So this probability is computed after observing B, so it is a **posterior probability**, which directly depends on the realisation of B. In order to compute it, we use $\Pr(B|A)$ which plays the role of the **likelihood function** of B. Finally, $\Pr(B)$ is called the **marginal probability** of B.

The writing of Bayes theorem can be given another form which is perhaps better known to you by noting an alternative decomposition of $\Pr(B)$:

$$\Pr(B) = \Pr(B \cap A) + \Pr(B \cap \bar{A}) = \Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})$$

where $\bar{A}$ is the complement of A, so that we can write:

$$P(A|B) = \frac{P(B|A)P(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})}$$

More generally, let us suppose that $\{A_i\}$ is a partition of the set of all possible events, then

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_j \Pr(B|A_j)\Pr(A_j)}.$$

So Bayes theorem is a learning mechanism using conditional probabilities. Its name comes from a British reverend, *Thomas Bayes* who lived during the eighteen century (1763). His theorem was rediscovered and generalised by the French mathematician *Pierre-Simon de Laplace* (1820). In his first edition of 1939, Jeffreys (1961) has proposed an axiomatisation of this theorem and has built the theory of statistical inference using it.

4.2 Bayesian inference

For statistical inference, we consider a random variable X, most of the time taking continuous realisations. We shall call these realizations our *sample space*. We shall assume that the sample space is equipped with a particular structure of σ-field so that we can define a probability measure over it. This measure is indexed by a *parameter* θ belonging to a parameter space Θ. If we assume that Θ is dominated by a $\sigma-$ finite measure (this is a restriction), probabilities over the sample space X can be described by a density function:

$$p(x|\theta).$$

So a non-parametric approach is left aside for the while. Given the realisation $(x_1, x_2, \ldots, x_i, \ldots, x_N)$ of N independent values of X, we can write down the

likelihood function of this observed sample:

$$\ell(\theta; x) = \prod_{i=1}^{N} p(x_i|\theta).$$

Classical inference is looking for the value of θ that is the most likely to have produced the observed sample, assuming that given the data density, there is somewhere a true value of the parameter on which we want to get information given the observed sample. However, there might exist other realisations of X other samples. So one of the main concerns of classical statistician is to ask the question: *What would happen if our sample size were tending to infinity?*

Bayesian statisticians follow a different way. They have decided to equip the parameter space Θ with a probability structure so that a *prior probability* $\varphi(\theta)$ can be stated. This means that there is no unique true value of θ, but uncertainty around the possible values that θ could take. The object of inference is to reduce this uncertainty by learning from the observation of a realisation x of X with:

$$\varphi(\theta|x) = \frac{\ell(\theta; x) \times \varphi(\theta)}{p(x)}.$$

This is another writing of Bayes theorem which requires a careful inspection.

1. $\varphi(\theta)$ is the prior density of θ which describes our prior knowledge around the plausible values of θ.

2. $\ell(\theta; x)$ is the likelihood function of the sample and the common element with the classical approach.

3. $\varphi(\theta|x)$ is the posterior density of θ, which means how our prior knowledge of θ was revised by the observation of one realisation x of the random variable X. By realisation, we mean the observation of a sample of a given size.

Finally $p(x)$ is the predictive density of x, given by:

$$p(x) = \int \ell(\theta; x) \times \varphi(\theta)\, d\theta.$$

It gives the probability of observing a particular realisation of our sample x, given all the possible likely values of the parameter θ. But it also insures that the posterior density integrates up to one.

The predictive density requires the evaluation of a large integral (the dimension of θ), but in fact this evaluation is rarely necessary. $p(x)$ is required for finding the integrating constant of the posterior density (the quantity necessary to insure that a density integrates to one). If $\ell(\theta; x)$ and $\varphi(\theta)$ belong to well-known families, the integrating constant of the posterior density $\varphi(\theta|x)$ can be recovered analytically. So Bayes theorem in this case can be simplified to:

$$\varphi(\theta|x) \propto \ell(\theta; x) \times \varphi(\theta),$$

which mean that the posterior density is proportional to the product of the likelihood function times the prior. This is the type of presentation adopted by Lindley (1971) for instance.

4.3 The likelihood principle

We come now to a very important interpretation of Bayes theorem. We lean on θ by experience, the experience here being the observation of a realisation of the sample. This means that we work conditionally on that observed sample and that this sample is given and unique. All the information we have is contained in the likelihood function and our revised information is described by the posterior density which gives a small sample result. *Once it is realised and observed, the sample is no longer a random variable*, so there is no problematic like what would have happened if we had observed something else? This means that Bayesian statisticians are not concerned by asymptotic theory. This does not mean however that they are not concerned by large sample approximations of the posterior density when the prior is dominated by the sample. But it leads to a different interpretation of the usual way of reporting inference results.

The fact that all the information is contained in the likelihood function led Jacques Drèze in his reply to Eric Sowey (Drèze 1983) to prefer using the Bayesian approach for teaching econometrics to economists because it does not require introducing side materials such as asymptotic theory, central limit theorem, asymptotic normality and so on. It makes justice to the available data series that might not be long, at least in macro-econometrics.

4.4 Credible sets

Once we have derived a posterior density $\varphi(\theta|x)$, what can we say? We can compute a posterior mean, a posterior standard deviation. More generally,

let us consider a function $g(\theta)$. We want to compute:

$$\int g(\theta)\varphi(\theta|x)\,d\theta.$$

If $g(\theta) = \theta$, we have the mean, if $g(\theta) = \theta^2$, we have the un-centered moment of order two. And so on. This computation can be done analytically in some cases, but more generally numerically, using simulation techniques that we shall detail in later on. But how to interpret a mean and a standard deviation? We have usually in mind a Normal distribution and are happy if the mean is twice the standard deviation. Why? It is better to define a credible set. A credible set is a region C of the parameter space such that its normalised surface is equal to a given level α, most of the time 0.90, 0.95 or 0.99. Formally, we have for $\alpha = 0.90$:

$$\Pr(\theta \in C) = \int_C \varphi(\theta|x)\,d\theta = 0.90.$$

But this set C is not unique. We can then look for the set which has the smallest area. This is the *Highest Posterior Density Interval* or HPDI.

The probability of C is perfectly defined and logical, once we have the posterior density of θ. We do not have the same property for classical confidence intervals. A classical estimator is a function of the sample noted $\hat{\theta}(x)$. This estimator has a distribution which is a function of the sample. That makes the whole difference.

Let us take the example of a normal sample $x \sim N(\theta, \sigma^2)$ where σ^2 is known. The sample mean is an estimator for θ and is distributed as $\bar{x} \sim N(\theta, \sigma^2/n)$. A classical confidence interval for θ at the α level is built as:

$$IC_\alpha = \bar{x} - t_i\frac{\sigma}{\sqrt{n}} < \theta < \bar{x} + t_i\frac{\sigma}{\sqrt{n}}$$

where t_i is defined such that:

$$\int_{-t_i}^{t_i} \frac{1}{\sqrt{2\pi}} \exp^{-u^2/2}\,d\,u = \alpha,$$

where u is distributed according to the standardised normal density. The logical foundation IC is weak because θ is assumed fixed in the classical view. The probability level is determined with respect to the distribution of $\bar{x}$ so logically we should have instead:

$$\theta - t_i\frac{\sigma}{\sqrt{n}} < \bar{x} < \theta + t_i\frac{\sigma}{\sqrt{n}},$$

which is a credible set for $\bar{x}$ with respect to the distribution of this estimator and of the sample.

Let us take an example found in Hoogerheide et al. (2009). It concerns the 95% HPD region for the average real GNP growth rate in the normal model with known variance, based on 24 quarterly observations from 1970 to 1975. Applying the above formulae yields $\theta|y \sim N(2.92, 0.91)$, so that the 95% HPD region for θ is:

$$[2.92 - 1.96 \times 0.91, 2.92 + 1.96 \times 0.91] = [1.14, 4.70],$$

which can be visualized in Figure 1. Because the normal density is a symmet-

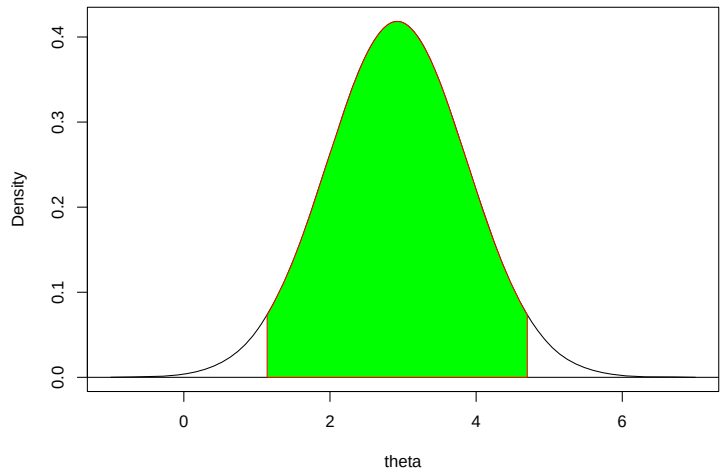

Figure 1: *The 95% HPD region for the average real GNP growth rate θ in normal model with known variance, based on 24 quarterly observations from 1970 to 1975.* Source: Hoogerheide et al. (2009)

ric unimodal density, the HPD is easy to find and identical to the classical interval.

A HPD region can be used to compare models in an asymmetric way, finding evidence against the null model. In the above example if the null model is a zero growth rate implying $\theta = 0$. It is rejected because $\theta = 0$ does not belong to a HPD interval.

4.5 Comparing models using posterior odds*

For an observed sample x, several models can be proposed as an explanation. How to compare these models? Suppose we have two models M_1 and M_2, index by θ_1 and θ_2. For each model M_i, we can derive:

$$\varphi(\theta_i|x, M_i) = \frac{\ell(x; \theta_i, M_i) \times \varphi(\theta_i|M_i)}{p(x|M_i)}$$

We can try to derive the posterior probability of each model, using the initial Bayes theorem given for discrete sets:

$$p(M_i|x) = \frac{p(x|M_i)p(M_i)}{p(x)}.$$

For comparing two models, a trick allows us to discard $p(x)$ by computing Bayes factors:

$$B_{12} = \frac{p(M_1|x)}{p(M_2|x)} = \frac{p(x|M_1)p(M_1)}{p(x|M_2)p(M_2)}.$$

An essential quantity is thus the marginal likelihood $p(x|M_i)$ which is obtained as:

$$p(x|M_i) = \int \ell(x; \theta_i, M_i)\varphi(\theta_i|M_i)d\theta_i.$$

It depends only on the prior and the likelihood function. It is in general difficult to compute this quantity, except in a number of simple cases, such as when comparing two linear regressions.

Model 1 is preferred to model 2 if $B_{12} > 1$. What is the degree of confidence of this decision rule? Jeffreys (1961) has provided what seems to be a classification, but which is in fact rough descriptive statement about standards of evidence in scientific investigation according to Raftery (1995).

$1 \leq B_{21} \leq 3$	not worth more than a bare mention
$3 \leq B_{21} \leq 10$	the evidence is positive
$10 \leq B_{21} \leq 100$	the evidence is strong
$B_{21} > 100$	the evidence is decisive

$p(M_1|x)$ is called the predictive density of model M_1 and is usually not very easy to compute, except in very particular cases. We shall come back to this point later on.

In this approach, models are treated in a symmetric way, they are simply compared and there is no privileged model. There is no null hypothesis.

4.6 Sufficient statistics

For making inference, should we keep the whole sample, i.e. all the observations or can we rely simply on a summary of that sample, say $t(x)$ which has a smaller size that the complete sample. This is the case if we do not loose information, which transcribed in term of posterior densities means:

$$\varphi(\theta|x) = \varphi(\theta|t(x)).$$

Because all the sample information is contained in the likelihood function, $t(x)$ is a sufficient statistics when the likelihood function can be factorised in a certain way (see e.g. Bauwens et al. 1999, chapter 2):

Theorem 1. *A necessary and sufficient condition for $t(x)$ to be a sufficient statistics for θ is that it is possible to factorise the likelihood function as:*

$$\ell(\theta; x) = h(x) \times k(\theta; t(x)).$$

A consequence of this theorem is that we can use the kernel $k(\theta; t(x))$ instead of the complete likelihood function, which is also another justification for the sign $\propto$ used above in Bayes theorem. We have sufficient statistics whenever the data density belongs to the exponential family, the definition of which requires a particular factorisation:

$$f(x|\theta) = h(x) \exp \sum_j u_j(x)\phi_j(\theta).$$

In this case the sufficient statistics are given by $t_j(x) = \sum_{i=1}^n u_j(x_i)$. To fix ideas, let us now indicate distributions that belong to the exponential family and distributions that do not:

1. Members of the exponential family: Normal, Gamma, Weibull, χ^2, beta, ...
2. Not in the exponential family: Student, uniform (as its supports depends on a parameter), mixtures of distributions.

We now give some examples of sufficient statistics for a series of generating processes

1. In a Bernouilli process, $x \sim B(p)$, sufficient statistics are given by the number of success $\sum x_i$ and the sample size n.
2. In the normal process, $x \sim N(\mu, \sigma^2)$, the sufficient statistics are n, $\sum x_i$ and $\sum x_i^2$.

3. For the Poisson process with parameter λ and for the exponential process with parameter θ, the sufficient statistics are n and $\sum x_i$.

4. For the gamma process, $x \sim G(\nu, \theta)$, the sufficient statistics are n, $\prod x_i$ and $\sum x_i$.

4.7 Natural conjugate priors

Natural conjugate priors are very convenient because they combine nicely with the likelihood function, leading to analytical results. Let us suppose that we have observed a sample x of size n coming from a data density belonging to the exponential family. Let us suppose that we have split the sample into two sub-samples x_1 and x_2. Is it possible to relate inference on the complete sample x with inference on the two sub-samples x_1 and x_2. The answer is **yes** because we can combine the sufficient statistics of sample x_1 with the sufficient statistics of sample x_2 so as to obtain sufficient statistics of the complete sample x. This is possible most of the time only if $p(x|\theta)$, the data density belongs to the exponential family. For instance if $x \sim N(\mu, \sigma^2)$ and if the sample sizes of x_1 and x_2 are n_1 and n_2 with $n_1 + n_2 = n$, then $\bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2)/n$.

Natural conjugate priors are related to the exponential family. Recalling the tale of the two samples in the *normal sampling process*, we now consider that our observations are contained only in x_2 and that x_1 is not observed but corresponds to an hypothetical sample. We can still compute sufficient statistics for this *hypothetical sample* and they will corresponds to the parameters indexing our prior density. To take an example, suppose that $x \sim N(\mu, \sigma^2)$ with σ^2 known. Sufficient statistics are $\sum_i x_i$ and n. Let us call x_0 an hypothetical sample of size n_0. By assumption $\bar{x}_0 \sim N(\mu_0, \sigma^2/n_0)$. We can deduce that the natural conjugate prior for μ will be a normal density with:

$$\varphi(\mu) = f_N(\mu|\mu_0, \sigma^2/n_0) \propto \sigma^{-n_0-1} \exp -\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2$$

while the likelihood function is:

$$\ell(x; \mu) \propto \sigma^{-n} \exp \frac{n}{2\sigma^2} \sum_i (x_i - \mu)^2.$$

Combining the prior together with the likelihood function, we get after some algebra:

$$\varphi(\mu|x) \propto \sigma^{-(n_*+1)} \exp \frac{n_*}{2\sigma^2} \sum_i (\mu - \mu_*)^2,$$

with:

$$n_* = n_0 + n, \qquad \mu_* = (n_0\mu_0 + n\bar{x})/n_*.$$

The posterior expectation of μ is a weighted mean of the prior mean and the sample mean where the weights are the respective sample size, hypothetical sample for the prior and observed sample for the sample mean.

4.8 Non informative priors

What happens if n tends to infinity or if n_0 tends to zero in the previous example? In the first case, the prior will be dominated by the sample as $\lim_{n\to\infty} \mu_* = \bar{x}$.

In the second case, the prior information will become weaker and weaker till it has the degenerate shape of a non-informative prior. More precisely:

$$\lim_{n_0\to 0} \varphi(\mu) \propto 1.$$

There are various ways of deriving a non-informative prior, and this is one of them. It is due to Novick (1969). A non-informative prior is obtained when taking a natural conjugate prior and letting its prior parameters go to their boundary values, the limit of their domain of definition. In the normal process with both mean and variance unknown, usual priors are the same as those derived from this principle.

$$\varphi(\mu|\sigma^2) \propto 1, \qquad \varphi(\sigma^2) \propto 1/\sigma^2.$$

We note that μ is a *location* parameter which translates a distribution over its support. The usual non-informative prior is proportional to 1.0 for location parameters. The second parameter, σ^2 is a *shape* parameter and its non-informative prior is different. It is proportional to the inverse of the shape parameter. This is illustrated in Figure 2.

The second important principle which can be used to define a non-informative prior is the *invariance principle* of Jeffreys (1961). This principle says that when there is little information, the prior should be independent of the way the model is parameterised. Jeffreys shows that this principle leads to taking a non-informative prior as:

$$\varphi(\theta) \propto |I(\theta)|^{1/2},$$

where $I(\theta)$ is the information matrix defined as:

$$I_{ij} = -\mathrm{E}\frac{\partial^2 \log \ell(\theta; x)}{\partial\theta_i \partial\theta_j}.$$

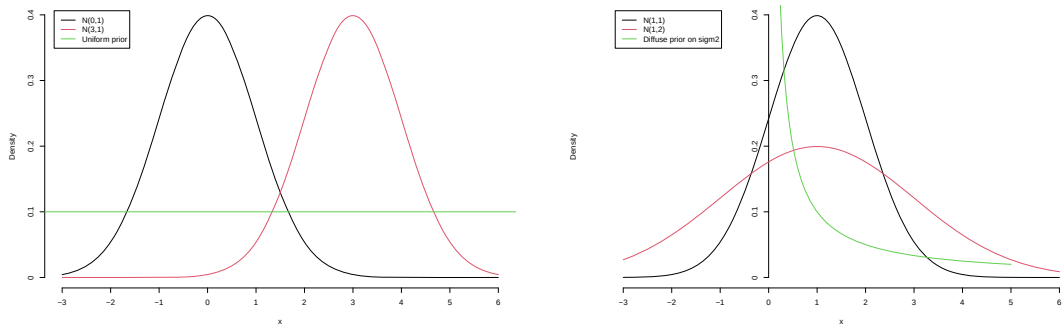

Figure 2: Diffuse prior for location and scale parameters

In the case of a normal process with known variance as above, the information matrix is equal to 1 so that the Jeffreys prior in this case is equal to the non-informative limit of the natural conjugate prior. This case was simple because the dimension of θ was one. In the multivariate case, the Jeffreys prior leads to prior densities that are different from the limit of the natural conjugate prior and leads to paradoxes. The usual prior is $\varphi(\mu, \sigma^2) \propto 1/\sigma^2$ as derived from the results above, while the Jeffreys prior would lead to $\varphi(\mu, \sigma^2) \propto 1/\sigma^3$. There is a difference in the exponent. So in general it is recommended to apply the Jeffreys principle separately to each parameter. For more discussion on this topic, see chapter 4 of Bauwens et al. (1999).

5 The linear regression model

It is convenient to start any econometric lecture with the linear regression model. Let us consider a random variable Y and a sample of n observations of it, noted in a matrix form $y' = [y_1, y_2, \cdots, y_n]$. The convention to write vectors being column vector, the prime sign here means the transpose. We could model y according to a Normal process with mean μ and variance σ^2 as before. But we are interested instead of *modelling the conditional expectation* of y, given the observation of k exogenous or explanatory variables noted X in a matrix form. If we detail this notation, we have:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

The regression model implies $\mathrm{E}(y|x) = X\beta$ or more commonly

$$y = X\beta + u,$$

where u_i, $i = 1, \ldots, n$, is an error term independently distributed as a Normal with zero mean and variance σ^2. We first note that in order to be able to write the likelihood function, we have to make a distributional assumption of the error term when the classical principle of ordinary least square does none, except for testing. For obtaining the OLS estimator, it is enough to assume that the u_i are IID and independent of the regressors X. Here, on top of that we have to make a distributional assumption. The likelihood function is:

$$\ell(y; \beta, \sigma^2) \propto \sigma^{-n} \exp -\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

This multivariate normal density is for the while a function of y. We have to develop the quadratic form so as to consider it as a function of β and σ^2. After some calculus, we get

$$\ell(y; \beta, \sigma^2) \propto \sigma^{-n} \exp -\frac{1}{2\sigma^2}(s + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}), \tag{1}$$

where

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y, & (2) \\ s &= y'y - y'X(X'X)^{-1}X'y \\ &= y'y - \hat{\beta}'X'X\hat{\beta}. & (3) \end{aligned}$$

This factorisation is interesting for finding the natural conjugate prior:

$$\varphi(\beta, \sigma^2) = \varphi(\beta|\sigma^2) \times \varphi(\sigma^2).$$

Remark:

> At this point, we have two options in the literature. Some authors like Bauwens et al. (1999) prefer to keep this parametrisation and thus to have an inverted gamma2 prior on σ^2. Another branch of the literature, such as Koop (2003), prefers to adopt another parametrisation in term of precision $h = 1/\sigma^2$ and in this case the prior on h is a gamma2.

5.1 Prior densities on σ^2 or on h

We have now to detail a class of distributions having a positive support and that are used to represent prior opinions. The first example is the gamma distribution. If $X \sim G(\nu, s)$ then its density is:

$$f(x|\nu, s) = C_G \times x^{\nu-1} \exp(-x/s),$$

where ν represents the degrees of freedom and s the scale parameter. C_G is the constant of integration such that the density integrates to one. The mean is νs and the variance νs^2. The χ^2 distribution corresponds to a particular parametrisation of the gamma density. If $X \sim \chi^2(\nu)$ then $X \sim G(\nu/2, 2)$. Its expectation is ν and its variance 2ν. The Gamma2 distribution is the one that will be interesting for devising a prior on $h = 1/\sigma^2$. If $h \sim G_2(\nu, s)$ then $h \sim G(\nu/2, 2/s)$. We see easily that it is related to the χ^2 as $h/s \sim \chi^2(\nu)$. The density of h is:

$$f(h|\nu, s) = C_{G2} \times h^{(\nu-2)/2} \exp -\frac{\nu h}{2s}.$$

Its expectation is ν/s and its variance $2\nu/s^2$.

For devising a prior on σ^2, one has to operate a transformation which leads to the definition of a new distribution, the inverted Gamma2, IG_2. If $\sigma^2 \sim IG_2(\nu, s)$ then its density is:

$$f(\sigma^2|\nu_0, s_0) = C_{IG} \times (\sigma^2)^{-(\nu_0+2)/2} \exp -\frac{s_0}{2\sigma^2} \tag{4}$$

The expectation and the variance are:

$$\mathrm{E}(\sigma^2) = \frac{s_0}{\nu_0 - 2}, \qquad \mathrm{Var}(\sigma^2) = 2\frac{s_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)}.$$

Note that these moments exist only for $\nu > 2$ and $\nu > 4$ respectively.

The gamma2 integrates to one (is a density) provided $\nu > 0$. And also $s > 0$, because a $s = 0$ would mean that the density is flat. So we have these two conditions. A way to obtain a non-informative prior is to take these two parameters at their limiting or boundary values. In this case:

$$\varphi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Finally, we notice that the IG2 density given in (4) appears in a part of the likelihood function (1). So it is a natural conjugate prior for the linear regression model.

5.2 Prior on β

β is a parameter that can take in theory any value, it is not restricted to positive values, so the Normal density can be used. However, its formulation is slightly different from usual as we need a conditional normal, conditional on σ^2. We have:

$$\varphi(\beta|\sigma^2) = f_N(\beta|\beta_0, \sigma^2 M_0^{-1}) \propto (\sigma^2)^{-1} \exp -\frac{1}{2\sigma^2}(\beta - \beta_0)' M_0 (\beta - \beta_0) \tag{5}$$

The prior expectation is given by β_0:

$$\mathrm{E}(\beta|\sigma^2) = \beta_0, \tag{6}$$

which give a direct interpretation to this parameter. The conditional prior variance is:

$$\mathrm{Var}(\beta|\sigma^2) = \sigma^2 M_0^{-1},$$

so that M_0 is a conditional prior precision matrix. This prior is natural conjugate for the linear regression model as we can recognise part of it in the likelihood function (1).

This conditional prior can be marginalizes so as to obtain a Student density implying:

$$\varphi(\beta) = \int f_N(\beta|, \beta_0, \sigma^2 M_0^{-1}) f_{IG2}(\sigma^2|s_0, \nu_0) d\sigma^2,$$

leading to:

$$\varphi(\beta) = f_t(\beta|\beta_0, M_0, s_0, \nu_0) \propto [s_0 + (\beta - \beta_0)' M_0 (\beta - \beta_0)]^{-(\nu_0+k)/2}.$$

The marginal prior moments are:

$$\mathrm{E}(\beta) = \beta_0 \qquad \mathrm{Var}(\beta) = \frac{s_0}{\nu_0 - 2} M_0^{-1}.$$

So in this way the interpretation of the parameters is simple. We identify clearly what are the parameters coming from the prior on the variance and the parameters coming from the prior on the regression coefficients.

But there is a redundancy among the parameters of this writing. We have four parameters, when only three are strictly necessary. So the Student density can be noted in different ways. For instance:

$$f_t(\beta) \propto \left(1 + (\beta - \beta_0)' \frac{S^{-1}}{\nu_0} (\beta - \beta_0)\right)^{-(\nu_0+k)/2},$$

which is also equivalent to

$$f_t(\beta) \propto \left(\nu_0 + (\beta - \beta_0)' S^{-1} (\beta - \beta_0)\right)^{-(\nu_0+k)/2}.$$

In this writing, S does not corresponds to the variance covariance matrix of β. The variance-covariance matrix of β is obtained as $S \times \nu_0/(\nu_0 - 2)$. For small degrees of freedom, the difference can be important.

5.3 Combining likelihood and prior

It is easy, but tedious to combine the conditional normal prior (5), the inverted gamma prior (4) and the likelihood function (1). Computational details are provided in Bauwens et al. (1999, page 58). The final result appear quite logical as:

$$\varphi(\beta, \sigma^2|y) \propto (\sigma^2)^{-(\nu_*+k+2)/2} \exp -\frac{1}{2\sigma^2}, (s_* + (\beta - \beta_*)' M_* (\beta - \beta_*). \qquad (7)$$

with:

$$\begin{aligned} M_* &= M_0 + X'X & (8)\\ \beta_* &= M_*^{-1}(M_0\beta_0 + X'X\hat{\beta}) & (9)\\ s_* &= s + s_0 + \beta_0' M_0 \beta_0 + \hat{\beta} X'X \hat{\beta} - \beta_*' M_* \beta_* & (10)\\ \nu_* &= \nu_0 + n & (11) \end{aligned}$$

In fact, this posterior density can be decomposed into the product of a marginal posterior density in σ^2, which is an inverted gamma2:

$$\varphi(\sigma^2|y) = f_{IG2}(\sigma^2|\nu_*, s_*).$$

The marginal posterior density of β is a Student density with:

$$\varphi(\beta|y) = f_t(\beta|\beta_*, M_*, s_*, \nu_*).$$

Once we have obtained these results, several options are possible. We can report the analytical posterior moments of σ^2 and β:

$$\mathrm{E}(\sigma^2|y) = \frac{s_*}{\nu_* - 2}, \quad \mathrm{E}(\beta|y) = \beta_*, \quad \mathrm{Var}(\beta|y) = \frac{s_*}{\nu_* - 2} M_*^{-1}. \qquad (12)$$

We can also *simulate random numbers* from these posterior densities, because interesting quantities might be just transformations of these parameters. In this case, it can be quite difficult to derive the analytical distribution of these transformations. But it is very easy to obtain draws from the posterior distribution of this transformation. We simply have to transform the draws obtained for β and σ^2 and these transformations will correspond to draws from the posterior distribution of the required transformation. We shall give an example in the next section with the Gini coefficient.

5.4 Empirical example: the Gini coefficient

The Gini coefficient is an index designed to measure inequality or dispersion in an income distribution. An empirical income distribution is formalized by

the sequence of observations $x_1, \cdots, x_n$. The Gini index is at value in [0,1]. The value 0 correspond to perfect equality, which means that everybody gets the same amount, the mean of the distribution, μ. The value 1 corresponds to perfect inequality where one individual gets everything, $n \times \mu$ and all the others have zero.

There are various ways of computing a Gini index. It is formally defined as the mean of all absolute income differences:

$$G = \frac{1}{2 \times n^2 \times \mu} \sum_{j=1}^{n} \sum_{i=1}^{n} |x_i - x_j|,$$

As there are only $n(n-1)/2$ different pairs in a sequence of n observations, this formula can be simplified into:

$$G = \frac{1}{n \times (n-1) \times \mu} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} |x_i - x_j|.$$

Because this formula involves two sums, it can be cumbersome to apply, so an alternative formulation was proposed by Deaton (1997, page 139). It relies on order statistics which are just observations ordered by increasing order. One way is to define ρ_i the rank of observation i and gives $\rho_j = 1$ if x_j is the maximum of the sample $x_{[n]}$ and $\rho_j = n$ if x_j is the minimum of the sample $x_{[1]}$. So $\rho_i = n + 1 - i$ and:

$$G = \frac{n+1}{n-1} - \frac{2}{n(n-1)\mu} \sum x_{[i]}(n+1-i).$$

This expression can be simplified into:

$$G = \frac{2}{n} \frac{\sum x_{[i]} \times i}{\sum x_{[i]}} - \frac{n+1}{n}.$$

In fact $(\sum x_{[i]} \times i)/\sum x_{[i]}$ is a regression coefficient. So Ogwang (2000) proposed to estimate the Gini as a by-product of a regression of x on its ranks and consequently to obtain a standard deviation for this estimate using the regression:

$$x_{[i]} = \alpha + \beta i + \epsilon_i.$$

The Gini coefficient is obtained as:

$$G = \frac{n^2 - 1}{6n} \frac{\hat{\beta}}{\bar{x}}.$$

But the randomness of G is here a function of the ratio of $\hat{\beta}$ and $\bar{x}$. So this solution is not very convenient in a classical framework and no simple formula was given by Ogwang (2000) to compute the standard error of this estimator.

Giles (2004) promoted another regression which corresponds to the computation of the mean of i, using a regression of i over a constant term, but taking into account heteroskedasticity in the error term, so that when correcting for heteroskedasticity the regression becomes:

$$i\sqrt{x_{[i]}} = \theta\sqrt{x_{[i]}} + u_i\sqrt{x_{[i]}}.$$

This is a weighted regression **without a constant term**. As in this regression $\hat{\theta} = \sum ix_i / \sum x_i$, the Gini coefficient is given directly by:

$$G = 2\frac{\hat{\theta}}{n} - \frac{n+1}{n}.$$

This time the variability of G depends only on the variability of $\hat{\theta}$ and is measured by $4\text{Var}(\hat{\theta})/n^2$. So both in a classical and in a Bayesian framework, this procedure is going to provide better results.

In this first empirical example, we shall consider an income variable drawn from a household survey, the Family Expenditure Survey in the UK, data collected in 1996. There are 6,042 observations. The sample Gini is 0.298, using the `Gini` command of the package `ineq` of `R`. But we do not have any information of the standard deviation of this estimate. We consider only the method of Giles to gain some information on the posterior density of the Gini coefficient and compare it to classical results. Using a regression, we find a result is identical to the previous case, and the estimated standard deviation is 0.007092548.

The question is of course to find the posterior distribution of the Gini. We have to proceed by simulation, using the results for for the posterior distribution of a linear regression model. We know this is a Student. With 10,000 draws, we get a posterior expectation of 0.2975283 and a standard deviation of 0.007124823. The classical approach has produced Gini = 0.297463 Standard error = 0.007092548, which are very closed values. However, a 95% HPD interval is [0.2835, 0.3115], while a classical 95% confidence interval, using the normal approximation would provide [0.2836, 0.3114], which is slightly smaller. This is an illustration of Bayesian results. In most cases, confidence intervals are slightly greater their classical counterparts.

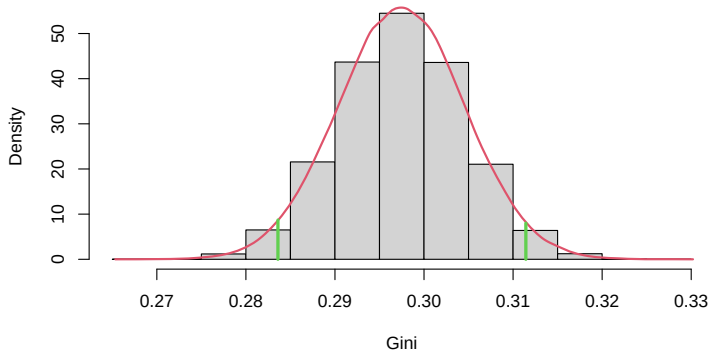

Figure 3: Posterior density of the Gini coefficient for the UK 1996

6 Some traditional simulation methods

Random numbers on a computer are not truly random numbers, but they look like random numbers. In fact they are generated from a deterministic chain:

$$X_t = (aX_{t-1} + b) \bmod m,$$

which produces a sequence of integer numbers between 0 and m. The value of m is machine determined, usually $m = 2^{31} - 1$. The starting value X_0 is called the seed. There are rules for determining optimally a and b. Here optimality means reaching the maximum length of a sequence. The sequence is finally divided by m so as to obtain a sequence of rational numbers between 0 and 1. The obtained sequence is at best uniformly distributed. It has to be transformed so as to obtain the desired distribution. Methods are as follows for univariate and multivariate distributions.

6.1 The inverse transform method

The cumulative distribution function $F(\cdot)$ of a random variable X is a monotonous increasing function between zero and one defined by:

$$F(x) = \Pr(X \leq x),$$

and the random variable $U = F(X)$ has a uniform distribution. So that a useful method for generating values for X is to use:

$$X = F^{-1}(U).$$

For some distributions, the inverse has an analytical form so the method is easy to implement. Examples are the exponential, the Weibull, the Pareto, the logistic, the Cauchy. For other distributions, we can always invert numerically the distribution for a given grid. But the method becomes costly.

6.2 The rejection method

The rejection method is a very powerful method for univariate distributions. The idea is quite simple. Suppose we want to draw from a complicated distribution $f(x)$ and that we have another distribution $q(x|\alpha)$ from which it is easy to draw random numbers and which is not too different from $f(x)$. Usually, $q(x|\alpha)$ is called the candidate function. Think for example the logistic and the Gaussian distribution. We build from $q(x|\alpha)$ an envelope to $f(x)$ with $c \times q(x|\alpha)$. We then draw a number from $q(x|\alpha)$ and we have to decide if it belongs or not to $f(x)$. The draw x_i is accepted if:

$$f(x_i) \leq c \times q(x_i|\alpha) \times U_i,$$

where U_i is a random draw from a uniform. The expected rate of acceptance is $1/c \leq 1$. So there is an interest to choose an optimal c, or in other words the best envelope. This method is used for instance to draw Gaussian random numbers.

6.3 Multivariate transformations

There are not many simple methods to draw random numbers from a multivariate distribution. Linear transformation are a useful tool for transforming a standardised random variable into a random variable with the same distribution, but with a specified mean and variance. The Choleski decomposition is particularly useful for this purpose. As a matter of fact, if $X \sim N(0, 1)$ in the univariate case, then

$$y = aX + b \sim N(b, a^2).$$

Let us now consider the multivariate case, where $X \sim N(0, I_p)$, where I_p is the identity matrix of dimension p. Then

$$y = AX + b \sim N(b, AA').$$

This time A is a square matrix of dimension $p \times p$ and b a column vector of dimension p. In the univariate case it was easy to find the value of a. We just had to take the square root of a^2. For a matrix, the operation

is more complicated. We have to decompose the matrix. If we want to draw random numbers from a multivariate normal distribution with zero mean and variance-covariance matrix Σ, we have to decompose this matrix. The Cholesky decomposition finds a lower triangular matrix L such that $\Sigma = L \times L'$, provided that Σ is positive, definite, symmetric. In `R`, this is provided by the `chol` function. This other solution is to compute the eigen values and eigen vectors of Σ. In this case $\Sigma = ADA'$ where D is a diagonal matrix with eigne values on its diagonal and A is the matrix of corresponding eigen vectors. This matrix is orthogonal which means that $A'A = AA' = I_p$. In `R`, the function `eigen` creates an object which contains eigen values and eigen vectors.

6.4 On a computer

Most softwares and in particular `R` offer directly implemented methods. For univariate distributions, these methods are directly available in `R` with `runif`, `rnorm`, `rt`, `rchisq` for instance. For multivariate distributions, specific packages have to be loaded with `mvtnorm` for multivariate normal and student distributions.

7 Conclusion

In this first lesson, I have introduced the basic notions necessary to understand what is Bayesian inference and how it is (slightly) different from classical inference. Bayesian inference is in a way the investigation of conditional probabilities and they are revised by observations.

I have taken specific examples that I will continue to use in the next chapters, but of course centered on the theme of poverty measurement. The word simulation will be particularly important in these lectures. Most of the concepts we will be interested in are functions of the parameters of the income distribution. Using simulations will be a particular elegant solution for finding the posterior distribution of these functions.

As I indicated in my foreword, several books are available which are concerned with Bayesian statistics and Bayesian econometrics. For basic principles, it is worth reading chapters 1 and 2 of Jackman (2009). The rather technical chapter 1 of Bauwens et al. (1999) is devoted to the relation between decision theory and Bayesian inference. What I have covered here is more developed in chapter 2 of Bauwens et al. (1999). For a good introduction to the classical econometrics of poverty and inequality, the book by Deaton (1997) is freely available on the web site of the World Bank:

https://elibrary.worldbank.org/doi/abs/10.1596/0-8018-5254-4

For poverty measurement, read our article Lubrano and Xun (2023). This lecture will follow exactly the sections of this paper.

References

Bane, M. J. and Ellwood, D. T. (1986). Slipping into and out of poverty: the dynamics of spells. *Journal of Human Ressources*, 21(1):1–23.

Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, Oxford.

Biewen, M. and Jenkins, S. P. (2006). Variance estimation for generalized entropy and Atkinson inequality indices: the complex survey data case. *Oxford Bulletin of Economics and Statistics*, 68(3):371–383.

Cappellari, L. and Jenkins, S. P. (2004). Modelling low income transitions. *Journal of Applied Econometrics*, 19(5):593–610.

Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics*, 150(1):30 – 40.

Deaton, A. (1997). *The Analysis of Household Surveys*. The John Hopkins University Press, Baltimore and London.

Drèze, J. H. (1972). Econometrics and decision theory. *Econometrica*, 40(1):1–18.

Drèze, J. H. (1983). Comment nonspecialist teaching of econometrics: A personal comment and personalistic lament. *Econometric Reviews*, 2(2):291–299.

Formby, J. P., Smith, W. J., and Zheng, B. (2004). Mobility measurement, transition matrices and statistical inference. *Journal of Econometrics*, 120(1):181–205.

Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: Some further results. *Oxford Bulletin of Economics and Statistics*, 66(3):425–433.

Hoogerheide, L. F., van Dijk, H. K., and van Oest, R. D. (2009). Simulation-based Bayesian econometric inference: principles and some recent computational advances. In Belsley, D. A. and Kontoghiorghes, E. J., editors, *Handbook of Computational Econometrics*. John Wiley and Sons, Ltd.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley, Chichester, UK.

Jeffreys, H. (1939, 1961). *Theory of Probability*. Oxford University Press, third edition.

Jenkins, S. P. and Lambert, P. J. (1997). Three I's of poverty curves, with an analysis of UK poverty trends. *Oxford Economic Papers*, 49(3):317–327.

Koop, G. (2003). *Bayesian Econometrics*. Wiley, Chichester, England.

Kuchler, B. and Goebel, J. (2003). Incidence and intensity of smoothed income poverty in European countries. *Journal of European Social Policy*, 13(4):357–369.

Lindley, D. V. (1971). *Bayesian Statistics, A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Lubrano, M. and Xun, Z. (2023). The bayesian approach to poverty measurement. In Silber, J., editor, *Research Handbook on Measuring Poverty and Deprivation*. Edward Elgar Publishing Ltd.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1):29–51.

Ogwang, T. (2000). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics*, 62:123–129.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Ravallion, M. and Chen, S. (2003). Measuring pro-poor growth. *Economics Letters*, 78:93–99.