Econ. 517, Econometric Theory I Fall 2001 Chris Sims

Models, Likelihood, Prior/Posterior, Standard Normal Linear Model*

1. Why "models" and "parameters"?

1.1. **The radical Bayesian view.** Models are a misleading construct. In this view, inference is always related to decision-making. We have a vector $[X, Y]$ of quantities we have not observed and are uncertain about. To make good decisions, we must start with a probability distribution, say defined by the density $p(x, y)$, over these uncertain quantities. Then we observe X. Since we still have not seen Y, we now need a probability distribution over Y for decision-making, and to be consistent we should form it as the pdf $p(y \mid x)$. No parameters or models have had to enter our discussion. There is just what we see, X, and what we don't see, Y, and inference is just going from $p(x, y)$ to $p(y \mid x)$.

1.2. **The radical non-Bayesian view.** It is inappropriate, at least in scientific discourse, to apply the notion of probability to purely subjective uncertainty. If β is a physical constant, we may not know its value, but it nonetheless has a single, unchanging value that is not in any sense "random". It is pointless to put a probability distribution on it. In such a case we might well be confident that an observable variable Y_t has for every t a probability distribution defined by a pdf $p(y; \beta)$. (Note that I did not write $p(y \mid \beta)$. That would suggest that β is a random variable we are conditioning on.) Knowing $p(y; \beta)$ and observing a sequence of i.i.d. Y_t's may provide us with information about β's value. Various functions of $\{Y_t,\ t = 1, \dots, T\}$ may turn out to have the property that they lie close to β with high probability. But the probability attaches to the functions of $\{Y_t\}$, not to β.

1.3. **The relaxed Bayesian view.** People find models a useful way to communicate statistical results, and we need to understand why. We start, like the radical Bayesian, with a division of unknowns into those to be observed, X, also called data, and those that will remain unobserved Y. We divide Y further as $Y = [\beta, Z]$, where β are parameters and Z are something else, which might sometimes be called "nuisance parameters", other times "unobservable components". The **model** is $p(x, z \mid \beta)$. In order to carry out inference, we have to form the **posterior pdf** $g(z, \beta \mid x)$. To do this we need to form the full joint pdf for (X, Z, β), so we need what is called a **prior pdf** for β — a marginal pdf $q(\beta)$ that can be combined with the model to form a joint pdf as $h(x,z,\beta) = p(x,z \mid \beta)q(\beta)$. From there we form the posterior by the usual rule for forming a conditional pdf from a joint one, so we compute

$$g(z,\beta \mid x) = \frac{p(x,z \mid \beta)q(\beta)}{\int p(x,z \mid \beta)q(\beta)\, d\beta\, dz}$$

Why bother with all this terminology, when what we end up doing is just what the radical Bayesian does? The reason is that it can be useful to distinguish between components of our uncertainty that are widely agreed upon, or at least widely regarded as interesting possibilities, by our audience, and those components that our audience disagrees widely about. The former components make up the model, the latter make up the prior. We try to present results so that they will be useful to an audience that may have widely differing priors.

Isn't this just what non-Bayesian inference does — analyze data without using subjective elements like prior distributions over parameters? No. A relaxed Bayesian approach does avoid making conclusions sensitive to a particular prior, but it preserves the view that results are ultimately to be combined with a prior to make decisions. Where possible, it tries to present results that summarize the shape of the function $p(x,z \mid \beta)$ as a function of β and z with x fixed — the **likelihood function**. While in many leading cases what is reported by classical statisticians can be interpreted as summarizing the shape of the likelihood, this is not true in some important econometric models — nonstationary time series models for example.

Note that most discussions of this subject leave out Z, and thereby make β, the parameter vector, coincident with Y, what we do not observe. We will do the same from here on in these notes. The presence of unobservable Z's that are not parameters can create difficult problems for inference and for reporting of results, but they have to be dealt with in the context of specific models.

1.4. **The Likelihood Principle.** In our discussion here, the idea that we should report the shape of the likelihood emerges naturally from a version of a Bayesian perspective. It is possible to derive from axiom systems roughly this same conclusion, though we will not do so here. This conclusion that everything useful that the data has to say about the parameters is contained in the likelihood function is what is known as the **likelihood principle**.

1.5. **Identification.** Suppose our model were $p(x \mid \beta)$, but it happened that $p(x \mid \beta)$ was the same for all β. In other words, the model implies that the distribution of X does not depend on β at all. Then of course the likelihood would always be completely "flat", equal to a constant function of β for every value of x. This is the most extreme version of a situation where β is not **identified**. The data in this case provide us with no information about β at all.

Less extreme cases:

- $p(x \mid \beta)$ does not depend on β for some values of x but for others it does. In this case, if we observe an X value for which the likelihood is flat, we can say that β is not identified in this sample.
- However classical statisticians would usually say that what counts is whether $p(x \mid \beta)$, regarded as a function of x over the whole range of possible x's, depends on β. In this case, if we can draw repeated samples of x, we will eventually get one for which the likelihood is not flat. This is what is most commonly meant by saying that "β is identified", if there are no qualifying phrases attached. The concepts below, of local identification and of identification of individual parameters, also can be interpreted as applying to particular samples or instead to the behavior of the whole $p(\cdot \mid \beta)$ function.
- It could be that $p(x \mid \beta)$ does depend on the vector β, but does not depend on some components of the vector. Then we could say that β_2, for example, is not identified even though β_j is identified for $j \neq 2$.
- It could be that the likelihood, though not flat everywhere, is flat in some neighborhood of a point β^*. Then we say that β is **locally** unidentified at β^*.

2. Bayesian Mechanics for the Standard Normal Linear Regression Model

The SNLM often denoted by the equation $Y = X\beta + \varepsilon$, asserts the following conditional pdf for the vector of Y data conditional on the matrix of X data and on the parameters β, σ^2:

$$p(\underset{T\times 1}{Y} \mid \underset{T\times k}{X}) = \varphi(Y - X\beta; \sigma^2 I) = (2\pi)^{-T/2}\sigma^{-T}\exp\left(\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right) \tag{1}$$

The most common framework for Bayesian analysis of this model asserts a prior that is flat in β and $\log\sigma$ or $\log\sigma^2$, i.e. $d\sigma/\sigma$ or $d\sigma^2/\sigma^2$. However, there are arguments in favor of other improper priors as a starting point, most prominently for using $d\sigma/\sigma^{k+1}$.[1] We will assume the prior has the form $d\sigma/\sigma^p$, then discuss how the results depend on p.

2.1. Marginal for σ^2.

We let $u(\beta) = Y - X\beta$ and denote the least squares estimate as $\hat\beta = (X'X)^{-1}X'Y$. Also $\hat u = u(\hat\beta)$. Then the posterior can be written, by multiplying

[1] This is the prior to which the reasoning behind **Jeffreys priors** leads. Jeffreys himself, though, favored the $d\sigma/\sigma$ prior for this model. You are not expected for this course to learn how Jeffreys priors are derived.

(1) by σ^{-p} and rearranging, as proportional to

$$
\begin{aligned}
&\sigma^{-T-p}\exp\left(-\frac{\hat{u}'\hat{u}+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right)d\beta\,d\sigma\\
&=\sigma^{-T-p}\exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2}\right)\left(\sigma^k\left|X'X\right|^{-\frac{1}{2}}\right)\sigma^{-k}\left|X'X\right|^{\frac{1}{2}}\exp\left(-\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right)d\beta\,d\sigma\\
&\qquad\propto\sigma^{-T-p+k}\left|X'X\right|^{-\frac{1}{2}}\exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2}\right)\varphi\big(\beta-\hat{\beta};\sigma^2(X'X)^{-1}\big)\,d\beta\,d\sigma\\
&\qquad\qquad\propto v^{(T+p-k)/2}\exp\left(\frac{\hat{u}'\hat{u}}{2}v\right)d\beta\,\frac{dv}{v^{3/2}}\,, \qquad (2)
\end{aligned}
$$

where $v=1/\sigma^2$. Integrating this expression w.r.t. β and setting $\alpha=\hat{u}'\hat{u}/2$ gives us an expression proportional to

$$
v^{(T+p-k-3)/2}\exp\left(-\frac{\hat{u}'\hat{u}}{2}v\right)dv\propto\alpha^{(T+p-k-1)/2}v^{(T+p-k-3)/2}e^{-\alpha v}dv\,,
$$

which is a standard $\Gamma((T+p-k-1)/2,\alpha)$ pdf. The prior arising from the multivariate Jeffreys analysis, $p=k+1$, therefore gives a $\Gamma(T/2,\alpha)$ pdf for v, regardless of k. The prior more usually called a Jeffreys prior, $d\sigma/\sigma$, produces a $\Gamma((T-k)/2,\alpha)$ distribution for v. The number $T-k$ is what is in this model called the **degrees of freedom**. Note that unless there are positive degrees of freedom, the $X'X$ matrix will not be invertible, the prior times the likelihood will therefore not be integrable in β, and the derivation we have just given does not go through. Because it is $v=1/\sigma^2$ that has the Γ distribution, we say that σ^2 itself has an **inverse-gamma** distribution. Since a $\Gamma(n/2,1)$ variable, multiplied by 2, is a $\chi^2(n)$ random variable, some prefer to say that $\hat{u}'\hat{u}/\sigma^2$ has a $\chi^2(T-k)$ distribution, and thus that σ^2 has an inverse-chi-squared distribution.

2.2. **Marginal on β.** Start with the same rearrangement of the likelihood (2), and rewrite it as

$$
v^{(T+p-3)/2}\exp\left(-\frac{1}{2}u(\beta)'u(\beta)v\right)dv\,d\beta\,.
$$

As a function of v, this is proportional to a standard $\Gamma\big((T+p-1)/2,u(\beta)'u(\beta)/2\big)$ pdf, but here there is a missing normalization factor that depends on β. When we integrate with respect to v, therefore, we arrive at

$$
\left(\frac{u(\beta)'u(\beta)}{2}\right)^{-(T+p-1)/2}d\beta\quad\propto\quad\left(1+\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{\hat{u}'\hat{u}}\right)^{-(T+p-1)/2}d\beta\,.
$$

This is proportional to what is known as a multivariate $t_n\big(0,(\hat{u}'\hat{u})/n\big)$ pdf, where $n = T + p - k - 1$ is the degrees of freedom. It makes each element of β_i an ordinary univariate $t_n(\hat{\beta}, s^2_\beta)$, where $s^2_\beta = s^2(X'X)^{-1}_{ii}$ and $s^2 = \hat{u}'\hat{u}/n$. As you will see later in the course, the sampling distribution of $(\hat{\beta} - \beta)/s_\beta$, considered as random across Y's with β and σ^2 fixed, is $t_n(0,1)$, which, if we take $p = 1$, exactly matches the implied posterior distribution of the same expression when considered as varying randomly with β while $\hat{\beta}$ and s^2_β remain fixed. Thus the statistics computed from the data can be analyzed with the same tables of distributions from either a Bayesian or non-Bayesian perspective.