DERIVATION OF MDD FOR FERNANDEZ/LEY/STEEL

This is a standard derivation, integrating the posterior by recognizing that it has the form of a normal-inverse-gamma distribution. It nonetheless could be useful to go through it if you haven't seen this kind of argument before. Also, there are some conceptual issues around the flat priors on σ^2 and the constant term α that I initially found confusing, and thus may be worth bringing out.

Each of the models FLS consider is a normal linear regression that asserts

$$y \mid X \sim N(X\beta, \sigma^2) . \tag{1}$$

They use a prior, on each model, in which

$$\beta \mid \left\{\sigma^2, X\right\} \sim N(0, \sigma^2(gX'X)^{-1}) . \tag{2}$$

They say the σ^2 and α parameters are "common to all the models", but they do not mean that they have the same value in all the models. Within each model, σ^2, α are unknown parameters, unrelated in the prior to the parameters of the same name in the other models. One can imagine that beliefs about σ^2 with one set of regressors would be correlated with beliefs about that parameter with other sets of regressors. But that is not allowed for in the FLS setup.

They use an improper prior for α, σ^2, proportional to $1/\sigma$, i.e. flat in α and $\log \sigma^2$.

Since flat priors are not integrable, multiplying them by any constant leaves their implications for inference unchanged. But in model comparison based on integrating the posterior to get marginal data density, rescaling a flat prior rescales the mdd. So generally in model comparison one can't use flat priors — results will be nonsense. But when, as here, there is in each model being compared a parameter for which we can use the same, flat, improper prior, the model comparison results are still meaningful. So long as the flat priors are rescaled the same way in each model, rescaling does not affect model comparisons.

This approach has pitfalls, though. For example, in this FLS paper's setup one might expect that if the model turns out to be big, with many variables getting large coefficients, the residual variance σ^2 might be smaller. So if we used proper inverse-gamma priors on σ^2, we might make the scale larger on the larger models. Then, if we let the scales grow toward infinity at the same rate, we would end up with flat priors on all the models, but the ratios of the heights of the densities would go to a constant other than one. In other words, we would be using "flat" priors of different scales, and this would affect odds ratios.

FLS transform the data by removing the sample means from X. This makes the likelihood factor, so that if the prior made α and β independent, the independence is preserved in the posterior. Of course this also changes the interpretation of the constant term, but since the transformation makes α an estimate of the expectation of the sample mean of y in every model, it makes the common flat prior more, not less, plausible.

Date: September 29, 2018.

With the $1/\sigma$ flat prior and the g-prior on β, and using X for the demeaned data, the posterior pdf is

$$\sigma^{-(k+1)}(2\pi)^{-k/2}g^{k/2}\left|X'X\right|^{\frac{1}{2}}\exp\left(-\frac{g}{2\sigma^2}\beta'X'X\beta\right)\cdot$$
$$\sigma^{-n}(2\pi)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\hat{u}'\hat{u}+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})+n(\alpha-\hat{\alpha})^2)\right). \quad (3)$$

Here $\hat{u}$ is the least squares residual vector and $\hat{\beta}$ and $\hat{\alpha}$ are the OLS estimators.

As a function of α, this is proportional to a $N(\hat{\alpha},\sigma^2/n)$ pdf, so integrating over α removes this term from the exponent and inserts a factor $\sqrt{2\pi}\sigma n^{-\frac{1}{2}}$. Then we "complete the square" for the terms in the exponent involving β, observing that they can be expressed as

$$\frac{-1}{2\sigma^2}\left((\beta-\hat{\beta}(1+g)^{-1})'X'X(1+g)(\beta-\hat{\beta}(1+g)^{-1})+\hat{\beta}'X'X\hat{\beta}g/(1+g)\right). \quad (4)$$

This makes the posterior, as a function of β, proportional to a $N\left(\hat{\beta}/(1+g),\sigma^2((1+g)X'X)^{-1}\right)$ pdf. Note that the factor introduced by integrating over β cancels with most of the g-prior term. After integrating out both β and α, then, we have

$$\sigma^{-n}n^{-\frac{1}{2}}\sqrt{2\pi}\left(\frac{g}{1+g}\right)^{k/2}\exp\left(-\frac{1}{2\sigma^2}\left(\hat{u}'\hat{u}+\hat{\beta}'X'X\hat{\beta}\frac{g}{1+g}\right)\right). \quad (5)$$

As a function of σ^2, this expression is proportional to an inverse-gamma density with $n/2-1$ degrees of freedom. Also, observe that (assuming y has also had sample mean removed)

$$\hat{u}'\hat{u}+\hat{\beta}'X'X\hat{\beta}=y'y\,. \quad (6)$$

Now we integrate over σ^2 and drop factors that don't change across models, to arrive at

$$\left(\frac{g}{1+g}\right)^{k/2}\left(\frac{y'yg}{1+g}+\frac{\hat{u}'\hat{u}}{1+g}\right)^{\frac{n}{2}-1}. \quad (7)$$

FLS have a different exponent on the second factor: $(n-1)/2$ instead of $\frac{n}{2}-1$. It might make sense to treat the "flat" prior on α as having a height proportional to $1/\sigma$. This would be the limit of conjugate priors with variances going to infinity. If FLS did this, that would explain the discrepancy.