Is 'Which model . . . ?' the right question?

Nicholas T. Longford

SNTL, Leicester, England, and Universitat Pompeu Fabra, Barcelona, Spain,
NTL@SNTL.co.uk

1 Introduction

In a typical data-modelling exercise, we consider a range of models, apply an algorithm for selecting one of them, and then quote the estimates and associated quantities, such as estimated standard errors, assuming that the selected model is valid; we ignore the process by which the model has been selected. Such statements are misleading because of the conditioning on the selected model. The appropriate statements should be conditional only on *a priori* settings, such as the collection of the models considered. The selection process is not ignorable — it has an impact, often profound, on the sampling distribution of the estimator.

In *synthetic estimation*, no single model is selected, but the estimators based on the candidate models are combined. Although difficult to apply universally, its application in some simple settings highlights the weaknesses of the established model-selection procedures. We conclude that whenever there is uncertainty about the appropriate model we should not select a model for describing the analysed data set and then apply it in all subsequent inferences. Instead of the estimator based on a single (target-specific) model, their convex combinations should be considered. This is in accord with Bayes factors [4]. However, coefficients (weights) assigned by our approach to the alternative models depend on the target of estimation or prediction (a parameter, a function of the parameters or the realisation of a random variable).

Although the sampling distribution of a combination $\tilde{\theta} = \sum_m b_\theta^{(m)} \hat{\theta}_m$ can be established when the coefficients $b_\theta^{(m)}$ are known, the unconditional distribution, recognising that the coefficients $b_\theta^{(m)}$ are estimated, can be established only by simulations. This should not be regarded as a drawback. Usually, the unconditional distribution of an estimator based on a selected model cannot be established analytically either.

The next section defines some terms that enable a clearer and more concise formulation of our criticism. Section 3 gives the general definition of a

synthetic estimator and explores its properties in the case of two candidate estimators. Section 4 illustrates synthetic estimation on an example.

2 Preliminaries

A model is defined as any class of distributions $\mathcal{D}$. Model $\mathcal{D}$ is said to be valid for a random vector $\mathbf{y}$ if it contains the joint distribution of $\mathbf{y}$. Instead of the (unconditional) distribution of $\mathbf{y}$ we may consider its conditional distribution given the values of a set of covariates (matrix $\mathbf{X}$). The *estimation process* is the collection of all the operations applied to the data $\mathcal{X}$ between data generation and a statement of the form

(`estimate, estimated standard error`),

or $\left(\hat{\theta}, \hat{s}^2 = \widehat{\mathrm{var}}(\hat{\theta})\right)$. We are concerned with estimation processes that consist of two steps: *model selection* and *estimation* based on the selected model. The purpose of model selection is to reduce the originally specified model $\mathcal{D}_0$, assumed to be valid, to a submodel $\mathcal{D}^*$ that is also valid, but estimators based on this submodel are more efficient. For example, the original model $\mathcal{D}_0$ may be given by a multidimensional space of parameters $\boldsymbol{\Theta}$ and its various submodels by projections of $\boldsymbol{\Theta}$ to lower dimensions.

Any data-dependent (stochastic) model-selection process results in a submodel that is not valid with certainty because a replication of the data-generation and selection processes may yield a different selected model. As a consequence, any statements that are conditional on the validity of the selected model are problematic because they confuse unconditional distributions, appropriate for *a priori* selected models, with distributions that are conditional on the selection process. This issue is closely related to model uncertainty [1].

Throughout, we regard the mean squared error, $\mathrm{MSE}(\hat{\theta}; \theta) = \mathrm{E}\{(\hat{\theta}-\theta)^2 \mid \theta\}$ for estimator (or predictor) $\hat{\theta}$ of the target θ, as the criterion for estimation (prediction). That is, estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ for estimating θ if $\mathrm{MSE}(\hat{\theta}_1\,; \theta) < \mathrm{MSE}(\hat{\theta}_2\,; \theta)$. MSE usually depends on some parameters (even on θ itself), and then it is estimated, and there may be more than one estimator. When the MSE does depend on a parameter, $\hat{\theta}_1$ may be more efficient than $\hat{\theta}_2$ only for some values of θ. We consider a set of models $\mathcal{D}_m$, $m = 0, 1, \ldots, M$, and assume that model $\mathcal{D}_0$ is valid. In a typical setting, models $\mathcal{D}_1, \ldots, \mathcal{D}_M$ are all submodels of $\mathcal{D}_0$, although we do not assume this in the general development.

Our target θ may be any function of the distribution of $\mathbf{y}$, unconditional or with a specified conditioning, so that we make no distinction between estimation (of model parameters) and prediction, which may involve terms that are represented in the model by random variables. Each model m is associated with an estimator $\hat{\theta}_m$; we assume that it has some desirable properties, such

as small or no bias and efficiency, but only when $\mathcal{D}_m$ is valid and none of its submodels are. For example, $\hat{\theta}_m$ may be the maximum likelihood estimator of θ under model m. We refer to $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_M$, as *single-model based estimators*. A model $\mathcal{D}_{m^*}$ is said to be *minimal* valid if it is valid and none of its submodels among $\mathcal{D}_0, \ldots, \mathcal{D}_M$ are valid.

Let I_m be the dichotomous variable that indicates the choice of model m in a particular model-selection process. These indicators add up to unity, $I_0+I_1+\cdots+I_M = 1$, because one of the models is selected for every conceivable data set. Denote by $\mathcal{M}$ the selected model; $\mathcal{M} = m$ when $I_m = 1$. The *selected-model based estimator* is

$$\hat{\theta}_{\mathcal{M}} = I_0\hat{\theta}_0 + I_1\hat{\theta}_1 + \cdots + I_M\hat{\theta}_M \,. \tag{1}$$

This formulation as a *mixture* shows that the distribution of $\hat{\theta}_{\mathcal{M}}$ depends not only on the selected model but on all the models that have a positive probability of being selected. Since the indicators I_m and the estimators $\hat{\theta}_m$ are mutually dependent, the distribution of their mixture $\hat{\theta}_{\mathcal{M}}$ cannot be established analytically even in some simple cases when the probabilities $p_m(\theta) = \mathrm{P}(I_m = 1 \,|\, \theta)$ and the joint distribution of $\hat{\theta}_m$ are known.

We consider the distributions of $\hat{\theta}_m$ under the single *a priori* nominated model 0. Therefore, we work with the distributions of each $\hat{\theta}_m$ not under model m, but when a different (more complex) model 0 applies. In particular, $\hat{\theta}_0$ is (almost) unbiased, whereas the other estimators may be biased. We should not disqualify a biased estimator because its sampling variance may be so small that its MSE is lower than for any other estimator.

Model selection is commonly interpreted as a search for a narrower valid model, so that the estimator based on it would be unbiased and would have smaller sampling variance than under a wider model. In view of (1), this is misleading. By re-interpreting model selection as a trade-off between inflated variance and bias due to lack of model validity, we reject the commonly held view that, under appropriate regularity conditions, the maximum likelihood estimator with a narrow valid model is (nearly) efficient. After all, the estimator based on a narrower invalid model is more efficient if the reduction of the sampling variance exceeds the increment of the squared bias.

If all the estimators $\hat{\theta}_m$ are unbiased and I_m are mutually independent from them, $\hat{\theta}_{\mathcal{M}}$ is unbiased. But $\hat{\theta}_{\mathcal{M}}$ may be biased when I_m and $\hat{\theta}_m$ are dependent. In conventional statements (inferences) we say, or imply, that $\hat{\theta}_{\mathcal{M}}$ is unbiased. This is in general incorrect. More precisely, we want it to be understood that if we have selected the appropriate model, then $\hat{\theta}_{\mathcal{M}}$ is unbiased. This statement is not valid either because the distribution of $\hat{\theta}_{m^*}$, given that model m^* is selected, and appropriately so, need not coincide with the *unconditional* distribution of $\hat{\theta}_{m^*}$. After all, even when model m^* is valid it would not be selected in all replications; different models would be selected for a highly selective subsample of the values of $\hat{\theta}_{m^*}$.

The sampling variance $\mathrm{var}(\hat{\theta}_{\mathcal{M}})$ is conventionally estimated by the conditional variance of $\hat{\theta}_{\mathcal{M}}$ given $\mathcal{M}$ *and* assuming that model $\mathcal{M}$ applies; that is,

model $\mathcal{M}$ has been selected appropriately. This estimator is

$$\hat{s}^2_{\mathcal{M}} \;=\; \widehat{\mathrm{var}}(\hat{\theta}_{\mathcal{M}}) \;=\; I_0\hat{s}^2_0 + I_1\hat{s}^2_1 + \cdots + I_M\hat{s}^2_M\,, \tag{2}$$

where $\hat{s}^2_m = \widehat{\mathrm{var}}(\hat{\theta}_m)$ is an estimator of the sampling variance of $\hat{\theta}_m$ assuming that model m applies and is selected unconditionally (that is, $\mathcal{M} \equiv m$). We assume that each $\hat{s}^2_m$ is (approximately) unbiased when model m applies, but not necessarily otherwise. Even if each $\hat{s}^2_m$ is unbiased unconditionally (assuming only that model 0 is valid), the estimator in (2) is biased, even when I_m and $\hat{s}^2_m$ are mutually independent. Intuitively, when we are not certain about the model and the estimators based on the candidate models do not coincide, the sampling variance is bound to be greater than if the appropriate model were identified with certainty. The penalty for less information is greater sampling variation, but the conventional estimation of MSE, or s^2, has no means of reflecting it.

3 From choice to synthesis

In Bayesian model averaging [4], the single-model based estimators are combined, with coefficients proportional to the posterior probabilities of the models. Model uncertainty can be addressed by EM algorithm or data augmentation, in which the appropriate model is regarded as the missing information. The M-step of this algorithm also combines the single-model based estimators. Motivated by these two approaches, we study convex combinations of the single-model based estimators, but seek to minimise the MSE of the combination directly. Let

$$\tilde{\theta}(\mathbf{b}) \;=\; (1-b_+)\hat{\theta}_0 + \mathbf{b}^\top\hat{\boldsymbol{\theta}}\,,$$

where $\mathbf{b}$ is the $M \times 1$ vector of weights assigned to the respective models $1,\ldots,M$, b_+ their total, and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\ldots,\hat{\theta}_M)^\top$ the vector of the single-model based estimators (with $\hat{\theta}_0$ omitted). The argument $\mathbf{b}$ of $\tilde{\theta}$ is essential because we will consider different values of $\mathbf{b}$. As $\mathbf{b}$ will turn out to be a function of the target θ, it should be indexed by θ. However, we consider only one target θ, and so we prefer to avoid a clutter of subscripts over the more rigorous notation $\mathbf{b}_\theta$. Let $\mathbf{d}$ be the vector of biases of $\hat{\boldsymbol{\theta}}$. We assume that model 0 is valid, so $\hat{\theta}_0$ is unbiased. Without the assumption that a model is valid, the problem is ill-defined. Denote $\mathbf{V} = \mathrm{var}(\hat{\boldsymbol{\theta}})$ and $\mathbf{C} = \mathrm{cov}(\hat{\boldsymbol{\theta}}, \hat{\theta}_0)$. By $\mathbf{1}$ we denote the column vector of ones of length implied by the context. For instance, $b_+ = \mathbf{b}^\top\mathbf{1}$.

If $\mathbf{d} = \mathrm{E}(\hat{\boldsymbol{\theta}}) - \theta\mathbf{1}$, $V_0 = \mathrm{var}(\theta_0)$, $\mathbf{C}$ and $\mathbf{V}$ are known, minimising the MSE of $\tilde{\theta}(\mathbf{b})$ is straightforward. We have

$$\mathrm{MSE}\{\tilde{\theta}(\mathbf{b})\} = (1-b_+)^2V_0 + \mathbf{b}^\top\mathbf{V}\mathbf{b} + 2(1-b_+)\mathbf{C}^\top\mathbf{b} + \left(\mathbf{b}^\top\mathbf{d}\right)^2$$

$$= V_0 - 2\mathbf{b}^\top(V_0\mathbf{1} - \mathbf{C}) + \mathbf{b}^\top(V_0\mathbf{1}\mathbf{1}^\top - \mathbf{C}\mathbf{1}^\top - \mathbf{1}\mathbf{C}^\top + \mathbf{V} + \mathbf{d}\mathbf{d}^\top)\mathbf{b}$$
$$= V_0 - 2\mathbf{b}^\top\mathbf{P} + \mathbf{b}^\top\mathbf{Q}\mathbf{b}\,, \qquad (3)$$

with $\mathbf{P} = \mathrm{cov}(\hat{\theta}_0\mathbf{1} - \hat{\boldsymbol{\theta}}, \hat{\theta}_0)$ and $\mathbf{Q} = \mathrm{E}\left\{(\hat{\theta}_0\mathbf{1} - \hat{\boldsymbol{\theta}})^\top(\hat{\theta}_0\mathbf{1} - \hat{\boldsymbol{\theta}})\right\}$. This is a quadratic function of $\mathbf{b}$, and its minimum is attained at the root of its vector of first-order partial derivatives:

$$\frac{\partial \mathrm{MSE}(\tilde{\theta})}{\partial \mathbf{b}} = 2(\mathbf{Q}\mathbf{b} - \mathbf{P})\,,$$

that is, at $\mathbf{b}^* = \mathbf{Q}^{-1}\mathbf{P}$, if $\mathbf{Q}$ is non-singular. The minimum attained is $\mathrm{MSE}\{\tilde{\theta}(\mathbf{b}^*)\} = V_0 - \mathbf{P}^\top\mathbf{Q}^{-1}\mathbf{P}$. When $\mathbf{Q}$ is singular, the optimal vector of coefficients $\mathbf{b}^*$ is not unique, but the same minimum would be attained after discarding one or several estimators in $\hat{\boldsymbol{\theta}}$.

Assuming that $\mathbf{b}^*$ is known, the *ideal synthetic* estimator $\tilde{\theta}(\mathbf{b}^*)$ is more efficient than either of the constituent (candidate) single-model based estimators $\hat{\theta}_m$ because these estimators correspond to the extreme choices of $\mathbf{b}$ in (3): $\mathbf{b} = \mathbf{0}$ (the vector of zeros), and $\mathbf{b} = \mathbf{e}_m$, $m = 1, \ldots, M$, the indicator vector for its mth element. In practice, $\mathbf{b}^*$ depends on unknown parameters, and so it has to be estimated. Unlike $\tilde{\theta}(\mathbf{b}^*)$, the *synthetic* estimator $\tilde{\theta}\left(\hat{\mathbf{b}}^*\right)$ may be less efficient than some of the constituent estimators $\hat{\theta}_m$.

The properties of $\tilde{\theta}\left(\hat{\mathbf{b}}^*\right)$ are difficult to explore, as are the properties of the selected-model based estimator $\hat{\theta}_{\mathcal{M}}$. In the next section, we compare $\tilde{\theta}\left(\hat{\mathbf{b}}^*\right)$ and $\hat{\theta}_{\mathcal{M}}$ when there are only two alternative models; $M = 1$.

3.1 'Choice between' *vs.* 'combination of'

When $M = 1$, $\mathbf{P}$ and $\mathbf{Q}$ are scalars, and the weight assigned to model 1 is

$$b^* = \frac{V_0 - C}{V_0 + V - 2C + d^2}$$

and

$$\mathrm{MSE}\left\{\tilde{\theta}(b^*)\right\} = V_0 - \frac{(V_0 - C)^2}{V_0 + V - 2C + d^2}\,,$$

where V, C, b and d are the respective univariate versions of $\mathbf{V}$, $\mathbf{C}$, $\mathbf{b}$ and $\mathbf{d}$. The synthetic estimator coincides with $\hat{\theta}_0$ when $V_0 = C$. When $(\hat{\theta}_0, \hat{\theta}_1)$ has a bivariate normal distribution, this condition can be interpreted as $\hat{\theta}_1 = \hat{\theta}_0 + \delta$ for a random variable δ independent of $\hat{\theta}_0$. That is, $\hat{\theta}_1$ is formed by adding white noise to $\hat{\theta}_0$. In this case, $\hat{\theta}_1$ contains no information additional to $\hat{\theta}_0$, so $\hat{\theta}_1$ is redundant. Besides, $\mathrm{var}(\hat{\theta}_0) < \mathrm{var}(\hat{\theta}_1)$. However, the conditions $\mathrm{var}(\hat{\theta}_0) < \mathrm{var}(\hat{\theta}_1)$ and $d \neq 0$ do not imply that $\tilde{\theta} = \hat{\theta}_0$. A simple counterexample arises when $\hat{\theta}_0$ and $\hat{\theta}_1$ are independent ($C = 0$); then $\tilde{\theta}$ is the combination of $\hat{\theta}_0$ and

$\hat{\theta}_1$ with weights proportional to their precisions (reciprocals of their MSEs). When $C < V_0$, the *biased* estimator $\hat{\theta}_1$ contributes to the improvement over the unbiased estimator $\hat{\theta}_0$. In contrast, the ambition of a model selection procedure is merely to use the better of the candidate estimators. Below, as well as in simulations in Section 4.1, we show that model selection falls well short of this goal. The synthetic estimator coincides with the biased estimator $\hat{\theta}_1$ only when $C = V + d^2$.

The selected-model based estimator $\hat{\theta}_{\mathcal{M}}$ has the bias

$$\begin{aligned}\mathrm{E}(\hat{\theta}_{\mathcal{M}}) - \theta &= p_0\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0) + p_1\mathrm{E}(\hat{\theta}_1 \,|\, \mathcal{M}=1) - \theta \\ &= p_1\mathrm{E}(\hat{\theta}_1 - \hat{\theta}_0 \,|\, \mathcal{M}=1)\,,\end{aligned}$$

derived from the identity $\theta = \mathrm{E}(\hat{\theta}_0) = p_0\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0) + p_1\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=1)$. Its MSE is

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_{\mathcal{M}}) &= p_0\mathrm{var}(\hat{\theta}_0 \,|\, \mathcal{M}=0) + p_1\mathrm{var}(\hat{\theta}_1 \,|\, \mathcal{M}=1) \\ &\quad + p_0p_1\left\{\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0) - \mathrm{E}(\hat{\theta}_1 \,|\, \mathcal{M}=1)\right\}^2 \\ &\quad + p_1^2\left\{\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0) - \mathrm{E}(\hat{\theta}_1 \,|\, \mathcal{M}=1)\right\}^2 \\ &= p_0\mathrm{var}(\hat{\theta}_0 \,|\, \mathcal{M}=0) + p_1\mathrm{var}(\hat{\theta}_1 \,|\, \mathcal{M}=1) \\ &\quad + p_1\left\{\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0) - \mathrm{E}(\hat{\theta}_1 \,|\, \mathcal{M}=1)\right\}^2 . \qquad (4)\end{aligned}$$

The combination of the variances, $p_0\mathrm{var}(\hat{\theta}_0 \,|\, \mathcal{M}=0) + p_1\mathrm{var}(\hat{\theta}_1 \,|\, \mathcal{M}=1)$ is estimated by its sample version $\hat{s}^2_{\mathcal{M}} = I_0\hat{s}^2_0 + I_1\hat{s}^2_1$ with bias even when $\hat{s}^2_1$ is unbiased, although the bias is usually not substantial. However, the last term in (4) makes a sizeable contribution to the bias of $\hat{s}^2_{\mathcal{M}}$, unless the conditional expectations $\mathrm{E}(\hat{\theta}_0 \,|\, \mathcal{M}=0)$ and $\mathrm{E}(\hat{\theta}_1 \,|\, \mathcal{M}=1)$ are similar or model 0 is selected very frequently. This points to a weakness of model selection: when the conditional means $\mathrm{E}(\hat{\theta}_m \,|\, \mathcal{M}=m)$ differ and the reduced model 1 is quite likely to be selected (p_1 is large), $\hat{\theta}_{\mathcal{M}}$ is both biased and its MSE is underestimated.

We cannot assess from (4) in general whether and when a parameter is estimated after model selection at least as efficiently as by one of the constituent estimators. The discussion is greatly simplified when the indicators I_0 and $I_1 = 1 - I_0$ are independent of the constituent estimators $\hat{\theta}_0$ and $\hat{\theta}_1$, e. g., when I_1 and $(\hat{\theta}_0\,, \hat{\theta}_1)$ are based on independent (data) sources. Then

$$\mathrm{MSE}(\hat{\theta}_{\mathcal{M}}) \;=\; p_0V_0 + p_1(V + d^2)\,,$$

a convex combination of $\mathrm{MSE}(\hat{\theta}_0)$ and $\mathrm{MSE}(\hat{\theta}_1)$. Hence, the selected-model based estimator cannot be more efficient than both its constituents. It is little comfort that it is superior to the worse of them. Examples can be constructed

in which the random selection between models 0 and 1 is superior to the selection by the appropriate hypothesis test.

Since the coefficient b^* has to be estimated, the synthetic estimator $\tilde{\theta}(\hat{b}^*)$ does not necessarily outperform both $\hat{\theta}_0$ and $\hat{\theta}_1$. Of course, its properties depend on how b^* is estimated. Suppose we substitute for b^* an 'incorrect' value $b^\dagger \in (0,1)$. The estimator $\tilde{\theta}(b^\dagger)$ is more efficient than both $\hat{\theta}_0$ and $\hat{\theta}_1$ when $\mathrm{MSE}\{\tilde{\theta}(b^\dagger)\} = V_0 - 2b^\dagger(V_0 - C) + b^{\dagger\,2}(V_0 + V - 2C + d^2)$ is smaller than both V_0 and $V + d^2$. The solution of these two inequalities is

$$2b^* - 1 \; < \; b^\dagger \; < \; 2b^* \, .$$

If $b^\dagger \in (0,1)$, only one of these inequalities is relevant; the first when $b^* > \frac{1}{2}$, and the second otherwise. Although they suggest that a modicum of error in setting $b^\dagger$ is tolerated, the consequences of a minor error $|\, b^\dagger - b^* \,|$ can be serious when b^* is close to zero or unity.

In a typical setting, $\hat{\theta}_0$ is an unbiased estimator based on a valid model, possibly with too many parameters, and $\hat{\theta}_1$ is a possibly biased estimator based on a submodel which may be invalid. So $\hat{\theta}_1$ has a smaller variance. Then $V_0 - C > 0$, and so $b^* > 0$ and $\mathrm{MSE}\{\tilde{\theta}(b)\}$ is a decreasing function of b in the right-hand-side neighbourhood of $b = 0$. Hence, we can always improve on $\hat{\theta}_0$ by using $(1-b)\hat{\theta}_0 + b\hat{\theta}_1$ with a small positive b; note the connection with ridge regression [3]. If $\tilde{\theta}(1)$ is obviously inefficient because of an overwhelming bias, we can protect our inference about θ by underestimating b^*, preferring the error $\hat{b}^* < b^*$ to its converse.

We can combine M estimators in stages; first combining $\hat{\theta}_0$ with $\hat{\theta}_1$ to obtain $\tilde{\theta}_{01}$, then combining $\tilde{\theta}_{01}$ with $\hat{\theta}_2$, and so on. We refer to this as step-wise synthesis. We do not recover the synthetic estimator formed by the $M+1$ candidate estimators $\hat{\theta}_m$ directly because the relative sizes of the coefficients in $\mathbf{b}^*$ are altered with the inclusion of a new estimator $\hat{\theta}_m$ in direct synthesis, but they are fixed in the step-wise synthesis.

A model-selection procedure is coherent *if* the selection is correct with certainty. Whether we select from $M+1$ models directly, or by pairwise comparisons, the result would be the same. However, when each elementary selection is subject to the two types of error, the coherence is lost, and the details of the selection algorithm matter. By the same token, if the coefficients in $\mathbf{b}^*$ have to be estimated, the advantage of combining more single-model based estimators may be reversed by introducing more 'noise', as $\mathbf{b}^*$ has more components and involves more parameters and their functions.

3.2 What is better?

[8] and [9] demonstrated on simple examples of prediction with ordinary regression that the synthetic estimator $\tilde{\theta}(b^*)$, although not uniformly more efficient than $\hat{\theta}_{\mathcal{M}}$, does not have the glaring deficiencies of the selected-model based estimator. The synthetic estimator with known coefficients b^*

gives a lower bound for the MSE of $\tilde{\theta}$, and the smaller of the MSEs of $\hat{\theta}_0$ and $\hat{\theta}_1$ gives a lower bound for $\mathrm{MSE}(\hat{\theta}_{\mathcal{M}})$. The fact that $\mathrm{MSE}\{\tilde{\theta}(b^*)\} < \min\left\{\mathrm{var}(\hat{\theta}_0),\, \mathrm{MSE}(\hat{\theta}_1)\right\}$ does not warrant the conclusion that synthetic estimation is (uniformly) more efficient in all settings.

Selected-model based estimation has a potential for improvement by choosing better algorithms (rules) for selection, whereas synthetic estimation may benefit from more efficient estimation of the coefficient b^*. This can be strengthened by incorporating external information, a source that is much more difficult to exploit in model selection. A quantum improvement in model selection is achieved by combining the single-model based estimators according to the quantity of evidence in support of each model [5] — this is *model averaging*. Synthetic estimation exceeds this standard in two aspects: instead of weighting by the quantity of evidence, minimum MSE is the arbiter for weighting, and the weight assigned to each model is specific for the target; synthesis can be described as *estimator averaging*. See [8] for an example of a patently inappropriate model being assigned a large weight.

Synthetic estimation is more difficult to implement with complex models when analytical expressions for the bias d and (co-)variances V_0, V and C are not available. Further difficulties arise when there are many candidate models, because a large number of terms involved in the matrix $\mathbf{Q}$ and vector $\mathbf{P}$ have to be evaluated or estimated, and the system of linear equations $\mathbf{Qb} = \mathbf{P}$ solved. We need not be concerned about the ill-conditioning or singularity of $\mathbf{Q}$ because that merely indicates that different convex combinations of the single-model based estimators have similar or identical MSEs. A much more serious concern is the impact of the uncertainty about $\mathbf{P}$ and $\mathbf{Q}$ on the estimated vector of optimal coefficients $\hat{\mathbf{b}}^* = \hat{\mathbf{Q}}^{-1}\hat{\mathbf{P}}$. This problem can be addressed, in principle, by inflating the diagonal and reducing the off-diagonal elements of $\mathbf{Q}$.

The dimensionality of the problem, given by the number of candidate models, can be reduced by step-wise synthesis, with models organised in groups. In the first step, models are combined within groups, and in the second the within-group synthetic estimators are combined. Of course, some efficiency is lost as the flexibility in combining the single-models based estimators is reduced. This approach is yet to be explored as a practical alternative to step-wise procedures involving several model-selection steps.

The established application of selected-model based estimators is grossly misleading in many settings and, just as for synthetic estimation, its properties can be established only by simulations. The wider the range of candidate models the greater the impact of conditioning and the greater the bias of the selected-model based estimator as well as of the conventionally reported estimator of its MSE. In simple settings, examples can be found in which model selection is detrimental — $\hat{\theta}_{\mathcal{M}}$ is less efficient than both its constituent estimators; see [8]. A theoretical weakness of $\hat{\theta}_{\mathcal{M}}$ is that the process of model selection is not informed by the use and purpose of the selected-model based

estimator. In this aspect, synthetic estimation is more flexible. Although 'different weighting for different targets' may disagree with our instincts regarding the uniqueness of the most appropriate model, we should value efficient estimation (or a direct effort to achieve it) higher.

4 Example

We compare selected-model based prediction with synthetic prediction on an example of logistic regression with a single covariate. The analysed data originate from an information technology experiment; the data custodian has requested their identity and the context of the data not to be disclosed. The outcome y is a dichotomous variable, in response to the stimulus represented by a single covariate x. The data comprise $n = 1000$ records, with x approximately uniformly distributed in $[0, 10]$; its values are integer-multiples of 0.1. The logistic regression yields the fit $0.0070 + 0.0464x$ with the standard error for the slope 0.0221. The conventional interpretation of the t-ratio $0.0464/0.0221{=}2.10$ is that the fitted model cannot be reduced to $\mathrm{P}(y = 1) = \text{const}$. The likelihood ratio is equal to 4.44, leading to the same conclusion. A graphical summary of the data is given in Figure 1 in terms of the proportions of 'successes' ($y = 1$) within the one-point bands [0,1), [1,2), ..., [9,10] of values of x. These proportions are represented by crosses $\times$; the various lines in the plot are discussed below. The bands contain between 91 and 109 observations. The standard deviation of 100 independent binary outcomes with probability 0.5 each is 0.05, so the deviations from the logistic regression (the thick dashes in Figure 1) do not contradict it.

Our targets are the probabilities of success, $p(x) = \mathrm{P}(y = 1 \mid x)$, for the integer values $x = 0, 1, \ldots, 10$. We assume that the logistic regression on x is valid, but consider as an alternative the constant-probability model which predicts $p(x)$ by the sample proportion $\overline{y} = (y_1 + \cdots + y_n)/n$. For synthetic estimation, we require the following quantities:

$$
\begin{aligned}
V_0 &= \mathrm{var}\{\hat{p}(x)\} \doteq [p(x)\{1 - p(x)\}]^2 \, \mathbf{x} \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{x}^\top \\
V &= \mathrm{var}(\overline{y}) = \frac{1}{n^2} \sum_{i=1}^{n} p_i (1 - p_i) \\
C &= \mathrm{cov}\{\hat{p}(x), \overline{y}\} \doteq \frac{1}{n} p(x)\{1 - p(x)\} \mathbf{x} \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{1} \\
d &= \mathrm{E}(\overline{y}) - p(x) = \overline{p} - p(x) \,, \qquad (5)
\end{aligned}
$$

where $\mathbf{x} = (1, x)$, $\mathbf{X}$ is the regression design matrix comprising the columns $\mathbf{1}$ and the values of x_i, $i = 1, \ldots, n$, $\mathbf{W} = \mathrm{diag}_i\{p_i(1-p_i)\}$ is the diagonal matrix of the iterative weights, $p_i = p(x_i)$, and $\overline{p} = (p_1 + \cdots + p_n)/n$ is the average probability over the design points x_i. The approximations in (5) are obtained

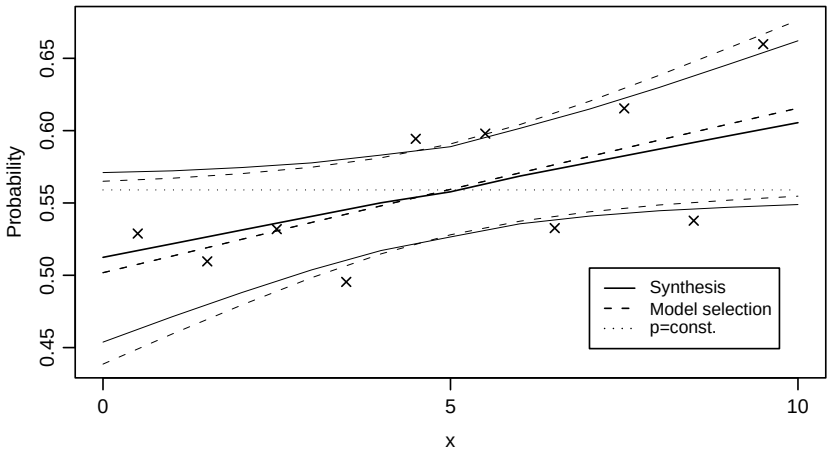

Fig. 1. Selected-model and synthetic prediction. The predictions are represented by thicker lines and the upper and lower pointwise 95% confidence limits by thinner lines.

by the Taylor expansion. In our case, the expansion is quite precise because the curvature of the logit function at the anticipated values of $p(x)$ is only slight, as is evident from Figure 1. The quantities in (5) are estimated naively. From them we obtain the estimates of the coefficients b^* listed in Table 1. They are in a narrow range, (0.177, 0.194), except for prediction at $x = 5$. For $x = 5$, very close to the mean of the values of x in the data ($\overline{x} = 4.98$), the two single-model based predictions almost coincide, so the coefficient b^* is immaterial. The difference between the predictions $\hat{p}_{01}$ and $\tilde{p}$, around 1% for x equal to 0 and 10, although not dramatic, is substantial in the particular context.

The conventionally estimated standard error of the selected-model based prediction, $\sqrt{\widehat{\mathrm{var}}(\hat{p}_{01})}$ in Table 1, is between 0–7.5% greater than the naively estimated root-MSE (rMSE) for the synthetic prediction, $\sqrt{\widehat{\mathrm{MSE}}(\tilde{p})}$. Both these estimators of standard error are optimistic; one because of conditioning on the selected model, and the other because it ignores the uncertainty about b^*. The model-selection based and synthetic predictions and their pointwise 95% confidence limits are drawn in Figure 1.

The simulations described in the next section indicate that $\widehat{\mathrm{var}}(\hat{p}_{01})$ underestimates its target $\mathrm{MSE}(\hat{p}_{01})$ by more than $\widehat{\mathrm{MSE}}(\tilde{p}) = \mathrm{MSE}(\tilde{p} \,|\, b^* = \hat{b}^*)$ underestimates $\mathrm{MSE}\{(\tilde{p}(\hat{b}^*)\}$. Thus, the differences in the precisions of $\hat{p}_{01}$ and $\tilde{p}$ are somewhat greater than implied by estimated standard errors in Table 1.

Table 1. Selected-model based and synthetic prediction.

	Selected-model		Synthesis		
x	$\hat{p}_{01}$	$\sqrt{\widehat{\text{var}}(\hat{p}_{01})}$	$\tilde{p}$	$\sqrt{\widehat{\text{MSE}}(\tilde{p})}$	$\hat{b}^*$
0	0.5018	0.0316	0.5124	0.0293	0.1857
1	0.5133	0.0269	0.5218	0.0252	0.1861
2	0.5249	0.0227	0.5313	0.0216	0.1865
3	0.5365	0.0191	0.5407	0.0185	0.1875
4	0.5480	0.0166	0.5501	0.0165	0.1935
5	0.5594	0.0157	0.5577	0.0156	1.0000
6	0.5708	0.0167	0.5686	0.0165	0.1877
7	0.5821	0.0191	0.5779	0.0185	0.1825
8	0.5934	0.0224	0.5872	0.0213	0.1805
9	0.6045	0.0263	0.5964	0.0247	0.1788
10	0.6155	0.0304	0.6055	0.0283	0.1770

4.1 Simulations

The properties of the selected-model based predictor $\hat{p}_{01}(x)$ and the synthetic predictor $\tilde{p}(x;\hat{b}^*)$ can be established by simulations which evaluate the empirical rMSEs of the predictors of $p(x)$ as functions of x for a plausible logistic regression. The four predictors, based on the logistic regression, the sample proportion, the selected model and the synthetic predictor, are about equally efficient at $x = 5$, the mean of the values of the regressor. With increasing distance from $\overline{x}$, the differences increase. The selected-model based predictor is more efficient than the synthetic predictor for small values of the slope β_1 (for about $\beta_1 < 0.015$). For steeper slopes β_1, the synthetic estimator is superior. The difference in rMSE for a given x first increases, till about $\beta_1 = 0.04$, and then decreases till about $\beta_1 = 0.08$. From then on, the differences in rMSEs based on model 0, selected model and the synthesis are very small, although the sample proportion (model 1) becomes less and less efficient. This is an expected outcome because, with increasing slope, any model selection would gravitate towards model 0, and synthesis also tends to assign increasing weight to model 0. In summary, when the selected-model based estimator is more efficient than the synthetic estimator it is so only narrowly, whereas for some values of β_1 synthesis is substantially more efficient.

The reported rMSE underestimates the rMSE of the selected-model based estimator, and the rMSE of the ideal synthesis underestimates the rMSE of the synthesis with $\hat{b}^*$. However, the underestimation by the ideal-synthesis rMSE is much smaller; details are omitted. The results are very similar to those in [8], where a similar simulation exercise is described for prediction with a simple regression model. With our logistic regression model, the dependence

of the iterative weights $p_i(1-p_i)$ on the linear predictor $\mathbf{x}\boldsymbol{\beta}$ is very weak because all the predicted probabilities are in the proximity of 0.5.

5 Conclusion

We have described an approach to dealing with model uncertainty, in which no model selection is used but the estimators based on the candidate models are combined. As a consequence, even when we are certain about the model, some of its submodels, even if invalid, may yield more efficient estimators. The derivation of the synthetic estimator suggests that we should abandon the maximum likelihood under the most parsimonious valid model as the standard for efficient estimation. The standard aspired to by synthetic estimation is the most efficient convex combination of a (general) valid model and of its submodels. It is a higher standard and it incorporates rewards for better prior information, when a narrower range of models is identified *a priori*.

In summary, 'Which model?' is the right question in many settings, but a response by a model, whichever way it is selected, is inappropriate. Pretending that we have identified the model correctly causes a distortion in the conventional inferential statements. We ignore uncertainty, whatever its source, at our peril; see [6].

Small-area estimation is a successful application of synthesis outside the realm of modelling; see [2] and [7].

References

1. Draper, D. N. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society* Ser. B **57**, 45–98.
2. Fay, R. E., and Herriot, R. A. (1979). Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.
3. Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–67.
4. Hoeting, J., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 381–417.
5. Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.
6. Lindley, D. V. (2000). The philosophy of statistics (with comments). *Journal of the Royal Statistical Society* Ser. D (*The Statistician*) **49**, 293–337.
7. Longford, N. T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society* Ser. A **162**, 227–245.
8. Longford, N. T. (2003). An alternative to model selection in ordinary regression. *Statistics and Computing* **13**, 67–80.
9. Longford, N. T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician.* Springer-Verlag, New York.