# S-values and All Subsets Regressions

Edward E. Leamer[1]

UCLA

Abstract

This paper compares and contrasts Bayesian variable-exclusion methods proposed by Fernando Ley and coauthors with methods proposed by Raftery and Sala-i-Martin et al. and with the s-values proposed by myself. A distinction is drawn between estimation uncertainty which is the focus of Ley's research and model ambiguity which arises in Ley's work and is the focus of my own recent proposal. The discussion is organized around the prior covariance matrix, which needs to be diagonal to support all-subsets regressions. The basic question that addressed here is: What aspects of the prior covariance matrix can be taken as known, what aspects can be estimated and what aspects require a sensitivity analysis because they are neither known nor estimable. When diagonality is in doubt, we are more-or-less forced into a model ambiguity sensitivity mode because the data are never rich enough credibly to estimate the full prior covariance matrix. When diagonality is assumed, the data evidence, though very limited, can help to estimate the diagonal elements, but this literature has not yet produced a compelling conventional treatment which will necessarily include both estimation uncertainty and model ambiguity as they relate both to the diagonal values and to the rest of the prior covariance matrix. But there has been a lot of progress.

[1] Prepared with the research assistance of Shekhar Mittal. Very useful comments by Mark Steel are gratefully acknowledged.

# 1 Introduction

Fernando Ley has been a coauthor on a series of important papers regarding Bayesian variable-inclusion methods for linear regression. Fernandez, Ley and Steel(2001a) offer a Bayesian study of the determinants of economic growth in a cross section of countries with large numbers of correlated explanatory variables and not-so-many observations, which is the kind of setting in which the data evidence can be too weak to estimate many of the regression coefficients with tolerable accuracy, if the all-inclusive model is estimated with unconstrained ordinary least squares. There is consequently a need to supplement the data evidence with prior information to the effect that the coefficients are probably small, perhaps small enough that some variables can be dropped from the equation. If these priors are properly deployed, estimates are shrunk toward zero and have greater accuracy than the estimates of the all-inclusive model, if the priors are correct.

This literature has suggested multiple ways of analyzing this type of data, including my own recent advice, Leamer(2014), which proposes a conventional measure of model ambiguity which I have called an s-value, to accompany a t-value and a p-value. The purpose of this paper is to compare and contrast the Bayesian alternatives to “all-inclusive” regression which have been proposed and studied by Fernandez, Ley and Steel(2001a) and by Sala-i-Martin, Doppelhofer and Miller (2004), by Ley and Steel(2009) and by Leamer(2014). These four papers are the focus because they all illustrate their proposals with a study of the same country-growth data set, but I will refer to the larger literature as appropriate.

The conventional choice for the prior mean in this literature is zero. After that, in importance, come the second moments. Thus the principal question that organizes this discussion is “What is a wise conventional choice of a prior covariance matrix?” A lot of the energy of these papers focusses on higher order moments, which are important, but I think not so susceptible to conventional thinking. I do agree that the assumption of a normal distribution for a prior is not a wise convention, but I am not convinced which of the alternatives is the best choice That is a somewhat elliptical way of saying that the treatment of the variable inclusion probability as an estimable hyperparameter by Ley and Steel(2009) seems very much on the right track, though I prefer the convention that the inclusion probability is one.

Leamer(2014) is explicitly proposing a Bayesian context-minimal method to be used in diverse settings, at least as a first pass. With less rhetorical commitment, FLS and SDM and LS are also offering Bayesian alternatives to all-inclusive regression to be used in diverse settings. FLS, SDM, LS and L are thus all claiming (1) that the priors that are the foundations of their analyses more accurately capture “your” prior than the diffuse prior underlying unrestricted OLS estimates, and also (2) that explicit priors allow the design of variable inclusion methods that are superior to the ad hoc specification searches which are the main alternative, including stepwise (aka unwise) regression.

Though all four papers use Bayesian approaches to studying the determinants of economic growth in a cross section of countries, they make conflicting conclusions. FLS(2001, abstract) conclude, "In contrast with Levine and Renelt(1992), our results broadly support the more "optimistic" conclusion of Sala-i-Martin (1997b), namely that some variables are *important* regressors for explaining cross-country growth patterns." SDM(2004, p. 833) similarly conclude "Our main results support Sala-i-Martin (1997a, b) rather than Levine and Renelt (1992): we find that a good number of economic variables have *robust* partial correlation with long-run growth." However, Leamer (2014, abstract) offers an opposite conclusion, "In contrast with the conclusion of Sala-i-Martin, Doppelhofer and Miller, I do not find many of these coefficient estimates to be *sturdy*, meaning their signs are ambiguous." A similar conclusion was reached by Ley and Steel(2009) who describe the dependence of their estimates of posterior variable-inclusion probabilities on how the inclusion probability (θ) and estimation shrinkage rate (g) are handled.

The three adjectives in italics (important, robust and sturdy) leave unclear exactly what the writers are referring to. The first step toward understanding these different conclusions is to be explicit about the two phases of a data analysis, (1) estimation and (2) sensitivity analysis, associated respectively with two kinds of errors that afflict the estimated regression coefficients: (1) statistical uncertainty as measured by t-values and p-values, and (2) model ambiguity revealed by a sensitivity analysis that perturbs features of the model (or prior) to determine if the estimates change very much, exemplified by Leamer's(2014) s-values.

Both model ambiguity and model uncertainty are traditionally explored with ad hoc specification searches in which regressions with different subsets of explanatory variables are computed, leading to, for example, Sala-I-Martin's (1997) "I just ran two million regressions.[2]" Theoretically, model uncertainty has been approached classically with a study of the statistical properties (e.g. mean squared errors) of biased estimators that result when, for example, statistically insignificant variables are omitted. No conventional approach has emerged from this literature probably because the stepwise estimators that are routinely used do not yield smaller mean squared than unconstrained OLS for all parameter values, and this classical approach has no way of trading worsened MSE for some parameter values for improved MSE for other values. Thus enters the conversation Bayesian language which expresses a willingness to trade worsened MSE at *a priori* unlikely parameter values for improved MSE at *a priori* likely parameter values. This doesn't make the trade-off problem disappear. It only provides a language which can, in principle, improve the conversation. We Bayesians have to make suggestions of prior distributions that our audience can live with, and explain why alternatives should be avoided. We need to do what we can to draw in an understandable way the borderline that separates the circumstances in which a method works well fromthe

[2] My extreme bounds analysis, Leamer(1978), tops that number by exploring analytically an infinity of regressions.

circumstances in which there is a better approach. Ultimately, the choice of method is not a matter of mathematics or Monte Carlo studies of properties of estimators. It's a matter of wisdom and understanding, maybe more than you and I possess.

The one thing that this literature agrees on is that the prior distribution is located at zero. There are many settings in economics in which the coefficients are expected to equal one or to sum to one, and in those cases a transformation of variables will change the prior mean to zero, though you have to convince your audience as well as yourself of the wisdom of this step.

While the location of the prior is pretty straightforward, the rest is not. I like to begin with the normal-normal case, a normal likelihood and a normal prior. This produces a familiar posterior mean which is a matrix weighted average of the unconstrained all-inclusive OLS estimator and the zero vector, with weights equal to the sample and prior precision matrices. Generally, if the parameters of a model can be estimated with tolerable accuracy, we need to worry only about how properly to describe the statistical uncertainty, but if the data evidence is very weak, then we need to study model ambiguity by perturbing the values of the parameters (or hyperparameters) to discover how much they matter. If we treat the full prior covariance matrix as a set of unknown parameters, we are forced into a model ambiguity study because the sample provides only one noisy measure of the k-dimensional coefficient vector, which provides hardly any information about the full k×k prior covariance matrix.

The problem of estimating the prior covariance matrix becomes more manageable if the dimensionality is reduced by restricting some of the parameters to take on pre-assigned values. This literature is based implicitly on the assumption that the prior covariance matrix is diagonal. This may not seem like a variable-selection setting, but an important theorem of Leamer and Chamberlain (1976) discussed below expresses the usual posterior mean of the regression coefficient vector when the prior covariance matrix is *diagonal* as a weighted average of the $2^k$ regressions found by including subsets of the k variables in all possible combinations. Very importantly, this result makes clear what is the informational basis all of these variable-inclusion algorithms – *a priori* independence of the parameters. If you are not comfortable with this assumption for studying your data set, at least as an approximation, don't try omitting variables in an accidentally chosen coordinate system. That will make your inferences accidental too. (Parenthetically, this diagonality assumption is very much in doubt with the growth regression studied here because, among other issues, the setting has multiple measurements of the same basic concepts. Rather than seeking the one measurement that does the best job by omitting variables, it may be better to seek the linear combination that works best for each basic concept.)

The commitment to a coordinate system in which to omit variables is absolute in FLS, LS and SDM, but Leamer's sensitivity analysis relaxes that commitment by allowing non-diagonal prior covariance matrices.

The next step is to think about the relative sizes of the k prior variances. Is one coefficient more likely to be close to zero than another? The answer to this question depends on the scales used to measure the variables. Leamer(2014) proposes using explanatory variables that are normalized to have unit variance, in which case the coefficients become standardized beta-coefficients. This, he suggests, supports the use of a prior covariance matrix proportional to the identity matrix, which Leamer(2014) uses to trace out a one-dimensional curve of posterior means (e.g. the ridge trace) and to define the range of perturbation of the prior covariance matrix in his proposed model ambiguity analysis. Estimates when the prior covariance matrix is proportional to the identity matrix can be expressed as weighted averages of principal component regressions, sometimes used to reduce the dimensionality of a regression model as an alternative to omitting variables. In other words, the Bayesian foundation for all subsets regression is a diagonal prior covariance matrix, while the Bayesian foundation for principal compenents regression is a prior covariance matrix proportional to the identity matrix. This Bayesian language thus turns something accidental like the choice of coordinate system and the choice of scales into something that is understandable based on the nature of the prior information.

Generally, if the prior covariance matrix is assumed to be proportional to a known matrix, then there are k observations (estimated coefficients) that could be used to estimate the proportionality factor, which would be enough observations to treat this as a statistical uncertainty problem not a model ambiguity problem if k is large. An example is the proposed estimation of the g-scalar by Ley and Steel(2012) and by Liang et. al.(2008).

Zero and infinity are two numbers that stay the same as the scale changes. Thus a way to obtain scale-invariant results is to assume that the diagonal elements of the prior covariance matrix can take on only one of these two values, corresponding respectively with including and excluding the variable. This is the SDM approach. A g-prior produces a Bayes estimate which is a weighted average of the unconstrained OLS estimate and the zero vector, both of which are invariant to scale changes, as is their average provided that g is selected in a way that is scale invariant. The mixtures of g-priors of Ley and Steel(2012) for the same reason produce scale-invariant estimates. I was surprised to discover that these mixtures of g-priors can be written in the usual matrix weighted form with the prior precision matrix equal to the sum of two matrices, one proportional to the sample precision matrix and the other a diagonal matrix.

Lurking in the background of these Bayesian averages of the all-subsets regressions is a variable-inclusion model that determines if the prior variance is zero or infinity. If without thinking much you decide to treat every model as if it were equally probable, that's equivalent to having a variable inclusion probability $\theta$ equal to ½, which might seem "neutral" or "uninformed" but is actually extremely prejudicial against both small and large models.

A prior with mass equal to $(1- \theta)$ at zero and with the remaining mass $\theta$ distributed with mean zero and variance $\gamma^2$ has a variance equal to $\theta\ \gamma^2$. This prior variance has to be expected to be reasonably small if regression coefficients in large models are to be estimated accurately. It's the combination of large models (large θ) and diffuse priors (large $\gamma^2$) that causes the problem. My approach has one model with all variables included, θ=1, but uses a proper prior distribution with a prior variance that diminishes as the model size increases. SDM in contrast use a diffuse prior for the coefficients but consider only values of θ less than ½.

Thus the first critical step is choice of coordinate system in which to constrain coefficients to equal zero. The next critical step is to decide which world we live in: lots of variables with small effects or a few variables with large effects. LS sidestep this second critical question by their proposal to "let the data speak" by treating θ and g as estimable parameters, but it seems best to accept the fact that the data voice is very soft with regard to the choice of prior covariance matrix, and what we think we hear the data saying has to be very much influenced by the amplifier of assumptions that we have placed beside the data. That reality leads to model ambiguity worries.

Model ambiguity formally enters the Bayesian analysis with the dual admission that priors matter and that priors are impossible to define precisely. All four papers discussed here use a Bayesian approach, explicitly allowing the prior to matter and to affect measures of statistical uncertainty. FLS do not perturb their recommended prior distribution, and thus do not address the model ambiguity problem. SDM do perturb their recommended prior distribution, in a limited one-dimensional way discussed below, and they do not find much model ambiguity, while LS conduct a related but two-dimensional sensitivity analysis, and uncover an unsettling amount of sensitivity. Thus when SDM refer to "important" variables and when FLS refer to support for SDM, this is mostly a reference to a prior-dependent measure of statistical uncertainty while the L quotation is all about model ambiguity, as are model sensitivity conclusions in LS and also the extreme bounds analysis used by Levine and Renelt.

In summary, the limited information that comes from a large-k small-n regression needs to be wisely allocated among the regression coefficients since standard errors with unconstrained OLS are often uncomfortably large. Our competitors (shamelessly) offer a push-button approach: with a push of a button you can estimate the OLS and with a few more clicks you concentrate the information on a subset of coefficients by dropping some variables here and there. To compete against this standard practice, we need to automate as much as possible, and thus keep the number of clicks on the keyboard to a competitive few. Fernando Ley, his coauthors Mark Steel and Carmen Fernandez, Adrian Raftery, myself and others are all working to achieve the goal of a semi-automated analysis that should work well in a broad set of circumstances. My basic contribution has been to distinguish clearly statistical uncertainty from model ambiguity, and to suggest a kind of interval arithmetic that produces an interval of equally good estimates (or t-values). When the interval is too wide to be useful, the data should be said to leave us confused. In other words, my contribution to econometrics is confusion. Confusion appears in the analysis of FLS, SDM and LS when they treat the variable inclusion probability and the regression coefficient shrinkage rates as ambiguous parameters.

We are not there yet, but we are making progress.

These ideas are discussed in more detail in the sections to follow. The notation of the normal linear regression model with a normal prior distribution is presented in Section 2. The reader is reminded in Section 2 that a Bayes estimator with a diagonal prior covariance matrix is also a weighted average of the $2^k$ regressions. Section 3 reviews the logic underlying the s-values presented more fully in my companion paper Leamer(2014). Section 4 explores the Bayesian "hypothesis-testing" framework used by FLS and SDM, sometimes called "all-subsets" regression, in which the overall estimate is a weighted average of the $2^k$ alternative estimates with different subsets of the k variables included in the model. I offer an explanation in Section 5 why I like my weights better. Section 6 compares and contrasts my inferences from the cross-country growth data with the inferences suggested by SDM(2004). And Section 8 offers some concluding remarks of a promotional nature.

## 2 The Statistical Assumptions

The data are assumed to consist of an $(n\times 1)$ "dependent" variable vector **y** and an $(n\times k)$ "explanatory" variable matrix **X.** It is assumed that the vector **y** conditional on **X** is normally distributed with mean $\mathbf{X}\beta$ and covariance matrix $\sigma^2\mathbf{I}$, were $\beta$ is a $(k\times 1)$ vector of unknown "regression coefficients" that link **y** with **X**, and where $\sigma^2$ is a scalar equal to the variance of each element of **y** given $\mathbf{X}\beta$. The ordinary least squares estimate of $\beta$ is then

$$\mathbf{b} = (\mathbf{X'X})^{-1}\,\mathbf{X'y} = \mathbf{N}^{-1}\,\mathbf{r}\,, \qquad \text{where } \mathbf{N}= \mathbf{X'X}, \text{ and } \mathbf{r}= \mathbf{X'y}$$

The corresponding sampling variance and precision matrices are

$$\text{Variance}(\mathbf{b}) = \sigma^2(\boldsymbol{X'X})^{-1} = \boldsymbol{H}^{-1} \qquad \text{Precision}(\mathbf{b}) = \mathbf{H} = \mathbf{(X'X)}/\,\sigma^2$$

Prior to the observation of ***y*** and ***X*** it is assumed that the analyst and possibly the analyst's audience have the opinion that the vector of coefficients $\beta$ is probably close to zero. This vague statement allows for doubtful variables that probably have small coefficients and also allows for similar variables with coefficients that are probably about the same.

This state of mind is approximated with a normal prior distribution for $\beta$ with mean vector **0** and variance matrix $\boldsymbol{V_1}$. Then the posterior mean is the familiar matrix weighted average of the OLS estimate ***b*** and prior mean ***0***, with weights proportional to the sample and prior precision matrices:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{V}) = \left(\boldsymbol{H} + \boldsymbol{V}_{\mathbf{1}}^{-\mathbf{1}}\right)^{-\mathbf{1}}\boldsymbol{Hb}. \tag{1}$$

In much of this discussion, I duck the mathematical complications created by an uncertain residual variance $\sigma^2$, and replace it with the OLS unbiased estimator. This ignores the effect that the prior on the coefficients might have on the estimate of $\sigma^2$, larger because the error sum of squares becomes larger as the coefficients are shrunk toward zero, but smaller because the degrees of freedom adjustment becomes less, as it would if one dropped one or more variables. I am pretty confident that the conclusions drawn from a data set using the methods I propose would be only slightly different if a completely appropriate treatment of uncertain $\sigma^2$ were used.[3]

# 3 Three Special Cases

The literature that is being discussed here has implicitly been discussing three different prior precision matrices: diagonal, proportional to the identity and proportional to **H** (the g-prior). Theorems described below link a diagonal prior precision to all-subsets regressions and an identity prior precision matrix to principal component regression.

## 3.1 Diagonal Prior Precision Matrix and the $2^k$ regressions

SDM, FLS and LS propose solutions to the regression estimation problem that are weighted averages of the $2^k$ estimates formed when subsets of the variables are excluded. This creates computational difficulties since for k =67 in the data they study the number of alternative models is 1.48E+20. That is 148 quintillion, if you are interested.[4] This seems like an entirely different approach from the matrix weighted average (1) of OLS and the vector zero. However, a result of Leamer and Chamberlain (1976) expresses the Bayes estimate with a *diagonal* prior precision matrix as a weighted average of the $2^k$ regressions with weights reported in Theorem 1 below. This supports a rather different way of thinking about the $2^k$ regressions and also allows computation of weighted averages of the $2^k$ regressions without doing a quintillion calculations.

**Theorem 1**: The matrix weighted average $\hat{\boldsymbol{\beta}} = (\boldsymbol{H} + \boldsymbol{D_1})^{-1}\boldsymbol{Hb}$, with $\mathbf{D}_1$=diag($d_1$, $d_{2, \ldots,}$ $d_k$) a nonnegative diagonal matrix, can be written as a weighted average of the $2^k$ estimates found by omitting variables in all combinations:

[3] A different approach to estimating $\sigma^2$ adopted by FLS is constrain the prior variance to be proportional to the residual variance $\sigma^2$ in which case $\sigma^2$ drops out of the expression (1). The conditioning on the residual variance $\sigma^2$ has the peculiar feature that you do not know how sure you are about the probable sizes of the coefficients until you see the data, because the data informs you about the value of $\sigma^2$. The priors I have proposed do not have this conditioning on $\sigma^2$, but on this matter, a difficult choice must be made. One can go the FLS route which is a mathematically correct use of a peculiar prior, or one can remove the conditioning of the prior on $\sigma^2$ and use an approximation to the posterior distribution that I have used with $\sigma^2$ set equal to the unbiased estimate from the full model, or one can remove the conditioning of the prior on $\sigma^2$ and do the numerical integration that is necessary to avoid this approximation. The problem with the third route is that it is difficult for estimation but nearly impossible for sensitivity analysis.

[4] SDM approximate this weighted average by sampling from the set of models and they report that a sample size of 89 million models is enough to approximate the posterior distribution with adequate accuracy.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{H} + \boldsymbol{D_1})^{-1}\boldsymbol{Hb} = \sum_I w_I \boldsymbol{b}_I$$

$$w_I = \left(\prod_{i \in I} d_i\right) |\boldsymbol{H}_{JJ}| / |\boldsymbol{H} + \boldsymbol{D_1}| \qquad (2)$$

$$\sum_I w_I = 1$$

where *I* is a subset of the first k integers selecting the omitted variables, *J* is the complementary set selecting variables to be included, $\boldsymbol{b}_I$ is the constrained OLS estimate with $\hat{\beta}_i = 0$ for $i \in I$, and $\boldsymbol{H}_{JJ}$ is the sample precision matrix of the *included* variables conditional on the exclusion restrictions, with the understanding that when *I* is the full set of integers from 1 to k the determinant of the null matrix is one: 1≡det($\boldsymbol{H}_J$) and when *I* is the empty set the product of the diagonals is set to one, $\left(\prod_{i \notin I} d_i\right) = 1.$[5]

## 3.2 Identity Prior Precision Matrix and principal component regressions

Another result in Leamer and Chamberlain(1976) expresses the posterior mean when the prior precision matrix is proportional to the identity matrix as a weighted average of the k+1 principal component regressions.

**Theorem 2**: The matrix weighted average $\widehat{\boldsymbol{\beta}} = (\boldsymbol{H} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Hb}$ can be written as a weighted average of the $2^k$ estimates found by omitting variables in all combinations:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{H} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Hb} = \sum_{\boldsymbol{j=0}}^{\boldsymbol{k}} w_{\boldsymbol{j}}\, \boldsymbol{c_j}, \qquad \sum_{\boldsymbol{j=0}}^{\boldsymbol{k}} w_{\boldsymbol{j}} = \boldsymbol{1}$$

where $\mathbf{c}_j$ is the jth principal component point formed by “dropping” from the equation the j principal components of **H** with the smallest roots.

[5] The determinant of a precision matrix can be called the generalized precision. The weight on one of the $2^k$ regressions per this theorem is proportional to a kind of generalized precision of the coefficients equal to the product of the generalized *prior* precision of the *excluded* variables times the generalized *sample* precision of the coefficients of the *included* variables conditional on the exclusions, thus relying on the data for the included variables and the prior for the excluded variables.

## 3.3 Mixtures of g-priors

A g-prior proposed by Zellner (1986) is like a previous data set with the same covariance structure as the current data but with zero estimates for all the coefficients. With the prior precision matrix equal to ***gH***, the posterior mean is a weighted average of only two of the $2^k$ estimates: the unconstrained OLS estimate ***b*** and the vector zero: $\widehat{\boldsymbol{\beta}}(\boldsymbol{V}) = (\boldsymbol{H} + \boldsymbol{gH})^{-1}\boldsymbol{Hb} = \boldsymbol{b}/(\mathbf{1} + \boldsymbol{g})$. If g=1/n the prior has a covariance structure that is the same as one sample observation, and the weight on the OLS estimate b becomes n/(1+n). I cannot think why this would be an appropriate prior distribution, except: it's mathematically convenient. Sometimes mathematical convenience alone is not a compelling argument. Certainly no economist that I know would compare the unconstrained OLS with the zero vector only. This g-prior sits unused on the theorist's shelf.

However, with *mixtures* of g-priors proposed by Fernandez, Ley and Steel(2001b), Ley, Steel and Fernandez(2012) and Liang et.al. (2008) all $2^k$ estimates have non-zero weights. For that reason, these mysterious methods of estimation have some hope of being deployed but the designers need to come clean about the prior distribution that can justify the approach. In that regard it is troubling that one cannot see in the matrix weighted average (1) any logic for an estimate for model *I* to use the corresponding g-prior: $\widehat{\boldsymbol{\beta}}_I(\boldsymbol{V}) = (\boldsymbol{H_I} + \boldsymbol{gH_I})^{-1}\boldsymbol{H_I b_I}$ where $\boldsymbol{H_I}$ refers to the appropriate submatrix of ***H*.** This makes the prior variance for a particular coefficient dependent on which other variables are included. But this property isn't so unusual. A sample variance is influenced in the same way by variable exclusions, just as would any prior covariance be adjusted if information were obtained that a coefficient is exactly zero. This suggests that the mixtures of g-priors have added to the sample information *two* different pieces of information: (1) a previous sample with the same covariance structure but with zero estimates for all the coefficients and (2) some additional information that each of the coefficients is small. This can be made transparent by writing the matrix weighted average that produces the weighted g-estimator[6]:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{V})(1+g) = (\boldsymbol{H} + g\boldsymbol{H} + \boldsymbol{D_1})^{-1}\boldsymbol{Hb}(1+g). \qquad (3)$$

As suggested, the prior precision matrix in this expression is composed of two pieces: one piece of evidence that the coefficients are small with precision matrix $\boldsymbol{gH}$ and another independent piece of evidence that the coefficients are small with diagonal prior precision matrix $\boldsymbol{D}_1$.[7] This seems strange to me. Can anyone help me understand this?

[6] Replace ***H*** in (2) with $\boldsymbol{H} + \boldsymbol{gH}$ to obtain the mixture of g-estimates.

[7] Note that (3) is describing the estimation of **β** (1+g) which has to be divided by (1+g) to get the estimate of **β,** in other words shrunk toward zero. That is the mathematical symptom of the fact that the g-prior estimate for each of the $2^k$ models is a weighted average of the zero vector and

# 4 Model Ambiguity: S-values

The prior precision matrix in the formula for the posterior mean (1) can be treated as uncertain or as ambiguous. By the word “uncertain” I refer to a setting in which the prior covariance matrix is treated as a parameter to be estimated, which requires a compelling prior distribution for this hyperparameter matrix. By the word “ambiguous” I refer to a setting in which evidence and prior do not determine a compelling estimate and a sensitivity analysis is required to determine if ambiguity in the prior covariance matrix really matters.

The sensitivity analysis proposed by Leamer(2014) that is summarized by the s-values is based on the assumption that, after standardizing the variables to have unit variances, the prior covariance matrix of the coefficients is not known exactly but can be bounded from above and below by matrices proportional to the identity matrix

$$v_L^2 \boldsymbol{I} \leq Var(\boldsymbol{\beta}) \leq v_U^2 \boldsymbol{I}$$

where $\boldsymbol{I}$ is the k×k identity matrix and $v^2$ is the variance that applies to all coefficients. The identity matrix is the only logical choice when the context is completely unknown and the variables are standardized. The interval of prior covariance matrices in the previous expression includes non-diagonal matrices to allow for the possibility that if the context were fully understood not all coefficients would have the same prior variances and some covariances would be nonzero.

With data coming from the regression process $y_t = \boldsymbol{x}_t{}'\boldsymbol{\beta} + \varepsilon_t$ the conditional variance of the dependent variable is $\sigma_y^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} + \sigma_\varepsilon^2$ where $\boldsymbol{\Sigma}_{xx}$ refers to the covariance matrix of the explanatory variables. With standardized variables $\sigma_y^2 = 1$, the residual variance $\sigma_\varepsilon^2$ is equal to the fraction unexplained 1-$R^2$, and from this we obtain the result that the fraction explained, $R^2$, is the generalized beta-coefficient, $R^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}$.

To help think about choices for the upper and lower bounds for the prior covariance matrix, the quadratic form $\boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}$ can be written as $trace(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}) = trace(\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}\boldsymbol{\beta}')$, and we can replace matrix $\boldsymbol{\beta}\boldsymbol{\beta}'$ with it's prior expectation $E(\boldsymbol{\beta}\boldsymbol{\beta}') = Var(\boldsymbol{\beta})+ E(\boldsymbol{\beta})E(\boldsymbol{\beta}') = Var(\boldsymbol{\beta}) = v^2\boldsymbol{I},$ to obtain, using standardized $\boldsymbol{X}$, $E(R^2) = trace(\, v^2\boldsymbol{\Sigma}_{xx}) = v^2k$. Thus we have the prior variance of each beta-coefficient equal to the expected $R^2$ divided by the number of parameters, k, $v^2 = E(R^2)/k$ .

Finally, to select a range of prior variances, we select a range of expected $R^2$.

the constrained OLS estimate, which means that all of the $2^k$ models contribute g/(1+g) of their weight to the zero vector, which is thus favored when g is large.

$$v_L^2 = \frac{\min \mathrm{E}(R^2)}{k} \leq v^2 \leq \frac{\max \mathrm{E}(R^2)}{k} = v_U^2$$

I have proposed taking the maximum expected $R^2$ to be 1.0 and the minimum of 0.1. In addition, I proposed splitting this wide interval into two sub-intervals, pessimistic from 0.1 to 0.5 and optimistic from 0.5 to 1.0. I have also proposed a treatment of favorite variables with larger prior variances on the favorites and smaller prior variance on the others, thus allowing the favorite coefficients more room to roam from zero while keeping the others closer to zero. The widest interval from 0.1 to 1.0 is proposed as a nearly context-free analysis, while the others are steps in the direction of letting the context matter.

# 5 Statistical Uncertainty: Bayesian Weights for $2^k$ Models

To turn this from a problem of model ambiguity into a problem of estimation uncertainty, we would need a compelling prior distribution for the diagonal prior precision matrix $\mathbf{D}_1$. One possibility that appeals to me is to assume that the inverses of these diagonal elements are drawn from a gamma distribution which means that the (marginal) prior distribution for the coefficients is a product of independent Student distributions.[8] A gamma distribution tends to yield either small or large variances and might be approximated with a two-valued distribution for the prior variances: zero and infinity. This gives rise to the $2^k$ models, each with a different subset of included variables.

The posterior probability of one of these $2^k$ models is proportional to the prior probability times the marginal likelihood formed from the predictive density integrated with respect to the prior distribution for the coefficients. The word "marginal" refers to a weighted likelihood with the prior density as weights. The likelihood function of the unrestricted model may be great at the maximum, but it is miniscule in the tails, and consequently the prior weights can have a huge effect on the marginal likelihood depending on how much prior weight is put in the tails of the likelihood function. SDM and FLS offer two different solutions for the choice of the weights. FLS propose a particular proper prior for the coefficients (g-prior) while SDM use an improper limiting spike and slab form. Taking the limit is a very delicate mathematical operation for familiar reasons discussed in an appendix.

## 5.1 Hyperparameter: variable inclusion probability θ

It is not just a prior for the regression coefficients that is needed. One also needs to describe how the models are formed. Many Bayesian treatments of model selection implicitly or

[8] Carvalho, Polson and Scott (2010) propose a half-Cauchy distribution with scale parameter τ for each of the prior standard errors, mixed with another half-Cauchy for τ with scale parameter σ, the standard error of the signal. They show that "the horseshoe estimator corresponds quite closely to the answers obtained by Bayesian model averaging under a point-mass mixture prior." In a sense, this is numerical confirmation of Theorem 1.

explicitly take the probabilities of each of the $2^k$ models to be equal, as if a random process selected variables to include with probability ½. Frankly, this is not likely to capture the state of prior opinions in many circumstances. It seems better to treat the inclusion probability θ as a parameter as suggested by Mitchell and Beuchamp(1988). Then one can either carry out a sensitivity analysis as in Sala-i-Martin, Doppelhofer and Miller (2004) or an estimation exercise as in Ley and Steel(2009).

A model with p included variables and k-p excluded variables has prior probability $(\theta)^p(1-\theta)^{k-p}$ . These prior probabilities favor models with close to $p=\theta k$ included variables. The posterior weight on one of the $2^k$ models is this prior weight times a sample dependent measure of the quality of the fit. This leads to the SDM criterion: $\mathrm{SDM}=\mathrm{n}^{-p/2}\left(ESS_p\right)^{-n/2}(\theta)^p(1-\theta)^{k-p}$ . SDM do not take θ as known, but instead perform a sensitivity analysis by sweeping out a set of results by letting θ vary from 5/67 to 28/67 in a study of the determinants of economic growth with k=67 explanatory variables.[9]

LS study the SDM case when the inclusion probability θ is taken as known, and also what I think is the more relevant case of θ drawn from a (two-parameter) beta distribution with mean equal to a/(a+b). The second case is called by LS the “random” θ model which first suggested to me that the inclusion probability θ was not the same for all $2^k$ models, when in fact there is only one value of θ.[10]

Since SDM use a diffuse prior for the regression coefficients in their $2^k$ models, the larger models will inevitably have large standard errors and small t-statistics, which won’t matter if the prior probability of the large models is sufficiently small. Indeed, the reason these authors see so much estimation accuracy in their estimates is because they have chosen a prior that strongly favors small models, as explained below. If, however, one uses the g-prior as done by Ley and Steel (2009), estimates with small posterior standard errors for the large models can be

[9] This is analogous to my sweeping out a set of estimates as the expected prior $R^2$ varies from 0.1 to 1.0. Both ways produce a one-dimensional set of estimates. In contrast, s-values come from a k-dimensional set of estimates of the k-dimensional coefficient vector. More on similarities and differences below.

[10] If one were observing binomial trials, “a” would be the number of successes, “b” the number of failures and a+b the number of trials. LS suggest setting “a” equal to one, and then the mean is 1/(1+b). This means that the prior is more concentrated if the mean is small, since the effective sample size 1+b needs to be great to make the mean small. In other words as b declines, two things are happening: smaller models are favored and the range of probable model sizes is shrinkage. For this reason, I am inclined to think that both a and b should be selected, perhaps with the mean a/(a+b) set to ½ and with a+b <2, meaning a U-shaped distribution favoring zero and one.

obtained if the g weight on the prior is large enough. The tradeoff between g and θ is studied Taplin and Raftery (1994, Section 5.2), by Ley and Steel(2009) and by Eicher, Papageorgiou and Raftery(2011). For estimation accuracy of the coefficients, one needs a prior that the coefficients are small: Small inclusion probability θ or included parameters with small coefficients. (large g)

The top panel of Figure 1 illustrates the counts of different models of each model size and the probabilities of these model sizes per SDM when k=67 and kbar=5, and the bottom panel illustrates the probability of drawing a specific model relative to the probability of a model with no variables included. The value in this lower panel at p=10 is 1.1E-11, meaning that the model with no variables is ten to the eleven more likely than one of the models with 10 variables. That seems like a hugely strong prior to me. It is this very strong prior that is necessary to shift the distribution of counts of models in the top panel into a probability distribution with mean at kbar/k = 5/67. Basically, since the inclusion probability is less than one-half, models with fewer number of variables are favored, strongly so relative to very large models.

Figure 2 provides the same information when the value of kbar=28, the maximum that is reported by SDM. Here the shift from the counts to the probabilities in the upper panel is less, and indeed there would be no shift if kbar=67/2. In that case, the analysis reverts to the FLS and Raftery(2005) approach with equal probability assigned to each of the $2^k$ models. Even that case is heavily prejudiced against the model with all the parameters included, since there is only one of these fully-loaded model but with k=67 there are 1.48E+20 other models. This means if each model had equal probability, the prior probability of the model with all 67 variables would be 1/1.48E+20 =6.78E-21. But, in addition, a kbar=5 (out of 67) favors the smaller models, and this reduces the prior probability of the model with all 67 variables included to 3.05E-76. A value of kbar=28 also favors the smaller models, but not so much. In that case the prior probability of the model with all 67 variables is 4.10E-26. Statistical insignificance is a special problem for the larger models, and this piror preference for the smaller ones surely elevates the resulting statistical significance of the estimates of the regression coefficients.

To settle on a context-free way of dealing with the problem of too little data to estimate accurately a large set of coefficients, you have to decide which world you live in. Do you live in a world with a few dominant effects and all the others so small they can be treated as zeroes? Then you want an analysis that sorts variables into two buckets: a small bucket of included variables and a large bucket of excluded variables. That's the SDM and Raftery world. Or do you live in a world of lots of small effects, with no meaningful way to separate the included from the excluded variables? Then you want to shrink all the coefficients toward zero but with no attempt to shrink some all the way to zero. That's the world in which the s-values apply. But even in this world, the shrunken estimate takes the weighted $2^k$ form. It's just different weights, to be discussed in the next section.

A question that might help to decide which world you live in is: As the sample precision gets larger and larger, is there some point at which you would decide just to do unconstrained OLS with all the variables included, or do you want to look at models with subsets of variables even then? If you are looking at regressions with omitted variables only to overcome deficiencies in the data evidence, then you are in my world. If you really think that some of the coefficients could be exactly zero, you should be exploring that possibility no matter how informative the data may be. Then you are in the SDM/Raftery world.

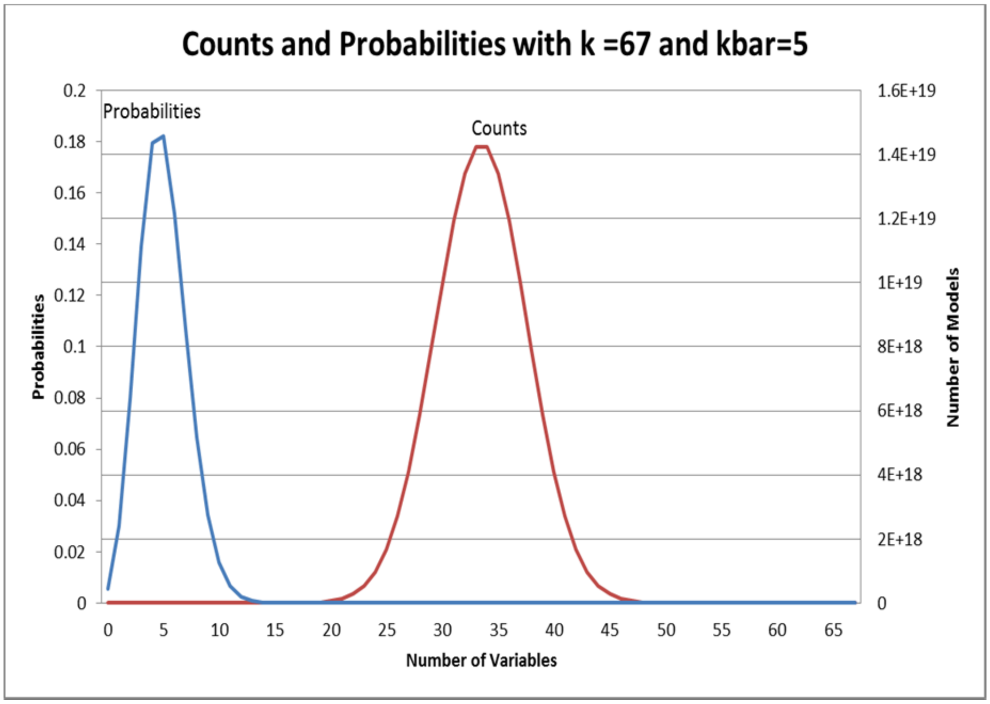
If I were going down the SDM/Raftery route, I might prefer the probabilities reported in the top panels in Figure 1 or Figure 2 to be uniform, meaning that there is just as good a chance of all 67 variables included as 33 variables included. To accomplish that it would take the probability of a specific model to be inversely proportional to the counts. In other words, the probability of 1/67 assigned to models with 33 coefficients has to be spread evenly over the 1.42E+19 models with 33 variables. This cannot be accomplished with a fixed value of $\theta$ but, as pointed out to me by Mark Steel this is what happens if $\theta$ is uniformly distributed. (a=1,b=1 in the notation of LS), so that's my choice and the choice also of Ley and Steel(2009).

## 5.2 FLS Model Probabilities

The FLS g-prior produces an estimate of the coefficient vector $\boldsymbol{\beta}_p$ for a model with p included variables that is a weighted average of only two vectors: the zero vector and the OLS vector with the p included variables. The weight on the zero vector is equal to g/(1+g). With the recommended value of 1/g =max($n$,$k^2$) the weight on the zero vector is thus 1/(1+ max($n$,$k^2$) ). This is replicated for all $2^k$-1 models[11] which thus augments the effective prior probability of the zero vector to $2^{-k}$ + (1-$2^{-k}$)/(1+ max($n$,$k^2$).

[11] The minus one refers to the model with no variables included

Figure 1 SDM Model Probabilities With k=67 and kbar=5



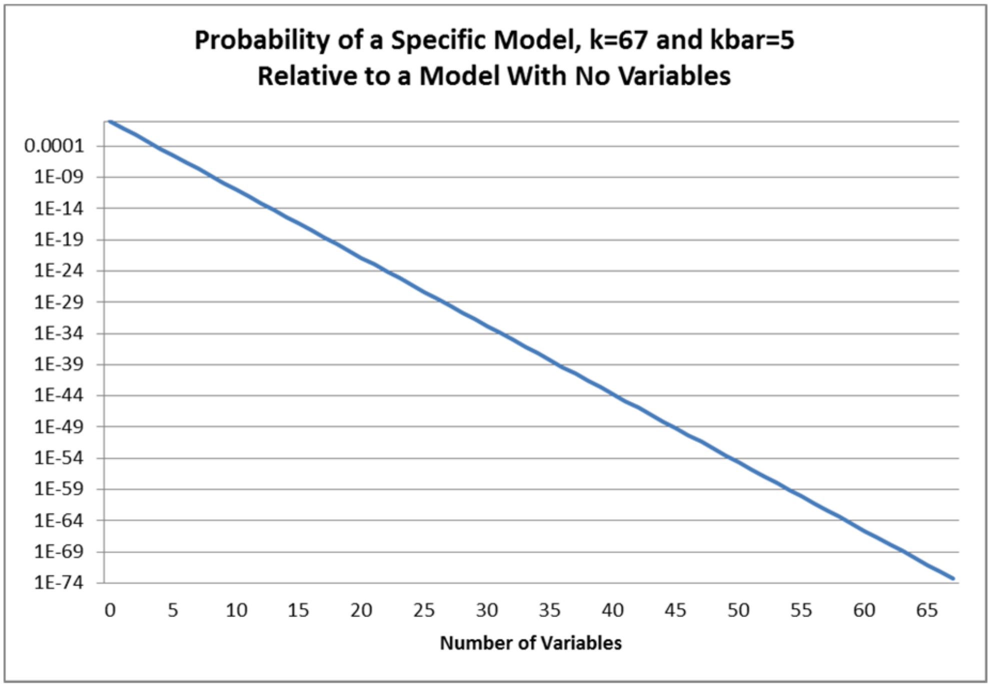
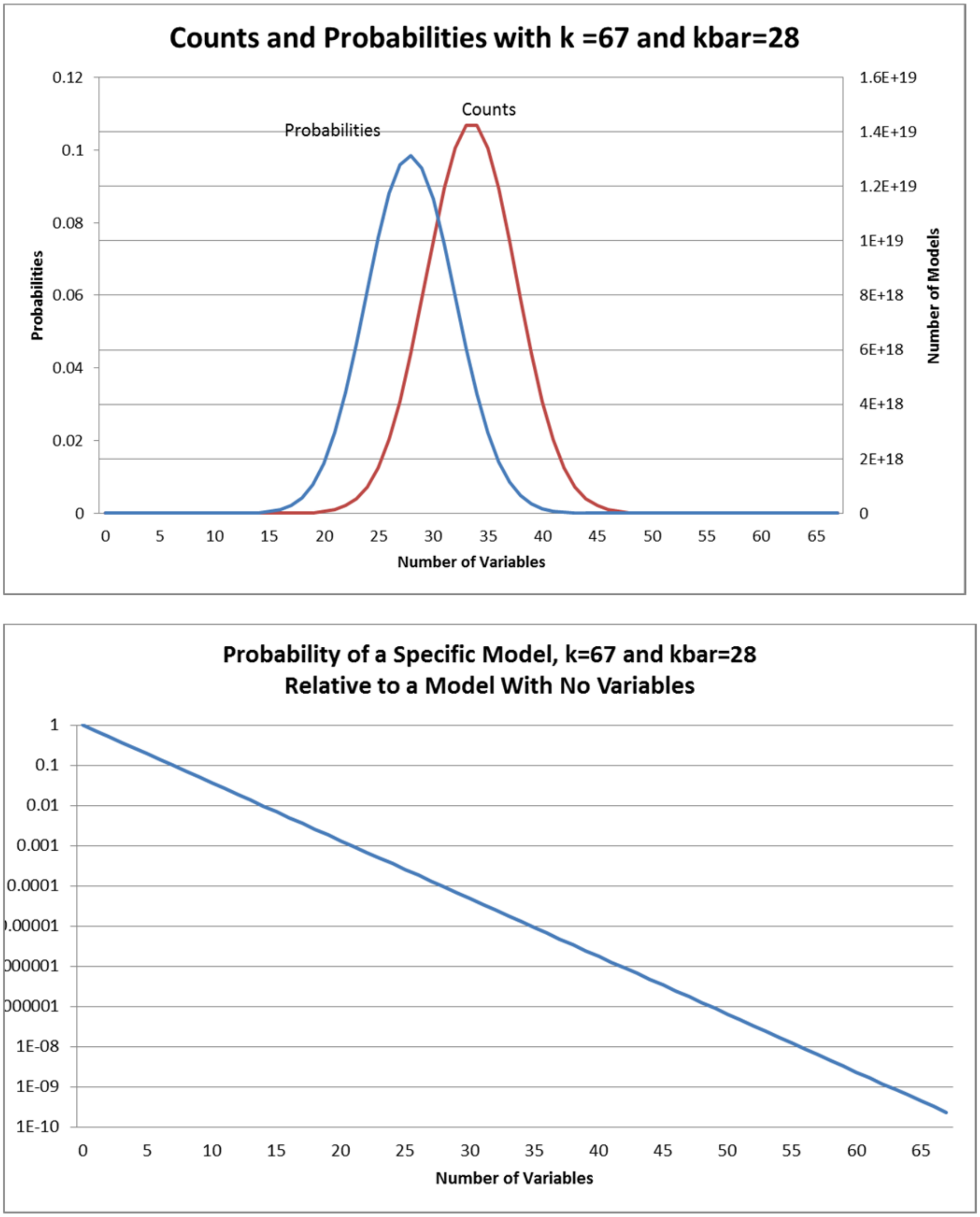


**Figure 2** **SDM Model Probabilities With k=67 and kbar=28**



## 6 Two different sets of weights for the $2^k$ regressions

While both SDM and Leamer have the $2^k$ regressions in either the foreground or the background, the approaches deploy very different weights, which have potentially big differences as the sample size grows. In Leamer(1978) I used the word “interpretative search” to refer to a setting in which an analyst is trying to improve the accuracy of the estimates by omitting doubtful variables, something that would be done only in small samples when the data evidence is too weak to allow accurate estimation of all the regression coefficients. And I used the word “model selection” search to refer to a setting in which the analyst entertains the possibility that some of the regression coefficients are exactly zero, something that merits testing even in a large sample. SDM and Raftery are recommending a model selection search while I am recommending an interpretative search.

In the notation of equation (1), $\boldsymbol{H}_{JJ} = \boldsymbol{X_p'X_p/\sigma^2}$ and $d_i = \upsilon_i^{-2}$. With this notation, the weight on a model per (2) is proportional to

$$w_I \propto \left(\prod_{i\in I} \upsilon_i^{-2}\right)\sigma^{-2p}|\boldsymbol{X_p'X_p}| \propto \left(\prod_{i\in J} \upsilon_i^{2}\right)\sigma^{-2p}|\boldsymbol{X_p'X_p}|$$

which uses the fact that $\left(\prod_{i\in I}\upsilon_i^2\right)\left(\prod_{i\in J}\upsilon_i^2\right) = \left(\prod \upsilon_i^2\right)$ is constant across models. Furthermore, if we assume as above that the prior variance is the same for all variables, we obtain the weights that apply to the “interpretive search” which has zero prior probability that a coefficient is exactly zero

$$w_I \propto \nu^{-2(k-p)}\sigma^{-2p}|\boldsymbol{X_p'X_p}| \propto \nu^{2p}n^p|\boldsymbol{X_p'X_p}/n\sigma^2| \qquad (3)$$

This contrasts with the limiting version of the weight for the “hypothesis testing search” in which there is positive probability of zero coefficients:

$$\lim_{\nu_p^2\to\infty} f_p(\boldsymbol{y}|\sigma^2) \propto \nu_p^{-p}n^{-p/2}|\boldsymbol{X}_p'\boldsymbol{X}_p/n\sigma^2|^{-\frac{1}{2}}\exp\left(-n\frac{ESS_p/n}{2\sigma^2}\right) \qquad (4)$$

These two sets of weights bear a puzzling inverse relationship with each other. A large prior variance favors the larger models in an interpretive search, because that means letting the data speak even if the data are weak. But a large prior variance favors the smaller models in an hypothesis testing search because the higher variance dictates an evaluation of models not where the maximum of the likelihood function occurs but instead where coefficients are extreme, which is where the likelihood values inevitably are small.

A large value of the conditional sample precision $|\boldsymbol{X}_p'\boldsymbol{X}_p/\sigma^2|$ is a reason to favor the model in an interpretative search because it means the data can overcome the doubt, but the same element works against the model in an hypothesis testing search apparently because it takes a better fit to keep a model afloat if it benefits from a rich sample of explanatory variables.

As sample size grows, if the explanatory variables are drawn from a stationary distribution the determinant $|\boldsymbol{X}_{\boldsymbol{p}}'\boldsymbol{X}_{\boldsymbol{p}}/n\sigma^2|$ will converge to some constant and the interpretative search weight (3) will be dominated by the $n^p$ term, which is greatest for the largest value of p. In other words, the choice of the unconstrained least squares estimate when the sample size is large is baked into the formula. Formula (4) doesn't have the same property. As sample size grows $\frac{ESS_p/n}{\sigma^2}$ converges to one for the true model and one for any other model that includes all the variables in the true model and some larger number for all other models. For the set of models that all have the same asymptotic ESS, the term that dominates the relative weights is $n^{-p/2}$ which picks the model with the smallest number of variables. (the true model). For the other models with larger asymptotic ESS, the log of the sample-size dependent part of (4) is $-\frac{p}{2}\log(n) - \frac{n}{2}\left(\frac{ESS_p/n}{\sigma^2}\right)$. Multiplying this by n/2 produces the expression$-p\,\frac{\log(n)}{n} - \left(\frac{ESS_p/n}{\sigma^2}\right)$. Since log(n)/n converges to zero as n converges to infinity, this expression is asymptotically determined by the measure of fit $\left(\frac{ESS_p/n}{\sigma^2}\right)$, which is best for the true model.

In simpler language, the model choice is consistent. Thus if my view that all the variables belong is right, the SDM/Raftery approach will converge to exactly the same estimate as mine – the unconstrained OLS estimate. For that reason, I do not think this discussion of what happens as sample size grows is really helpful in selecting between the two approaches. What does matter? Aside from the apparent inverse relationship between the two formulae, the thing that stands out is that the weights (4) depend on the quality of the fit of each of the $2^k$ models as measured by the error sum of squares, $ESS_p$, while that term is entirely absent from (3). This is a consequence of the fact that the estimator that underlies the weights in (3) is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{H} + \boldsymbol{D_1})^{-1}\boldsymbol{Hb}$, which is a *fixed* weighted average of the OLS ***b*** and the vector zero, while SDM allow the elements of ***D₁*** to be data dependent though taking on one of two extreme values: either zero (include the variable) or infinity(exclude it).[12]

In other words, SDM take an estimation approach and treat the diagonal elements of $\boldsymbol{D_1}$ as a set of parameters to be estimated, while I replace the diagonal $\boldsymbol{D_1}$ with a nondiagonal prior precision matrix $\boldsymbol{V_1^{-1}}$ and take a model ambiguity approach which shows how much the estimates change as the prior covariance matrix is varied within a plausible range. Model ambiguity and sensitivity analysis enters the SDM framework entirely through the choice of kbar: the expected number of included variables

[12] Incidentally, I am inclined to think that the 2k model structure would work better if the two choices for the prior variance were small and large, like 0.1/k and 0.5/k when standardized variables are used. More on this in another paper.

## 6.1 Two different kinds of sensitivity levers: Prior Covariance matrix and Prior Inclusion Probability

A sensitivity analysis designed to uncover model ambiguity problems needs to perturb a parameter or hyper-parameter about which the data are not very informative. The prior covariance matrix $\boldsymbol{V}_1$ is an obvious candidate since the vector of true coefficients is only one draw from this distribution and it impossible to estimate a covariance matrix with only a single observation. My s-values are derived from an interval of prior covariance matrices.

If the prior covariance matrix is the diagonal $\boldsymbol{D}_1^{-1}$ that still leaves just one observation to estimate each of the k prior variances. Raftery, FLS and SDM make estimation of $\boldsymbol{D}_1^{-1}$ feasible by assuming that each diagonal element can take on only one of two values – zero or "very large".

The parameter that is used for the SDM sensitivity analysis is $\bar{k}$, the expected number of included variables. By varying $\bar{k}$ SDM sweep out a one-dimensional curve of estimates, which is analogous to my one-dimensional curve of Bayes estimates as the prior expected $R^2$ is varied from zero to one.

While I find Sala-i-Martin et al (2004) sensitivity analysis intriguing, I have two comments contrasting their analysis with mine. First of all, the task of choosing an interval of values for $\bar{k}$ bewilders me, and I cannot think of a conventional range other than $0 < \bar{k}$ , while the task of choosing a minimum and maximum expected $R^2$ seems doable, and has led to my specific suggestion regarding conventional ranges of prior distributions. But I may only be fooling myself.

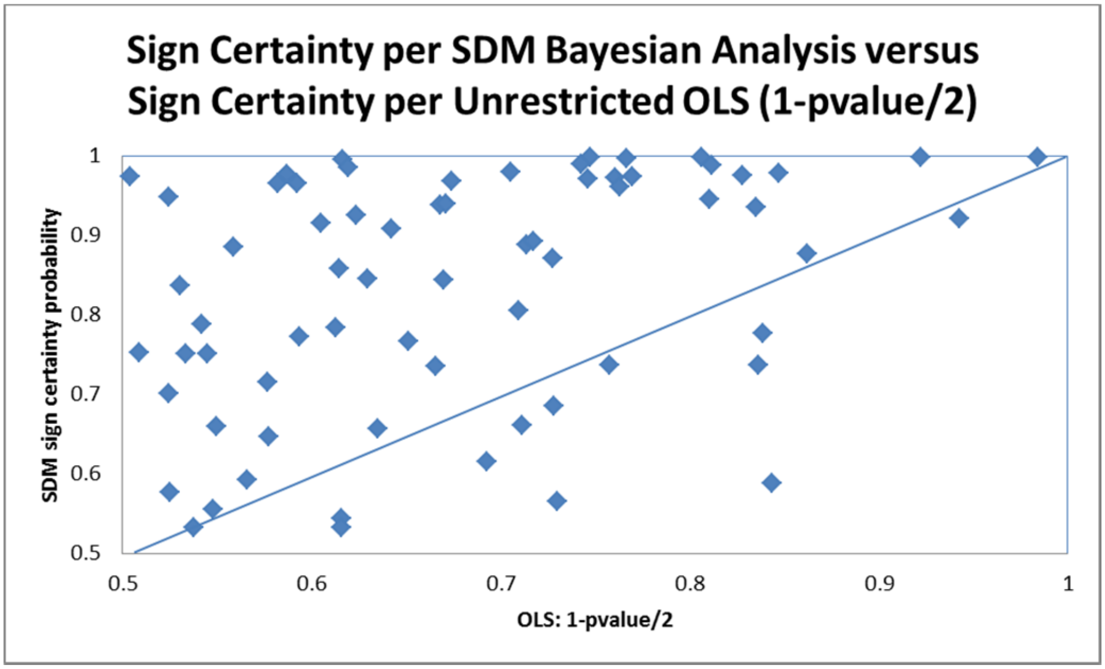
Second, my family of proposed prior distributions is multidimensional with an uncertain ambiguous prior covariance matrix. I am inclined to think that the model ambiguity has a distinctly multidimensional character. To express the point differently, I do not impose a coordinate system for omitting variables and I instead allow complex prior correlations among the coefficients consistent with a prior covariance matrix that is bounded from below by a matrix that is proportional to the identity matrix (with the variables scaled to have unit variance).

# 7 Leamer or SDM

Regardless of the words, in the end, these two puddings need to be tasted. I think you would find either pudding to have a better taste than the gruel currently provided by ad hoc treatments of this problem.

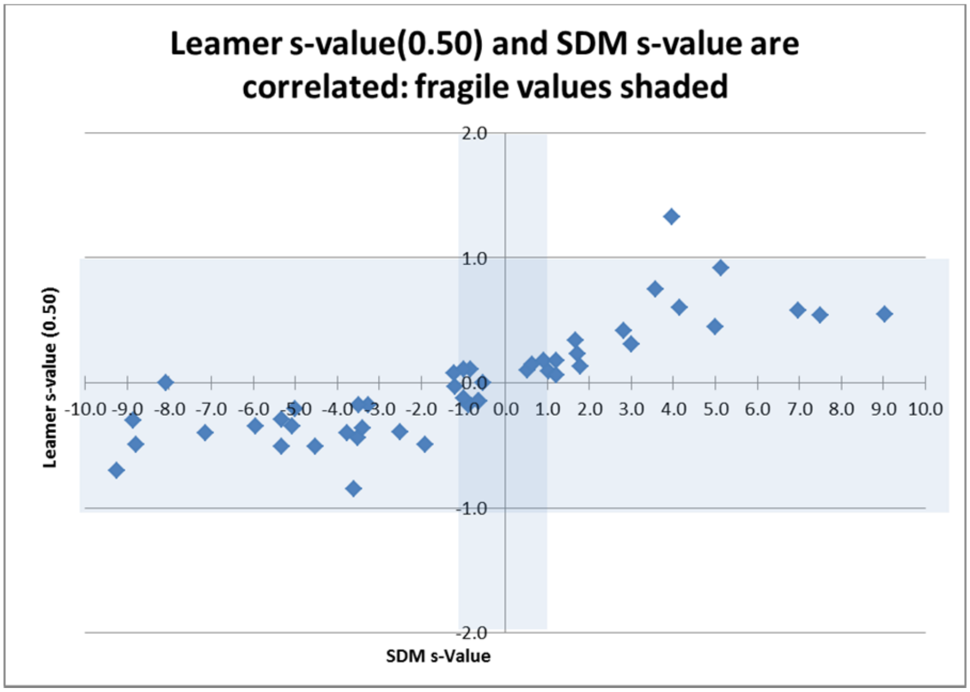
To understand SDM better, Table 1 contrasts the sampling uncertainty measured by OLS with all 67 variables with the sampling uncertainty recorded by SDM with kbar=7. The first column contains the sampling "probability" that the sign is the same as the OLS estimate equal to one minus the p-value divided by two. The column headed "BACE sign certainty" is the posterior probability that the coefficient has the same sign as the posterior mean (or zero), allowing for the all-subsets structure for the prior. SDM provide a lot of big sign certainty numbers compared with unrestricted OLS. This is a consequence of the fact that the sample evidence has been concentrated on a few regression coefficients by the assumption that there only about seven out of 67 variables that have non-zero coefficients as described in the previous section. Figure 3 compares the SDM sign certainty probabilities with the OLS sign certainty. The 45 degree line separates cases in which SDM is more certain than OLS. Most of the results are above the line.

Figure 3 OLS versus SDM sign certainty



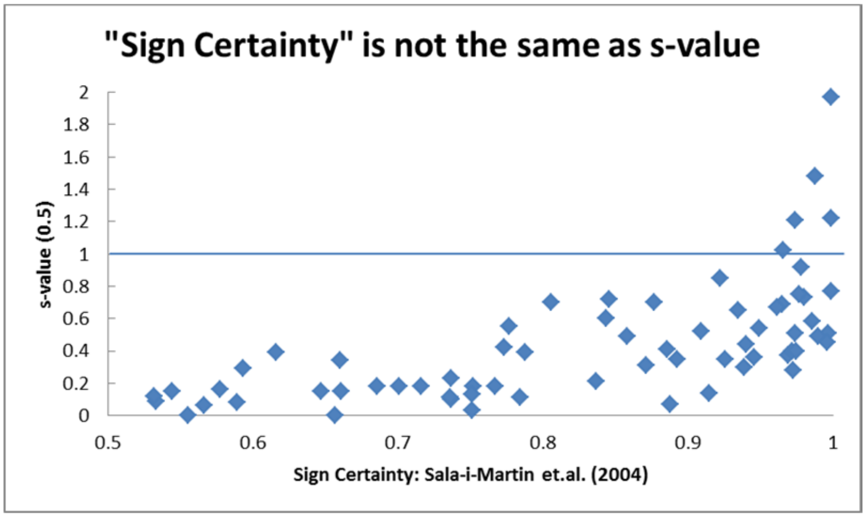
SDM do not make explicit reference to model ambiguity problems but they do trace out the estimates as kbar is varied from 5 to 28. The estimates that they report for kbar=5 and kbar=28 are reported in Table 2. These are used to compute the implied SDM s-values equal to the average of these two estimates divided by half the difference between the max and the min. The variables in this table are ordered by the absolute value of the SDM s-values from smallest to largest, and a line separates the few with s-values less than one from the many withs-values greater than zero. Only eight coefficients change sign between kbar=5 and kbar=28. The others are judged to be sturdy. The last column of Table 2 are my s-values at optimistic the 0.50 level. The few above one in absolute value are highlighted in the display, Figure 4 is a scatter diagram that compares the SDM s-values with the Leamer s-values, with the regions of s-values less than one shaded. Although there is an association between these s-values, the SDM assessment of the value of these data is much more favorable than mine.

**Figure 4 Leamer s-values versus SDM s-values**



The model ambiguity results reported by Sala-i-Martin et. al. involve varying the expected number of included variables and this doesn't produce estimates that vary much. So they argue that the sampling uncertainty is small, and the model ambiguity is very small. I have a different interpretation. Figure 5 makes the point by contrasting their sign certainty, many bigger than 90%, with my s-values, mostly less than one, signifying a fragile sign

**Figure 5 Scatter of Sala-i-Martin et al sign certanty and my s-values**



## 8 Concluding Comments

The most important point that I share with LS, Raftery, SDM and FLS and other contributors to this literature is that methodical soundness is an important goal, maybe an essential one. I believe that the only way an economist analyzing non-experimental data can be methodologically sound is to be explicit about the prior information that is being utilized. Priors play a role, one way or another, and one of the great advantages of a Bayesian approach is that the method is sound and the prior explicit. Explicit is not the same as understandable, and understandable is the real goal. Making it understandable is a critical step toward getting it right. This paper amounts to a discussion about the kinds of priors that are both understandable and that might be deployed in broad sets of circumstances.

I propose that the journey toward understanding begin with the posterior mean of the regression coefficient vector associated with a given prior covariance matrix $\mathbf{V}_1$ which takes the familiar matrix weighted form. With that as the starting point there are three items to work on: What aspects of $\mathbf{V}_1$ can be taken as known, what aspects can be estimated and what aspects require a sensitivity analysis because they are neither known nor estimable. Much of this literature implicitly makes the assumption that $\mathbf{V}_1$ is a diagonal matrix, thus imposing the restriction that the prior covariances are all zero. The message of Theorem 1 above is that this is tantamount to choosing a parameterization in which to omit variables. This still leaves the k diagonal elements either to estimate or to perturb. The parameterization can be further reduced by assuming a common variable inclusion probability θ and a common prior variance $v^2$ for regression coefficients of included variables. This makes the prior variances all equal to θ $v^2$ which makes the prior covariance matrix proportional to the identity, connected to principal component regression per Theorem 3. This makes the inferences scale-dependent, which is wisely avoided with the conventional standardization of the scales to make the sample variance all one. Raftery/SDM produce scale invariant estimates because of their assumption that the conditional-on-inclusion prior variance $v^2$ infinite, and LS use the g-prior to produce scale-invariant estimates. Leamer(2014) deals with this scale issue by using standardized variables with unit variances.

I understand the potential importance of higher order moments of the prior, but am inclined to think our audience is likely to be confused and even bewildered after choosing the first and second moments. But on that point, much of what is written in this paper may inappropriately suggest a sharp dividing line between priors that have atoms of mass at zero (FLS, LS and SDM and Raftery) and priors that have central tendencies at zero but no mass points (Leamer). All these approaches do their best to approximate prior opinions, and produce estimates that at least are methodologically sound in contrast to the ad hoc methods used by many. My opinions are probably best captured with a prior density with a steeper central tendency than a normal distribution and flatter tails, which means that the prior gravitational force pulling estimates of coefficients toward zero should be powerful when the estimates are small but weak when the estimates are large. Expressed in words, I think the coefficients are pretty small but if they are not, then I don't really have much of an opinion. These beliefs might be captured adequately with a prior that is mixture of normals, some with small variances and some with large. FLS, LS and SDM use a mixture, including one normal with a zero variance which is how the atom of mass at zero is created. I have argued that this approximation goes awry when the sample information becomes great, but for small samples it might work just fine, and might be better than what I am proposing, which is a normal prior, not a mixture of normals.

# 9 Appendix: Weighted Likelihoods with Diffuse Piors

Consider a model with p included variables denoted by the *nxp* matrix of explanatory variables **X**$_p$. Further assume that the p regression coefficients come from a normal distribution with variance proportional to the *pxp* identity matrix $\upsilon_p^2\boldsymbol{I}_p$. Then the unconditional (predictive) distribution of the vector ***y*** has mean vector ***0*** and covariance matrix $\boldsymbol{U} = \upsilon_p^2\boldsymbol{X}_p\boldsymbol{X}_p' + \sigma^2\boldsymbol{I}_n$:

$$f_p(\boldsymbol{y}|\sigma^2) = (2\pi)^{-n/2}|\boldsymbol{U}|^{-1/2}\exp(-\frac{1}{2}\boldsymbol{y}'\boldsymbol{U}^{-1}\boldsymbol{y}) \qquad (\boldsymbol{1})$$

$$\boldsymbol{U}^{-1} = \left(\nu_p^2\boldsymbol{X}_p\boldsymbol{X}_p' + \sigma^2\boldsymbol{I}_n\right)^{-1} = \sigma^{-2}\boldsymbol{I}_n - \sigma^{-2}\boldsymbol{X}_p\left(\sigma^{-2}\boldsymbol{X}_p'\boldsymbol{X}_p + \nu_p^{-2}\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}_p'\sigma^{-2}$$

$$|\boldsymbol{U}| = |\upsilon_p^2\boldsymbol{X}_p\boldsymbol{X}_p' + \sigma^2\boldsymbol{I}_n| = |\sigma^2\boldsymbol{I}_n|\left|\nu_p^{-2}\boldsymbol{I}_p + \sigma^{-2}\boldsymbol{X}_p'\boldsymbol{X}_p\right|/|\nu_p^{-2}\boldsymbol{I}_p| = \sigma^{2n}\nu_p^{2p}\left|\nu_p^{-2}\boldsymbol{I}_p + \sigma^{-2}\boldsymbol{X}_p'\boldsymbol{X}_p\right|$$

where $|\boldsymbol{U}|$ is the determinant of the matrix ***U***. This marginal likelihood is what multiplies the prior probability of each model to compute the posterior probability. Unfortunately, the marginal likelihood depends on that had-to-choose prior parameter $\upsilon_p^2$. A traditional way of getting rid of that dependence is to let the prior become infinitely diffuse by setting $\upsilon_p^2$ equal to infinity, which produces Bayes estimates of regression coefficients the same as the OLS estimates. Some Bayesians imagine that this represents the kind of “ignorance” that “should” be the start of a scientific investigation, since this apparently lets the data speak, unencumbered by the opinions of the analyst. But the opinion represented by this distribution “spike and slab” prior is actually quite informative. It asserts that a coefficient is certainly large in the sense that the probability that it lies in any finite interval is zero. One can get away with that preposterous statement if the data are generated by a normal linear regression process with a known set of variables because the tails of the likelihood function decline so sharply that the data are able to overcome completely this wild prior opinion. To express this in a more constructive way, in the usual regression setting with a known set of explanatory variables, it doesn’t much matter if you think that $\upsilon_p^2$ is large, very large or very very large or infinite. They all imply about the same inferences from the data in a normal linear regression setting: just use OLS. Unfortunately, this lack of sensitivity logic doesn’t apply to the model selection problem since a prior that has a mass point at zero cannot be otherwise diffuse for technical reasons discussed next, but rhetorically because it causes impossible dissonance between the perfect knowledge that a coefficient is zero and the perfect knowledge that the coefficient is larger in absolute value than any finite number.

Technically, we can see what happens when we try to make the prior diffuse by exploring the limit of the marginal likelihood as $\upsilon_p^2$ converges to infinity. The limit of the quadratic form is

$$\lim_{\nu_p^2\to\infty}\boldsymbol{y}'\,\boldsymbol{U}^{-1}\boldsymbol{y} = \sigma^{-2}\boldsymbol{y}'\left(\boldsymbol{I}_n - \boldsymbol{X}_p\left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\right)\boldsymbol{y} = \frac{ESS_p}{\sigma^2}$$

$$ESS = \boldsymbol{y}'\left(\boldsymbol{I}_n - \boldsymbol{X}_p\left(\boldsymbol{X}_p'\boldsymbol{X}_p\right)^{-1}\boldsymbol{X}_p'\right)\boldsymbol{y}$$

The limit of the determinant is

$$\lim_{\nu_p^2\to\infty} |\boldsymbol{U}|^{-\mathbf{1/2}} = \sigma^{-(n-p)}|\boldsymbol{X}_p'\boldsymbol{X}_p|^{-1/2}\nu_p^{-p}$$

Thus the limiting form of the marginal likelihood is

$$\lim_{\nu_p^2\to\infty} f_p(\boldsymbol{y}|\sigma^2) \propto \nu_p^{-p}\sigma^{-(n-p)}|\boldsymbol{X}_p'\boldsymbol{X}_p|^{-1/2}\exp(-\frac{ESS_p}{2\sigma^2})$$

To rid this expression of the $\sigma^{-(n-p)}$ term, it is common to use the conjugate prior assumption that the prior variance $\upsilon_p^2$ is proportional to the residual variance $\sigma^2$ ,as if the prior information came from some previous sample from the same process: $\upsilon_p^2 = \lambda_p^2\sigma^2$. Then $\nu_p^{-p}\sigma^{-(n-p)} = \lambda_p^{-p}\sigma^{-n}$. Since the term $\sigma^{-n}$ is common to all models, we can drop it from the expression, just as we dropped the other constants.

More critically, this limit has a serious problem which is discussed in Leamer(1978, 108-114): This marginal likelihood in the limit is dominated by the term $\nu_p^{-p}$ which converges to zero at a rate that increases with p the number of parameters. *This means that the models with the smallest number of parameters are infinitely better than any other models.* The only way to correct for this it to let the prior variance $\upsilon_p^2$ converge to infinity in a way that is dependent on p, $\lambda_p^p = \lambda_k^k$, where k is the number of parameters in the largest model. To express this differently, we need to keep constant the ratios of the products of the variances of models of different size. Thus using $\upsilon_p = \lambda_p\sigma = \lambda_k^{k/p}\sigma$ we find that $\nu_p^p\sigma^{(n-p)}$ =$\left(\lambda_k^{k/p}\sigma\right)^p \sigma^{(n-p)} = \lambda_k^k\sigma^n$ where $\lambda_k^k$ is the same number for every model, and can be suppressed in the formula. Then we get the limiting expression

$$\lim_{\nu_p^2\to\infty} f_p(\boldsymbol{y}|\sigma^2) \propto |\boldsymbol{X}_p'\boldsymbol{X}_p|^{-1/2}(\sigma^2)^{-n/2}\exp(-\frac{ESS_p}{2\sigma^2})$$

With $h = \sigma^2$ having the "uninformative" prior distribution $h^{-1}$ *dh,* we can marginalize this parameter from the likelihood using the fact that a gamma distribution integrates to one to obtain:

$$\int (h)^{n/2}\exp\left(-\frac{ESS_p}{2}h\right)h^{-1}\,d\,h \propto \left(ESS_p\right)^{-n/2}$$

This produces the limiting marginal likelihood reported by Leamer(1978, page 111)

$$\lim_{\nu_p^2\to\infty} f_p(\boldsymbol{y}) \propto |\boldsymbol{X}_p'\boldsymbol{X}_p|^{-1/2}\left(ESS_p\right)^{-n/2}$$

We can capture the sample size effect on $|\boldsymbol{X}_p'\boldsymbol{X}_p|^{-1/2}$ by writing $\mathbf{X}_p'\mathbf{X}_p$=n $\mathbf{S}_p$ where n is the sample size and $\mathbf{S}_p$ is the sample covariance matrix of the explanatory variables which, if drawn from a stationary distribution will converge to the true covariance matrix as the sample size n increases.

$$\lim_{\nu_p^2 \to \infty} f_p(\boldsymbol{y}) \propto |\boldsymbol{S}_p|^{-1/2} \mathrm{n}^{-p/2} \left(ESS_p\right)^{-n/2}$$

Finally, we can hope that $|\boldsymbol{S}_p|$ doesn't vary enough across models to worry about. If the explanatory variables are drawn from a stationary distribution $\boldsymbol{S}_p$will converge to the true covariance matrix as the sample size n increases. If that is the case, the sample size dependent components of this expression dominate, and we obtain the asymptotic marginal likelihood discussed in Leamer(1978,p.11) and used by Sala-i-Martin et. al.(2004)

$$\lim_{n \to \infty} f_p(\boldsymbol{y}) \propto \mathrm{n}^{-p/2} \left(ESS_p\right)^{-n/2}$$

Two times the negative of the logarithm of this expression is the Schwarz(1978) criterion $\mathrm{p}\log(\mathrm{n}) + \mathrm{n}\log(ESS_p)$.

I have repeated the derivation of this result to reissue warnings regarding its use. The assumption of a conjugate prior has no connection with the reality that I live in, and the neglect of the generalized variance $|\boldsymbol{S}_p|$ seems mistaken. With two standardized variables, this determinant is one minus the squared correlation between the two variables. When this is kept in the formula, a two-variable model with a high correlation between the explanatory variables is given a helping hand, since it cannot be expected to fit as well as a model with two uncorrelated variables. [13] That seems sensible to me, but I am not sure because another way to obtain the weighted $2^k$estimate discussed below produces the opposite result.

[13] Incidentally, the formula breaks down if the covariance matrix of explanatory variables is singular (determinant zero), since the diffuse prior assumption that underlies the formula is inappropriate when the data are completely uninformative about a linear combination of parameters. Then, even with an infinite sample, the prior matters.

## 10 References

Akaike, H., 1977. "On entropy maximization principle". In: Krishnaiah, P.R. (Editor). Applications of Statistics, North-Holland, Amsterdam, pp. 27–41.

Barro, Robert J. “Economic Growth in a Cross Section of Countries.” *The Quarterly Journal of Economics*, 106 ( No. 2 1991): 407-43.

Brock, W.A., Durlauf, S.N. and West, K.D. (2003). Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 1:235-322.

Brock, William A. and Steven N. Durlauf (2001) “Growth Empirics and Reality,” *The World Bank Economic Review* , Vol. 15, No. 2 (2001) , pp. 229-272

Carvalho, Carlos .M., Nicholas G. Polson, and James G Scott, (2010. “The horseshoe estimator for sparse signals.” *Biometrika* 97, 465–480.

Chamberlain, Gary and Edward E. Leamer (1976) , “Matrix Weighted Averages and Posterior Bounds,” *Journal of the Royal Statistical Society*, B, 38, 73-84.

Durlauf, Steven N.(2000) Review of Growth, Inequality, and Globalization: Theory, History, and Policy. by Philippe Aghion; Jeffrey G. Williamson, *Journal of Economic Literature* , Vol. 38, No. 3 (Sep., 2000) , pp. 637-638

Eicher, T.S., Papageorgiou, C., Raftery, A.E., (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics 26*, 30–55.

Fernández, Carmen , Eduardo Ley and Mark F. J. Steel(2001a), “Model Uncertainty in Cross-Country Growth Regressions,” *Journal of Applied Econometrics*, Vol. 16, No. 5 (Sep. - Oct., 2001), pp. 563-576

Fernández, Carmen , Eduardo Ley and Mark F. J. Steel(2001b), “Benchmark priors for Bayesian model averaging.” *Journal of Econometrics* 100, 381–427.

Hoerl, Arthur E. and Robert W. and Kennard (1970), “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics,* Vol. 12, No. 1 (Feb 1970), 55-67.

Kalli, Maria and Jim E. Griffin (2013), “Time-varying sparsity in dynamic regression models, *Journal of Econometrics*, Available online 8 November 2013

Kruskal, William (1978) “Tests of Significance” in William H Kruskal and Judith M. Tanur ed. (1978) *International Encyclopedia of Statistics*, The Free Press: New York, pp. 944-958

Leamer, Edward E. (1975) "A Result on the Sign of Restricted Least Squares Estimates," *Journal of Econometrics,* 3 (1975), 387-390.

Leamer, Edward E. (1978) *Specification Searches: Ad Hoc Inference With Nonexperimental Data*, New York: Wiley.

Leamer, Edward E. (1982), “Sets of Posterior Means with Bounded Variance Priors,” *Econometrica,* Vol. 50, No. 3 (May 1982), pp. 725-736.

Leamer, Edward E. (2014), “S-values: Conventional measures of the sturdiness of regression coefficients,” working paper.

Leamer, Edward E.  and Gary Chamberlain (1976) "A Bayesian Interpretation of Pretesting," with G. Chamberlain, *Journal of the Royal Statistical Society*, Series B, 38 (No. 1, 1976), 85-94.

Ley, Eduardo, Mark F. J. Steel and Carmen Fernández, (2009), “On the effect of prior assumptions in Bayesian model averaging with applications to growth regression.” *Journal of Applied Econometrics* 24, 651–674.

Ley, Eduardo, Mark F. J. Steel and Carmen Fernández, (2012), Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* , 171, 251–266.

Ley, Eduardo, Mark F. J. Steel (2009), “On the Effect of Prior Assumptions on Bayesian Model Averaging with Applications to Growth Regression,” *Journal of Applied Econometrics*, 24: 651-674.

Levine, Ross and David Renelt (1992) “A Sensitivity Analysis of Cross-Country Growth Regressions,”  The American Economic Review, V 82, Issue 4 (Sept), 942-963.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. (2008).Mixtures of g-priors for Bayesian variable selection. J. Am. Stat. Assoc.,130(481):410–423,

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83:1023-1032.

Raftery, A.E. (1995). Bayesian model selection for social research. *Sociological Methodology*, 25:111{163.

Rodrik, Dani (2012)“Why We Learn Nothing from Regressing Economic Growth on Policies,” Seould Journal of Economics, Vol 25, No. 2, 137-

Sala-I-Martin,  Xavier X. (1997a), “I Just Ran Two Million Regressions” , *The American Economic Review,* Vol. 87, No. 2, Papers and Proceedings . (May, 1997), pp. 178-183.

_______________(1997b) "I Just Ran Four Million Regressions." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6252, November .

Sala-i-Martin, Xavier, Gernot Doppelhofer and Ronald I. Miller (2004) , "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *The American Economic Review*, Vol. 94, No. 4 (Sep., 2004), pp. 813-835

Schwarz, Gideon E. (1978). "Estimating the dimension of a model". Annals of Statistics 6 (2): 461–464. doi:10.1214/aos/1176344136. MR 468014.

Taplin RH, Raftery AE. (1994). "Analysis of agricultural field trials in the presence of outliers and fertility jumps." *Biometrics* 50: 764–781.

Zellner, A. and Siow, A. (1980) Posterior odds ratios for selected regression hypotheses. In Bayesian Statistics: Procee dings of the First International Meeting held in Valencia (Spain) , (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 585{603. Valencia: University Press.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. The Statistician , 32, 23–34

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti , (eds. P . K. Goel and A. Zellner), pp. 233-243. North-Holland/Elsevier.

Ziliak, Stephen, and McCloskey, Deirdre, (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, University of Michigan Press, 2009.

# 11 Tables

Table 1 SDM Measures of Statistical Significance compared with OLS

TABLE 2 BASELINE ESTIMATION FOR All 67 VARIABLES

Source: Sala-i-Martin, Doppelhofer and Miller (2004)

| | | | | BACE |
|---|---|---|---|---|
| | | | Posterior | sign |
| | | OLS | inclusion | certainty |
| | | 1-p/2 | prob. | prob. |
| | | | | kbar=7 |
| | | | (1) | (4) |
| 1 | East Asian dummy | 0.7475 | 0.823 | 0.999 |
| 2 | Primary schooling 1960 | 0.9225 | 0.796 | 0.999 |
| 3 | Investment price | 0.984 | 0.774 | 0.999 |
| 4 | GDP 1960 (log) | 0.8065 | 0.685 | 0.999 |
| 5 | Fraction of tropical area | 0.767 | 0.563 | 0.997 |
| 6 | Population density coastal 1960's | 0.6165 | 0.428 | 0.996 |
| 7 | Malaria prevalence in 1960's | 0.7425 | 0.252 | 0.99 |
| 8 | Life expectancy in 1960 | 0.6195 | 0.209 | 0.986 |
| 9 | Fraction Confucian | 0.8115 | 0.206 | 0.988 |
| 10 | African dummy | 0.7055 | 0.154 | 0.98 |
| 11 | Latin American dummy | 0.674 | 0.149 | 0.969 |
| 12 | Fraction GDP in mining | 0.8475 | 0.124 | 0.978 |
| 13 | Spanish colony | 0.7465 | 0.123 | 0.972 |
| 14 | Years open | 0.587 | 0.119 | 0.977 |
| 15 | Fraction Muslim | 0.761 | 0.114 | 0.973 |
| 16 | Fraction Buddhist | 0.77 | 0.108 | 0.974 |
| 17 | Ethnolinguistic fractionalization | 0.5045 | 0.105 | 0.974 |
| 18 | Government consumption share 1960' | 0.828 | 0.104 | 0.975 |
| 19 | Population density 1960 | 0.5925 | 0.086 | 0.965 |
| 20 | Real exchange rate distortions | 0.5825 | 0.082 | 0.966 |
| 21 | Fraction speaking foreign language | 0.763 | 0.08 | 0.962 |
| 22 | (Imports + exports)/GDP | 0.5245 | 0.076 | 0.949 |
| 23 | Political rights | 0.668 | 0.066 | 0.939 |
| 24 | Government share of GDP | 0.8355 | 0.063 | 0.935 |
| 25 | Higher education in 1960 | 0.8105 | 0.061 | 0.946 |
| 26 | Fraction population in tropics | 0.6715 | 0.058 | 0.94 |
| 27 | Primary exports in 1970 | 0.624 | 0.053 | 0.926 |
| 28 | Public investment share | 0.9425 | 0.048 | 0.922 |
| 29 | Fraction Protestant | 0.6425 | 0.046 | 0.909 |

| | | | | |
|---|---|---|---|---|
| 30 | Fraction Hindu | 0.605 | 0.045 | 0.915 |
| 31 | Fraction population less than 15 | 0.7275 | 0.041 | 0.871 |
| 32 | Air distance to big cities | 0.714 | 0.039 | 0.888 |
| 33 | Government consumption share deflated with GDP prices | 0.7175 | 0.036 | 0.893 |
| 34 | Absolute latitude | 0.758 | 0.033 | 0.737 |
| 35 | Fraction Catholic | 0.5305 | 0.033 | 0.837 |
| 36 | Fertility in 1960's | 0.6515 | 0.031 | 0.767 |
| 37 | European dummy | 0.616 | 0.03 | 0.544 |
| 38 | Outward orientation | 0.559 | 0.03 | 0.886 |
| 39 | Colony dummy | 0.615 | 0.029 | 0.858 |
| 40 | Civil liberties | 0.63 | 0.029 | 0.846 |
| 41 | Revolutions and coups | 0.862 | 0.029 | 0.877 |
| 42 | British colony | 0.67 | 0.027 | 0.844 |
| 43 | Hydrocarbon deposits in 1993 | 0.5935 | 0.025 | 0.773 |
| 44 | Fraction population over 65 | 0.73 | 0.022 | 0.566 |
| 45 | Defense spending share | 0.8365 | 0.021 | 0.737 |
| 46 | Population in 1960 | 0.7095 | 0.021 | 0.806 |
| 47 | Terms of trade growth in 1960's | 0.509 | 0.021 | 0.752 |
| 48 | Public education spending/GDP in 1960's | 0.839 | 0.021 | 0.777 |
| 49 | Landlocked country dummy | 0.5245 | 0.021 | 0.701 |
| 50 | Religion measure | 0.545 | 0.02 | 0.751 |
| 51 | Size of economy | 0.7115 | 0.02 | 0.661 |
| 52 | Socialist dummy | 0.542 | 0.02 | 0.788 |
| 53 | English-speaking population | 0.7285 | 0.02 | 0.686 |
| 54 | Average inflation 1960-1990 | 0.613 | 0.02 | 0.784 |
| 55 | Oil-producing country dummy | 0.5335 | 0.019 | 0.751 |
| 56 | Population growth rate 1960-1990 | 0.538 | 0.019 | 0.533 |
| 57 | Timing of independence | 0.577 | 0.019 | 0.716 |
| 58 | Fraction of Land Area Near Navigable Water | 0.635 | 0.019 | 0.657 |
| 59 | Square of inflation 1960-1990 | 0.666 | 0.018 | 0.736 |
| 60 | Fraction spent in war 1960-1990 | 0.548 | 0.016 | 0.555 |
| 61 | Land area | 0.525 | 0.016 | 0.577 |
| 62 | Tropical climate zone | 0.693 | 0.016 | 0.616 |
| 63 | Terms of trade ranking | 0.5775 | 0.016 | 0.647 |
| 64 | Capitalism | 0.8435 | 0.015 | 0.589 |
| 65 | Fraction Orthodox | 0.55 | 0.015 | 0.66 |
| 66 | War participation 1960-1990 | 0.566 | 0.015 | 0.593 |
| 67 | Interior density | 0.616 | 0.015 | 0.532 |

**Table 2** **Leamer s-value and apparent SDM s-value**

SDM Table 4

Posterior Means Conditional on Inclusion with Different Prior Model Sizes

Ranked by implicit s-value = (Max+Min)/(Max-Min)

| | | | | | SDM | s-val |
|---|---|---|---|---|---|---|
| Order | Rank | Variable | kbar5 | kbar28 | "s-val" | (0.50) |
| 1 | 34 | Absolute latitude | 0.000169 | -5.4E-05 | 0.5 | 0.10 |
| 2 | 60 | Fraction spent in war 1960-1990 | -0.00256 | 0.000801 | -0.5 | 0.00 |
| 3 | 51 | Size of economy | -0.00073 | 0.000166 | -0.6 | -0.15 |
| 4 | 37 | European dummy | -0.00138 | 0.006152 | 0.6 | 0.15 |
| 5 | 63 | Terms of trade ranking | -0.00418 | 0.000894 | -0.6 | -0.15 |
| 6 | 54 | Average inflation 1960-1990 | -8.2E-05 | 0.000008 | -0.8 | 0.11 |
| 7 | 53 | English-speaking population | -0.00484 | 0.000311 | -0.9 | -0.18 |
| 8 | 47 | Terms of trade growth in 1960's | 0.035425 | -0.00154 | 0.9 | 0.18 |
| 9 | 59 | Square of inflation 1960-1990 | -1E-06 | 0 | -1.0 | 0.11 |
| 10 | 67 | Interior density | 0 | -4E-06 | -1.0 | -0.12 |
| 11 | 56 | Population growth rate 1960-1990 | 0.059271 | 0.001145 | 1.0 | 0.09 |
| 12 | 50 | Religion measure | -0.00441 | -0.00041 | -1.2 | -0.03 |
| 13 | 57 | Timing of independence | 0.001258 | 0.000121 | 1.2 | 0.18 |
| 14 | 44 | Fraction population over 65 | 0.011873 | 0.118943 | 1.2 | 0.06 |
| 15 | 64 | Capitalism | -0.00009 | -0.00089 | -1.2 | 0.08 |
| 16 | 65 | Fraction Orthodox | 0.007559 | 0.001914 | 1.7 | 0.34 |
| 17 | 45 | Defense spending share | 0.052439 | 0.013697 | 1.7 | 0.23 |
| 18 | 55 | Oil-producing country dummy | 0.003559 | 0.000995 | 1.8 | 0.13 |
| 19 | 7 | Malaria prevalence in 1960's | -0.0173 | -0.0054 | -1.9 | -0.49 |
| 20 | 62 | Tropical climate zone | -0.00326 | -0.0014 | -2.5 | -0.39 |
| 21 | 43 | Hydrocarbon deposits in 1993 | 0.00022 | 0.000461 | 2.8 | 0.42 |
| 22 | 31 | Fraction population less than 15 | 0.045099 | 0.0226 | 3.0 | 0.31 |
| 23 | 36 | Fertility in 1960's | -0.00593 | -0.01117 | -3.3 | -0.18 |
| 24 | 25 | Higher education in 1960 | -0.07373 | -0.04013 | -3.4 | -0.36 |
| 25 | 49 | Landlocked country dummy | -0.00151 | -0.00273 | -3.5 | -0.18 |
| 26 | 26 | Fraction population in tropics | -0.01201 | -0.00668 | -3.5 | -0.44 |
| 27 | 14 | Years open | 0.012831 | 0.007235 | 3.6 | 0.75 |
| 28 | 28 | Public investment share | -0.04923 | -0.08696 | -3.6 | -0.85 |
| 29 | 13 | Spanish colony | -0.01102 | -0.00641 | -3.8 | -0.40 |
| 30 | 1 | East Asian dummy | 0.023633 | 0.014105 | 4.0 | 1.33 |
| 31 | 42 | British colony | 0.004059 | 0.00248 | 4.1 | 0.60 |
| 32 | 17 | Ethnolinguistic fractionalization | -0.01192 | -0.0076 | -4.5 | -0.51 |
| 33 | 35 | Fraction Catholic | -0.00663 | -0.00995 | -5.0 | -0.21 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 34 | 6 | Population density coastal 1960's | 0.000009 | 0.000006 | 5.0 | 0.45 |
| 35 | 27 | Primary exports in 1970 | -0.01215 | -0.00815 | -5.1 | -0.35 |
| 36 | 12 | Fraction GDP in mining | 0.036381 | 0.053946 | 5.1 | 0.92 |
| 37 | 5 | Fraction of tropical area | -0.01476 | -0.01009 | -5.3 | -0.51 |
| 38 | 66 | War participation 1960-1990 | -0.00075 | -0.00109 | -5.3 | -0.29 |
| 39 | 33 | Government consumption share deflated with GDP prices | -0.03065 | -0.04302 | -6.0 | -0.35 |
| 40 | 8 | Life expectancy in 1960 | 0.000812 | 0.000608 | 7.0 | 0.58 |
| 41 | 18 | Government consumption share 1960's | -0.047 | -0.03544 | -7.1 | -0.40 |
| 42 | 22 | (Imports + exports)/GDP | 0.009305 | 0.007118 | 7.5 | 0.54 |
| 43 | 58 | Fraction land area near navig. water | -0.00287 | -0.00369 | -8.1 | 0.00 |
| 44 | 39 | Colony dummy | -0.0052 | -0.00414 | -8.8 | -0.49 |
| 45 | 23 | Political rights | -0.00177 | -0.00141 | -8.9 | -0.30 |
| 46 | 48 | Public education spending/GDP in 1960's | 0.127413 | 0.159089 | 9.0 | 0.55 |
| 47 | 41 | Revolutions and coups | -0.00726 | -0.00902 | -9.2 | -0.70 |
| 48 | 20 | Real exchange rate distortions | -0.00008 | -6.6E-05 | -10.4 | -1.02 |
| 49 | 4 | GDP 1960 (log) | -0.00825 | -0.00998 | -10.5 | -0.77 |
| 50 | 16 | Fraction Buddhist | 0.022501 | 0.019788 | 15.6 | 1.21 |
| 51 | 52 | Socialist dummy | 0.00414 | 0.004706 | 15.6 | 0.39 |
| 52 | 29 | Fraction Protestant | -0.01097 | -0.00973 | -16.6 | -0.52 |
| 53 | 9 | Fraction Confucian | 0.055184 | 0.050184 | 21.1 | 1.48 |
| 54 | 21 | Fraction speaking foreign language | 0.006864 | 0.0075 | 22.6 | 0.67 |
| 55 | 19 | Population density 1960 | 0.000012 | 0.000013 | 25.0 | 0.69 |
| 56 | 3 | Investment price | -8.3E-05 | -8.9E-05 | -28.7 | -1.97 |
| 57 | 40 | Civil liberties | -0.00726 | -0.00679 | -29.8 | -0.72 |
| 58 | 30 | Fraction Hindu | 0.016548 | 0.017318 | 44.0 | 0.14 |
| 59 | 38 | Outward orientation | -0.00332 | -0.00318 | -46.7 | -0.41 |
| 60 | 15 | Fraction Muslim | 0.012361 | 0.012863 | 50.2 | 0.28 |
| 61 | 24 | Government share of GDP | -0.03445 | -0.03511 | -104.1 | -0.65 |
| 62 | 10 | African dummy | -0.01416 | -0.01391 | -110.1 | -0.73 |
| 63 | 2 | Primary schooling 1960 | 0.025899 | 0.025605 | 175.2 | 1.22 |
| 64 | 11 | Latin American dummy | -0.01203 | -0.01207 | -573.6 | -0.37 |
| 65 | 32 | Air distance to big cities | -1E-06 | -1E-06 | na | 0.07 |
| 66 | 46 | Population in 1960 | 0 | 0 | na | 0.70 |
| 67 | 61 | Land area | 0 | 0 | na | -0.16 |