David Giles

Bayesian Econometrics

9. Model Selection - Theory

- One nice feature of the Bayesian analysis is that we can apply it to drawing inferences about entire *models*, not just parameters.
- Can't do this with frequentist approach, especially if models are *non-nested*!
- We can put a prior on the *Model Space*, apply Bayes' Theorem, and then get posterior information about the competing models.
- Once we assign a Loss Function, we can then choose a model among the competing ones, so as to minimize posterior expected loss.
- Alternatively, we can use the *posterior probabilities* associated with each of the competing models as weights - create a weighted average of the results from each model. Bayesian Model Averaging.

- *e.g.*, M1: $y = X\beta + \varepsilon$; M2: $y = Z\gamma + u$
- Classical methods for choosing between these models can lead to conflicting outcomes - *e.g.*, the Cox Test (& extensions such as J-Test).
- They are virtually useless when it comes to more than 2 models at once.
- Our Bayesian Framework:

  We already have

  (i) a sample space, Y, with a joint data density, $p(\boldsymbol{y} \mid \boldsymbol{\theta})$

  (ii) a parameter space, $\Omega = \{\boldsymbol{\theta}\}$, and a prior density, $p(\boldsymbol{\theta})$

  We'll generalize the density in (i) to $p(\boldsymbol{y} \mid \boldsymbol{\theta_i}, M_i)$

  We'll generalize (ii) to $\Omega_i = \{\boldsymbol{\theta_i}\}$, for the *i'th model* (*i* = 1, 2, ...*m*), with an associated prior density, $p(\boldsymbol{\theta_i} \mid M_i)$.

We'll add a Model Space, $M = \{M_i\}_{i=1}^{m}$, with an associated prior mass function, $p(M_i)$; i = 1, 2, …., m. (m can be countably infinite.)

- We could write this mass function on the model space more completely as

  $p(M_i|\,\boldsymbol{\theta_i})$, where

  $$0 \le p(M_i|\boldsymbol{\theta_i}) \le 1 \quad ; \quad i = 1, 2, ..., m.$$

  $$\sum_{i=1}^{m} p(M_i|\boldsymbol{\theta_i}) = 1$$

- A *potential* difficulty with this last property is that we have to specify the model space exhaustively; and the "True Model" (DGP) has to be one of the competing models.
- We'll see later how this issue can be dealt with quite easily.
- Now let's put all of this together.

- We can define two densities that are generalizations of what we have already:

Conditional Data Density:

$$p(\boldsymbol{y} \mid M_i) = \int_{\Omega_i} p(\boldsymbol{y}, \boldsymbol{\theta_i} \mid M_i) d\boldsymbol{\theta_i} = \int_{\Omega_i} p(\boldsymbol{y} \mid \boldsymbol{\theta_i}, M_i) p(\theta_i \mid M_i) d\boldsymbol{\theta_i}$$

(*multi-dimensional integrals, again*)

Marginal Data Density:

$$p(\boldsymbol{y}) = \sum_{i=1}^{m} p(\boldsymbol{y} \mid M_i) p(M_i)$$

(Only the last of these results requires that we have exhaustively specified the model space.)

- Now we're ready to apply **Bayes' Theorem** to get the Model Space.

- The Posterior Probability for Model *i* is:

$$p(M_i \mid \boldsymbol{y}) = p(M_i)p(\boldsymbol{y} \mid M_i)/p(\boldsymbol{y})$$

$$\propto p(M_i)p(\boldsymbol{y} \mid M_i)$$

where the normalizing constant is $[p(\boldsymbol{y})]^{-1} = [\sum_{i=1}^{m} p(y \mid M_i)p(M_i)]^{-1}$.

- Note that the calculation for the posterior *probability* for Model *i* will be incorrect if the model space is not properly specified.

- However, even in the latter case, we can still make pair-wise comparions between the competing models.

- Specifically, we compute the Bayesian Posterior *Odds* in favour of one model over another.

- The Prior Odds in favour of Model *i* over Model *j* are $p(M_i)/p(M_j)$.

- The corresponding Bayesian Posterior Odds (BPO) are:

$$\text{BPO}_{ij} = [p(M_i \mid \boldsymbol{y})/p(M_j \mid \boldsymbol{y})] = \frac{p(M_i)p(\boldsymbol{y} \mid M_i)/p(\boldsymbol{y})}{p(M_j)p(\boldsymbol{y} \mid M_j)/p(\boldsymbol{y})}$$

Or, $\text{BPO}_{ij} = \left[\frac{p(M_i)}{p(M_j)}\right] \times \left[\frac{p(\boldsymbol{y} \mid M_i)}{p(\boldsymbol{y} \mid M_j)}\right]$

(Prior odds) ("Bayes factor")

- We can use the BPO to compare 2 models, even if the model space is incomplete.
- If, in fact, the model space *is complete*, then we can get the individual *posterior probabilities*:

e.g.: $[p(M_1 \mid \boldsymbol{y})/p(M_2 \mid \boldsymbol{y})] = 0.2$ and $[p(M_1 \mid \boldsymbol{y})/p(M_3 \mid \boldsymbol{y})] = 4$

Then,

$$p(M_2 \mid \boldsymbol{y}) = 5p(M_1 \mid \boldsymbol{y})$$

$$p(M_3 \mid \boldsymbol{y}) = 0.25p(M_1 \mid \boldsymbol{y})$$

$$p(M_1 \mid \boldsymbol{y}) = 1 - p(M_2 \mid \boldsymbol{y}) - p(M_3 \mid \boldsymbol{y})$$

and so,

$$p(M_1 \mid \boldsymbol{y}) = 0.16 \;\; ; \;\; p(M_2 \mid \boldsymbol{y}) = 0.80 \;\; ; \;\; p(M_3 \mid \boldsymbol{y}) = 0.04$$

A Decision Rule:

- Now use the Bayes' principle of "Minimum Expected Loss" (MEL) to help us to select between alternative models.
- Let L_{ij} (≥ 0) when M_i is the "True Model", but we choose M_j.
- $L_{ii} = 0$; $i, j = 1, 2, \ldots, m$. $L_{ij} \neq L_{ji}$, in general.

  So:

True

Selected	M_1	M_2
M_1	0	L_{21}
M_2	L_{12}	0

- When we choose M_1, the Posterior Expected Loss is:

$$E[L(M_1)|\, \boldsymbol{y}] = L_{11}\, p(M_1 \mid \boldsymbol{y}) + L_{21}\, p(M_2 \mid \boldsymbol{y}) = 0 + L_{21}\, p(M_2 \mid \boldsymbol{y})$$

- When we choose M_2, the Posterior Expected Loss is:

$$E[L(M_2)|\, \boldsymbol{y}] = L_{12}\, p(M_1 \mid \boldsymbol{y})$$

- Using the **MLE Rule** we will choose M_1 over M_2, iff

$$E[L(M_1)|\, \boldsymbol{y}] < E[L(M_2)|\, \boldsymbol{y}]$$

i.e., iff $[p(M_1 \mid \boldsymbol{y}) / p(M_2 \mid \boldsymbol{y})] > (L_{21}/L_{12})$

(BPO$_{12}$)

- If the Loss Function is *symmetric* choose M_1 over M_2, iff **BPO$_{12}$ > 1**.

- Can make pair-wise choice *without individual posterior probabilities*.

Some other results

- Can apply these ideas to *any* models. In econometrics, examples include: basic regression models; regression with non-standard assumptions; systems of equations; *etc*.
- If the models are "nested", and if we have proper priors for the parameters in each model, then BPO $\longrightarrow$ LR as $n \to \infty$.
- AIC, SIC, *etc*, can be interpreted as functions of the BPO.
- If we have regression models that are non-nested, with equal numbers of parameters, the BPO / MEL rule becomes equivalent to a "maximize R^2" rule as the prior information becomes increasingly "diffuse".

A simple example

- Suppose that $y \sim N[\theta\ , 1]$ and we have just <u>one observation</u>.
- We want to choose between $H_1: \theta = 1$ and $H_2: \theta = -1$.
- $BPO_{12} = \frac{p(\theta=\theta_1)}{p(\theta=\theta_2)} \times \frac{p(y\ |\theta=\theta_1)}{p(y\ |\theta=\theta_2)}$.
- In our case, the "Bayes factor" is

$$\frac{p(y\ |\theta = 1)}{p(y\ |\theta = -1)} = \frac{exp\left\{-\frac{1}{2}(y-1)^2\right\}}{exp\left\{-\frac{1}{2}(y+1)^2\right\}}$$

$$= exp\left\{-\frac{1}{2}(y^2 - 2y + 1 - y^2 - 2y - 1)\right\} = e^{2y}$$

- If we have equal prior probabilities, and a symmetric loss function, we'll choose H_1 if $e^{2y} > 1$. That is, if $y > 0$.
- Similarly, we'll choose H_2 if $e^{2y} < 1$. That is, if $y < 0$.
- If $y = 0$, we'll be indifferent between the 2 hypotheses, *a posteriori*.
- Does this make sense? (Of course!) And we have just one observation.
- Suppose we draw $y = 0.5$, and we have prior odds of "1"; and $L_{12} = L_{21}$.
- Then $BPO_{12} = e^1 = 2.718$, and $p(H_1 \mid y) + p(H_2 \mid y) = 1$.
- So, $p(\theta = 1 | y) = 0.73$; and $p(\theta = -1 \mid y) = 0.27$.
- If $y = 1$, then $p(\theta = 1 | y) = 0.88$; and $p(\theta = -1 \mid y) = 0.12$; *etc*.
- Experiment with different prior odds, and asymmetric losses.

- How does this compare with what a frequentist would do?
- Let $H_0 = H_1$ and $H_A = H_2$, Choose $\alpha = 5\%$.
- $Z = (y-1)/1$. Reject H_0 if $Z < -1.645$. That is, if $y < -0.645$.
- $y = -0.645$ corresponds to $BPO_{12} = e^{-1.29} = 0.275$.
- This implies that $p(\theta = 1 | y) = 0.784$; and $p(\theta = -1 \mid y) = 0.216$, if we have equal prior probabilities.
- If $BPO_{12} = 0.275$, and we have equal prior probabilities for the 2 hypotheses, what loss structure would "match up" with the frequentist's 5% significance level?
- Reject $H_1: \theta = 1$ if $BPO_{12} < (L_{21}/L_{12})$. We'd need $(L_{21}/L_{12}) = 0.275$.
- $L_{12} = 3.636 L_{21}$.
- $Loss[Choose\ H_1 \mid H_2 True] = 3.636 \times Loss[Choose\ H_2 \mid H_1 True]$.