Data Splitting Strategies for Reducing the Effect of Model Selection on Inference

Julian J. Faraway
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Abstract

When an appropriate model for data is not completely known, the data is often used to select a model. Very often inference is then made from the selected model assuming that it had been known from the beginning. Estimates of the error of predictions or other quantities associated with that model take account of the uncertainty about the parameters of the model, but not the uncertainty about the model itself. Such error estimates tend to be too small, especially when the model uncertainty dominates the parametric uncertainty. Models are usually selected on the basis of fit, so typically the data fit the selected model rather well thus making the error seem small. In data splitting, one part of the data is used solely for model selection and the other part for inference thus hopefully avoiding the over-optimism induced by using the same data to both select and estimate the parameters of a model. Data splitting is easy to implement and thus is an attractive alternative to complex methods of adjusting for the effect of model selection on inference.

Three tasks need to be performed - model selection, prediction and error assessment. We investigate different strategies for allotting the two parts of the data between these three tasks. We devise a new graphical method for jointly assessing prediction accuracy and error estimates called an honesty plot. The plot can be used to show actual coverage of confidence intervals of any given nominal level. Variable selection, Box-Cox transformation and more complex simulation experiments are used to compare the various strategies. The performance of data-splitting is found to be no better than using all the data for both selection and inference.

Keywords: Model uncertainty, cross-validation, model building, prediction, subset selection.

1 Introduction

When the same data are used to select both the model and estimate the parameters of that model, there is a danger that the predictions made by that model will be more optimistic than they should be. To expand on this, consider that in many cases a suitable model for the data is not completely known, and various data-analytic techniques will be used to try to find a satisfactory model. Even in cases where a good model is supposedly known *a priori*, it is considered good practice to check that the data do indeed fit that model, which raises the possibility that we may change the model. It is common practice, once a good model has been found, to pretend that this had been known all along, and to conduct the inference allowing for the uncertainty about the parameters but not that about the model. This means that the assessed variability tends to be lower than it really should be because the effects of model uncertainty have not been taken account of. Chatfield (1995) provides a good overview of the effects of model uncertainty on statistical inference.

Usually, it is not through ignorance or lethargy that the effects of model uncertainty on inference are not accounted for – it is quite difficult to do it honestly. Almost all the theory of statistical inference rests on the model being known, whereas in practice the model is often not known and is instead selected using both the data and prior and/or contextual knowledge. Within the frequentist paradigm, one possible approach is to embed the model selection process within a larger model and thereby view model selection as the estimation of nuisance parameters. For example, the selection of the index of transformation in the Box-Cox transformation could be regarded as the estimation of an additional parameter which could then be allowed for using standard inferential techniques. When the model selection amounts to selecting the order or form of the model, this program becomes more difficult to apply though not impossible. Another idea is to bootstrap the whole data analysis. If we view the data analysis itself as part of the estimation then we may have some hope of getting honest estimates of variability using the bootstrap technique; see Faraway (1992)

One possible Bayesian approach is similar to the one described first above. The model selection process is embedded within some larger model determined by the selection procedures being used, and the standard Bayesian techniques are then applied. By using a hierarchical framework, the Bayesian approach is generally less cumbersome than the equivalent frequentist method. One problem is the large class

of models that must usually be considered, which poses difficulties in assigning meaningful priors as well as a substantial computational burden.

The disadvantages are generally shared between both approaches. Most of the methods proposed are rather complicated and require specialised software for each different kind of model and model selection method. Glance through the manual for a standard statistical package and imagine each procedure generalised to take account of the effects of model uncertainty. The size and complexity of such packages would be increased by an order of magnitude. Note that the methods for model uncertainty assessment are generally complex and could not be readily implemented by the general user, unlike ideas such as the bootstrap.

There is another more important defect in these methods. They generally require that the model selection procedure be fully specified in advance. Much data analysis is iterative in nature — the next action often depends on the results of the previous action. Furthermore, graphics form an important part of data analysis and such techniques are extremely difficult to characterise exactly. So in practice, data analysis used for model selection is quite complex and yet cannot be precisely specified in advance. For the type of methods described above to really work we would need to be able to pre-specify our data analysis — that is essentially write a program that does data analysis reliably. Nobody is at all close to doing this. So in effect we do not escape the pre-specified model paradigm — all we might achieve is the use of a more flexible model.

Draper (1995) (and to some extent Hill (1990)) tries to get round this problem of having to prespecify the model by suggesting that we follow the usual data-analytic techniques, and then when we have found a good model (or models) we expand this model, embedding it in a richer family. We then assign priors to the bigger family of models and proceed in the usual Bayesian manner. The frequentist could also take a model expansion approach. The main problem is in knowing in which direction to expand the model. Hopefully, the data analysis will be suggestive, but it is no small difficulty to decide how to do this.

Given these difficulties, it is natural to return to the origin of the problem, namely that the data are being used twice, to find the model and then to make inference from that model. The pessimist claims that this is trying to do too much with the data and that, having used the data to find a good model, we should await the arrival of fresh data and use that to do the inference using the model we found with the original data. Of course, fresh data are not always immediately available, so the idea of data splitting is to divide the data into two parts (not necessarily of equal size), using the first to select the model and the second to do the inference. Since model selection and statistical inference from a known model are well understood and can now be done in isolation from one another, we seem to avoid the complexities of the model uncertainty approaches discussed above. We do not have to prespecify our data-analytic method or worry about what we were thinking about when we selected the model, all provided that we keep the second part of the data in a sealed box and do not look at it until we have selected the model using the first part. We are free to use the whole range of exploratory data analytic techniques without needing to characterize them exactly. Thus data splitting seems to be a simple and attractive alternative to the above.

It is not always possible to do data splitting. We need to be able to divide the data into two samples which can be regarded as independent or exchangeable samples from the same population. For example, if we suspect correlated errors in regression data, then data splitting would be difficult if not impossible. Time series can be split on time but this is a different situation from that discussed below – time series cannot be split randomly and the motivation for splitting may be the assessment of time homogeneity. In other cases, splitting the data may not leave enough data in either or both parts to find and estimate the parameters of the model. This would happen when the number of variables is large relative to the number of cases.

Splitting is also not always necessary or appropriate. Some statistical investigations involve the search for a model which represents some underlying truth. In this case, model selection is the end in itself and splitting may not be needed. I shall use my models as a means to making predictions of future observables. The truth or correctness of the particular model selected need not be considered, only its predictive performance.

Splitting the data is intended to give the effect of having new data, but one should realize that this is less than true. The real test of a model comes when it is applied to truly new data, collected at a different time under perhaps different conditions. The second part of the split data set cannot reveal biases in the sampling process used to collect the original data. Only new data collected under different conditions can hope to do this - see Hirsch (1991). When truly new data do become available, an assessment of exchangeability with the original data needs to be made before proceeding. In the split sample case, this assessment is not needed.

I discuss the history and various forms of data splitting in Section 2. The performance of these data splitting strategies is compared in Section 3 and the conclusions are in Section 5.

2 Data Splitting Strategies

A history of data splitting can be found in Stone (1974). The dangers of using the same data to both select and fit the model have been known for many years and data splitting is a simple

technique for dealing with it that was practical to use when computational costs were high. These same high computational costs also meant that the amount of exploratory data analysis to find good models was much lower in the past. Because less data analysis was possible, models would tend to fit the data less well, and so the data splitting would tend to reveal the bias in the chosen model rather than the variability due to the model selection process. Detecting bias might lead one to modify the model thus invalidating the independence of the second sample thereby effectively abandoning data splitting. Hence there was less incentive to split the data in the past. Nowadays, extensive data analysis is easy and commonplace, and so the tendency is to overfit, making variance, not bias, more of a concern. However, detecting a larger variance using data splitting, than the selected model would suggest, is expected and would not tempt one to change the chosen model. Thus data splitting seems more relevant and useful now than in the past.

Stone's main interest was in the use of cross-validation to select models. Data splitting could be regarded as a rather crude form of cross-validation but that is certainly not our purpose here. We scrupulously wish to hold out a sample of the data that will not be used for model selection in any way. Mosteller and Tukey (1977, p. 37) also discuss data splitting as a form of cross-validation, but focus on it as form of model selection which we wish to avoid here.

There are three different tasks we need to perform:

1. Selection of the model.
2. Estimation of the parameters of the selected model. Point predictions can then be made.
3. Assessment of the variability in the predictions.

Conceivably we could split the data three ways and use a different part for each of the above tasks. This has been suggested in passing by Miller (1990, p. 13) and is, perhaps, what Mosteller and Tukey (1977) meant by double cross-validation. I investigated this strategy and found it clearly inferior to any of those discussed below, so henceforth I restrict attention to splitting the data into two parts.

Which parts of the data should be used to perform the three tasks above? Most previous authors have used the first sample (sometimes called the *estimation* sample) to select the model and also estimate the parameters of that model. The second sample, called the *validation* sample is then used to assess the performance of the selected model. Call this strategy A and the naive strategy where the same data is used for all tasks, strategy N. Validation is not, as Stone (1974) says, a good descriptive word, since the second sample is not used to determine the validity or correctness of the selected model, merely to assess its predictive performance. Picard and Cook (1984), Picard and Berk (1990) and Roecker (1991) make this division of labour. These authors focus on the estimation of an average (over a subset of the predictor space) mean squared error of prediction. Certainly, sample splitting in this manner might enable a better estimate of this quantity, although it seems that in practice one would want assessments of variability in particular predictions, and the overall summary measure might be a number without context or use. Another common thread is that all three articles concentrate on variable selection in regression models, whereas (as these authors admit) the applicability of data splitting is much wider. Special techniques can be used to estimate the mean squared error of prediction when attention is restricted to just variable selection in regression. However, in general we cannot expect such techniques to be available, so we avoid them when using strategy A in the comparisons to follow.

Data splitting can be viewed as artificially providing new data. If one really did have new data, one might well use the model selected by past experience and fit the data to that model. This suggests a second data splitting strategy (B), where the first sample is used only to select the model and the second sample is used both to estimate the parameters of the model leading to predictions and to assess the variability in those predictions. This is the approach mentioned by Miller (1990, p. 13) and Hurvich and Tsai (1990) and actually implemented by Faraway (1992). Interestingly, none of the proponents of these two approaches seems to have considered the other.

A third strategy (C) can also be motivated by the arrival of new data. One might retain the originally selected model, but merge the new data with the old and re-estimate the parameters. Thus the first sample would be used to select the model, and the combined sample would be used to estimate the parameters and to assess variability. Note that the independence of the model selection and model assessment has now been lost, which might be expected to cause some over-optimism in our assessment of model performance. The advantage is that all the data are being used to estimate the parameters thus hopefully reducing the variability in those parameter estimates.

Strategy	Model Selection	Estimation	Variance Assessment
A	Part 1	Part 1	Part 2
B	Part 1	Part 2	Part 2
C	Part 1	All	All
N	All	All	All

Table 1: Division of Labour for Data Splitting Strategies

One aspect of model performance is accuracy of prediction, which we might measure by mean square error (MSE), for example. Strategies A, B and C do not use all the data

for selecting the model, so that we cannot expect them to find as good a model as if we had used all the data. Similarly, A and B only use a portion of the data for the estimation of the parameters, which would lead to greater variability (although maybe less bias) than if all the data were used. Thus, it is unreasonable to expect strategy A or B to outperform strategy N in terms of predictive MSE.

Indeed, the purpose of data splitting is to obtain better estimates of the variability of predictions, and the price one pays is that the actual variability (which one is trying to estimate) of the predictions will tend to be higher. Thus we need to measure not just the performance of point prediction but also the variance assessment.

Let the prediction of the future response Y_0 for a given value of the covariates x_0 be $\hat{Y}_0(x_0)$ and its estimated standard error be $se(\hat{Y}_0(x_0))$. We can measure the point performance of our methodology by looking, for example, at the values of $\hat{Y}_0(x_0) - Y_0$ in the long run, but this does not measure the accuracy of our variability assessment. Confidence intervals for the predictive values can be constructed and we can assess how the actual coverage of the intervals compares with the nominal levels. We can simply record the proportion of future observed values that fall in the, say, 95% confidence intervals. However, it would be more instructive to see the actual coverage for a whole range of nominal levels. To this end define the predictive z-statistic

$$z = (\hat{Y}_0(x_0) - Y_0)/se(\hat{Y}_0(x_0)).$$

We might hope that, in the long run, z is centered around 0, but this in itself is not enough. If $\mathrm{Var}\, z$ is much greater than 1, this would indicate general underestimation of variability, whereas $\mathrm{Var}\, z$ much less than one indicates overestimation of error. If normality holds, then we can compare the observed z's to the standard normal to assess what I shall term the *honesty* of our predictions. In other words, $\Phi(z)$ should then be Uniform(0,1). The actual coverage of the confidence intervals for predictions can be assessed by comparing the empirical distribution of the $\Phi(z)$ to the standard uniform. In particular, given a set of m observed z_j, $j = 1, \ldots, m$, we can plot $l_j = j/(m+1)$ against $\Phi(z_{(j)}) - l_j$. In Figure 1, I show several such curves for z's randomly generated from different distributions. I call these *honesty plots*. In practice where one might be assessing the performance of predictions against the observed values, one would usually just plot the points, but since $m = 1,000$ here is large, lines are better.

I have smoothed the curves a small amount due to the inherent roughness of the empirical processes. This would not be necessary for just one curve but it does make it possible to distinguish several curves on the same plot. Looking at the case where the variance of the prediction is underestimated and taking the worst case, we observe that at a probability of 0.2 the discrepancy is about -0.15. So if the predictions were well-calibrated, i.e. the mean and variance were on average correct, we would expect about 20% of the observed $\Phi(z)$'s to be less than 0.2 but in this case about 35% are. So for a central confidence interval with $100 - 2 \times 20$=60% nominal confidence, there would be only about $60 - 2 \times 15$=30% actual coverage. If distributions other than normal were expected, this could all be suitably adjusted. Another approach to the assessment of predictive variability is the log-scoring method of Good (1952)

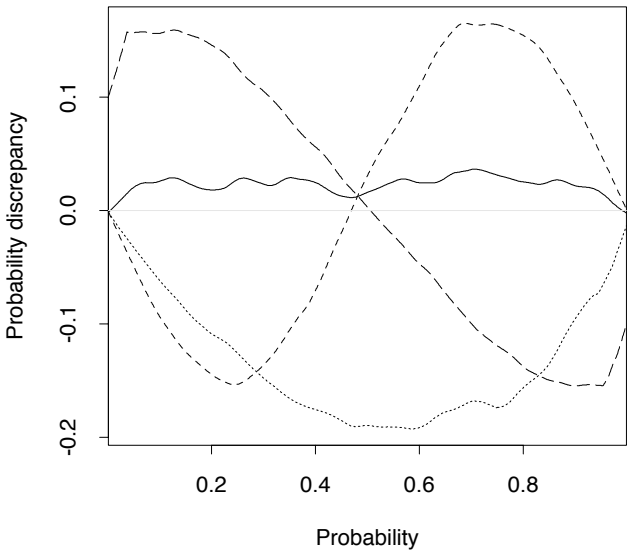

Figure 1: Honesty plots for known z: $N(0,1)$ (well-calibrated predictions) — solid, $N(-0.5,1)$ (true value underestimated) — dotted, $N(0,4)$ (variance underestimated) — short dashes, $N(0,1/4)$ (variance overestimated) — long dashes.

I have focussed on prediction here since predictions can be made in the original scale of the response and are thus directly interpretable. The meaning and existence of parameters may change from model to model and so it is not easy or even necessarily meaningful to assess variability in parameter estimates except in relation to the particular model chosen. A good example of this occurs with the Box-Cox transformation — see Hinkley and Runger (1984)

3 Simple Model Selection

In this section, I consider a two simple examples where the model that is used to make a prediction is first selected using some data-dependent procedure. In both cases, I would not recommend using data splitting in practice since the model selection procedures are relatively simple and well defined and there is the possibility of applying various whole-data

methods for adjusting the inference for the model selection effect. Data splitting would be more appropriate where the model selection procedure is either complex or dependent on graphical methods which are difficult to precisely define. Nevertheless, if data splitting cannot work well in these simple circumstances, then we can have little confidence in its value in more complicated situations.

3.1 Regression Variable Selection

Let $X_i \sim U(0,1)$ and ϵ_i be independent and identically distributed $N(0,\sigma^2)$ for $i = 1, \ldots n$. Let $Y_i = \alpha + \beta X_i + \epsilon_i$. Suppose we wish to predict Y, given some x_0, but suspect that there may be no linear relationship between X and Y and so test the hypothesis that $\beta = 0$. The test we use is based on the usual least squares based t-statistic $t = |\hat{\beta}|/se(\hat{\beta})$. Thus our predicted value is

$$\hat{Y}(x_0) = \begin{cases} \hat{\alpha} + \hat{\beta} x_0 & \text{if } t > c \\ \bar{Y} & \text{if } t \le c \end{cases},$$

where c is some specified critical value.

We consider the prediction of the mean response — i.e. the average value of future observed values of Y for given x_0. This means that we can compare our prediction to a known true value, and it also amplifies the model selection effects over what we would observe if the problem were to predict a single future value. Of course, the prediction of a single future value is more often the objective, but my primary aim is to compare data splitting strategies and the prediction of the mean response is more convenient for this purpose.

Thus the naive estimated variance of $\hat{Y}(x_0)$ is

$$\hat{\mathrm{Var}}\,\hat{Y}(x_0) = \begin{cases} \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right) & \text{if } t > c \\ \frac{\sum (Y_i - \bar{Y})^2}{n(n-1)} & \text{if } t \le c \end{cases},$$

where $\hat{\sigma}^2$ is the usual least squares regression estimate of σ^2.

This is a mere cartoon of a real statistical analysis. In practice, residuals would be examined for evidence of departure from assumptions and the possibility of transformations of the variables investigated. Various graphical examinations of the data would also be made. Thus even in this simple regression setting, most statisticians would carry out a rather complex and difficult to characterize procedure to select their model. Furthermore, in this case where we suspect that β is close to 0, then the procedure above is not to be recommended - a shrinkage type estimator or Bayesian approach could be used according to taste.

Several authors have worked on the variable selection effect on regression inference, including Freedman, Navidi, and Peters (1988), Kipnis (1991), Pötscher (1991) and Raftery, Madigan, and Hoeting (1993), and it would seem that as long as one is prepared to confine one's model selection purely to the inclusion or exclusion of a single variable then there is some hope of reasonably adjusting the inference for the model selection effect without resort to data splitting. The general problem of how to adjust the inference for variable selection still awaits a completely satisfactory solution.

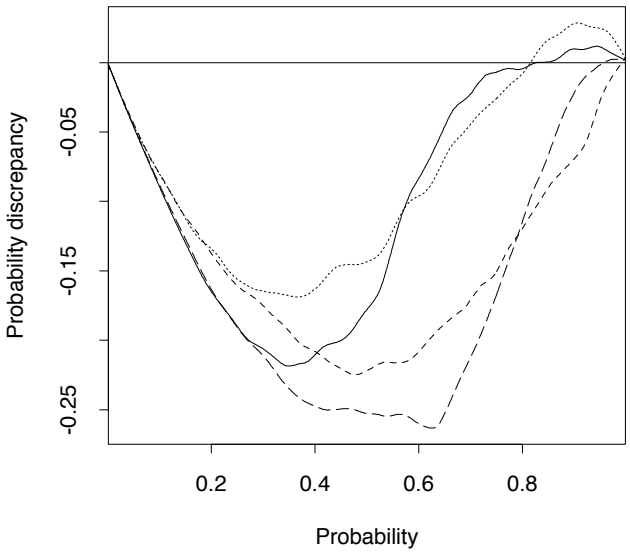

Figure 2: Honesty plot of $l_j = j/1001$ against $\Phi(z_{(j)}) - l_j$ where $j = 1, \ldots, 1000$ for data splitting strategies N — solid, A — dotted, B — short dashes, C — long dashes.

Data splitting strategies A,B & C require the data to be split into two parts of size n_1 and n_2. Snee (1977) discusses ways in which the data can be split in a balanced manner, but Roecker (1991) found that this was only a small improvement over random splitting in the variable selection setting. In more complex situations it may be difficult to be clever about splitting the data in a balanced way, so we will use random splitting for convenience and versatility.

We use the subscripts 1 and 2 to distinguish the regression estimates derived from the first and the second sample. Thus

$$\begin{aligned} \hat{Y}_A(x_0) &= \hat{\alpha}_1 + \hat{\beta}_1 x_0 \text{ or } \bar{\mathrm{Y}}_1 \\ \hat{Y}_B(x_0) &= \hat{\alpha}_2 + \hat{\beta}_2 x_0 \text{ or } \bar{\mathrm{Y}}_2 \\ \hat{Y}_C(x_0) &= \hat{\alpha} + \hat{\beta} x_0 \quad \text{or } \bar{\mathrm{Y}} \end{aligned}$$

where in all cases I choose the first definition when $t_1 > c$ and the second otherwise. The estimated variances are

$$\begin{aligned} A &: \hat{\sigma}_2^2 \left(\frac{1}{n_1} + \frac{(x_0 - \bar{X}_1)^2}{\sum_1 (X_i - \bar{X}_1)^2} \right) \text{ or } \frac{\sum_2 (\mathrm{Y_i} - \bar{\mathrm{Y}}_2)^2}{\mathrm{n_1}(\mathrm{n_2} - 1)} \\ B &: \hat{\sigma}_2^2 \left(\frac{1}{n_2} + \frac{(x_0 - \bar{X}_2)^2}{\sum_2 (X_i - \bar{X}_2)^2} \right) \text{ or } \frac{\sum_2 (\mathrm{Y_i} - \bar{\mathrm{Y}}_2)^2}{\mathrm{n_2}(\mathrm{n_2} - 1)} \end{aligned}$$

$$C \quad : \quad \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right) \text{ or } \frac{\sum (\mathrm{Y_i} - \bar{\mathrm{Y}})^2}{\mathrm{n(n-1)}},$$

where $\sum_1$ and $\sum_2$ indicate sums over the first and second parts of the data respectively.

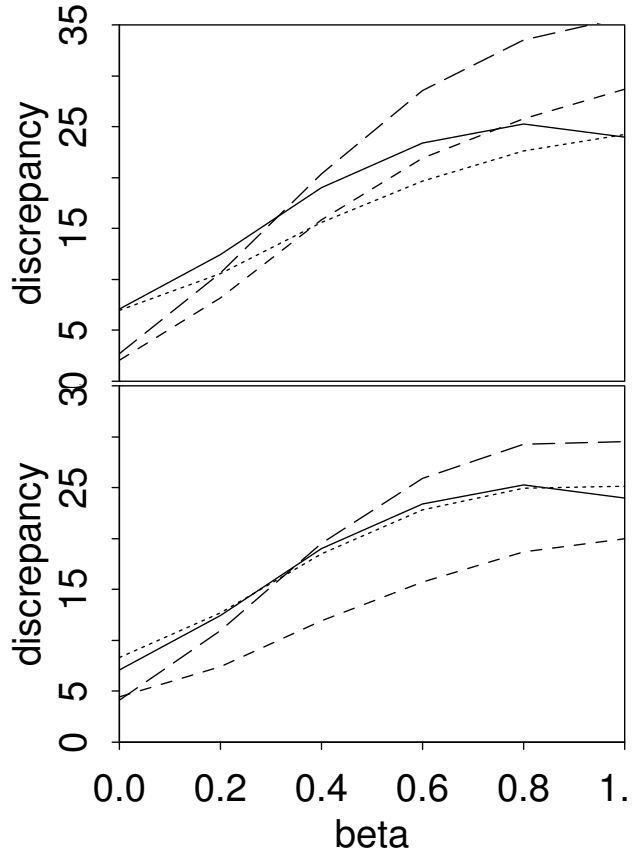

Figure 3: Maximum discrepancy(x100) for strategies N — solid, A — dotted, B — short dashes, C — long dashes with β varying. $f = 0.5$ in top pane and $f = 0.75$ in bottom pane

The properties of $\hat{Y}(x_0), \hat{Y}_A(x_0), \hat{Y}_B(x_0), \hat{Y}_C(x_0)$ can be accurately investigated by simulation. Note that even with a simple set-up like this, it is very difficult to determine the exact distributions of the estimators. Asymptotic approximations are not much help since the model selection effect vanishes with increasing sample size.

First I investigated the point performance for all four strategies. Not surprisingly, I found that the whole data strategies, C and N, performed significantly better. These results are expected — we do not split the data with the objective of getting better point estimates.

To investigate the assessment of uncertainty computed the values

$$z_j = (\hat{Y}_j(x_0) - \alpha - \beta x_0) / \sqrt{\hat{\mathrm{Var}}\, \hat{Y}_j(x_0)}, \quad j = 1 \ldots m,$$

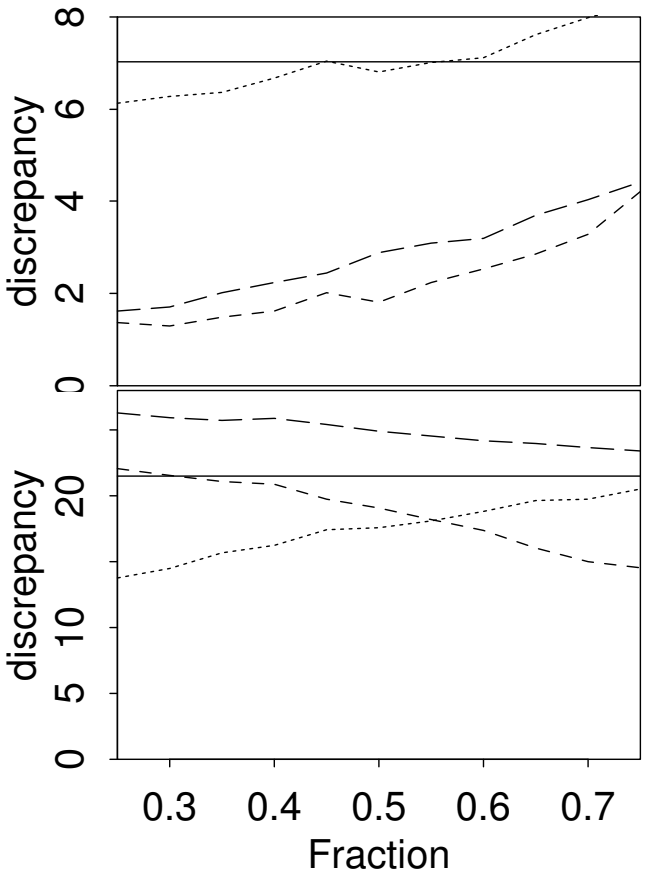

Figure 4: Maximum discrepancy(x100) for strategies N — solid, A — dotted, B — short dashes, C — long dashes with f varying. $\beta = 0$ in the top pane and $\beta = 0.5$ in the bottom

and plotted $l_j = j/(m+1)$ against $\Phi(z_{(j)}) - l_j$ in Figure 2 for the case $m = 1,000$ replications, with $n = 20, \alpha = 0, \sigma = 1, c = 1, x_0 = 1, \beta = 0.5, f = 0.5$. For each replication, a new set of X's are generated. A fixed value of c was used for all four strategies. The test for $\beta = 0$ should be regarded as a rudimentary variable selection method rather than a hypothesis test, so levels of significance are not of concern here. The normal distribution is not quite appropriate here, but it is difficult to determine what is. I have smoothed the plot a small amount again.

That the bulk of all curves lie below the line is indicative of the general underestimation of both the true value and the true variance of prediction in all strategies. In particular, one can see that the actual coverage of the confidence intervals will be far less than the nominal. In this example the naive strategy is performing almost as well as any of the other strategies, if not better than B and C particularly.

Is this true in general? Using a summary of performance

in the honesty plots such as the maximum distance between the curve and the zero line we can see how this summary varies as β (Figure 3) and and f (Figure 4) are varied, using $m = 40,000$ replications, with $n = 20, \alpha = 0, \sigma = 1, c = 1, x_0 = 1$, in Figures 3. As can be seen there are no clear winners or losers. I also tried varying other parameters but the general message was the same. Given the better point performance of the naive strategy, it seems that data splitting pays a price without delivering the reward.

3.2 Box-Cox transformation

Let $X_i \sim U(0,1)$ and ϵ_i be i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots n$. Let

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

We observe X and Y but not λ ($\lambda = 0$ is equivalent to $\log Y$), which we will determine using the Box-Cox method. In keeping with common practice, we select λ from a finite set of interpretable values, in this case {-2.0, -1.5, -1.0, -0.5, -0.25, 0.0, 0.25, 0.5, 1.0, 1.5, 2.0}. We simply pick the value in this set that gives the highest value of the likelihood. Then α and β are then estimated using least squares, without testing β as in the previous example.

So in this example, the model selection is just the determination of λ, and again additional data analysis would normally be done in practice. The model selection effects were investigated by Bickel and Doksum (1981). Some controversy was engendered over the interpretation of β when λ is variable – see Box and Cox (1982) and Hinkley and Runger (1984). Since we focus on prediction and not parameter estimation, these interpretational issues do not arise as long as one accepts that the estimation of λ using the Box-Cox method does potentially inject extra variability into the prediction. The prediction and variance estimates usually need to be transformed back to the scale of observation. The variance estimates are transformed using the delta method.

RMSE error for this example are much the same as for the variable selection. Naive and C dominate and the performance of A and B varies according to the parameter settings. In the plots of the maximum discrepancy in Figure 5, as β is varied, using $m = 40,000$ replications, with $\lambda = 0, n = 20, \alpha = 0, \sigma = 1, x_0 = 0.5$. We see that the naive strategy does a lot better than the rest, because all of the data are used to determine λ, which is the crucial aspect of this scenario. The use of the wrong λ leads to biased predictions. C dominates both A and B.

This example demonstrates one major danger of splitting the data — the model selection may be seriously damaged if only a fraction of the data are used for this purpose.

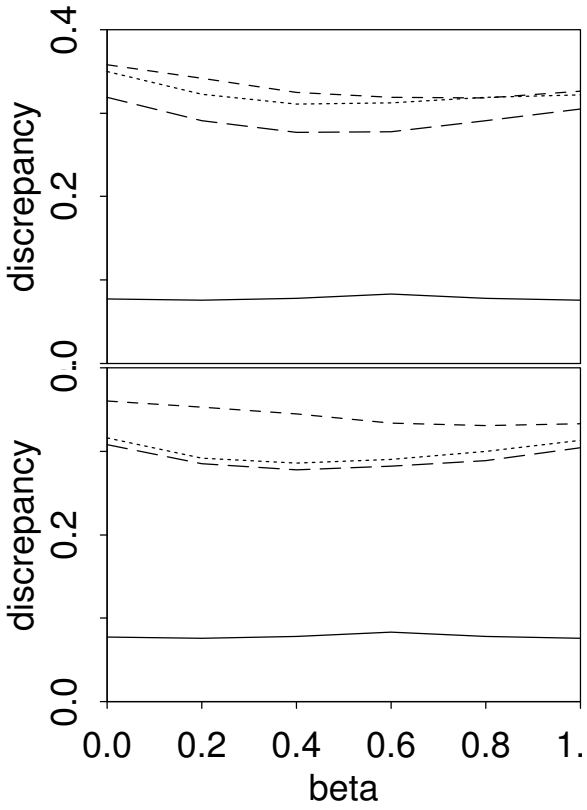

Figure 5: Maximum discrepancy using the Box-Cox transformation for data splitting strategies N — solid, A — dotted, B — short dashes, C — long dashes as β varies, $f = 0.5$ in top pane and $f = 0.75$ in bottom pane.

4 Complex Model Selection

As the data analysis becomes more complex, the number of models considered becomes much greater along with the possibility of substantial model uncertainty. Given that both frequentist and Bayes approaches to adjusting for model uncertainty become impractical as soon as more than one or two data analytic procedures are used, we might hope that data splitting would provide at least a crude assessment of true variability. We consider a regression problem where the data

are generated from the model

$$g(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(X_{ij}) + w_i \epsilon_i$$

where X_{ij} are independently distributed F_x, ϵ_i is distributed F_ϵ with $\beta_0 = 0$ and $\beta_i = 1,\ i \neq 0$.

Let us try to predict the mean response Y at the point $(0.2, 0.2, ..., 0.2)$. We will consider four scenarios which will be represented as modifications of the following default values: $n = 50, p = 5$, g & f_j identity functions, F_x & F_ϵ standard normal, $w_i = 1$.

Model Label	Description
Vanilla	Default values
Outlier	$F_\epsilon \sim \frac{2}{3}N(0,1) + \frac{1}{3}N(0,3^2)$
Nonlinear	$F_x \sim U(0,0.2)$, $g^{-1}(x) = e^{x/5}$
Hetero	$F_x \sim U(0,0.2)$, $w_i = \sum_{j=1}^{p} \beta_j f_j(X_{ij})$

Now in the previous example, the variable selection procedure was precisely specified so it was straightforward to simulate. In contrast, a regression analysis carried out by a Statistician will be quite complex. Several procedures will be used. Some procedures, typically diagnostic, involve the assessment of graphical figures. The ordering of the procedures is often not independent — the choice of the next action may depend on the outcome of the last. Such a regression analysis is almost impossible to specify precisely in a completely realistic manner. Nonetheless, since it will be impractical and unrepeatable to do a large number of data analyses by hand, I have attempted to program a somewhat realistic regression analysis. I did this in Faraway (1992) which should be consulted for more details but I will outline the procedures below.

I completely specify a regression data analytic action $R()$ so that $\mu' = R(\mu)$ where μ', μ are regression models/data and μ is arbitrary. The functions I have implemented functions are

1. Outlier check
2. Influential points check
3. Check for non-constant variance
4. Box-Cox transformation
5. Check for transformation on predictors
6. Variable selection by backward elimination
7. Variable selection by forward selection
8. Outlier/Influential Point restoration

The regression analysis consisted of these actions applied in the stated order. Ten thousand replications were used with a fraction $f = 0.5$ in every case. The results are shown in Figures 6, 7, 8, 9.

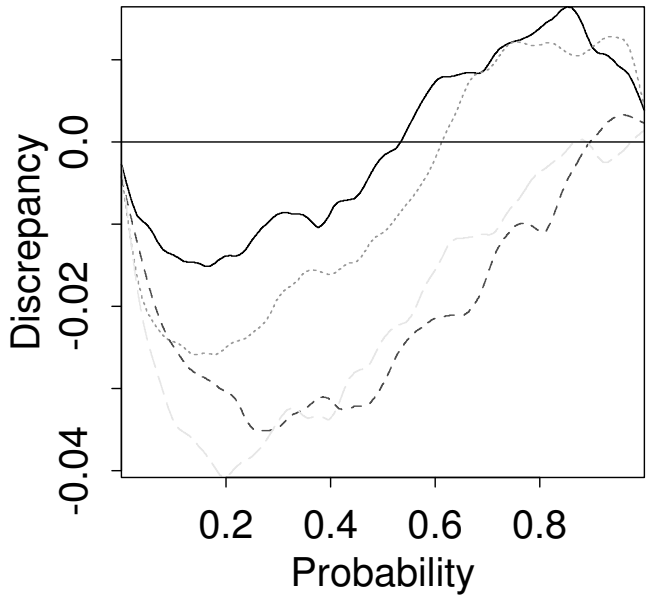

Figure 6: Scenario: Vanilla. Maximum discrepancy for strategies N — solid, A — dotted, B — short dashes, C — long dashes

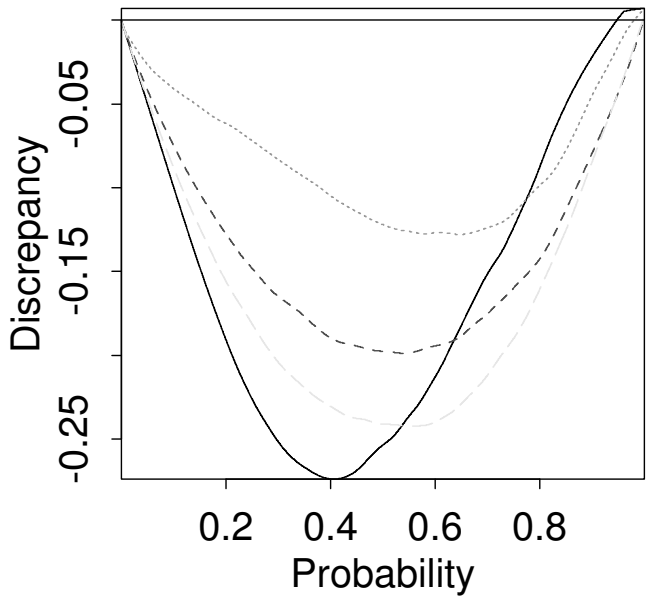

Figure 7: Scenario: Outlier. Maximum discrepancy for strategies N — solid, A — dotted, B — short dashes, C — long dashes

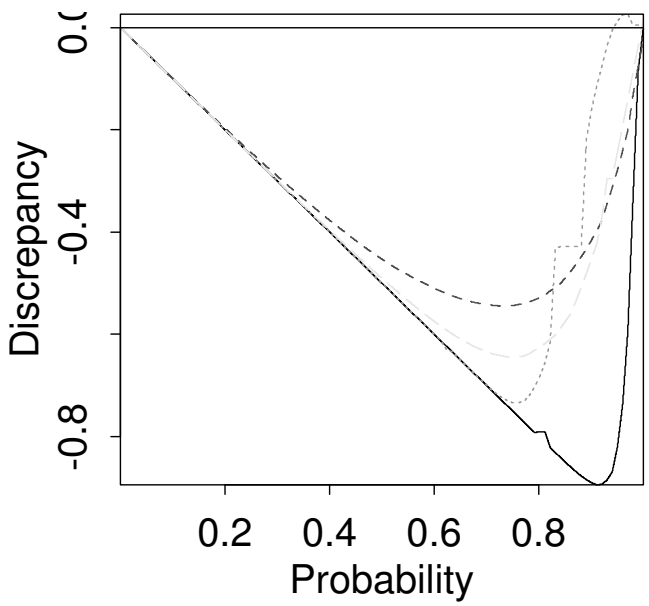

Figure 8: Scenario: Hetero. Maximum discrepancy for strategies N — solid, A — dotted, B — short dashes, C — long dashes

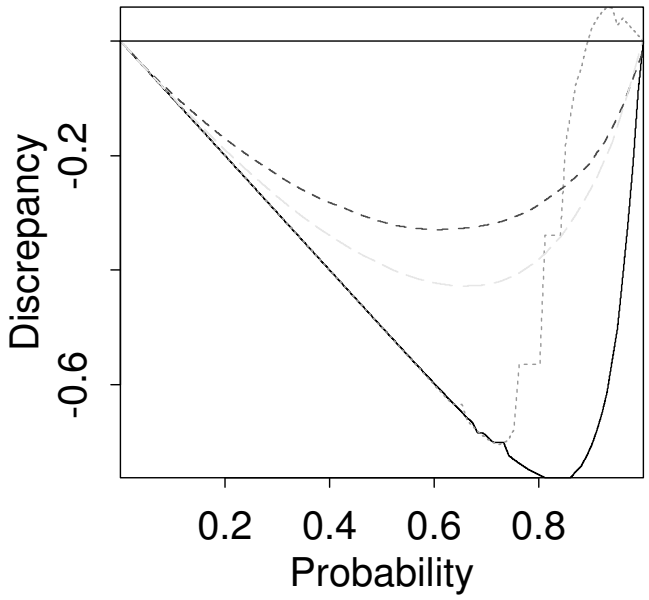

Figure 9: Scenario: Non-Linear. Maximum discrepancy for strategies N — solid, A — dotted, B — short dashes, C — long dashes

In the "Vanilla" case, where no data analysis was really necessary, we see that the naive strategy is best although the discrepancies are not substantial for any of the strategies. In the "Outlier" case, we see that strategy A was best with B beating C but the assessment of the naive strategy depends on how the discrepancy is judged. In the "Hetero" case, we see generally large discrepancies with the Naive strategy being the worst. However, the performance of the other three is in no way good considering that point accuracy has been sacrificed to obtain better variability assessment. In the "Nonlinear" case, we again see poor performance for all strategies with A and the naive being the worst.

In other simulations not shown here, the best strategy varied according to the situation. I do not believe it is possible to say that one strategy is better than another as far as assessing variability goes. On the other hand, point accuracy is surely degraded by data splitting. Even for given data, I do not think it is possible to determine the best strategy without access unknowable information like the "true" model and its parameters.

Throughout, I tried to see if there was some universally recommendable value of f. Without quite substantial knowledge of the true model and even its parameters, it is impossible to pick f optimally. Furthermore, the best f varies substantially from case to case.

5 Conclusion

Data splitting strategies A, B & C cannot be distinguished from one another in the sense of honesty of prediction, although C does tend to give better point performance. We see that data splitting sacrifices some accuracy in point estimates without the reward of greater honesty in prediction. In some cases the results (see the Box-Cox case) are significantly worse when data splitting is used. The situations we have directly investigated are relatively simple and not entirely realistic, but we cannot feel any confidence that it would work any better in more complex and realistic situations. Given that data splitting will most assuredly cost something in terms of predictive accuracy, without any guarantee of a return in predictive honesty, it seems difficult to recommend.

So it seems one is stuck with the complex approaches described in the introduction. For cases where the data-analytic process is well specified and not artificially limited — Madigan and Raftery (1994) would be a good example — then the "big model" approach can be used, utilising either frequentist or Bayesian methodology according to religious preference.

When the data analytic process is too complex or vague to specify completely, then one can use *a posteriori* model expansion as in Draper (1995) and try to do an honest job in choosing the direction of such expansion. An alternative

is Faraway (1992)'s bootstrap-based approach. This may be workable for preprogrammed data-analytic actions, but general application awaits more sophisticated data-analytic environments than are currently available.

These findings also have implications for a succession of data analyses. Strategies B and C can be compared to the situation where new data become available after a model has been discovered using the original data. Both B and C do not reselect the model, but given their performance above it seems that it would be better to start from scratch – pool all the data and reselect the model.

Acknowledgements

This work was started while visiting the Statistics Group at the University of Bath. My particular thanks to Chris Chatfield & David Draper.

References

Bickel, P. and K. Doksum (1981). An analysis of transformations revisited. *JASA 76*, 296–311.

Box, G. and D. Cox (1982). An analysis of transformations revisited, rebutted. *JASA 77*, 209–210.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A Statistics in Society 158*(3), 419–466.

Draper, D. (1995). Assessment and propogation of model uncertainty. *JRSS-B 57*, 45–97.

Faraway, J. (1992). On the Cost of Data Analysis. *Journal of Computational and Graphical Statistics 1*, 215–231.

Freedman, D., W. Navidi, and S. Peters (1988). On the Impact of Variable Selection in Fitting Regression Equations. In T. K. Dijkstra (Ed.), *Lecture Notes in Economics and Mathematical Systems*, pp. 1–6. Springer-Verlag.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological) 14*(1), pp. 107–114.

Hill, B. (1990). A theory of bayesian data analysis. In S. Geisser (Ed.), *In Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, pp. 49– 73. North Holland.

Hinkley, D. and G. Runger (1984). The analysis of transformed data (with discussion). *JASA 79*, 302–319.

Hirsch, R. (1991). Validation samples. *Biometrics 47*(3), 1193–1194.

Hurvich, C. and C.-L. Tsai (1990). The impact of Model Selection on Inference in Linear Regression. *American Statistician 44*, 214–217.

Kipnis, V. (1991). Evaluating the impact of exploratory procedures in regression prediction. *Comp. Stat. Data. Anal. 12*, 39–55.

Madigan, D. and A. Raftery (1994). Model Selection and Accounting for model uncertainty in graphical models using Occam's window. *JASA 89*, 1535–1546.

Miller, A. (1990). *Subset selection in regression*. Boca Raton, FL: CRC Press.

Mosteller, F. and J. Tukey (1977). *Data analysis and regression. A second course in statistics*. Reading, Mass: Addison-Wesley.

Picard, R. and K. Berk (1990). Data splitting. *American statistician 44*, 140–147.

Picard, R. and R. Cook (1984). Cross-Validation of Regression Models. *JASA 79*, 575–583.

Pötscher, B. (1991). Effects of model selection on inference. *Econometric Theory 7*(2), 163–185.

Raftery, A., D. Madigan, and J. Hoeting (1993). Model Selection and Accounting for Model Uncertainty in Linear Regression Models. Technical Report 262, Department of Statistics, University of Washington.

Roecker, E. (1991). Prediction error and its estimation for subset-selected models. *Technometrics 33*, 459–468.

Snee, R. (1977). Validation of regression models. Methods and examples. *Technometrics 19*, 415–428.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological) 36*, 111–147.