DUMMY OBSERVATION PRIORS REVISITED

CHRISTOPHER A. SIMS

ABSTRACT. TBW

I. INTRODUCTION

II. THEIL'S "MIXED ESTIMATION" IDEA

Most econometricians are familiar with Theil's (ref) "Mixed Estimation" idea, which he presented as a method for handling the problem of "multicollinearity" in regression. The idea is simple: add dummy observations to the sample that come out of the econometrician's head, reflecting uncertain prior knowledge about the model's parameters. While Theil did not label his idea as Bayesian, it obviously is Bayesian in spirit, and it can be given a Bayesian interpretation. In fact, adding dummy observations to the sample can be interpreted as nothing more or less than using a conjugate prior, since a conjugate prior is any prior that has the same form as the likelihood for additional observations.

In a regression model of the form

$$y(t) = X(t)\beta + \varepsilon(t) = x_1(t)\beta_1 + x_2(t)\beta_2 + \varepsilon(t)\,,\ t = 1, \ldots, T\,,\ , \tag{1}$$

(with ε i.i.d. $N(0, \sigma^2)$) we usually think of adding prior information that $\beta_1 \sim N(\mu_1, \sigma^2/\lambda_1^2)$ by adding the dummy observation

$$y(0) = \lambda_1 \mu_1\,,\ X(0) = [\lambda_1\ 0]\,. \tag{2}$$

If we add also a similar dummy observation for β_1, this is equivalent to using a prior on the β's that asserts

$$\beta \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \sigma^2 \begin{bmatrix}\lambda_1^{-2} & 0\\ 0 & \lambda_2^{-2}\end{bmatrix}\right)\,. \tag{3}$$

But this apparent equivalence between dummy observations and directly asserted priors breaks down if we ask what happens if, in addition to these beliefs about β_1 and β_2 separately, we also believe that $\beta_1 + \beta_2$ is approximately one (as might

Date: April 28, 2005.

be true, e.g., in a production function model). With a directly asserted prior, we can assert a marginal prior on β_1 and also a marginal prior, treated as independent of that on β_1, on *either* $\beta_1 + \beta_2$ or β_2 alone, but at that point we have finished the job. The two marginal priors imply a joint prior and therefore a marginal prior on both β_2 and the sum. But if we are thinking in terms of dummy observations, we can use all the results of our mental self-questioning, so that our list of dummy observations becomes

$$\begin{bmatrix} y(0) \\ y(-1) \\ y(-2) \end{bmatrix} = \begin{bmatrix} \lambda_1\mu_1 \\ \lambda_2\mu_2 \\ \lambda_3 \end{bmatrix}, \qquad \begin{bmatrix} X(0) \\ X(-1) \\ X(-2) \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ \lambda_3 & \lambda_3 \end{bmatrix}. \tag{4}$$

These dummy observations imply a joint distribution for β_1 and β_2 in which σ_i^2/λ_i^2 is no longer the variance of the marginal distribution of β_1. The implied unconditional prior variance for β_i will be smaller, because of the additional "observation" on the sum of coefficients. But this is not hard to allow for intuitively, and the dummy observation approach to setting up the prior is arguably easier to interpret. For example, if we want to imply symmetry in the joint distribution, it is clear that this can be done by setting $\lambda_1 = \lambda_2$, $\mu_1 = \mu_2$. Working with the prior covariance matrix directly it is also easy to impose symmetry, but perhaps not so easy to see how choices of prior covariance or precision matrices translate into beliefs about the standard error of the marginal prior on $\beta_1 + \beta_2$.

The fact that we can use as many mutually independent dummy observations as we like, whereas for a k-dimensional parameter vector we can generally specify only k independent marginal prior distributions on functions of the parameter vector, reflects a fundamental point about dummy observations — successive dummy observations change the prior as if they reflected successive error-ridden observations of functions of the parameter vector. Except in the linear case with no more dummy observations than parameters, this is quite different from specifying successive independent marginal prior distributions on the same list of functions of the parameter vector. As we will see below, the differences are especially important for high-dimensional and highly nonlinear models.

III. Nonlinearity

The dummy observation idea need not be confined to implementing conjugate priors or to linear models. We can build up a prior via mental observations on nonlinear functions of parameters. This can be helpful, because often it is easier to specify beliefs about nonlinear functions of parameters as if we had noisy observations on them than as if we knew their distribution. Technically, dummy observations on nonlinear functions of parameters treat Jacobian terms differently from the way they are treated in putting a prior distribution directly on the nonlinear

function. The usual form of this more general type of dummy observation will be a factor in the prior of the form $p(f(\beta) - \hat{f})$, where p is interpreted as proportional to the pdf of the "error" in the mental observation on $f(\beta)$ and $\hat{f}$ is interpreted as the observation itself. If instead of building up a prior this way we were to propose a sequence $q_i(f_i(\beta))$ of independent marginal pdf's for functions of β, we would have to use not only the q_i's, but also the Jacobian term $|\partial f / \partial \beta|$. Because of the effects of the Jacobian term on the shape of the resulting pdf, specifying a prior in terms of pdf's for f's can easily have unintended implications. Of course this is true in principle for priors built from dummy observations also, but the connection from dummy observation error pdf's p_i to the resulting shape of a prior on β is more direct, and thus often more interpretable.

IV. Normalizing for Model Comparison

Within the normal linear regression model, the shape of the posterior calculated by simply tacking the dummy observations onto the sample and tracing out the resulting "likelihood" is easily seen to exactly match that calculated by using the corresponding conjugate prior directly. There is a difference, however. The the conjugate prior has a scale factor that does not match the scale factor in the dummy-observation likelihood. For model comparison, this is important. In linear regression models, including sets of linear regressions with the same right-hand-side variable lists (like VAR's) correcting the scale factor can be done analytically, as the shape of the conjugate prior is of known (normal-inverse-Wishart) form.[1]

Though it is easy to make this correction, it is also easy to forget it, since the dummy observations work so easily without it when model comparison is not at issue. Note that the same issue comes up with respect to "training sample" priors, which can be interpreted as simply using the likelihood as a posterior, while weighting it by the inverse of the integral of the likelihood constructed from some initial part of the sample. Since training samples are used mainly in the context of model comparison, though, the proper construction of the weights is likely to be recognized as central.

V. VAR impulse responses

It is relatively easy to put priors on reduced form VAR models with conjugate-prior dummy observations, and only somewhat more difficult to handle priors on

[1] Software that does this for VAR's with arbitrary dummy observations, as well as dummy-observation versions of the "Minnesota Prior", is available on my website as R and matlab files `mgnldnsty.R`, `mgnldnsty.m` and `matrictint.R`, `matrictint.m`. The latter calculate the integrated density for a general multivariate regression likelihood, whereas the former specialize to VAR's and the Minnesota Prior.

contemporaneous coefficients in structural VAR models. Much of the VAR literature has used priors of these convenient forms. The monetary structural VAR literature, though, has obviously also used informal prior beliefs about what are reasonable shapes for impulse responses, and there has been some dissatisfaction with this aspect of the literature. A few have proposed ways to incorporate beliefs about impulse responses more formally — Uhlig (ref), Faust (ref), Canova (ref), Kocięcki and Dwyer (ref) among them. Kocięcki and Dwyer propose putting priors directly on the impulse responses and confront the problem that impulse responses are highly nonlinear functions of the regression parameters (and thus have important Jacobian terms) and that they are a higher-dimensional object than the parameter vector, so that only proper subsets, or sets of functions of, the impulse responses have nonsingular distributions. The other three authors simply sidestep these problems by proposing methods that, while producing interesting results, have no clear interpretation as Bayesian or any other kind of probability-based inference.

A related problem is the tendency, documented in earlier papers of mine (ref), for VAR estimates conditioned on initial observations to attribute implausible degrees of explanatory power to initial conditions that are estimated as very far from steady state. The Minnesota Prior does mitigate this tendency, but it only pushes the model toward having *some* unit roots, and in models with many lags or many variables, this leaves lots of room for other types of persistence, particularly for roots elsewhere on the unit circle. A direct way to control this problem would be to put priors on functions of impulse responses — for example asserting that sums of squared changes in the impulse responses after some horizon are small.

REFERENCES

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY
E-mail address: sims@princeton.edu