Stat Comput
DOI 10.1007/s11222-014-9522-9

Does data splitting improve prediction?

Julian J. Faraway

Received: 5 April 2013 / Accepted: 4 October 2014

Abstract Data splitting divides data into two parts. One part is reserved for model selection. In some applications, the second part is used for model validation but we use this part for estimating the parameters of the chosen model. We focus on the problem of constructing reliable predictive distributions for future observed values. We judge the predictive performance using log scoring. We compare the full data strategy with the data splitting strategy for prediction. We show how the full data score can be decomposed into model selection, parameter estimation and data reuse costs. Data splitting is preferred when data reuse costs are high. We investigate the relative performance of the strategies in four simulation scenarios. We introduce a hybrid estimator that uses one part for model selection but both parts for estimation. We argue that a split data analysis is prefered to a full data analysis for prediction with some exceptions.

Keywords Cross-validation · Model assessment · Model uncertainty · Model validation · Prediction · Scoring

1 Introduction

Predictions based on statistical models are sometimes disappointing. We do not expect predictions to be exactly correct so we compute measures of uncertainty but the observed outcomes may lay well outside these expressions of the expected variation. Some failures cannot be anticipated as unexpected results can occur because the system generating the observables has changed since we collected the data used to develop the model. Alternatively perhaps our data was not broad enough to include more extreme events leading us to fail to anticipate such outcomes in our predictive model. However, sometimes the problem lies not in the data but in the statistical methods used.

In many applications, we do not know or choose to specify in advance the particular form of model to be used. We frequently use the data to select the model. Once we have selected a model, we often proceed to estimate the parameters and make predictions with an assessment of variability that reflects the uncertainty about the parameters but not about the model. This lapse is the source of many statistician-generated failures with prediction. See Berk et al. (2009) and Wit et al. (2012) for recent discussion of these issues and, less recently, in Chatfield (1995) and Pötscher (1991).

How should predictions be judged? Suppose we are asked to predict y given x. We are provided with training data consisting of (x, y) pairs from which we develop a model and estimate parameters. We are now given new x and asked to make predictions $\hat{y}$. We then observe future y_0 and could judge our success by the closeness of $\hat{y}$ to y_0. Statisticians are aware of the problem of overfitting, namely using too complex a model resulting in a misleadingly good fit to the training data. We have become skilled at avoiding this problem by balancing fit against complexity using methods like AIC. Regularized and penalized regression are also among the popular solutions that avoid the problem of overfitting and enjoy good point prediction performance.

Now suppose we are required to express the uncertainty in our predictions, perhaps in the form of a prediction interval or more generally as a predictive distribution. As Statisticans, we take pride in our ability to provide such uncertainty assessments. To see how well we are doing, we might check whether 95 % of the observed future responses fall within their 95 % prediction intervals. More elegantly, we could use scoring

J. J. Faraway (✉)
Department of Mathematical Sciences, University of Bath,
Bath BA2 7AY, UK
e-mail: jjf23@bath.ac.uk

Published online: 29 October 2014

to judge how well the observed future responses match with the predictive distributions. All too often we find that our predictions are too optimistic because the observed future responses vary more than we allowed for. Avoiding overfitting is helpful but the primary source of the problem is the uncertainty assessments.

Using the same data to both select and estimate the parameters has long been recognised as generating overoptimistic inference but practitioners frequently do little to address it. One reason for this lack of action is the difficulty in doing anything about the problem. One approach is to integrate the model selection and parameter estimation. For example, the Box–Cox method selects the index of transformation of the response in a regression problem. Instead, we might regard this as parameter estimation rather than model selection and proceed accordingly as in Hinkley and Runger (1984). Another example is the LASSO method which succeeds in combining regression variable selection with estimation although getting the subsequent inference correct is problematic—see Belloni and Chernozhukov (2013). For more realistic complex model selection procedures that involve a combination of procedures and use both graphical and numerical approaches, a holistic approach to inference seems impractical. A Bayesian approach that assigns priors to models as well as parameters is possible as in Draper (1995) but the approach becomes unworkable unless the space of models is small. In many cases, the space of models cannot be reasonably specified before analysing the data. Another idea is to use resampling methods to account for model uncertainty as in Faraway (1992). However, this method requires the model selection process to be pre-specified and automated. It also requires that these processes be implementable completely in software which excludes the possibility of human judgement in model selection.

There is a simple solution—data splitting. We divide the data into two, not necessarily equal, parts. One part is used to select the model and the other part is used to estimate the parameters of the chosen model. The problem of overconfidence in the chosen fit is avoided because we use fresh data to estimate the parameters of the model. Data splitting can easily be used in a wide range of problems and requires no special software. The analyst is free to select the model using any procedure which can involve subjective elements and need not be specified in advance. Unfortunately, there is a drawback. Less data is used to select the model so we can expect some reduction in the probability we select a better model. Furthermore, less data is used to estimate the parameters of the model so we can expect some additional uncertainty in prediction. Will the gain from avoiding data re-use compensate for the loss in model selection and parameter estimation performance? An answer to this question is the purpose of this paper.

Stone (1974) provides an early history of data splitting. Dawid (1984) discusses the related but different issue of updating models as new data arrive sequentially. It may seem that having more data should improve the model but Meng and Xie (2013) show that this is not necessarily true. This is another example where using all the data symmetrically may not give the best results.

In Sec. 2, we discuss the uses of data splitting and the methods we use to evaluate its effectiveness. We propose a decomposition of prediction performance into model selection, parameter estimation and data resuse costs. In Sec. 3, we present simulations to explore the relative value of data splitting. Sec. 4 contains our conclusions.

2 Methods

2.1 Model validation

Data splitting has been used for a variety of purposes. Having chosen and estimated a model, we may wish to evaluate how well the model can be expected to perform in practice. Under this setting, the first part of the data is used for model selection and estimation and second part is used to generate a measure of how well the model will predict future observations. Data splitters Picard and Cook (1984), Picard and Berk (1990) and Roecker (1991) have an objective of estimating the average (over a subset of the predictor space) mean squared error of prediction. The aim for these authors is to obtain a better estimate of the quality of future predictions. Certainly, these methods will obtain more realistic estimates of this quality than using the naive estimate provided by the full data fit but this will come at a substantial price. Because less data is used for model selection and estimation, the prediction quality we are trying to estimate will be itself degraded.

In contrast to selection-estimation data splitting, where we hope the splitting might result in better prediction, validation data splitting will almost surely make predictions worse. It depends how useful we find the estimate of prediction quality as to whether this loss is worthwhile. In some cases, analysts are in competition to produce the best predictive model. For example, in the Netflix prize competition (Bell and Koren 2007) it was essential to hold back part of the data to evaluate the performance of the model since the competitors estimates of performance could not be trusted. Another example is internet business Kaggle which challenges analysts to develop the best predictive method (Carpenter 2011). In other cases, there may be some regulatory requirement to assess model performance.

However, unless there is some reason not to entirely trust the analyst, there are better ways to estimate predictive performance than the naive method or validation data splitting. Simple methods like tenfold cross-validation can be used

which, while not perfect, will give more realistic estimates of future model prediction performance. Indeed, more recent authors discourage the use of data splitting for this purpose. Steyerberg (2009), Schumacher et al. (2007) and Molinaro et al. (2005) all advise against data splitting although the latter two references investigate cases where the number of variables is large relative to the number of observations. In such cases one call ill-afford to give up observations for validation purposes.

This form of data splitting reserves part of the current data for validation purposes but the true test comes when genuinely new data arrive as Hirsch (1991) points out. Altman and Royston (2000) also points out the distinction between statistical and clinical validity. All this is true but we shall restrict our concern to the data we have at hand and do the best we can with that. Some authors have proposed splitting data into three parts where one part is used for selection, the second for estimation and third for validation. These include Miller (1990, p. 13), Mosteller and Tukey (1977) and Friedman et al. (2008). Some have used data splitting for quite different reasons—see Heller et al. (2009). We have focussed on the problem of prediction but questions also arise in hypothesis testing—see Cox (1975) and Dahl et al. (2008)

2.2 Scoring

We need a measure of quality for predictions. We might make a prediction $\hat{y}$ and subsequently observe y_0 and judge performance based on the difference between the two, averaged over repeated instances. Root mean squared error (RMSE) is the most commonly used measure of this type. RMSE uses only the point estimate but ignores any expression of uncertainty in the prediction such as standard errors or prediction intervals. Sometimes we care only about point performance and RMSE is sufficient. In such situations, there is no point in reserving any part of the data for improving expressions of uncertainty and typically all the data should be used for both model selection and parameter estimation unless validation is required. But if we do care about assessments of uncertainty, we need some way to measure this aspect of prediction performance.

Scoring is a method of judging the quality of a predictive distribution compared to actual observations. Suppose we specify a predictive distribution $f(y)$ and we subsequently observe y_0 then we assign a score of $-\log f(y_0)$. Log scoring was proposed by Good (1952) and has been used commonly in fields such as weather forecasting where we must often judge the accuracy of forecasts. The idea is illustrated in Fig. 1, where we see two predictive distributions that have the same mean. The solid-lined predictive distribution scores better when we observe A, a value close to the mean prediction. But the dashed-line predictive distribution scores more highly when we observe B which is further from the mean. Hence scoring will reward predictive distributions that express the uncertainty in prediction more faithfully. RMSE uses only point prediction and would not distinguish between these two distributions.

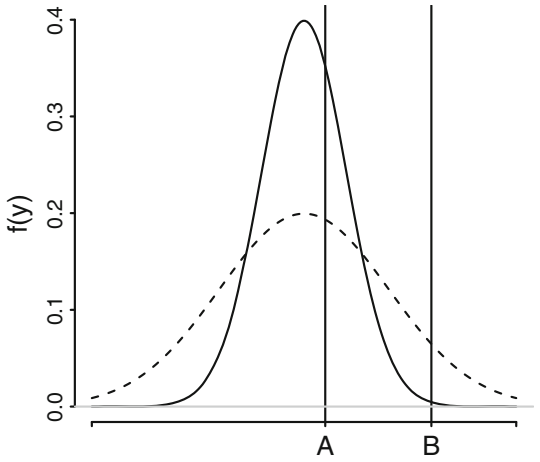

Fig. 1 Scoring performance for two predictive distributions compared for observed A and B

Other scoring rules have been proposed and discussed by Gneiting and Raftery (2007) but log scoring appears most appropriate for this task as it is analogous to a log likelihood for new observations and the estimators we use in the simulations either are or are closely related to maximum likelihood estimators. It is a proper scoring rule as defined by Parry et al. (2012).

Log scoring is also used for binary response prediction where we use the probability rather than the density. This also provides an intuitive way of interpreting a difference in scores. For example, a difference of 0.1 would represent a difference of about 10 % in predicting the correct result. Interpretation for continuous responses can be made using a difference in log-likelihoods with the discrete case providing some intuition for the interpreting the magnitude of these differences.

2.3 Decomposition of performance

Consider a data generating process $\mathcal{Z}$ which produces pairs (X, Y). We are given n draws from $\mathcal{Z}$ from which to construct a model that provides a predictive distribution for Y given a value from X. We will subsequently be given new values from X for which we are asked to construct predictive distributions for future values of Y. Suppose we entertain a set of models $\{M_i\}$ indexed by i that provide for such predictive distributions. The set of models might be infinite in number and might not contain the true data generating model. We have a model selection procedure S such that given data of size n, Z_n, $S(Z_n) = M_i$ indicating that given data Z_n, model M_i is selected. The parameters of M_i will also need to

be estimated by some other procedure which is distinct from S.

We evaluate prediction performance by finding the expected score, $E_{\mathcal{Z}}g(Z, M)$, where $g(Z, M)$ is the score for data Z and model M. The parameters of M will be estimated (or derived) using Z. The model may be specified seperately or it may be selected using Z. The expectation needs to be computed over two independent sets of draws from $\mathcal{Z}$. First, there are the observations Z_n used to select (perhaps) and estimate the parameters and second, there are the future draws from $\mathcal{Z}$ used to score the performance of the chosen model. To evaluate S for sample size n, we need to find $E_{\mathcal{Z}}g(Z_n, M = S(Z_n))$.

We now decompose this measure of performance into four parts, respectively, best possible performance, model selection cost, parameter estimation cost and data re-use cost as seen in Eq. 1. Let Z_∞ represent a black box with the ability to generate datasets of unlimited size from the generating process but not explicit knowledge of this process. Z_∞ allows the best possible parameter estimation. We set $S(Z_\infty) = M^*$ where M^* is the best possible model chosen from those available. If the true model is contained within the set of possible models considered and we use a consistent model selection procedure, then M^* will be the true model. When the true model is not found in M or when the model selection process is deficient in some crucial way, M^* is not the true model. In the true model case, the true parameters will be used in M^* provided the parameter estimation is consistent. In other words, we would have found $\mathcal{Z}$. Where M^* is not the true model, the parameters chosen will be the limiting values derived from the estimation process as $n \to \infty$. Also let Z_n^F be another sample independent of Z_n. Z_n^F is only used to estimate parameters as the model will already have been chosen.

There are some potential technical difficulties with the existence of limits involving the choice of models and estimation of the parameters of incorrect models. These can arise in situations where the set of feasible models grows with the sample size or for unusual model selection strategies. We shall not tackle these issues here because our findings are based on simulation and these theoretical problems do not arise. Our presentation in this section is heuristic.

$$\begin{aligned} E_{\mathcal{Z}}g\left(Z_n, M = S(Z_n)\right) &= E_{\mathcal{Z}}g\left(Z_\infty, M^*\right) \\ &+ E_{\mathcal{Z}}I(S(Z_n) = M)\left[g(Z_\infty, M) - g\left(Z_\infty, M^*\right)\right] \\ &+ E_{\mathcal{Z}}I(S(Z_n) = M)\left[g(Z_n^F, M) - g(Z_\infty, M)\right] \\ &+ E_{\mathcal{Z}}I(S(Z_n) = M)\left[g(Z_n, M) - g(Z_n^F, M)\right] \end{aligned} \tag{1}$$

Note that the terms on the RHS cancel so the equation is true in a trivial sense but our interest lies in the meaning of the terms. In general, these components would be very difficult to compute theoretically but are quite accessible by simulation. We present the simulation equivalents of these terms first before discussing the meaning. In the simulation, we can draw n_r samples Z_n^j of size n from Z for $j = 1, \ldots, n_r$. For each sample, we apply S to select and estimate a model. To compute the scores, we will draw future observation samples of size n_e from Z where $n_e >> n$. The performance can then be estimated using:

$$\begin{aligned} &\frac{1}{n_r}\sum_j g\left(Z_n^j, S\left(Z_n^j\right)\right) \\ &= \frac{1}{n_r}\left\{\sum_{\text{models } i}\ \sum_{j:S(Z_n^j)=M_i} g\left(Z_n^j, S\left(Z_n^j\right) = M_i\right)\right\} \end{aligned} \tag{2}$$

In computing g, we need to average over the large sample of future observations of size n_e. We have kept this implicit in the expressions for compactness. On the RHS, we have broken up the sum over the set of models selected. Let p_i be the proportion of cases where model M_i is selected, then we can estimate the four components using:

$$\begin{aligned} &\frac{1}{n_r}\sum_j g\left(Z_n^j, S\left(Z_n^j\right)\right) = g\left(Z_\infty, M^*\right) \\ &+ \sum_{Models\ i} p_i\left\{g\left(Z_\infty, M_i\right) - g\left(Z_\infty, M^*\right)\right\} \\ &+ \sum_{Models\ i} p_i\left\{g\left(Z_n^F, M_i\right) - g\left(Z_\infty, M_i\right)\right\} \\ &+ \frac{1}{n_r}\sum_j\left\{g\left(Z_n^j, S\left(Z_n^j\right)\right) - g\left(Z_n^F, S\left(Z_n^j\right)\right)\right\} \end{aligned} \tag{3}$$

In the simulations to follow, the true model lies among the feasible set of models so there is no difficulty in specifying M^*. In such circumstances, we can use the true parameter values for M^*. Computing $g(Z_\infty, M_i)$ is more difficult since we are being asked to give the best parameters for the wrong model. We can approximate this by using a very large rather than infinite sample to estimate this. In cases where M^* is not the true model, the same sort of approximation could be used. In the second and third terms, the summands do not depend on the data Z_n^j so I have dropped the sum over j. Nevertheless, we do need simulation to compute these terms.

The first term, $g(Z_\infty, M^*)$, represents the best that can be done if we use the best model and we have an unlimited supply of training data denoted by the Z_∞. We have no particular interest in this quantity but it is useful for scaling purposes since it represents a bound on prediction performance.

The second term represents the model selection effect where more weight (p_i) put on sub-optimal models will increase the score (where high scores are bad). When a con-

sistent model selection process is used, this term will tend to zero as the sample size increases. The difference between the full and split data values for this term will be of the same order as the rate at which p_i tends to zero or one accordingly. The magnitude of the difference will depend also on the relative sizes of $g(Z_\infty, M_i)$ over $\{M\}$. The full data strategy will be favoured most when there is only one good model that is hard to find. In many cases, the best model will be accompanied by several almost as good models—in such cases splitting will not lose by much. Although the split data strategy will almost always trail the full data approach, the difference will not be large excepting situations with larger numbers of variables and smaller numbers of cases where the splitting of the data may cause the model selection to fail entirely.

The third term represents the cost of parameter estimation weighted by the choices of model. We have used $g(Z^F, M_i)$ to represent the empirically expected score under model M_i where fresh estimation data is available. We could estimate this using only the replications where M_i is selected but in the simulations to follow we can get a better estimate by computing it for all replications. The full data strategy will almost always produce lower scores for this component than the split data approach. The difference between the two will typically be $O(n^{-1/2})$. Excepting the situation where the split data is insufficient to estimate the parameters, the difference between full and split is thus stochastically bounded.

The fourth term represents the cost of data reuse. For the split data approach, this term will have expectation zero because the part of Z_n^j used to estimate the model is independent of the selection process. Z_n^F is simply another such independent sample of the same size. For the full data strategy, it is difficult to say anything about the likely size of this term because of the complex relationship between model selection and estimation. As we shall see, this term can be large and can easily outweigh the advantages the full data had in selection and estimation.

Thus the full data strategy will outperform the split data strategy on both model selection and parameter estimation but the difference is bounded and well understood because it is simply the effect of sample size. In contrast, the split data strategy will beat the full data strategy on data reuse cost. This cost could be very large. One must distinguish the problem of overfitting from that of model uncertainty. A statistician who tends to overfit will pay more in model selection and parameter estimation costs but not necessarily more for data reuse—consider the example of fitting too high an order in a polynomial regression. The statistician who wisely balances fit and complexity may reduce selection and estimation costs, but will still be exposed to data reuse costs when using a full data strategy.

2.4 Estimators

We consider four estimators based on data Z which is randomly split into a model selection set Z_1 and parameter estimation set Z_2 where the model selection set has size $[fn]$ for fraction f.

- FD: Full data estimate. Z is used to both build/select the model and to estimate the parameters of that chosen model.
- SD: Split data estimate. Z_1 is used to build/select the model and Z_2 is used to estimate the parameters.
- SAFE: (Split Analysis, Full Estimate). Z_1 is used to build/select the model and Z is used to estimate the parameters
- VALID: Validation corrected estimate: Z_1 is used to build/select the model and estimate the parameters used to generate point predictions. Z_2 is used to generate a new estimate of the standard error to be used in the construction of predictive distributions.

The SAFE estimator is motivated by the decomposition of model performance. We see that the SD strategy will under-perform the FD approach for parameter estimation. But conditional on the model selection, the SAFE estimator will avoid this loss because all the data is used for estimation. On the other hand the FD strategy is vulnerable to heavy losses due to data re-use. Because the SAFE strategy withholds a portion of the data from model building, we have some protection against severe over-confidence because data not used in the model selection is used for the estimation. SAFE cannot mitigate any losses due to model selection.

The VALID estimator is motivated by the validation approach to data splitting. The validation sample may well reveal the overconfidence of our original predictions. We can form a new estimate of the standard error using

$$\hat{\sigma}^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 / \left([(1-f)n] - 1\right)$$

where the sum runs over Z_2. Such an approach is straightforward for location-scale type models like the Gaussian linear model. For the binary response regression example we consider here, a VALID estimator makes no sense but may be appropriate where there is a richer class of models for such responses, as might be found in some Bayesian approaches. The VALID estimator will be effective for strategies that tend to overfit but it is difficult to see how it might affect data resuse costs. The original predictive distributions might also be improved in other ways such as the one suggested by Dawid (1984).

In Little (2006), a suggestion is made that the model estimation should be Bayesian but the model assessment and

checking should be Frequentist. This fits well with a data splitting approach because these activities can use different parts of the data.

3 Simulations

It would be nice to explore the effects of data splitting mathematically. Unfortunately, this is very difficult. Finite sample calculations are required since asymptotically the issue becomes practically irrelevant. In Leeb and Pötscher (2005), a calculation is made for the simple regression problem where the model selection is the determination of the significance of a single predictor. They demonstrate that problems with uniform convergence can arise for post-model selection estimators. Unfortunately, it becomes impractical to perform such calculations for richer and more realistic model selection scenarios. Hence, we resort to simulation. We consider four different regression modelling scenarios. In the first, we look at transforming the response, in the second, selecting the predictors and in the third we consider an outlier rejection method. Our final example looks at variable selection for binary response regression. We have deliberately kept the modelling options simple so that we can carefully explore design choices that might affect the value of data splitting. In practice, model building is usually more sophisticated and combines several elements, both formal and informal.

3.1 Box–Cox

The Box–Cox method selects the index of transformation on the response in a linear regression model. The simulation setup is $X_i \sim U(0, 1)$ and ϵ_i i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots n$ with

$$Y_i^{\lambda} = \alpha + \beta X_i + \epsilon_i$$

We fix $\alpha = 0$ although the term will be estimated. We observe X and Y but not λ (where $\lambda = 0$ is equivalent to $\log Y$), which we estimate using the Box–Cox method. In keeping with common practice, we select λ from a finite set of interpretable values, in this case $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$. The α and β are then estimated using least squares. In this example, the model selection is just the determination of λ.

We ran a simulation with $n_r = 4,000$ replications for each run using a full factorial design varying over all combinations of $\sigma = 0.1, 1, 10$, $\beta = 0, 1$, $n = 18, 48$, true $\lambda = -0.5, 0, 0.5$ and training fractions of $f = 1/3, 1/2$ and $2/3$.

We computed the scores as follows: for each selected model, we generated 4,000 realisations from the known true model which were used to score the fitted models (as we do in all subsequent simulations). We find the predictive distributions needed to compute the score using the predicted values and corresponding standard errors. There are more sophisticated ways to the construct frequentist predictive distributions, see for example Lawless and Fredette (2005), but we have consistently opted for the straightforward t-distribution scaled by the predicted mean and variance. For more discussion of the justification for such intervals, see Xie and Singh (2013).

In Fig. 2, we plot the difference in scores (FD–SD) , (FD-SAFE) and (FD-VALID). We see that FD is superior to SD in all combinations. The sample size has by far the largest effect on the margin of difference. For $n = 48$, we see that the difference in score is a relatively small 3–5 %. All the other factors have little consistent effect on the difference on the score differences although a training fraction of 1/2 seems a good overall choice. The SAFE estimator is an improvement over the SD estimator in all cases and is often close in performance to the FD estimator, particularly if the 1/3 training fraction case is avoided. This might be expected as model selection cost would be largest in the $f = 1/3$ case. The VALID estimator falls between the SAFE and SD estimators in performance.

Figure 3 shows the estimated relative contribution as described in (3) for the FD case. We estimate the model selection, parameter estimation and data re-use costs and plot the proportion of the cost attributable to each part. We average over the three training fraction runs at each level of the other simulation variables as this fraction has no impact on the FD estimator. Data reuse costs are minimal and parameter estimation costs account for a large share of the costs. Situations where the model selection costs are higher would be situations where the SAFE estimator might be expected to make the least improvement over SD—there is some evidence from the plots to support this.

In the Box–Cox scenario, the FD strategy wins every time. There is very little model selection taking place with only five to choose from. With a small amount of data, the selection and estimation components favour the FD approach quite clearly and the data reuse costs are not significant. For the larger dataset, the difference between full and split was relatively small.

3.2 Variable selection

Consider a model:

$$y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

where ϵ_i is i.i.d. $N(0, \sigma^2)$, $X_{ij} \sim U(0, 1)$, $\text{cor}(X_i, X_k) = \rho \quad \forall i \neq k$ and $\alpha = 0$ but is estimated. We follow a stepwise AIC-based variable selection procedure as described by the `step` function of Venables and Ripley (2002).

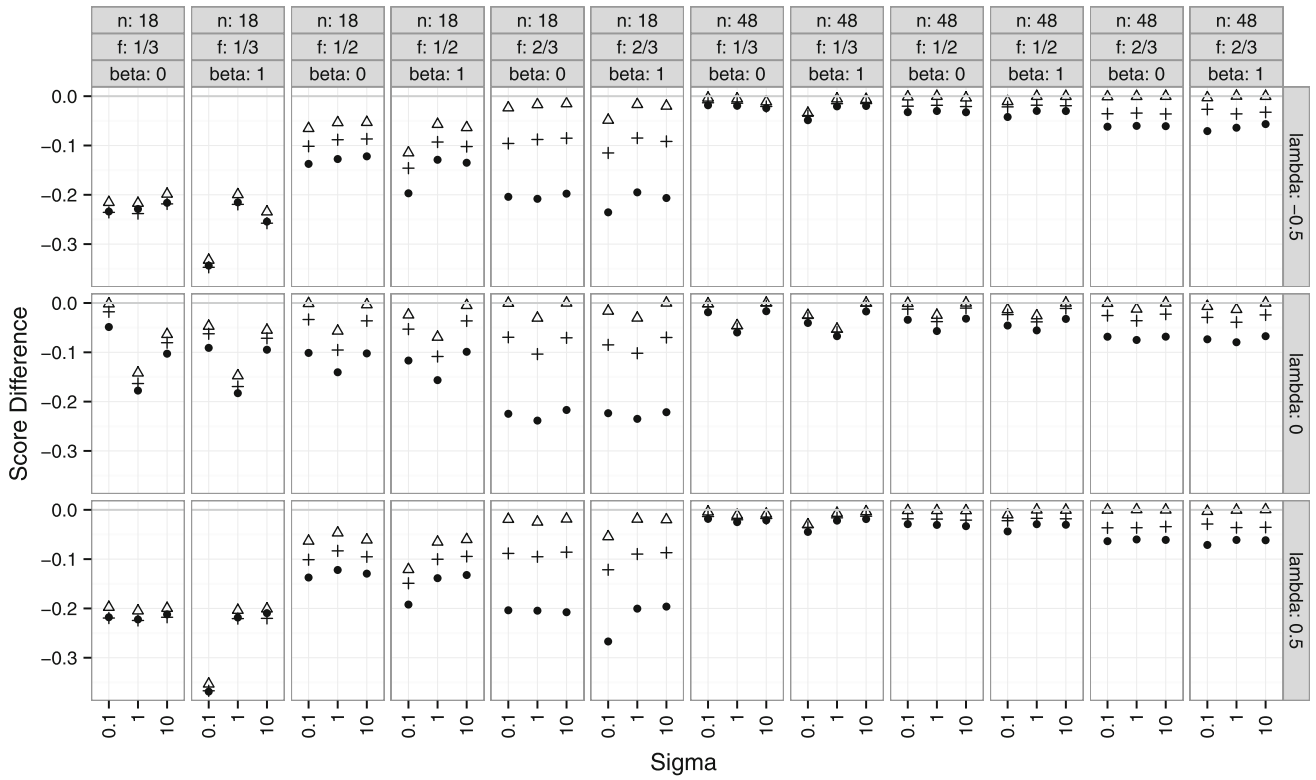

Fig. 2 Box–Cox simulation results showing the FD minus SD score as a solid dot, FD minus SAFE as an empty triangle and FD minus VALID as a plus symbol

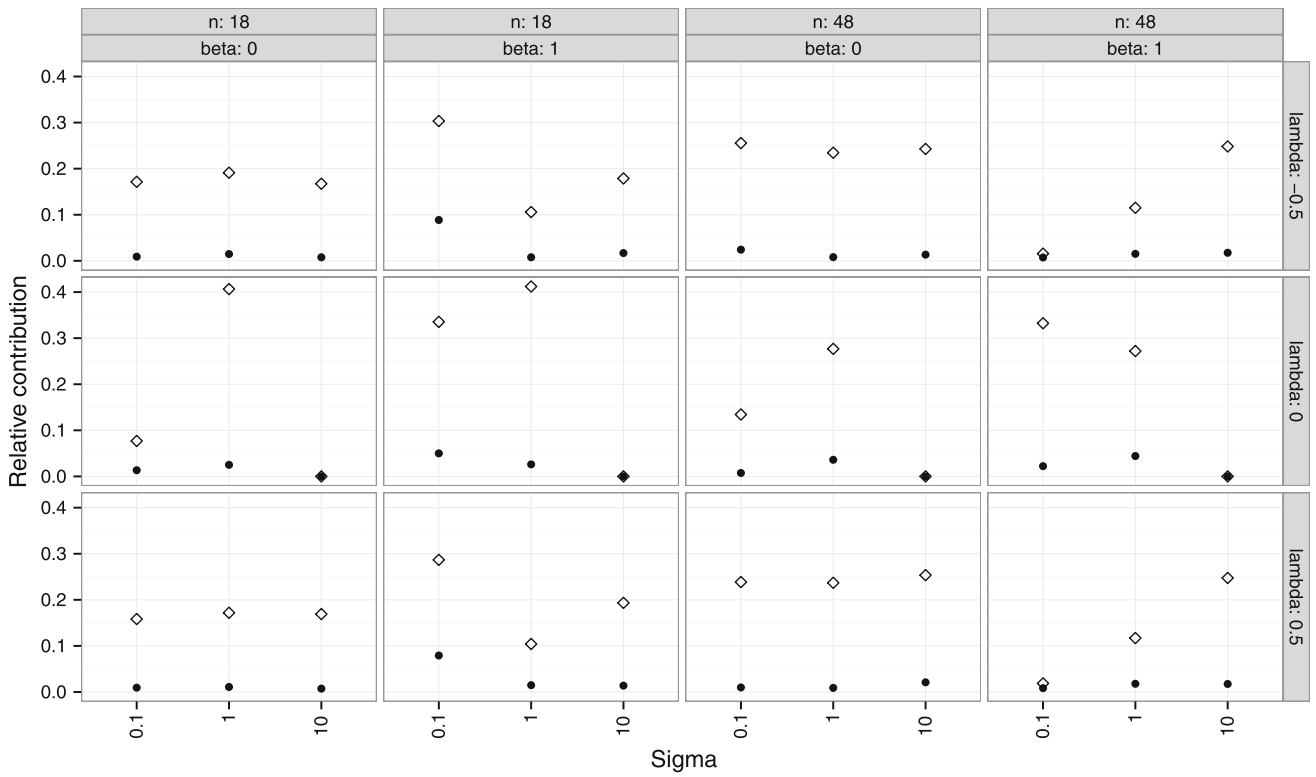

Fig. 3 Model performance decomposition for the Box–Cox scenario using FD. Solid dot represents data reuse relative contribution while open diamond represent data reuse plus the model selection contribution. Parameter estimation is the remaining part completing the fraction to one

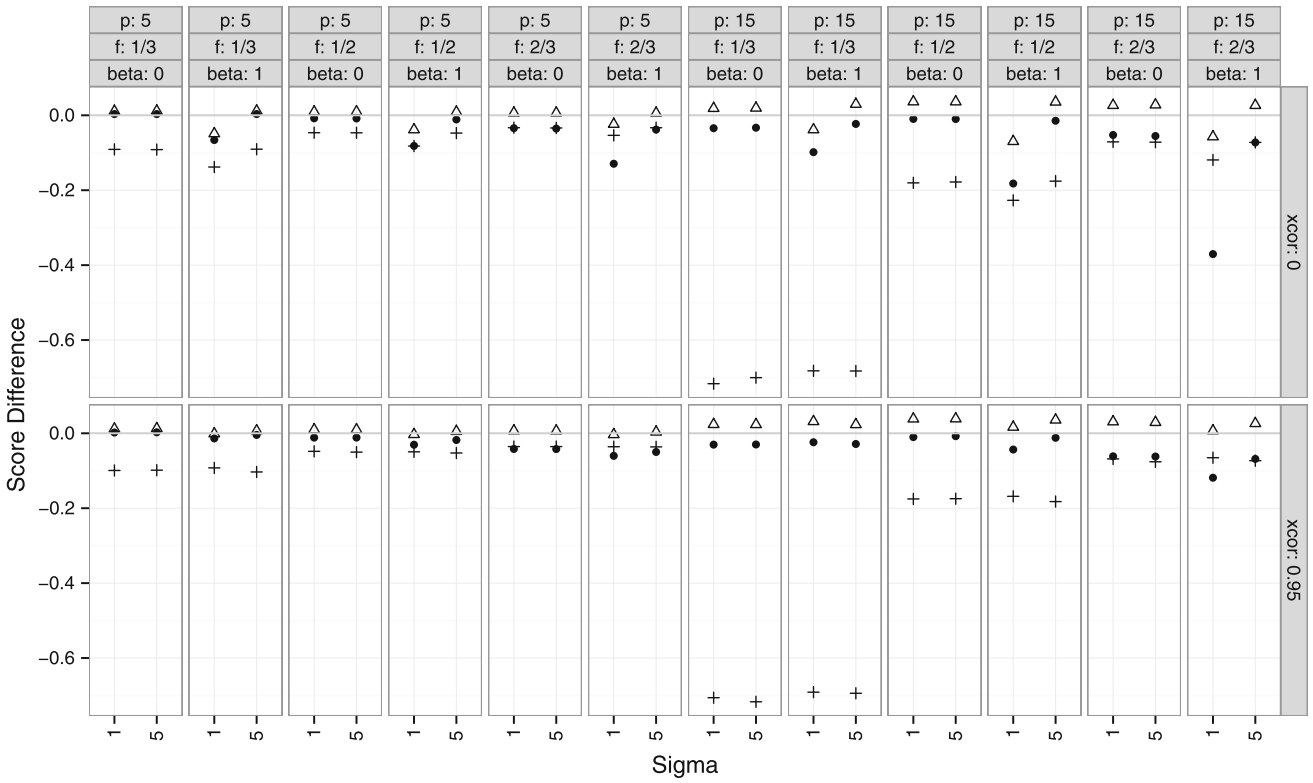

Fig. 4 Variable selection simulation results showing the FD–SD difference in scores as a solid dot, FD-SAFE difference as an open triangle and FD-VALID as a plus symbol

The simulation used 4,000 replications and $n = 60$ with a full factorial design over all combinations of $\sigma = 1, 5$, $\beta_i = 0, 1 \forall i$, $p = 5, 15$, $\rho = 0, 0.95$ and training fractions of $1/3$, $1/2$ and $2/3$. The results showing the average difference in score between the FD–SD analysis methods along with the FD-SAFE and FD-VALID differences are shown in Fig. 4.

In Fig. 4, we see that FD is always better than SD but the margin of difference is mostly small. The best choice training fraction varies although $f = 2/3$ is always worst. We see that the SAFE estimator is always better than the SD estimator and often outperforms the FD estimator. The VALID estimator can perform badly although it competes well with SD when $f = 2/3$.

Figure 5 shows the relative contribution as described in (3). This shows that the data reuse costs can be quite significant which allows for split strategies to be favoured. The higher model selection costs for the $\beta = 1, \sigma = 1$ combination are reflected in the reduced performance of the split strategies as seen in Fig. 4.

We extended the data analysis method to allow consideration of potential second order terms in the predictors (even though these are not present in the true model). Under these conditions, the SD strategy tends to outperform FD particularly as the number of potential models grows.

3.3 Outliers

In this scenario, we investigate a model building strategy that eliminates outliers. Consider a model with $X_i \sim U(0, 1)$ with $i = 1, \ldots n$. Let ϵ_i be i.i.d. t with degrees of freedom d. Let

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where the true value of $\alpha = 0$ although the model will estimate it. The model building strategy fits the model using least squares and computes the studentized residuals, r_i. Any case with $|r_i| > 3$ is deleted and the model is refitted. This case deletion process is repeated until all remaining $|r_i| < 3$. This is a crude procedure and one which we would not recommend, particularly when we can see the true model generating process. Nevertheless, it is representative of procedures that delete or down-weight cases that do not fit the proposed model well. We would prefer an approach to prediction to behave reasonably well even if the model building strategy is not the best.

We ran a simulation with 4,000 replications. We used a full factorial design running over all combinations of $n = 18, 48$, $\sigma = 1, 5$, $\beta = 0, 1$, $d = 3, \infty$ and training fractions of $1/3$, $1/2$ and $2/3$. We see the outcome in Fig. 6.

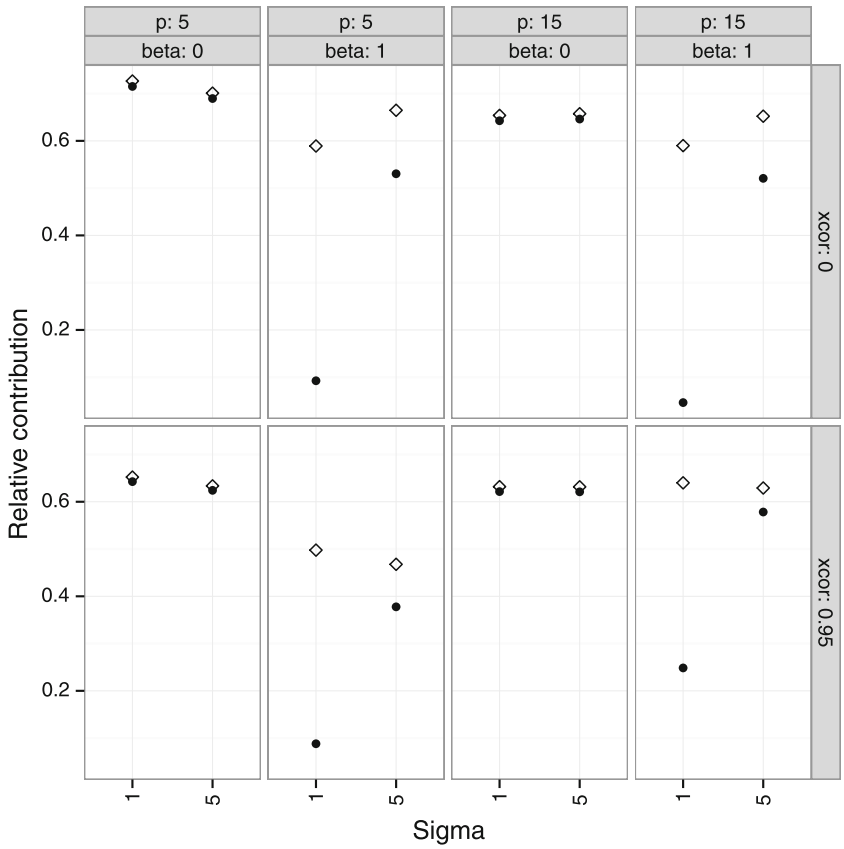

Fig. 5 Model performance decomposition for the variable selection scenario. Solid dot represents data reuse relative contribution while open diamond represents data reuse plus the model selection contribution. Parameter estimation is the remaining part

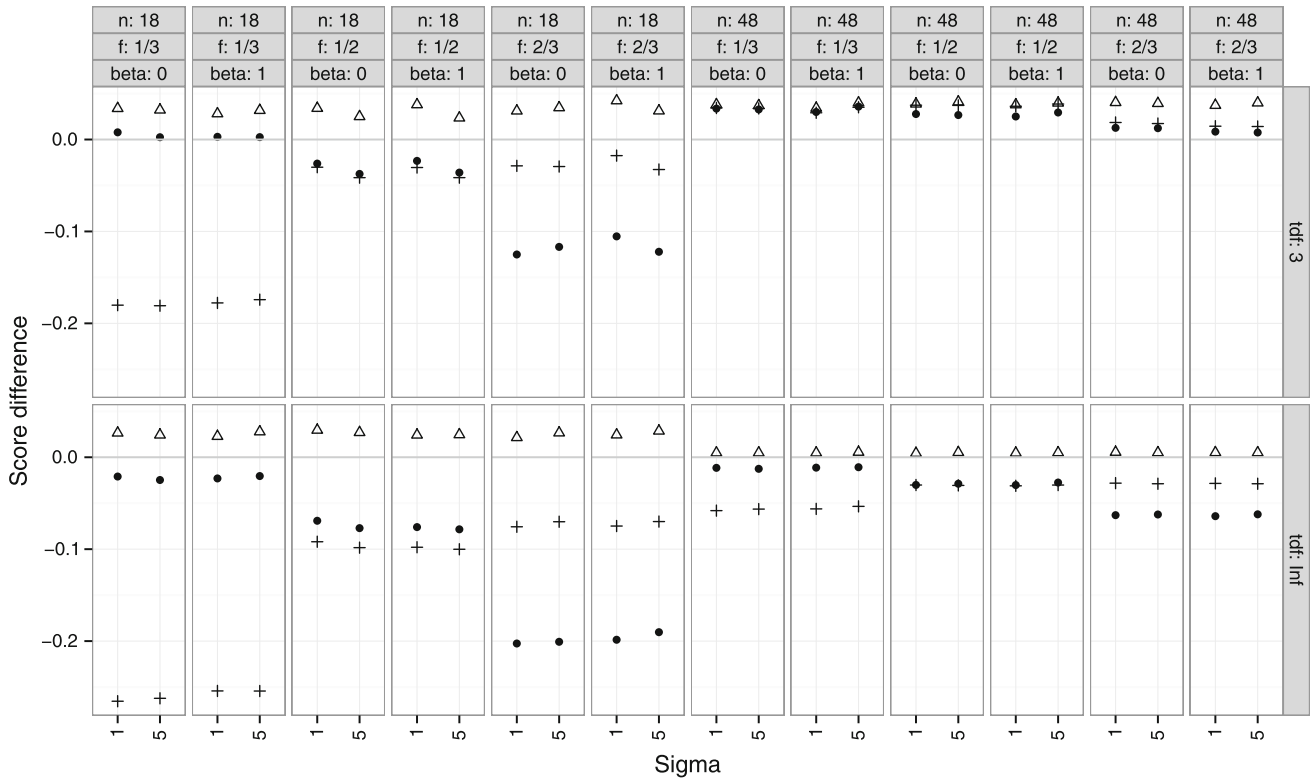

Fig. 6 Outlier simulation results showing the FD–SD score as a solid dot, FD-SAFE as a triangle and FD-VALID as a plus symbol

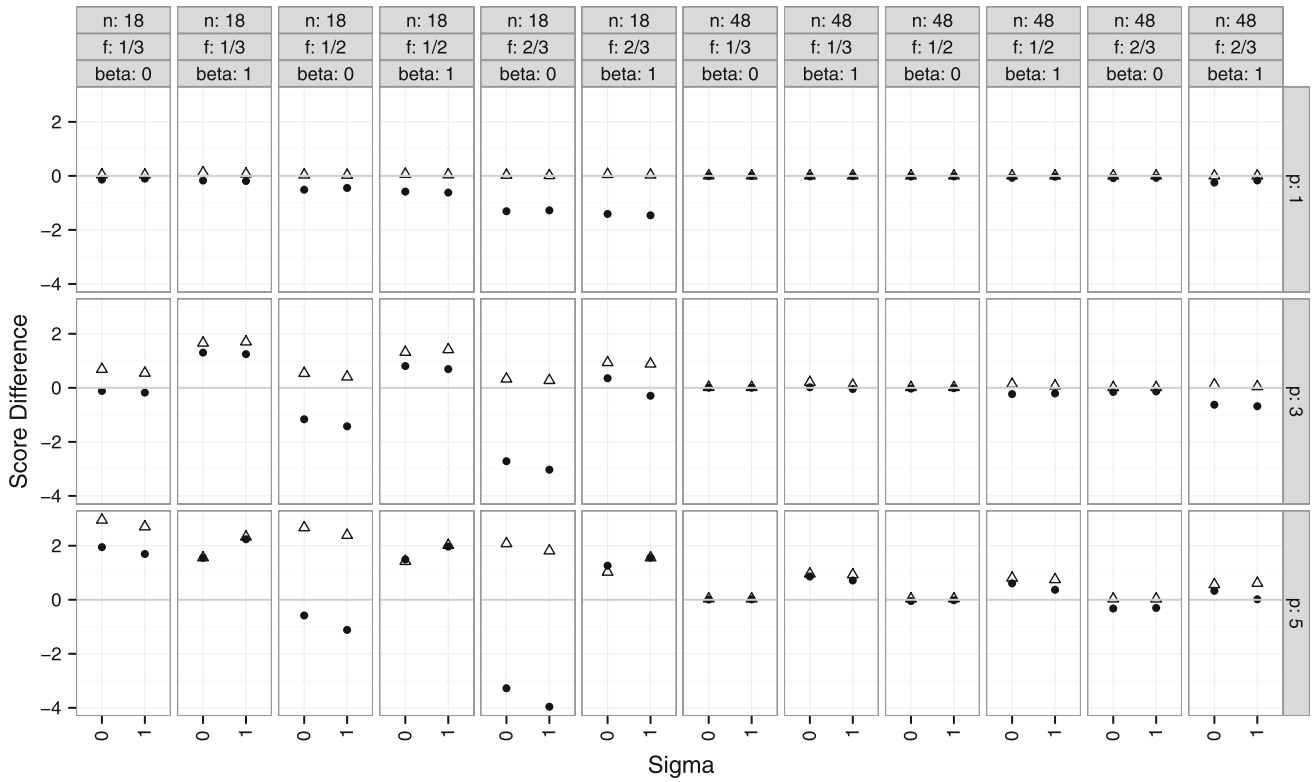

Fig. 7 Binary response simulation results showing the FD–SD score as a solid dot and FD-SAFE as a triangle

When the error is t_3 we see that the SD approach is superior for the larger sample but inferior when the error is t_∞ i.e. normal. In all cases, the smallest training fraction is preferred for SD but this is simply because the model selection process here just decides which points to include. This choice becomes irrelevant because the test data is used to estimate the model. The other factors have little effect on the outcome. We see that the SAFE estimator performs very well in this instance although the superiority over the FD strategy is due to the sensible decision to use all the data to estimate the model. The VALID estimators performs similarly to the SD estimator except that the preferred $f = 2/3$. The outlier deletion strategy is clearly artificial in this scenario but this does show how this type of data analytic procedure, which is often used in more plausible forms in practice, can have serious consequences for prediction performance

There is no model selection effect in this scenario as the model is fixed, only the data used vary. We can compute the decomposition in (3) and find that the data reuse costs dominate the parameter estimation comprising a large proportion the total contribution of these two components across the 48 simulation conditions (plot not shown).

3.4 Binary response

We finish with a variable selection problem in a binary response logistic regression model. Let $\eta_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$ for $i = 1, \ldots, n$ where ϵ_i is i.i.d. $N(0, \sigma^2)$ and $X_{ij} \sim U(0, 1)$. Now let $P(Y_i = 1) = \exp(\eta_i)/(1 + \exp(\eta_i))$.

The model building strategy is variable selection using stepwise AIC as implemented in Venables and Ripley (2002). We vary $p = 1, 3, 5$, $n = 18, 48$, $\sigma = 0, 1$, $\beta = 0, 1$ and training fractions of $1/3, 1/2$ and $2/3$. When $\sigma = 0$, we have the standard case while $\sigma = 1$ is a measurement error situation. We used 4,000 simulation replicates each over all factorial combinations of these 5 parameters.

The results are shown in Fig. 7. We see that the results are much more stable for n=48 than for n=18. For the simple $p = 1$ case there is very little model selection and the FD strategy is preferable. When there are more predictors, a split data strategy is preferred when there really are significant predictors ($\beta = 1$) but the FD approach works best when the null model is true. We see that the SAFE estimator is almost always an improvement over the SD estimator and sometimes better than the FD estimator depending on the scenario. Note the the VALID estimator makes no sense for this binary response model.

4 Conclusion

Our main conclusion is that, except in specific circumstances, a data splitting strategy is superior to an FD strategy for prediction purposes. In particular, SAFE is the best type of splitting strategy. Caution is necessary when trying to generalise from simulations which are necessarily limited in scope. But we must bear in mind that finite sample theoretical results of any generality are not available. Prediction from statistical models is a common activity and practical advice is necessary.

The simulation results are not unfavourable to the FD strategy but the simulations understate the relative value of an SD strategy. Simulations cannot simulate human judgement which is an important part of model building which implicitly increases the number of potential models substantially. Increasing model choice favours an SD strategy since data reuse costs are amplified. Admittedly, more models puts more emphasis on model choice but often the best model is accompanied by many close contenders so SD will typically not lose by much to FD. Our simulations have also shown SD improving relative to FD as model choice increases.

All four scenarios we have presented are relatively simple which favours the FD strategy. Furthermore, the model building strategy has been completely specified. Under such conditions we would tend to choose an FD strategy but even here, we see that the SD and particularly the SAFE estimator might be a better choice. In many real data analyses on observed, not simulated data, the model building strategy will be more complex and difficult or impossible to pre-specify. In such circumstances, a split data strategy is preferable. The decomposition of the model performance indicates that the loss in using a split data approach is limited for model selection and parameter estimation once we move away from the p close to n situations. On the other hand, the FD approach is susceptible to large losses due to data re-use costs. The size of these data resuse losses is very difficult to anticipate whereas the cost of using less data to select and estimate the model is much better understood. This suggests that a split strategy is safer than the FD strategy.

We make the following tentative recommendations—in regression problems with a continuous or binary response with independent observations (i.e. having no grouped, hierarchical or serial structure), the analyst should decide between a full and a split analysis based on the following considerations:

An FD analysis should be considered where the set of models to be considered use such a large number of parameters relative to the number of observations that SD would imperil selection and estimation. An FD analysis is also preferred when the model building step will be simple and involve a choice among a small set of possibilities. In situations where the analyst is prepared to pre-specify the exact form of the data analysis in advance, bootstrapping or other resampling techniques can be used to account for model uncertainty. An FD analysis is likely to be preferable in designed experiments where there is often limited model choice.

A split data analysis should be preferred when the analyst has no fixed conceptions about the model to be used for the data and plans to search for a good choice using a range of numerical and graphical methods. The SAFE estimate should be preferred to the SD estimate. If it becomes clear that there is insufficient data in the split datasets to find and estimate a model, then the analyst can always revert to the FD approach (but a switch from full to split cannot be allowed). The empirical evidence gives no clear choice about the data splitting fraction. With more knowledge of the data generating process, more specific guidance could be given but that same knowledge would also reduce or eliminate the need for model building. So we recommend a 50:50 split as the default choice.

More work is necessary before attempting recommendations for other types of response or data of a more structured form. It is also more difficult to make a recommendation for situations where an explanation of the relationship between X and Y is the goal rather than prediction. Certainly the same issues of overconfidence arise but interpretation of parameters explaining the relationship become problematic when the form of the model is changing.

References

Altman, D.G., Royston, P.: What do we mean by validating a prognostic model? Stat. Med. **19**(4), 453–473 (2000)

Bell, R., Koren, Y.: Lessons from the Netflix prize challenge. ACM SIGKDD Explor. Newsl. **9**(2), 75–79 (2007)

Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. Bernoulli **19**(2), 521–547 (2013)

Berk, R., Brown, L., Zhao, L.: Statistical inference after model selection. J. Quant. Criminol. **26**(2), 217–236 (2009)

Carpenter, J.: May the best analyst win. Science **331**(6018), 698–699 (2011)

Chatfield, C.: Model uncertainty, data mining and statistical inference. J. R. Statist. Soc. Ser. A **158**(3), 419–466 (1995)

Cox, D.: A note on data-splitting for the evaluation of significance levels. Biometrika **62**, 441–444 (1975)

Dahl, F., Grotle, M., Saltyte Benth, J., Natvig, B.: Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. Eur. J. Epidemiol. **23**(4), 237–242 (2008)

Dawid, A.: Present position and potential developments: some personal views statistical theory the prequential approach. J. R. Stat. Soc. Ser. A **147**, 278–292 (1984)

Draper, D.: Assessment and propogation of model uncertainty. J. R. Stat. Soc. Ser. B **57**, 45–97 (1995)

Faraway, J.: On the cost of data analysis. J. Comput. Gr. Stat. **1**, 215–231 (1992)

Friedman, J., Hastie, T., Tibshirani, R.: Elements Statistical Learning, 2nd edn. Springer, New York (2008)

Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. **102**(477), 359–378 (2007)
Good, I.J.: Rational decisions. J. R. Stat. Soc. Ser. B **14**(1), 107–114 (1952)
Heller, R., Rosenbaum, P.R., Small, D.S.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**(487), 1090–1101 (2009)
Hinkley, D., Runger, G.: The analysis of transformed data (with discussion). J. Am. Stat. Assoc. **79**, 302–319 (1984)
Hirsch, R.: Validation samples. Biometrics **47**(3), 1193–1194 (1991)
Lawless, J.F., Fredette, M.: Frequentist prediction intervals and predictive distributions. Biometrika **92**(3), 529–542 (2005)
Leeb, H., Pötscher, B.M.: Model selection and inference: facts and fiction. Econom. Theory **21**(01), 21–59 (2005)
Little, R.: Calibrated bayes. Am. Stat. **60**(3), 213–223 (2006)
Meng, X., Xie, X.: I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? Econom. Rev. **33**, 1–33 (2013)
Miller, A.: Subset Selection in Regression. CRC Press, Boca Raton (1990)
Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics **21**(15), 3301–3307 (2005)
Mosteller, F., Tukey, J.: Data Analysis and Regression. A Second Course in Statistics. Addison-Wesley, Reading (1977)
Parry, M., Dawid, A.P., Lauritzen, S.: Proper local scoring rules. Ann. Stat. **40**(1), 561–592 (2012)
Picard, R., Berk, K.: Data splitting. Am. Stat. **44**, 140–147 (1990)
Picard, R., Cook, R.: Cross-validation of regression models. J. Am. Stat. Assoc. **79**, 575–583 (1984)
Pötscher, B.: Effects of model selection on inference. Econom. Theory **7**(2), 163–185 (1991)
Roecker, E.: Prediction error and its estimation for subset-selected models. Technometrics **33**, 459–468 (1991)
Schumacher, M., Binder, H., Gerds, T.: Assessment of survival prediction models based on microarray data. Bioinformatics **23**(14), 1768–1774 (2007)
Steyerberg, E.: Clinical Prediction Models. Springer, New York (2009)
Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B **36**, 111–147 (1974)
Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)
Wit, E., Heuvel, E.V.D., Romeijn, J.W.: All models are wrong..: an introduction to model uncertainty. Stat. Neerl. **66**(3), 217–236 (2012)
Xie, M.G., Singh, K.: Confidence distribution, the frequentist distribution estimator of a parameter — a review. Int. Stat. Rev. **81**, 3–39 (2013)