PERSPECTIVE

The Problem of Overfitting

Douglas M. Hawkins*

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455

Received October 30, 2003

INTRODUCTION − THE PURPOSE OF QUANTITATIVE MODELS

Model fitting is an important part of all sciences that use quantitative measurements. Experimenters often explore the relationships between measures. Two subclasses of relationship problems are as follows:

• *Correlation problems:* those in which we have a collection of measures, all of interest in their own right, and wish to see how and how strongly they are related.

• *Regression problems*: those in which one of the measures, the *dependent variable*, is of special interest, and we wish to explore its relationship with the other variables. These other variables may be called the *independent variables*, the *predictor variables*, or the *covariates.* The dependent variable may be a continuous numeric measure such as a boiling point or a categorical measure such as a classification into mutagenic and nonmutagenic.

We should emphasize that using the words 'correlation problem' and 'regression problem' is not meant to tie these problems to any particular statistical methodology. Having a 'correlation problem' does not limit us to conventional Pearson correlation coefficients. Log-linear models, for example, measure the relationship between categorical variables in multiway contingency tables.

Similarly, multiple linear regression is a methodology useful for regression problems, but so also are nonlinear regression, neural nets, recursive partitioning and *k*-nearest neighbors, logistic regression, support vector machines and discriminant analysis, to mention a few. All of these methods aim to quantify the relationship between the predictors and the dependent variable. We will use the term 'regression problem' in this conceptual form and, when we want to specialize to multiple linear regression using ordinary least squares, will describe it as 'OLS regression'.

Our focus is on regression problems. We will use *y* as shorthand for the dependent variable and *x* for the collection of predictors available. There are two distinct primary settings in which we might want to do a regression study:

• *Prediction problems:* We may want to make predictions of *y* for future cases where we know *x* but do not know *y*. This for example is the problem faced with the Toxic Substances Control Act (TSCA) list. This list contains many tens of thousands of compounds, and there is a need to identify those on the list that are potentially harmful. Only a small fraction of the list however has any measured biological properties, but all of them can be characterized by chemical descriptors with relative ease. Using quantitative structure−activity relationships (QSARs) fitted to this small fraction to predict the toxicities of the much larger collection is a potentially cost-effective way to try to sort the TSCA compounds by their potential for harm. Later, we will use a data set for predicting the boiling point of a set of compounds on the TSCA list from some molecular descriptors.

• *Effect quantification:* We may want to gain an understanding of how the predictors enter into the relationship that predicts *y*. We do not necessarily have candidate future unknowns that we want to predict, we simply want to know how each predictor drives the distribution of *y*. This is the setting seen in drug discovery, where the biological activity *y* of each in a collection of compounds is measured, along with molecular descriptors *x*. Finding out which descriptors *x* are associated with high and which with low biological activity leads to a recipe for new compounds which are high in the features associated positively with activity and low in those associated with inactivity or with adverse side effects.

These two objectives are not always best served by the same approaches. 'Feature selection'—keeping those features associated with *y* and ignoring those not associated with *y* is very commonly a part of an analysis meant for effect quantification but is not necessarily helpful if the objective is prediction of future unknowns. For prediction, methods such as partial least squares (PLS) and ridge regression (RR) that retain all features but rein in their contributions are often found to be more effective than those relying on feature selection.

What Is Overfitting? Occam's Razor, or the principle of parsimony, calls for using models and procedures that contain all that is necessary for the modeling but nothing more. For example, if a regression model with 2 predictors is enough to explain *y*, then no more than these two predictors should be used. Going further, if the relationship can be captured by a linear function in these two predictors (which is described by 3 numbers—the intercept and two slopes), then using a quadratic violates parsimony.

Overfitting is the use of models or procedures that violate parsimony—that is, that include more terms than are necessary or use more complicated approaches than are necessary. It is helpful to distinguish two types of overfitting:

• Using a model that is more flexible than it needs to be. For example, a neural net is able to accommodate some curvilinear relationships and so is more flexible than a simple linear regression. But if it is used on a data set that conforms to the linear model, it will add a level of complexity without

* Corresponding author e-mail: doug@stat.umn.edu.

10.1021/ci0342472 CCC: $27.50
Published on Web 12/02/2003

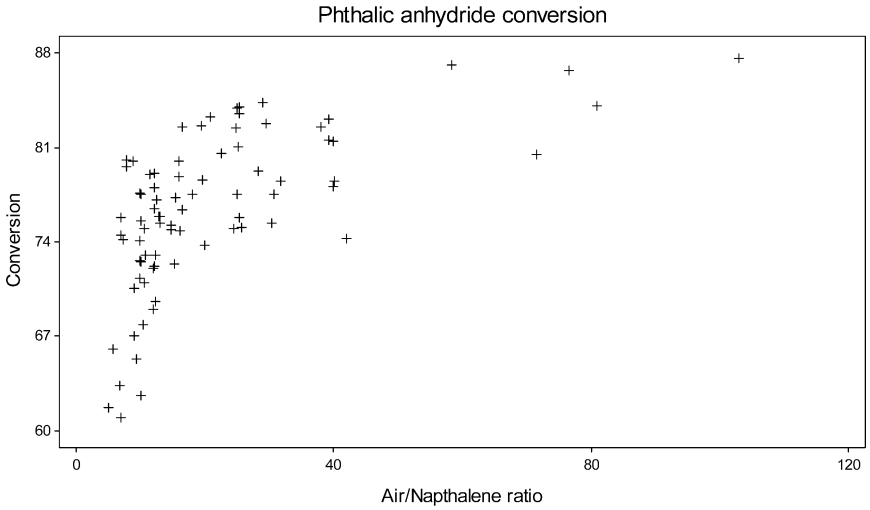

Figure 1.

any corresponding benefit in performance or, even worse, with poorer performance than the simpler model.

• Using a model that includes irrelevant components—for example a polynomial of excessive degree or a multiple linear regression that has irrelevant as well as the needed predictors.

Overfitting is undesirable for a number of reasons.

• Adding predictors that perform no useful function means that in future use of the regression to make predictions you will need to measure and record these predictors so that you can substitute their values in the model. This not only wastes resources but also expands the possibilities for undetected errors in databases to lead to prediction mistakes.

• In a feature selection problem, models that include unneeded predictors lead to worse decisions. In drug discovery for example, a mistaken decision to use the number of NH_2 groups in a QSAR model when this number is actually irrelevant will lead to wrongly ignoring compounds based on their irrelevant number of NH_2 groups. Valuable leads can be lost in this way.

• Adding irrelevant predictors can make predictions worse because the coefficients fitted to them add random variation to the subsequent predictions.

• The choice of model impacts its portability. At one extreme, a one-predictor linear regression that captures a relationship with the model is highly portable. Anyone anywhere can apply your model to their data simply by knowing two numbers—the values of the regression slope and intercept. At the other extreme are models that are not at all portable but can be effectively reproduced only by reusing the modeler's software and calibration data. A fundamental requirement of science is that one experimenter's results can be duplicated by another experimenter in another location, so it needs no emphasis that where both are valid more portable models are to be preferred to less portable.

A Toy Example. To show some of the key issues, we will use a simple data set, which was drawn to the author's attention by R. Dennis Cook. The data set, an oxidation experiment from Franklin et al.,[1] lists several variables; we will use as dependent the percentage mole conversion of naphthalene to phthalic anhydride. There are 80 cases. We will work with a single predictor: the air-to-naphthalene ratio in the charge.

Figure 1 is a scatterplot of these variables. A glance at the plot is enough to warn that a simple linear regression cannot capture the relationship. Modelers who favor linear regression respond by looking for a transformation of *x* and/or *y* that might straighten the relationship. A log transformation of *x* stretches the scale on the left and compresses it on the right, as this plot clearly calls for, and might be the first thing tried. The result is shown in Figure 2, which also has a fitted regression line for reference.

The log transformation seems fairly successful in making the plot linear, and we will work with it and from here on write *y* for the dependent phthalic hydride conversion fraction and *x* for the natural-log-transformed air/naphthalene ratio. Next, we split the data set into two samples by sorting the pairs in ascending order of *x* and separating out the odd-numbered and the even-numbered cases. This way of splitting ensures that the distribution of *x* values of the two samples is much the same.

We fitted various statistical models to the data. Each model was fitted separately to the odd-number sample and to the even-number sample, and its fit was assessed by applying it to both samples. Reapplying the fit to the same data set (the 'calibration set') that was used to fit it is called 'resubstitution' and is contrasted with 'holdout', which is applying it to the other data set that was not involved in the fitting.

The first fit is a simple linear regression of *y* on *x*. The intercept and slope and their standard errors and R^2 are below. Also shown are two mean squares of the difference between the actual *y* and its prediction. The resubstitution mean square bases its predictions on the regression fitted to that sample, and the holdout mean square bases its prediction on the regression fitted to the other sample.

	odd-number sample	even-number sample
intercept	58.33	59.19
(s.e.)	2.91	3.06
slope	6.27	6.16
(s.e.)	1.01	1.04
R^2	0.5036	0.4787
resubstitution mean square	15.80	18.06
holdout mean square	16.14	18.40

In both the odd- and even-number samples, the resubstitution and holdout mean squares are very similar, showing that the

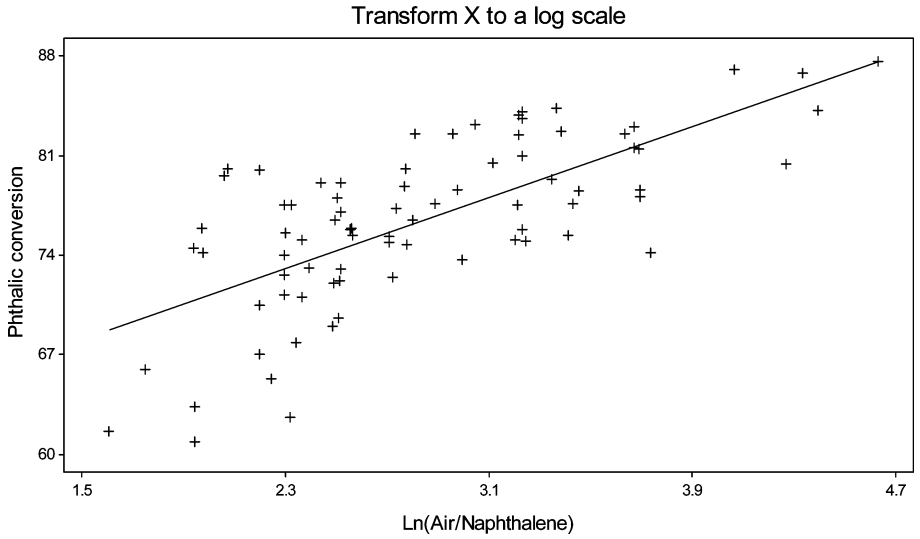

Figure 2.

same-sample and out-of-sample predictions are equally precise. This is a hallmark of a parsimonious model, though not a guarantee that the parsimonious model fits well.

Figure 2 showed that the log transformation had removed much of the curvature but perhaps not all. We now use two simplistic but illustrative alternative fitting methods to address this concern. We will evaluate the fit of each with three prediction mean squares—the two used above and augmented by a third:

• The leave-one-out (LOO) cross-validation mean square. Like the resubstitution mean square, this also reuses the calibration data as a test sample, but when predicting each calibration case LOO removes it from the sample and refits the model on the remaining cases, using the calibration sample itself to simulate the prediction of future unknowns.

We will put the LOO and holdout measures together under the label 'out-of-sample' measures to reflect the fact that they are computed using cases that were withheld from the selection and calibration of the model. This contrasts with the 'in-sample' measure given by resubstitution.

The first more flexible fit is a nearest neighbor approach—predict the y of any case to be the observed y of the experiment with closest x and, in the event of tied x's, the closest in time order**.** The three mean squares calculated using each of the half-samples are as follows:

mean square	odd-number sample	even-number sample
resubstitution	0	0
holdout	23.37	26.29
leave-one-out	25.49	32.89

Several comments can be made here. The resubstitution error is zero, since the calibration case most similar to any calibration case is itself, which predicts it perfectly. While extreme, this illustrates the general principle that the resubstitution estimate of any method overestimates its actual performance and that the more flexible the method the more overoptimistic the estimate.

Next, for both samples, the LOO mean square is similar to but higher than the holdout estimate. This is despite the fact that LOO is reusing the same data set while the holdout is not. The reason is that LOO is actually a more stringent test than holdout since a case's nearest neighbor in LOO in its own sample is on average slightly further away than is its nearest neighbor in the other sample. This illustrates the point that the LOO (which does not require a separate validation sample) gives a serviceable and slightly conservative estimate of precision.

Finally, the holdout mean squares of the nearest neighbor are both higher than those using the simple linear regression—23.27 vs 16.14 for fitting to the even-number sample and evaluating using the odd-number and 26.29 vs 18.4 for fitting to the odd-number sample and evaluating using the even-number. The nearest-neighbor approach therefore overfits the data. In this data set, using this more flexible model to address concerns about nonlinearity costs more than the nonlinearity it is meant to cure.

As an example of the other type of overfitting, we fitted a quintic polynomial by least squares. The coefficients and the t values testing whether the coefficient might be zero are as follows:

	odd-number sample		even-number sample	
	coeff	t	coeff	t
linear	2921	2.71	−725	−0.56
quadratic	−2132	−2.73	540	0.61
cubic	761	2.74	−190	−0.65
quartic	−132	−2.75	32	0.68
quintic	9	2.76	−2	−0.70

The coefficients of the even-half fit and the odd-half fit are utterly different to the point of not even having the same sign. Further, when fitting the quintic to the odd-number sample, the fifth power is significant at the 0.9% level, which would lead many analysts to think that a fifth degree polynomial fit was justified; fitting to the even-number sample however gave no significance above the linear term. This instability of the fitted model is disconcerting, given the close matching of the two samples it is fitted to.

The three error measures for these fits are as follows:

mean square	odd-number sample	even-number sample
resubstitution	11.62	16.78
holdout	17.64	42.84
leave-one-out	13.43	18.98

• The resubstitution mean squares are lower than they were for the linear regression, illustrating how adding even

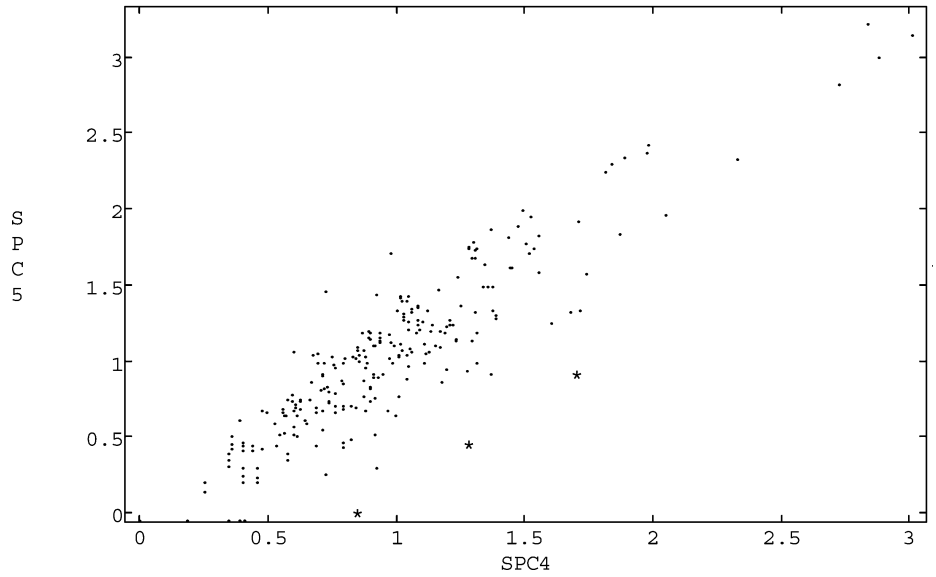

Figure 3.

nonsense terms to a regression improves the apparent fit to the data used in the calibration.

- The resubstitution and LOO error estimates agree much better than in the 1-NN case. This is because the quintic polynomial is less of an overfit than is the 1-NN.
- The holdout mean square for calibrating with the even-number sample and testing with the odd is 17.64, higher than was found using simple linear regression, illustrating that adding unnecessary predictors, while reducing the resubstitution error, can make holdout out-of-sample predictions worse.

The holdout mean square for fitting with the odds and verifying with the evens is enormous: 42.84. Most of this comes from the last two cases. These are extrapolations for the model calibrated using the odd-numbered cases, and the terrible predictions are an illustration of the way high degree polynomials explode outside the range in which they were fitted.

This toy example illustrates the features of overfitting—the fit as measured by resubstitution of the overfitted model is better than that of the more parsimonious model, but that measured by holdout is worse. And while the LOO gave a good picture of the ability of the model to generalize in five of the six settings illustrated, it did not capture how the overfitted regression would behave when used for severe extrapolation.

It is important not to lose sight of the fact that overfitting is not an absolute but involves a comparison. One model is an overfit if its predictions are no better than those of another simpler model. Thus to decide whether a model is overfitting, it is necessary to compare it with other simpler models. The original linear regression looked to be defensible if less than ideal. Both the 1-NN and the quintic gave worse out-of-sample predictions, whether measured by the independent holdout or by LOO cross-validation. This means that both these methods overfit this data set.

As this illustrates, the greater flexibility of the higher-degree polynomial and the 1-NN approach is both a blessing and a curse. It is a blessing in that it allows the fitted model to capture a richer variety of relationships. It is a curse in that it makes the fit more dependent on the individual observations, which adds random variability and can lead to worse predictions.

Interpolation and Extrapolation. Everyone knows that extrapolation is dangerous and undesirable. However it is not always obvious which predictions are extrapolations and which interpolations and different modeling methods break down in different ways when used to extrapolate.

To illustrate the first point, it is obvious that we are extrapolating if any of the predictors is outside the range of values covered by the calibration data. Extrapolations of this type are easy to detect. Multivariate extrapolations however are less easily seen. Figure 3 is a plot of ${}^4\chi_{PC}$ (axis labeled 'SPC4) versus ${}^5\chi_{PC}$ (axis labeled SPC5) for the TSCA boiling point data set. Note the three points are plotted with asterisks; even though their ${}^4\chi_{PC}$ and ${}^5\chi_{PC}$ are within the range of values spanned by the data set, they are multivariate extrapolations. It is not obvious from the plot, but these points are more than 4 standard deviations from the main cloud of points in the northwest to southeast direction.

Gnanadesikan[2] coined the label 'Type B multivariate outliers' for these extrapolations to contrast them with the 'Type A multivariate outliers' that stick out in one or more of the individual components and so are easily found by simple range checks on the data. With just two predictors of course, Type B outliers can easily be found in scatter plots of the predictors such as Figure 3, but with three or more predictors this is no longer effective. Numeric measures of extrapolation are therefore highly valuable.

Different methods incur different kinds of errors in extrapolation, so the modeling method also figures in the degree to which extrapolation is dangerous.

When linear modeling methods such as conventional regression extrapolate, their predictions increase or decrease without limit and without regard to any physical bounds the dependent might have. Extrapolations of linear models can be catastrophically bad. Inverse distance weighted average methods tend to give extrapolation that move toward the grand mean of the calibration data. While not as bad as an unbounded prediction, this is also a poor choice in settings where there is a strong predictive model. Nearest neighbor and recursive partitioning predict extrapolated cases using

the mean of some nearby calibration cases. This is generally better than using the grand mean and better than a linear fit for extreme extrapolation but worse for more moderate extrapolation.

Assessment of Fit - - Known Statistical Results. While high-degree polynomials are particularly prone to bad extrapolation, there is no method that is entirely free of suspicion when extrapolating. It is therefore helpful to use measures of multivariate compatibility of the predictions to recognize extrapolations where there are multiple predictors.

We can get at least a conceptual picture by looking at the case of a single-predictor linear regression using the 'nicest' text-book statistical model

$$y = \beta_0 + \beta_1 x + e$$

where β_0 and β_1 are the true intercept and slope, and the true residual e follows a Gaussian distribution with zero mean and constant variance σ^2. If you calibrate the regression using a sample of size n, the 'unusualness' of an arbitrary predictor value x (which may be the x of one of the cases, or some x where a future prediction is to be made) is measured by its 'leverage'

$$h = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

Leverage measures whether you are interpolating or extrapolating. The minimum possible value of h is $1/n$ and corresponds to the best possible degree of interpolation. Values of x that fit comfortably within the calibration set have small leverage, and those outside the set have large leverage. The leverage h carries over into the multiple predictor multiple regression setting with the same general interpretation but a more complicated formula. If you fit a multiple regression with p coefficients, the leverage of the cases in the calibration data set will average to exactly p/n. There is a common heuristic to classify a position x as low or high leverage according to whether their h is below or above $2p/n$.

Leverage is highly relevant to assessment of multiple regression fit because

- The resubstitution fitted residual of one of the cases used in the calibration will have a variance $\sigma^2(1 - h)$, where h is that case's leverage.
- Using the fitted regression to predict the y of some future case, or of a holdout case, gives a prediction error with variance $\sigma^2(1 + h)$.

This means that high leverage cases in the calibration data will conform to the fitted regression very (and unrealistically) well. But if you wish to use the fitted regression to make a prediction of a future case with that same leverage, the prediction is likely to be quite poor. Looking at the mean squared errors of prediction and the properties of the leverage h, this shows the well-known result that the resubstitution mean square in multiple regression has an expected value of $\sigma^2(1 - p/n)$ while with evenly spread cases, the out-of-sample mean squared error is, to a rough approximation, $\sigma^2(1 + p/n)$. If p/n is small, the two measures will be quite similar, but if p/n is large the resubstitution mean square will be much lower than the LOO. This is shown in the naphthalene data where the linear regression ($p = 2$) had similar values for the two measures, the ratio of the holdout to the resubstitution mean square being 1.02, very close to 1, but the quintic polynomial ($p = 6$) showed much bigger differences, the ratio being around 1.15.

Large values of h are a warning of extrapolation and unreliable prediction. For example, when fitting the quintic polynomial to the odd-numbered compounds in the naphthalene data set, the largest of the hold-out even-numbered compounds had a leverage of 15, so it is no surprise that it was predicted so badly.

While these neat algebraic properties do not carry over to other modeling methods, the concepts do. Any modeling method that gives unbiased estimates will give residuals whose variance is of the form $\sigma^2(1 - h)$, where h is a function of the case's position and of the flexibility of the method. A flexible method adapts well to each case in the data set. This implies that each case has a high leverage h. This is illustrated by the nearest neighbor fit to the naphthalene data. This fits one constant per observation, so $p = n$ and the resubstitution errors are all zero.

In general, when we apply a fitted model to some future case that was not in the calibration data set, the error of prediction will have a variance that we can conceptualize as $\sigma^2(1 + h)$ where h measures the atypicality of the case and is also a function of the modeling method. In the nearest-neighbor prediction, all cases have $h = 1$, so the prediction error variance is $2\sigma^2$.

A large difference between the resubstitution error variance and one obtained by either cross-validation or an independent holdout sample is a hallmark of a large h setting—the ratio of the number of tunable parameters to the sample size is large. This is not exactly the same thing as overfitting, though the two are very close. A model whose resubstitution error is much smaller than that of a holdout is suspect of being overfitted, but the converse is not true. An overfitted model could give quite similar values for the resubstitution and a holdout error. For example in the naphthalene data regression using a quadratic does not fit appreciably better than a linear function and so is an overfit, even though it gives quite similar resubstitution and holdout error variances.

Yaffe et al.[3] describe the use of a classification procedure to predict aqueous solubility of organic compounds. In a calibration set of size $n = 437$, the resubstitution errors have a standard deviation of 0.0045. An independent holdout data set of size 78 gives prediction errors with a standard deviation of 0.16, which is larger by a factor of 35. The vast difference between these two estimates of error is an indication that the modeling method has close to one parameter per observation. While this certainly raises the suspicion that the model overfits, a final judgment would have to be by whether there was some simpler model that described the data equally well or whether this high level of model flexibility really was required.

How To Assess the Fit of a Model. After a model has been fitted, there is the question of assessing its quality. This actually involves a number of elements, not all of which are widely recognized.

- Is the form of the model appropriate?

This issue is illustrated by the naphthalene data, with the phthalic anhydride as dependent, and the air/naphthalene ratio as predictor. One could blindly fit a linear regression on the untransformed scale of air/naphthalene. Fitted to one-half

and evaluated on the other, this gives mean squared prediction errors of 21.5 for each half. This is well above the figures we get with the transformed *x*, illustrating the fact that the natural scale of the predictor is less effective than the log scale. But there is a deeper concern—bias. Like the stopped clock which gives the right time twice a day, the natural scale regression is unbiased at two levels of air/naphthalene ratio—at about 5 and 80, but at all other values of the ratio its predictions are systematically biased.

Note that getting the right scale (natural or transformed) for a predictor is obviously important in 'global fit' methods such as multiple regression but matters even in more model-free methods such as the nearest neighbor since a case's neighbors on the log scale of a predictor may differ from those on the natural scale.

Texts on regression modeling like Cook and Weisberg[4] stress the value of graphical and numeric checks for curvature, outliers, extrapolations and the like, and it is a good idea in any modeling exercise to run graphic checks such as plots of the residuals versus the predicted values and the individual predictors to screen problems such as these.

- What are sensible measures of fit?

There is a tendency to use mean-square-based measures of fit of models, such as the residual standard deviation or some form of R^2. These measures are not particularly appropriate if the variability of the data is not constant. So, for example, if the variance of *y* increases with its mean (as is extremely common) measures involving mean squares (including R^2) tend to be overwhelmed by whatever is occurring at the high end of the *y* scale. Once again, this question is best dealt with by some graphical checks, possibly followed by a transformation. For example, if *y* has a constant coefficient of variation, then it may make sense to work with *y* on a logarithmic scale.

- Assuming that we have solved problems of model form and suitability of fit measures, how do we measure the fit?

The common situation is that we want to measure fit using a mean-squared-based measure such as the residual standard deviation or perhaps some form of R^2. Since prediction of future unknowns will lead to a prediction error with a variance of the generic form $\sigma^2(1 + h)$ where *h* is some leverage-like quantity reflecting random variability in the fitting procedure, it makes logical sense to use a mean square of prediction error as the measure of fit. Despite having a number of known deficiencies however, R^2 is the most widely used measure of fit, and so we will concentrate discussion on assessment using variants of R^2 with the general comment that the broad conclusions apply also to measures based on variance of prediction.

Let y_i be the value of the dependent observed on the *i*th of some collection of size *m* cases (we will call this set the 'assessment set') being used to assess the model fit. Write $\hat{y}_i$ for the prediction of y_i given by some model and $\bar{y}$ for the mean of the *y*'s. The sum of squared deviations from the mean is

$$\mathrm{SSD} = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

The sum of squares of the prediction errors is

$$\mathrm{SSP} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

We can define a squared correlation by

$$R^2 = 1 - \frac{\mathrm{SSP}}{\mathrm{SSD}}$$

Several choices of the assessment set come to mind.

- The resubstitution estimate obtained from plugging in the same data that were used to fit the model.
- Use of an independent holdout test sample that was not hitherto used in the modeling process.
- Some form of sample reuse in which the same cases are used both in calibration and as holdout test cases. The best-known of these methods is 'leave-one-out' or LOO cross-validation.

It has long been known that resubstitution always overestimates the quality of the fit and can be abysmal. In the case of multiple regression, there is enough known theory to make resubstitution workable: this is done by the standard degrees of freedom correction of estimating the variance from SSP/($n-p$) and not SSP/*n*. More generally, in the class of 'linear smoothers' (for example splines) in which each $\hat{y}_i$ is a linear combination of the observed *y*'s it is possible to work out the divisor for SSP that will remove the bias in the resubstitution estimate. However outside this very limited setting the resubstitution error is generally too biased to be worth using.

The use of an independent holdout sample is intellectually attractive. If the model was fitted paying no attention to the holdout sample, and can nevertheless predict the holdouts, then the model must generalize well and conversely. However getting reliable information from the independent holdout sample requires that it be large.

Another Example. A data set on the prediction of the boiling points of 1037 compounds from the TSCA inventory using 40 topostructural indices was discussed in Hawkins et al.[5] We will use this data set as a test bed for some experiments. Of the 40 indices, 10 had the same values in all but a handful of the compounds, and for purposes of our experiments we eliminated them from the data set and continued with the remaining 30. A multiple regression gives a model that predicts the boiling point with a standard deviation of 38 °C, and $R^2 = 0.83$. Table 1 shows the fitted coefficients. Those coefficients that are statistically significant at the 0.1% level are highlighted with three stars. These are predictors that are valuable in predicting boiling point regardless of what other predictors are used, and if predictors are to be removed these predictors should not be among them. The remaining predictors are indeterminate; it may be that they are irrelevant to predicting BP, or it could be that they are valuable but that their predictive power duplicates that of other predictors.

This model is better than it might seem. There are many pairs of compounds with identical values for all predictors; these different compounds with identical descriptors will necessarily get identical predicted boiling points whatever approach is used to set up the prediction. This means that the variation in boiling points of compounds with the identical values on all predictors puts a limit on the attainable predictive accuracy of any and all models. This limit turns

Table 1. Full Regression

term	coefficient	significance	term	coefficient	significance
constant	−1118.28		$^5\chi$	61.77	
I^W_D	−682.59		$^6\chi$	117.61	***
$\bar{I}^W_D$	−2513.46		$^3\chi_C$	−182.33	***
W	737.39		$^4\chi_{PC}$	−30.21	
I^D	3908.96	***	$^5\chi_{PC}$	5.47	
H^V	78.17		$^6\chi_{PC}$	43.89	
H^D	3554.99		P_0	−6503.95	***
$\overline{IC}$	34.89		P_1	312.22	
O	−22.57	***	P_2	308.04	
M_1	1796.82		P_3	−53.91	
M_2	−1292.80	***	P_4	−78.57	***
$^0\chi$	2401.96	***	P_5	−29.71	
$^1\chi$	1525.82		P_6	−26.79	
$^2\chi$	297.44		P_7	−13.49	***
$^3\chi$	313.56	***	J	−32.94	
$^4\chi$	137.12				

out to be $R^2 = 0.888$, a mathematical bound not far above the 0.83 attained by the multiple regression model.

Regression diagnostics (as outlined in ref 4) show some concerns—there are a few unusually influential cases and large residuals, but nothing so striking as to call for a different model. In fact (as shown in previous writing on this data set) we can get better fits using more sophisticated fitting methods such as partial least squares or ridge regression, but the plain multiple regression fits well enough to illustrate our main points.

Holdout Samples. This model's explained variance betters that of many in the literature. Can we recognize its quality from holdout samples? In view of the large sample size of 1037 compounds relative to the 30 predictors ($p/n = 0.03$) and the use of a rigid modeling method we will assume that the predicted value given by this full regression gives the 'true' value for y and that there is no fitting random variability to speak of.

Let us then take random samples of various sizes to assess the model fit. As these are samples taken from the pool of 1037, they are not really holdout samples, but this is not vital for two reasons—that they form just a tiny fraction of the pool and that in any event as partial resubstitutions they tend to overstate predictive ability.

We used random samples of size $m = 5$, 10, 15, 20 and 50, drawing 1000 samples of each size and computing SSD, SSP and R^2. The box and whisker plot in Figure 4 shows the spread of the R^2 values. The plot has been truncated at zero; 8% of the R^2 for $m = 5$ were negative along with a smattering of those for $m = 10$ and 15. Another quarter of the values for $m = 5$ were higher than the mathematically maximum attainable value of 0.89.

At all m values there is substantial random variability in the R^2 values, though as expected this reduces substantially with the size of the validation set.

The true R^2 is 0.83; the averages of the estimates from the different sample sizes are as follows:

m	5	10	15	20	50
mean R^2	0.62	0.74	0.78	0.80	0.82

This brings out another point—that the estimates of R^2 from these modest-sized holdout samples are downward biased.

The conclusion one draws from this is that small independent holdout samples are all but worthless in assessing model fit. They are not reliably able to recognize a good model when they see one, are substantially biased, and have large random variability.

Large holdout samples do not have this concern; the $m = 50$ holdouts had quartiles of 0.79 and 0.87 for their R^2 estimates, suggesting that using holdout samples at least this large should generally give acceptably accurate estimates of R^2.

Correct Use of the Holdout Sample. As an aside, a mistake sometimes made when using holdout samples is the use of measures of the quality of fit that involve some sort of postprocessing to match the 'raw' predictions coming out of the model to the actual values of the holdout samples. As you do not know the y value of real future unknowns, postprocessing of the predictions is impossible and measures that use postprocessing are invalid. This mistake can be quite subtle. For example Selwood et al.[6] discussed a QSAR to predict biological activity using 16 compounds to calibrate and an independent 15 holdout samples to validate. Figure 5 is a plot of the observed versus the predicted values using their fitted model.

Along with the points, the plot also shows separate regression lines of the two sets. The regression line for the calibration data is the 45° line through the origin, but that for the validation data is much different.

The authors note that three of the validation compounds are visual outliers and argue that their chemical structure is different than the remaining compounds and that they can validly be excluded. This sort of argument is generally somewhat suspect, in that after the fact it is always easy to see ways in which some cases differ from the others. It also begs the question of how we would have known not to apply the model to these three compounds if this had truly been a set of future unknowns. However even if we accept that argument and delete these three cases, the predictions are quite poor $-SSD = 7.28$, $SSP = 4.86$, $R^2 = 0.33$. The paper's opposite conclusion, that the overall fit is quite good, rests on a freshly fitted regression line including the validation compounds. This sort of postprocessing negates the status of the holdout sample and is invalid.

An extreme form of inappropriate measure is the correlation coefficient between the predicted and observed values in the validation sample. As the correlation coefficient presupposes that you will fit an intercept and slope to the data (which is obviously impossible in the prediction setting), this is an invalid measure that cannot be used. The correct use of the hold-out validation sample is through raw, untuned deviations between the observed values and those predicted by the model.

Sample Reuse Cross-Validation. There are settings in which one might feel comfortable in holding back 50 or 100 cases as a truly independent test set, but these settings are not the norm in QSAR work where more commonly sample sizes are small. This leads to the third possibility, of a sample reuse. In this, we repeatedly split out m of the cases for validation and fit the model using the remaining $n-m$ cases. The overall model fit is then assessed by putting together the hold-out predictions of the different groups of size m. Generally, we cycle through the cases so that each case gets to be in one of the validation groups. A common example of reuse is 'leave-one-out' (LOO) cross-validation in which $m = 1$, and each sample in turn is left out and predicted

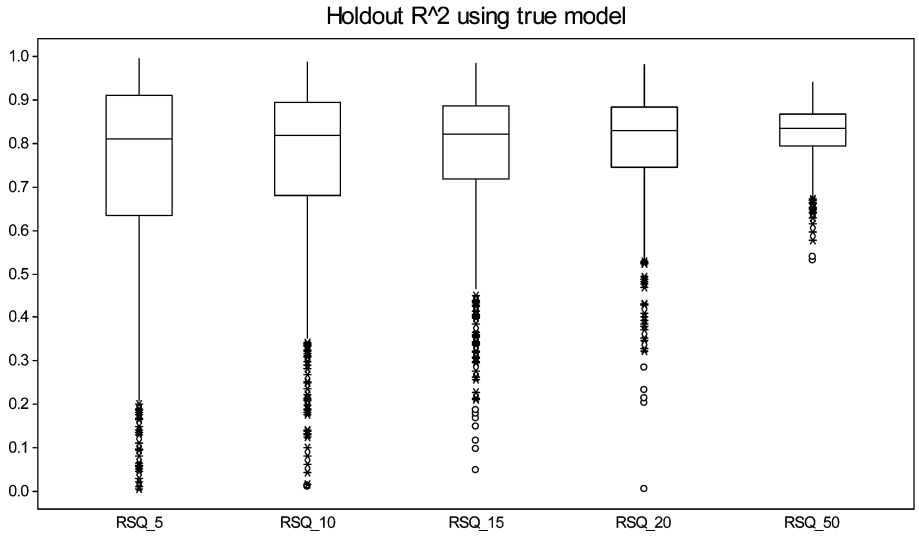

Figure 4.

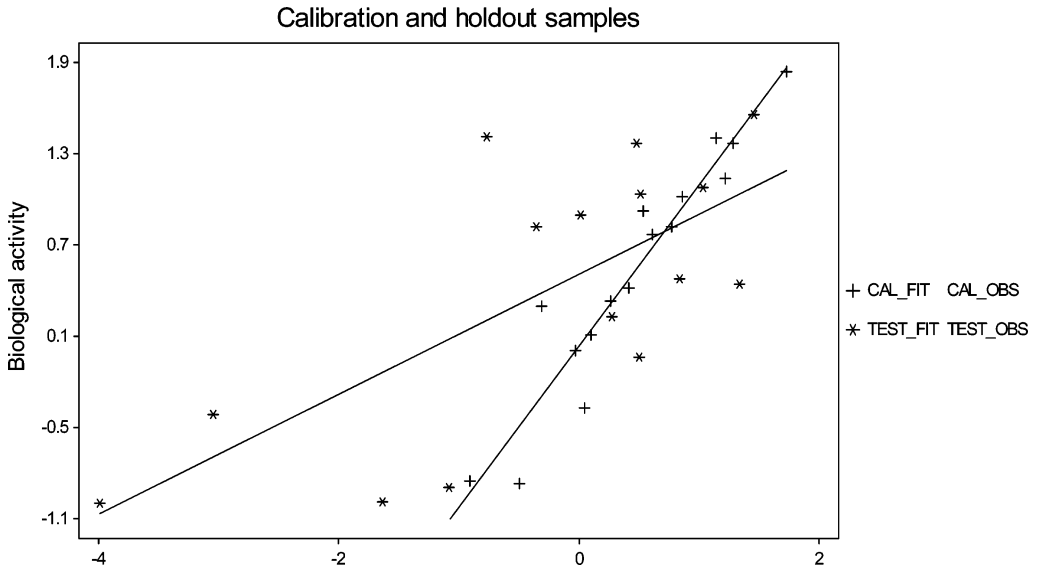

Figure 5.

from a model fitted to the remaining $n - 1$ cases. Another common example is 10-fold cross-validation in which $m = n/10$. Here the sample is randomly split into 10 groups. Ten model fits are performed, omitting each of the 10 groups in turn and predicting it from the cases in the remaining 9. At the cost of more computation but leading to intuitively better results, one can repeat a 10-fold cross-validation several times making different random splits into the 10 groups.

Leave-One-Out. There is a widespread perception in the modeling community that cross-validation (and in particular LOO) is a poor method of verifying the fit of a model. We can test this empirically using the TSCA boiling point data. We made a two-part experiment. In one, we repeatedly took a random calibration sample of 100 compounds and fitted the linear regression to predict boiling point, assessing the quality of the fit in the following two ways:

- Using the R^2 of an independent holdout sample of size 50
- Using the 'gold standard' of applying the fitted regression to all 937 remaining compounds in the set and calculating the true R^2.

In the second part, we merged the two data sets into a single data set of size 150 which we used for the calibration and then tested the fit using LOO cross-validation. The measures of fit for this step are as follows:

- The q^2 of the 150 cases
- The gold standard of applying this fitted regression to the remaining 937 compounds in the set and calculating the true R^2.

Each random sample therefore gives rise to two fitted models, each of which is assessed with two measures.

Some details need attention. As we are randomly sampling quite a small set of calibration compounds, there are probably going to be large extrapolations, something that linear regression does not do well. To handle these sensibly, when a prediction was to be made, the leverage of the predicted point was calculated. If the leverage was less than 2, then the regression prediction was used. If the leverage was bigger than 2, then the average of all calibration data was used as the prediction.

Using plain least squares multiple regression with this simple safeguard against wild extrapolation is chosen deliber-

ately, not because it is a cutting-edge technology but because it is so routine that a similar experiment could be performed by anyone with basic statistical computation capability.

The results of 300 such random samplings give the following distribution of the squared multiple correlations:

	calibrate using			
	150 cases and cross-validate		100 cases and validate with an independent 50	
	true R^2	q^2	true R^2	holdout R^2
mean	0.7213	0.7140	0.6103	0.5968
sd	0.0395	0.0699	0.0713	0.1472

The main features of this table are as follows:

- That the true R^2 of the models fitted using 150 compounds are higher than those fitted using 100 compounds, averaging 0.72 versus 0.61. They are also much less variable from one random calibration set to another, having a standard deviation hardly more than half the size. Of course that you can fit better and more stable models if you use more data is neither surprising nor new, so this observation merely confirms what one would expect.
- That both the q^2 values and the holdout R^2 values average close to the true quantities that they are estimating. Thus neither is biased.
- That the q^2 values vary substantially less than do the holdout R^2 values going from one random pick of calibration compounds to another. This is also a sample size issue; LOO's holdout sample of size 150 is treble that of the 50 independent holdout cases, so its much lower random variability is again not surprising.

Whether from the viewpoint of getting a better model or from that of more reliably assessing the quality of your model, the conclusion is that if you have 150 compounds available, you will do better using all of them to fit the model and then cross-validating to assess fit than if you were to use 100 of them to fit and hold out 50 to validate. By maximizing both the sample size used for fitting and that used for validation, LOO is better by both criteria than splitting the data into separate calibration and validation sets.

To explore this conclusion in a different setting, we chose a small (8 predictor) and less predictive set of features whose multiple regression in the full data set gives $R^2 = 0.57$ and repeated the experiment. This gave

	calibrate using			
	150 cases and cross-validate		100 cases and validate with an independent 50	
	true R^2	q^2	true R^2	holdout R^2
mean	0.5073	0.5016	0.4838	0.4614
sd	0.0315	0.0772	0.0454	0.1689

This set of runs duplicated all the features seen in the earlier set, confirming, in this more marginally descriptive multiple regression, that using all compounds for fitting and using cross-validation gives better regressions (higher true R^2) with equally reliable self-assessment (q^2 centers on the true R^2), and the assessment is more stable than that of the holdout (standard deviation of 0.077 less than half of 0.169).

k-Nearest Neighbors. The broad conclusions are not confined to least squares regression. Consider for example *k*-nearest neighbor prediction in which the boiling point of a compound is predicted to be the average of the boiling points of the *k* compounds closest to it in predictor space. There is one explicit parameter to fix in *k*-NN—the size *k* of the pool of neighbors to use in the prediction. The best *k* is commonly established using a cross-validation, but good results can be obtained much more simply by noting that the random error on predicting an unknown using the *k*-NN approach will be the random error in the mean of *k* observations plus the random deviation between the boiling point of the compound itself and the overall mean boiling point for a compound of that composition. Standard results on the variance of a difference then show that the prediction error has standard deviation $\sigma\sqrt{1+1/k}$. Unless *k* is 1 or 2, this standard deviation will be dominated by the leading '1+' term, and larger *k* will give little improvement in prediction precision, while greatly adding to the potential for bias from an overly large neighborhood. We therefore used a fixed value of $k = 4$ nearest neighbors in simulations of random calibration subsets of size 150 compounds used to predict the boiling point. The q^2 and true R^2 of 30 random sets gave

	true R^2	q^2
mean	0.7383	0.7350
sd	0.0196	0.0532

As with the least-squares regression modeling, there is no hint of systematic bias between the q^2 and the true R^2 values and no reason to doubt the veracity of q^2 as an indicator of the R^2 of the fitted model.

Use Multiple Regression or k-NN? This is not a methods comparison paper, but these results invite some comment on whether the multiple regression or the *k*-nearest neighbor is the better modeling method. Each has advantages and drawbacks. Both are easy to implement. The regression model wins on portability in that the coefficients of Table 1 allow it to be applied to any future data set, whereas applying *k*-NN would require access to the calibration data set. This simple *k*-NN matches or exceeds the prediction accuracy obtained using plain multiple regression, making it competitive as a prediction tool for this data set, particularly in view of its not needing much checking of model assumptions. Regression has the advantage of an explicit function relating the dependent variable to each predictor. In some settings—for example drug discovery—this information rather than the ability to predict future compounds may be the main objective, and this leads to favoring methods such as regression or recursive partitioning that assign value to predictors explicitly. In other settings where the objective is prediction this benefit may have no value and the simplicity of *k*-NN may prevail.

Feature Selection. Feature selection is a step carried out in many analyses of reducing an initial too-large set of predictors down to some smaller number that are felt to include all the predictors that matter. There are two reasons for doing a feature selection: if the fitted model is to be interpreted (as for example in drug design problems), then it is clearly advisable to avoid interpreting nonexistent effects, and this suggests trying to find and eliminate predictors that are not paying their way. The other reason is the hope that better predictions can be made from a model trimmed down to just the relevant terms. While both these reasons are valid, their end result is less so. Feature selection inevitably makes some errors of each type—selecting features that are not

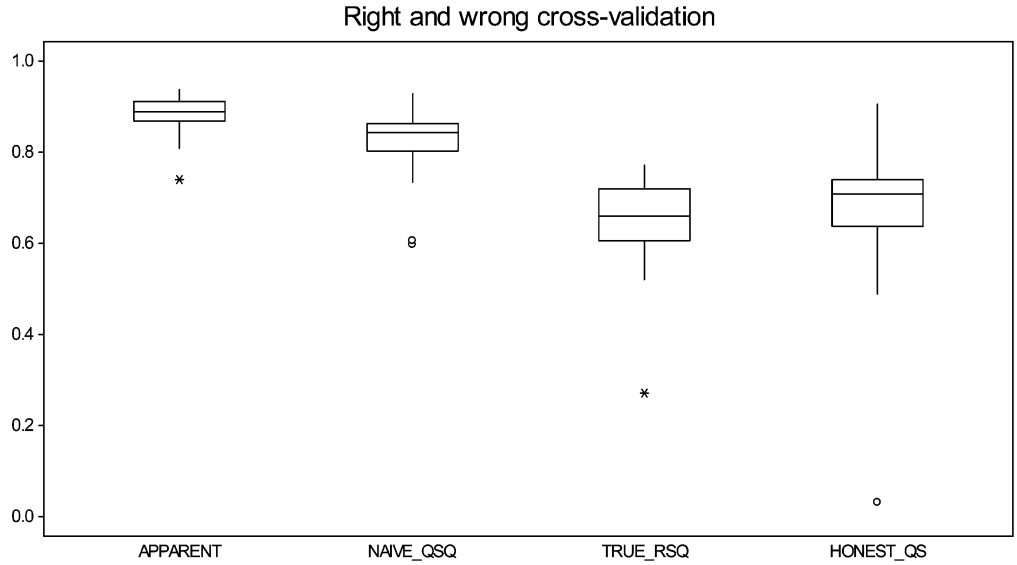

Figure 6.

actually relevant and omitting features that are—and overstates the value of the features that end up selected. On the prediction side, methods such as subset regression have fared poorly[7] in comparison with methods such as PLS and ridge regression that keep all predictors.

However putting this aside and focusing on the validation of fits, feature selection is a potent source of possible validation mistakes—as illustrated by the widely referenced Rencher and Pun[8]—but one that is often mentioned only in passing or overlooked altogether in publications.

The TSCA data set can be used to illustrate feature selection. We repeatedly took random samples of 50 calibration cases and looked for the best subset of predictors to use in a multiple regression. This sample size of 50—smaller than we used for the earlier experiments—was chosen because it is in the small-sample setting that one typically feels the greatest need to reduce the size of the feature set.

There are too many predictors (30) for 'all subsets' regression to be really workable in a repeated resampling setting, so we used a search with 10 000 random starts followed by refinement that seeks the best subset of each size from 3 through 14. While this search is not as reliable as an all-subsets search, it is much less likely than methods such as stepwise regression, forward selection or backward elimination to get trapped in a poor local optimum. We used as a final model the largest regression all of whose predictors had absolute *t* values in excess of 2, imitating common practice in regression subsetting.

The first five of these regressions all retained 4, 5 or 6 predictors. Their resubstitution R^2 and the true R^2 given by applying them to the remainder of the data set were as follows. Also shown is a 'naïve q^2' obtained by taking the features selected by the subset search and then refitting that regression omitting each case in turn and predicting it from the rest.

resubstitution R^2	true R^2	naïve q^2
0.86	0.73	0.82
0.89	0.74	0.83
0.87	0.72	0.84
0.88	0.72	0.83
0.94	0.69	0.93

Clearly the resubstitution R^2 has a large upward bias. Perhaps most striking is the last sample, where the resubstitution 0.94 exceeds the 0.888 mathematical upper bound attainable in this data set. It is also striking that, though the naïve q^2 is a bit smaller than the resubstitution R^2 in all five samples, it is far above the true R^2.

This might seem to indicate a failure of the LOO cross-validation. In fact it is a result of leaving the validation until too late in the process. Once the optimal variables for a calibration sample have been selected, most of the optimistic performance bias has been introduced, and all the cross-validation can do at that stage is remove the modest bias caused by tuning the coefficients to the calibration data.

A correct LOO cross-validation can be done by moving the delete-and-predict step *inside* the subset search loop. In other words, we take the sample of size *n*, remove one case, search for the best subset regression on the remaining *n*-1 cases and apply this subset regression to predict the hold-out case. Repeat for each of the cases in turn, getting a true hold-out prediction for each of the cases. Use these holdouts to measure the fit.

The result of doing this is summarized in Figure 6—a box and whisker plot obtained from repeating this sampling and fitting process 20 times. The four boxes are for the resubstitution R^2, the naïve q^2 obtained by taking the feature subset as given and just doing the leave-out for the final stage of fitting the coefficients, the true R^2, and the honest q^2 given by omitting the hold-out compound from the feature selection as well as the regression fitting phase.

The picture is quite clear. The resubstitution estimate of precision is hopelessly optimistic, and doing LOO without redoing the feature search hardly improves it. LOO cross-validation provides an honest picture of the predictive power of the subset model only when the hold-out step includes a fresh feature selection search.

As an aside on the errors of thinning out the features, the multiple regression shown in Table 1 using all the data found 10 highly significant predictors, indicating that regardless of which of the other predictors might be used, these 10 are clearly needed. Most of the subset fits ended up using 7 or fewer features, so that many important features were left out.

The effect of leaving them out is that features that were selected and that correlate with the omitted features in this data set are made to do double duty. Their coefficients reflect not only their own impact on BP but also a component coming from their attempt to 'explain' any correlated features that were erroneously left out of the model. The regression with the highest true R^2, 0.77, found in these 20 searches was

$$BP = 287 + 41\overline{IC} - 514M_2 - 102^4\chi - 297^3\chi_C + 78^6\chi_{PC} + 794P_2$$

This regression uses only two of the 10 'must have' predictors shown in Table 1, omitting the other eight and instead including four more marginal ones. Its coefficients are also substantially different than those seen in the full regression.

Wrapup on LOO. Statistical theory (summarized in ref 9) predicts that LOO will generally perform well, if somewhat conservatively in that it is predicted to somewhat understate the actual predictive performance of the fitted model. The experiments reported here are consistent with the theory. It is essential though that the leave-out encompass the entire model fitting operation and not just the last stage of fine-tuning coefficients.

LOO does however have two blind spots. If the compound collection is made up of a few core chemical compositions, each of which is represented by several compounds of nearly identical composition *x*, then the operation of removing any single compound will not be sufficient to get its influence out of the data set, because of the fraternal twin(s) still in the calibration. Under these circumstances, LOO will overstate the quality of the fit. This situation is quite common in drug discovery data sets, where each promising lead compound will be accompanied by a host of others with just minor differences in molecular structure and with generally very similar biological activity.

The opposite situation arises if the calibration data set consists of extreme compounds, each of which is an extrapolation from the set remaining if it is excluded. Under these conditions LOO will suggest a poor fit, whereas the fit may be good. A calibration data set comprising extreme compounds is in fact exactly what experimental design criteria such as D optimality or space filling attempt to find. As a straight line is better determined by points the further apart they are, using compounds at the edges of descriptor space makes for better future predictions in that future compounds are likely to be interpolates of the calibration set, and D-optimality seeks compounds as far out to the edges of descriptor space as it can find. However when any of the D-optimal points is omitted in LOO it is likely to be an extrapolation from the remaining sample, and the LOO criterion will suggest unrealistically poor model fit for the real future unknowns, which will mainly be interpolates.

Leave-Several-Out Cross-Validation. The main concern with LOO is its potential breakdown if it is applied to a collection containing two or more close analogues of each of some basic molecule types. It may also be unattractive in large data sets because of its computational cost, which is directly proportional to the sample size. A potential remedy for both these difficulties is a cross-validation in which not one but several compounds are left out and predicted using the remainder of the data.

A popular method of this type is 10-fold cross-validation, in which the available sample is split into 10 groups of equal size. A model is fitted to the entire sample, followed by 10 fresh fits. In each of these, one group is omitted, modeling performed on the other 9, and the holdout group is predicted. If there is some concern about the randomness involved in the ten-way split (for example because the sample size is modest), then the random splitting and fitting can be repeated.

Applying this method to the TSCA boiling point data set, we generated random data sets of size 150 for use in fitting and testing the 'high R^2' and the 'low R^2' models. In each run, we split the 150 compounds into 10 groups of size 15 and performed the 10-fold cross-validation. The results were as follows:

	high R^2 model	lower R^2 model
mean R^2	0.6930	0.4967
sd	0.0784	0.0726

In both settings, the 10-fold cross-validation gives very similar results, in terms of both the mean and the standard deviation of the R^2, to those given by LOO cross-validation. This further validates the LOO approach and suggests that, in the context of this data set, we do not have the problem of sets of near-identical chemicals that could derail LOO.

CONCLUSION

Overfitting of models is widely recognized as a concern. It is less recognized however that overfitting is not an absolute but involves a comparison. A model overfits if it is more complex than another model that fits equally well. This means that recognizing overfitting involves not only the comparison of the simpler and the more complex model but also the issue of how you measure the fit of a model.

More flexible models with more tunable constants can be expected to give better resubstitution performance than less flexible ones, whether or not they give better predictions in the broader population in which predictions are to be made. Resubstitution measures of performance therefore have little value in model comparison since they reflect model complexity rather than model predictive power.

Some sort of out-of-sample prediction is essential in getting an honest picture of a model's predictive ability. A resubstitution error mean squared error much smaller than the out-of-sample error mean squared error suggests high model flexibility and is one warning sign of overfitting. A firm diagnosis or refutation of overfitting though depends on seeing whether the complex model fits better than do simpler candidate models.

Where possible there is much to be said for having an independent holdout validation sample removed at the start and kept under lock and key to prevent impermissible peeking. However this independent validation set needs to be large to be trustworthy, and fitting a model to a minority of the data makes no intuitive sense. Therefore the full holdback approach is attractive only when there is a large sample available. This setting is a luxury that is not often available in QSAR modeling where usually a sample is precious and needs to be exploited to the full. Methods such as LOO and multifold cross-validation use all available

information for both calibration and validation. They should be basic tools in model fitting and assessment.

Figure 4 was generated using an independent holdout sample but is relevant to cross-validation also, using the sample size as the number of holdout samples. The calculations in this paper suggested that validating using at least 50 samples gave generally reliable results and that 20 was perhaps not too bad, but that using fewer samples (such as 5 or 10) was unreliable. This suggests the following guidelines:

- If no more than 50 samples are available, use all of them for the modeling and use LOO for validation, taking all necessary precautions to ensure in the reanalyses to predict each of the samples that the left-out sample does not in any way influence the fit.
- As an alternative to the LOO, it may be helpful to do a number of multifold cross-validations—for example making several distinct random ten-way splits and averaging the results of the 10-fold cross-validations.
- If the sample size is appreciable larger than this—up to several hundred—then again use all samples for the calibration and validate using a multifold (for example 10-fold) cross-validation.
- If huge samples are available, then the traditional 'one-half for calibration and one independent half for validation' paradigm becomes workable. But even here, it is generally helpful to do more than one fit—for example having fitted to the calibration half and tested on the validation half, to swap the roles of the two-half-samples and compare the results. Even here though it is better, absent some contraindication such as severe computational complexity, to use a multifold cross-validation.
- If the collection of compounds consists of, or includes, families of close analogues of some smaller number of 'lead' compounds, then a sample reuse cross-validation will need to omit families and not individual compounds.
- If the available sample is much below 50, then realize that it is impossible to fit or validate incontrovertibly generalizable models. This does not mean that there is no value in doing the fitting and cross-validation; just that appreciable uncertainty will inevitably remain after this is done. See Figure 4 for some evidence of the impact of smaller validation sample sizes on the reliability of the validation.

Returning to the overfitting question, justifying a more complex modeling approach entails showing that its additional complexity is necessary and that the same quality of fit cannot be obtained by simpler models. What are the benchmark simpler models? This is not entirely obvious. When predicting a continuous or a binary dependent variable, a straightforward least squares multiple regression would be an obvious choice where the problem is overdetermined, and perhaps partial least squares (PLS[7]) makes a natural benchmark where the problem is underdetermined. These benchmarks are helpful because they are standardized and supported by easily accessible software and therefore are easy for anyone to replicate as the basic principles of science require. As implied by the naphthalene data, this is a modest hurdle to leap, as common transformations such as going to a log scale may lead to equally simple but better-fitting models. Conversely however, complex models that cannot beat these simple ones are hard to justify.

There is also a need for attention to the domain of applicability of a model. No model can be expected to extrapolate successfully, yet it is not always obvious what predictions are extrapolations and what are interpolations. Simple single-variable range checks cannot do this reliably. Measures analogous to the leverage are important in multiple predictor problems.

We have not touched on other concerns such as outliers and missing data; these would require a full-length paper in their own right. Another necessary omission is an overview of useful modeling technologies. Hastie et al.[10] discuss the statistical aspects of a number of modern approaches.

ACKNOWLEDGMENT

The research reported in this paper was supported in part by Grant F49620-01-0098 from the United States Air Force. The author is grateful to Stanley Young and to the editors for a number of suggestions for improvement.

REFERENCES AND NOTES

(1) Franklin, N. L.; Pinchbeck, P. H. P.; Popper, F. A statistical approach to catalyst development, part 1: the effect of process variables in the vapour phase oxidation of naphthalene. *Trans. Institute Chem. Eng. 1956*, 34.
(2) Gnanadesikan, R. *Methods for statistical data analysis of multivariate observations*; Wiley: New York, 1997.
(3) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP Based on Quantitative Structure−Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177−1207.
(4) Cook, R. D.; Weisberg, S. *Applied Regression Including Computing and Graphics;* Wiley: New York, 1999.
(5) Hawkins, D. M.; Basak, S. C.; Shi, X. QSAR with Few Compounds and Many Features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663−670.
(6) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Crcognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.
(7) Frank, I. E.; Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109−135.
(8) Rencher, A. C.; Pun, C. P. Inflation of R2 in best subset regression. *Technometrics* **1993**, *22*, 49−53.
(9) Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; SIAM: Philadelphia, 1982.
(10) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, 2001.

CI0342472