TECHNICAL APPENDIX

This appendix includes some more precise details about computational aspects of the BACE procedure with particular emphasis on the sampling algorithm and convergence. Given the form of our prior distribution, the prior inclusion probability for each variable is $\bar{k}/K$ as described in the main text. Represent a model, M_j, as a length K binary vector in which a one indicates that a variable is included in the model and a zero indicates that it is not. Then:

$$\text{(A1)} \qquad P(M_j) = \left[\prod_{i=1}^{K} M_{ji}\frac{\bar{k}}{K}\right]\cdot\left[\prod_{i=1}^{K}(1 - M_{ji})\left(1 - \frac{\bar{k}}{K}\right)\right] = \left(\frac{\bar{k}}{K}\right)^{k_j}\cdot\left(1 - \frac{\bar{k}}{K}\right)^{K-k_j}$$

where k_j is the number of included variables in model j and M_{ji} is the i'th element of the M_j vector. The second equality in (A1) holds only in the case of equal prior inclusion probabilities for each variable, but the first equality is easily adapted to the case in which the prior inclusion probabilities may differ across variables. If the set of possible regressions is small enough to allow exhaustive calculation, one may substitute (A1) into (7) to calculate the posterior model probabilities and then use (8) and (9) to calculate the posterior mean and variance. For each term of the sum one calculates the appropriate OLS regression, gets the OLS parameter estimates for the β's and σ and the sum of squared errors. These allow the computation of the individual term in (8) and (9). Also the posterior probabilities allow the calculation of any other features of the posterior distribution which may be of interest based on the 2^K-term version of (3). As for the other quantities cited in this paper, the "sign-certainty statistic" is given by:

$$\text{(A2)} \qquad \text{sign certainty for } \beta_j = \sum_{j=1}^{2^K} P(M_j|y)P\left[sign(\beta_j) = sign\left(E(\beta_j|y)\right)|M_j,y\right]$$

The histograms for the posterior densities are calculated as follows. An initial run established the important range of the distribution of the estimates for each β. This was then split into 100 equal size bins for the histogram. Since for each regression the ratio of $\hat{\beta}$ to the estimated standard deviation of the error term is distributed $t(T\text{-}k\text{-}1)$ we can use a t-CDF to evaluate the amount of probability contained in each bin. This is then weighted by the posterior probability of the regression. Note that the calculation of these histograms is quite computationally intensive as with each regression we must make 100 times k calls to a t-CDF.

When we are sampling randomly from the space of possible models we want the limits of all of our quantities of interest to approach their true values as the number of sampled models approaches infinity. If we let the probability of sampling M_j be given by $P_s(M_j)$ then the weight attached to each regression must be adjusted by the inverse of the sampling probability. This is because as the number of sampled regressions approaches infinity the fraction of times a particular regression is run approaches its sampling probability, when in sums such as (8) and (9) each regression gets equal weight. Thus, with sampling the analog of (7) becomes:

where $m(i)$ represents the model index associated with the i'th randomly sampled model and N is

$$P(M_j|y) = \frac{\dfrac{P(M_j)}{P_s(M_j)} T^{-k_j/2} SSE_j^{-T/2}}{\left(\sum_{i=1}^{N} \dfrac{P(M_{m(i)})}{P_s(M_{m(i)})} T^{-k_{m(i)}/2} SSE_{m(i)}^{-T/2} \right)} \tag{A3}$$

the number of models sampled. This version of the weights can then be used to calculate sampling analogs of (8) and (9). The intuition for (A3) is that we are over-sampling some models so as usual we have to deflate observations by their sampling probabilities. (A3) is particularly easy to calculate when the sampling probabilities are equal to the prior probabilities in which case they cancel and need not even be computed. This is the sampling strategy discussed in the text of randomly selecting models by randomly including variables with their initial inclusion probability. So long as the sampling probabilities of all models are greater than zero all of the numerical approximations will be consistent.

Trial-and-error calculation indicated that for the present problem the prior-weight sampling was leading to slow convergence of the parameter estimates. This is because it samples many, many poorly fitting regressions which receive little weight in the averages. Instead we used the following procedure which we refer to as the "stratified sampling": we ran 100,000 regressions using the prior weight sampler and then adjusted the sampling inclusion probabilities to be equal to the posterior inclusion probabilities estimated from the initial sample. In order to guard against too much impact from errors made in the first 100,000 regressions we limited the sampling probabilities to lie in the interval [0.1, 0.85]. Some experimentation suggested that moderate changes in these bounds has little effect on the behavior of the algorithm. Again, since any set of sample inclusion probabilities will work asymptotically the choice of these parameters is not critical. Thus our stratified sampler over-samples "good" regressions.

We then need some way of judging whether or not the sampled analogs of (8) and (9) are approaching their limits. As always, convergence criteria are somewhat arbitrary. For the estimates reported in the paper we examined changes in the posterior means of the β's. First we normalized the coefficient estimates by the ratio of the standard deviation of y to the standard deviation of x. The standardization with respect to y is only to make the size of the convergence criterion easy to interpret. This transformation standardizes the β's into units of standard deviations of y per standard deviation of x. Then in order to declare that the estimates "converged" we looked at the change in the estimates of the normalized β's generated by sampling a further 10,000 regressions. When this change fell below 10E-06 for ten consecutive sets of 10,000 regressions the algorithm declared convergence. In addition, we also checked for convergence of the posterior inclusion probability with changes less than 10E-4. For our stratified sampling technique these parameter changes fall smoothly as a function of the number of regressions so that this criterion is reasonable. For the prior probability sampler this change is much less reliable with the occasional set of 10,000 having a large impact: we would not recommend the use of this sampler with this particular convergence criterion. For our baseline estimation with $\bar{k} = 7$ we also investigated the performance of the sampler and convergence criterion by performing a number of further runs with the same convergence criterion: these all converged around 80-90 million regressions; in addition we ran a sampler with 200 million stratified draws. Results were very similar: they suggest that the posterior inclusion probabilities

in table 2 are accurate to at least two decimal places, while the conditional β estimates are even more accurate. Estimates based on only two million or so regressions are even quite close to the 89 million regression baseline. This suggests that our methodology will create quite accurate approximations in reasonable computing times even with very large model spaces.

In the Bayesian Model Averaging statistics literature, which has used fully Bayesian estimates of individual models, the most popular sampling algorithm appears to be the MC^3 algorithm mentioned in the main text. We were resistant to using this algorithm because its mechanism, based on the Metropolis-Hastings criterion, is quite difficult to understand intuitively. In order, however, to both try to ensure that our stratified sampler is generating correct answers and to compare it to procedures in other work we created a test data set. This used all of the observations in our main data set but with only 20 variables rather than the full set of 67. This reduces the set of possible regressions to around one million which easily allows the precise calculation of the sums in (7), (8) and (9). We then performed sampling runs with 50,000 regressions each and calculated a weighted mean-squared error criterion for the posterior means of the β's with the weighting matrix being $(X'X)^{-1}$. By this criterion the stratified sampling algorithm was about four times as accurate as MC3, but for both accuracy was quite reasonable.