An Improved Approximation to the Estimation of the Critical *F* Values in Best Subset Regression

David W. Salt,[†,‡] Subhash Ajmani,[‡] Ray Crichton,[‡] and David J. Livingstone[‡,§]

Department of Mathematics, Buckingham Building, Lion Terrace, University of Portsmouth, Portsmouth, United Kingdom, Centre for Molecular Design, University of Portsmouth, United Kingdom, and ChemQuest, Sandown, United Kingdom

Received March 29, 2006

Variable selection methods are routinely applied in regression modeling to identify a small number of descriptors which "best" explain the variation in the response variable. Most statistical packages that perform regression have some form of stepping algorithm that can be used in this identification process. Unfortunately, when a subset of p variables measured on a sample of n objects are selected from a set of k ($>p$) to maximize the squared sample multiple regression coefficient, the significance of the resulting regression is upwardly biased. The extent of this bias is investigated by using Monte Carlo simulation and is presented as an inflation factor which when multiplied by the usual tabulated F ratio gives an estimate of the true 5% critical value. The results show that selection bias can be very high even for moderate-size data sets. Selecting three variables from 50 generated at random with 20 observations will almost certainly provide a significant result if the usual tabulated F values are used. An interpolation formula is provided for the calculation of the inflation factor for different combinations of (n, p, k). Four real data sets are examined to illustrate the effect of correlated descriptor variables on the degree of inflation.

1. INTRODUCTION

Without doubt, the most popular modeling method used is multiple linear regression where the response variable y is linked to a block of p independent variables by a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon \tag{1}$$

where ϵ is the usual random error term. At one time, variables were selected for inclusion in a regression model because there was strong justification for including them. This justification would have been based on some a priori idea or hypothesis that the chosen variables were the critical factors in establishing the variability in the response variable. Although this approach may still exist, more often today, researchers may have no prior knowledge as to the relative importance of the various independent variables. This is particularly true in the drug design industry where nowadays thousands of physicochemical properties are readily available for the construction of quantitative structure−activity relationship and quantitative structure−property relationship models[1−3]

As a consequence, many regressions are fabricated by selecting variables whose observed values make a "significant" contribution to explaining the variation in the observed values of the response variable. Various methods are available for this selection, for example, forward selection, backward elimination, stepwise selection, and best subset.[4] All these methods select the "best" p variables from a data set of n observations measured on k independent variables. "Best" usually means that the variables selected are those which maximize the squared sample multiple correlation coefficient R^2. As a consequence, the typical F test used to determine the level of significance of R^2 is biased except when $p = k$, that is, when all the variables available are included in the regression model. The situation where $p = k$ is largely of academic interest in drug design applications were k may be in the range of 50−200. Topliss et al.[5,6] demonstrated that the more independent variables (X) that are available for selection in a multiple linear regression model, the more likely a model will be found by chance. These authors recommended that in order to reduce the risk of chance correlations there should be a certain ratio of data points to the number of independent variables available. Unfortunately, this ratio was often misinterpreted as the number of data points to the number of independent variables in the final model, a practice that did very little if anything to reduce chance effects. The bias in the F test is illustrated in Figure 1 where the distribution of the F ratio calculated for best subset regressions on randomly distributed data simulated for $n = 20$; $p = 3$; and $k = 3−5$, 10, 25, and 50 is presented. For each value of k, 50 000 sets of random data were generated containing 20 observations on $k + 1$ variables (one y and k x's), and for each of these sets, the F ratio of the best subset (maximum R^2) was noted. The resulting 50 000 F ratios form an estimate of what is termed a sampling distribution. Sampling distributions determine the critical values of a test statistic for a given significance. For example, if you have a response variable y and three independent variables (x's) and you wish to assess the statistical significance of the regression of the y on these three x's, the critical value of F, for a significance level of

* Corresponding author e-mail: david.salt@port.ac.uk.
[†] Department of Mathematics, University of Portsmouth.
[‡] Centre for Molecular Design, University of Portsmouth.
[§] ChemQuest.

10.1021/ci060113n CCC: $37.00
Published on Web 12/02/2006

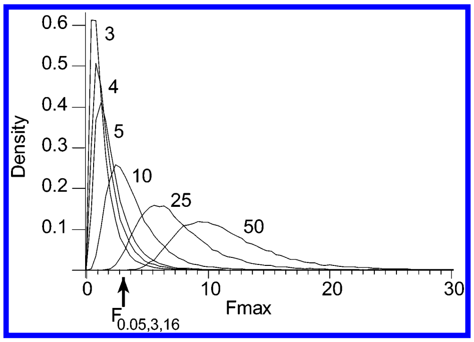

Figure 1. Plot of the distributions of the *F* ratios calculated for the best subset regressions on randomly distributed data simulated for $n = 20$; $p = 3$; and $k = 3-5$, 10, 25, and 50.

5% on the basis of $p = 3$ and $n - p - 1 = 20 - 3 - 1 = 16$ degrees of freedom, is 3.24. This value can be found in the usual *F* tables. It is marked by the arrow in Figure 1, and the area to the right of this point under the sampling distribution with $k = 3$ is 0.05. The interpretation of this is, if the data is random, then 95% of models found for a given $n = 20$ and $p = 3$ will have observed *F* ratios less than 3.24 and 5% will have ratios greater than or equal to this critical value. Now, if in a real situation a regression produces an observed *F* ratio $\geq$ 3.24, then the regression would be deemed significant because the probability of this happening when the data is random is only 0.05. However, for the other values of *k*, we see that the corresponding upper tail areas for the distributions calculated for $k = 4$, 5, 10, 25, and 50 increase with *k* and are respectively 0.10, 0.15, 0.48, 0.96, and $\approx$1. This means that, when constructing a regression model consisting of three variables selected from a set of four variables using best subset selection, if the critical value of 3.24 is still used, the actual significance level is 10%, that is, double the level intended. When $k = 5$ and the value of 3.24 is used, the actual significance level is 15%; when $k = 10$, this level rises to 48%, and with a pool of 50 variables from which to select three, the significance level is approximately unity; that is, the model will almost certainly be significant even though in the population the response variable is independent of the *x* variables. This is quite a sobering result.

This "selection" bias is little known to many users of multiple linear regression but has of course been widely discussed by the statistical community.[7-9] Copas,[7] for example, comments that, although nearly all statistical packages offer some form of variable selection, the usual least-squares properties are invalid when the variable subset is selected in a supervised manner. Unfortunately, the manuals and help windows that come with many of these programs talk about significance levels, *F* and *t* and so forth, with no indication of the problem. As a consequence, many users use these programs under a false sense of security. Miller,[10] while discussing the hypothesis testing of regression coefficients in stepwise variable selection methods, draws attention to the fact that the maximum *F* to enter does not follow an *F* distribution. This has also been discussed, among others, by Diehr and Hoflin,[11] Draper and Smith,[4] and Pope and Webster.[12]

Methods of overcoming inflated significance values arising from using ordinary *F* tables has received much attention in the literature, albeit in more mainline statistical journals.[7-9] However, there are papers to be found in other areas which also consider the problem. [13-16]Although these publications differ in fine detail, the main theme is one of using Monte Carlo simulation to generate the empirical distribution of some fit statistic like R^2 (the coefficient of determination) or the *F* ratio. These simulations are done for a variety of combinations of *k* (the number of variables available to select from), *p* (the number of variables in the final model), *n* (the sample size), and the method of model construction. In the earlier papers, the number of simulations was relatively small, on the order of 100,[11] but with advances in computing, figures up to 2000 are to be found.[8] After each simulation, the maximum R^2 ($R_{\max}{}^2$) or maximum *F* ($F_{\max}$) is saved and finally ordered smallest to largest. The 100α% significance level critical value of the distribution is the $100(1 - \alpha)$th percentile of these ordered values. The relationship between *F* and R^2 is given by

$$F = \frac{\nu_2 R^2}{\nu_1(1 - R^2)} \tag{2}$$

where $\nu_1 = p$ and $\nu_2 = n - p - 1$; therefore, given the value for one of them, the other can easily be found. If the random variables are independent normal with a zero mean and a standard deviation of one, then when $p = k$, the sample R^2 has a β distribution and the *F* ratio given in eq 2 has the usual *F* distribution with degrees of freedom, $\nu_1 = p$ and $\nu_2 = n - p - 1$. However, the question that needs to be answered is, what is the distribution of $F_{\max}$ (or $R_{\max}{}^2$) when $p < k$?

A number of papers have been published in order to shed light on the answer to this question. Rencher and Pun,[8] for example, in their work on stepwise regression, provide an argument based on extreme value distributions which leads to the following formula for the critical $R_{\max}{}^2$ values with a significance level (the probability of wrongly rejecting the null hypothesis of no model):

$$R_{\max}{}^2 = B^{-1}\left[1 + \frac{\log_e(1 - \alpha)}{(\log_e N)^{1.8N^{0.04}}}\right] \tag{3}$$

where B^{-1} is the inverse of the cumulative β function with parameters $p/2$ and $(n - p - 1)/2$, $N = k!/p!(k - p)!$, that is, the number of different models of size *p* that can be constructed from *k* and where the term $(\log_e N)^{1.8N^{0.04}}$ was found empirically to provide a better fit to their simulated $R_{\max}{}^2$ values.

Diehr and Hoflin[11] take a different approach to modeling their $R_{\max}{}^2$ values obtained from best subset regression. They used the power function

$$R_{\max}{}^2 = w(1 - \nu^p) \tag{4}$$

where *w* and *v* are constants whose values were found for given values *n*, *k*, and significance level α by forcing the function to pass through the two known values of $R_{\max}{}^2$ which exist for $p = 1$ and $p = k$. When $p = k$, the value of $R_{\max}{}^2$ can be obtained from eq 2 with the *F* being provided by the usual tabulated values. When $p = 1$, the values of $R_{\max}{}^2$ can be found once again from eq 2, but this time, the tabulated *F* value used is the one which corresponds to a significance level set to $\alpha^* = 1 - (1 - \alpha)^{1/k}$ instead of α. The critical values of the $F_{\max}$ distribution can be found similarly using the usual F tables with α^*. Diehr and Hoflin's[11] choice of power function (eq 4), like Rencher and Pun's[8] (eq 3), was based on fitting various types of functions and selecting the one which performed the best.

In this paper, we have extended the range of combinations of the input parameters (n, p, k) to those provided in previous work in this area. Also, as a direct result of the larger number of simulations performed at each combination (50 000), there is an improved precision of the critical $F_{\max}$ values. Rather than constructing a set of power functions to fit our simulated results as is the case with Diehr and Hoflin[11] or trying to mix empiricism with mathematical theory as with Rencher and Pun,[8] we have chosen to use nonlinear regression to construct an appropriate function to enable critical values to be obtained for given combinations of (n, p, k). Finally, we provide a Web address which allows the visitor to run our ultrafast simulator to obtain $F_{\max}$ values for their given combinations of (n, p, k) (www.cmd.port.ac.uk/cmd/fmax-main.shtml).

2. RESEARCH DESIGN

The Monte Carlo simulations were performed by generating n observations from a $k + 1$ dimensional multivariate normal population with a zero mean and identity covariance matrix. All subsets regression was performed for models of size p and the resulting maximum F ratio stored. This procedure was then repeated m times. Ordering the m F ratios allows the upper 100α% critical values of what we term the $F_{\max}(x;\, \upsilon_1,\upsilon_2,\, k)$ distribution. The values of p, k, and n used were $p = 1, 2, ..., 10$; $k = p, p + 1, ..., 15, 50, 100$, and 150; and $n = 20$, 50, 75, and 100.

3. REAL DATA SETS

Randomly generated data sets are useful up to a point, but the results obtained from them would be of little value unless they were corroborated with those using real data. To this end, four data sets have been used. The first is the well-known Selwood[17] set containing 53 descriptor values calculated for 29 compounds. The compounds are analogues of the mammalian electron transport inhibitor antimycin A_1, which had been tested in an in vitro assay against filarial worms in the search for treatments for diseases such as river blindness and elephantiasis. The descriptors were calculated using an in-house (Wellcome) computational chemistry program. The second data set we refer to as the Kappa[18] charge transfer set, which consists of measured values of charge-transfer complex association constants for a series of 35 monosubstituted benzenes. Physicochemical descriptors for these compounds were computed with another in-house (SmithKline Beecham) computational chemistry program. The original Kappa data set contains 31 descriptors, but two have been removed because of collinearity problems. The third data set is from the laboratories of the Centre for Molecular Design, University of Portsmouth,[19] and comprises an in vitro measure of insect toxicity of 19 synthetic pyrethroid analogues together with values for nine physicochemical descriptors. The fourth data set will be referred to as the Damborsky[20] anilines and phenol data set and contains the response variable log IGC50, which is a measure of toxicity, and nine physicochemical descriptors.

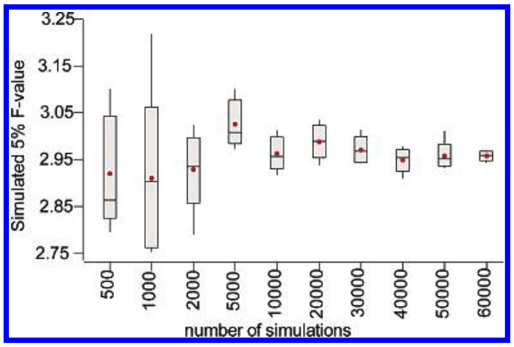

Figure 2. Plot of the 5% $F_{\max}$ for the combination (n, p, k) = (20, 5, 5) for a range of simulation lengths.

4. SIMULATION RESULTS

The Monte Carlo simulations are very computer-intensive particularly for some of the values of p and k near the upper end of their ranges. The decision on how many simulations to perform for each combination of (n, p, k) must be based on the tradeoff between the time it takes and the precision of the results. Figure 2 shows a plot of the 5% $F_{\max}$ for the combination (20, 5, 5) for a range of simulation lengths. As $p = k = 5$, these simulated $F_{\max}$ values are in fact estimates of the tabulated critical value 2.96 with 5 and 14 degrees of freedom. Five runs were performed at each run length, and as can be seen, the variability generally becomes less as the number of simulation run lengths increases. What also can be seen is that results based on hundreds and even the low thousands of simulations are quite unstable, implying that some of the earlier-published critical values are likely to be in error. On the basis of these experiments, we decided that 50 000 would provide a reasonable balance between speed and accuracy.

5. VARIABLE SELECTION INFLATION INDEX

The difficulty for analysts wishing to make use of these simulated results is that they are not exhaustive, and so it is unlikely that their particular set of n, p, and k has formed part of the simulation research design. This problem is handled as discussed above either by use of interpolation or by fitting a function to the simulation results. There are no hard and fast rules for the choice of the form of function to be fitted, and in practice, it is usually nothing more than a piece of empiricism. Wilkinson and Dallal,[9] for example, report the use of a model containing "twenty linear and nonlinear terms plus their interactions". Diehr and Hoflin,[11] on the other hand, adjusted for the upward bias in their theoretical calculations by selecting a function which "was found to fit well". We also have to follow this empirical road to providing a means for generating critical values based on our simulations. However, whereas previous authors have almost without exception concentrated on generating critical values of $R_{\max}{}^2$, our function (below) generates approximations to the 95th percentile of the $F_{\max}$ distribution. Early attempts to identify a relatively simple function proved to be unsuccessful, and as Wilkinson and Dallal[9] found, models with large numbers of terms were necessary to generate the required complexity. However, recognizing that the critical values of the $F_{\max}$ distribution are always greater than or equal to those of the corresponding F distribution, we can write

$$F_{\max,0.05} = F_{\nu_1\nu_2,0.05} \times I(\nu_1,\nu_2,k) \qquad (5)$$

where $F_{\nu_1\nu_2,0.05}$ is the usual tabulated 95th percentile of the F distribution for 5% significance with degrees of freedom $v_1(= p)$ and v_2 $(= n - p - 1)$, and where $I(\nu_1\nu_2,k)$ is an inflation index with the property that

$$I(\nu_1,\nu_2,k) = \begin{matrix} 1 \text{ for } p = k \\ >1 \text{ for } p < k \end{matrix}$$

The structure of this inflation index was then based on identifying the most parsimonious model that would provide the necessary precision and is given by

$$I(\nu_1,\nu_2,k) = N^d \qquad (6)$$

where

$$d = \Bigg[a_1 \ln(v_1 + v_2 + 1) + a_2 \ln(v_1) + a_3[\ln(v_1)]^2 + a_4 \ln(k) + a_5 \ln(v_2) + a_6 \ln(N) + \frac{a_7 \ln(v_2) + a_8 \ln(v_1)}{\ln(v_1 + v_2 + 1)} + a_9 \frac{\ln(v_1)}{\ln(k)} + a_{10}\Bigg]^2 \qquad (7)$$

$N = k!/p!(k - p)$, that is, the number of different models of size p that can be constructed from k variables and where the coefficients a_i, i=1, 2, ..., 10, are given in Table 1.

Observed values of the inflation index are given by dividing the simulated $F_{\max}$ values by the corresponding tabulated values. The fit of eqs 6 and 7 to the inflation index is illustrated in Figure 3 where in a plot of observed (simulated) against fitted (predicted) values we see very little deviation from a line of $y = x$.

We have tabulated $I(v_1,v_2,k)$ for k = 5, 10, 20, 50, 75, 100, and 150; p = 1−3, ..., 10; and n = 20, 50, and 100, and these are presented in Table 2. Inspection of the values in this table shows how the variable selection process upwardly biases the observed F ratios calculated for the best subset regressions. The extent of the bias clearly varies depending on the number of variables available for selection (k), the final model size (p), and the number of data points (n), and this was obviously going to be the case. Despite this expectation, it was quite a revelation to observe the extent of this bias even for moderately sized data sets. For example, a regression consisting of three variables selected from a set of 20 measured on 50 observations would have an observed F ratio in excess of 7.28, that is, a value 2.59 times the usual tabulated value $[F_{v_1,v_2,0.05} \times I(v_1,v_2, k)]$, to achieve statistical significance at the 5% level. For the same situation but with a selection set of variables of 75, this critical value would be 11.72, that is, over 4 times (4.17) the value most people unknowingly believe they should be using.

Scanning the entries in Table 2, it can be seen that the inflation index I varies enormously. To see more readily how the inflation index varies according to the values of (n, p, k), it has been plotted (Figure 4) for a range of n [= 20(1)-100] and p [= 1(1)10] for four values of k (= 10, 15, 20, and 50). In Figure 4a (k = 10), we observe that, with the exception of the region around the lowest values of p and n, I decreases monotonically with both n and p, and at $p = k$, I = 1. With k = 15 (Figure 4b), the characteristics of the plot change dramatically. As p increases, I rises to a maximum before again falling to unity. It can be seen that this behavior persists for k = 20, when the maximum now occurs at higher values of p. At k = 50, there is no maximum over the range of values plotted. The explanation for this behavior lies with N, the number of models that can be generated for given values of p and k. As p increases for $p < k/2$, N increases, and for $p < N/2$, N decreases, and it is this movement in N which drives the changing profile of the I surface plots.

Table 1. Values of the Coefficients in eq 9

a_1	a_2	a_3	a_4	a_5
0.657 578	0.182 325	−0.028 287	0.021 956	−0.643 342
a_6	a_7	a_8	a_9	a_{10}
−0.000 566	2.373 840	−0.130 264	0.155 937	−3.048 618

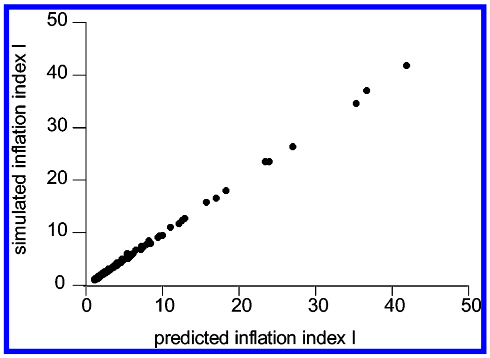

Figure 3. Plot of the observed and fitted inflation index calculated from eqs 6 and 7.

To assess the ability of the interpolation formula given in eqs 5−7 to provide values of the inflation index for values of (n, p, k) not used to construct the relationship, a random selection of 15 (n, p, k) combinations were made. The 15 simulations were performed, and the resulting values of the inflation index together with those calculated from eqs 5−7 are presented in Figure 5. As can be seen in this figure, there is very good agreement between the two sets of values, thus providing evidence that eqs 5−7 provide a reliable means of calculating the inflation index which, together with the usual tabulated F values, enables the 5% critical values of $F_{\max}$ to be determined.

The discussion so far has dealt entirely with random variables, that is, variables whose pairwise correlations are zero. However, in a real situation, the so-called independent variables are likely to be correlated. Work by McIntyre et al.[13] using stepwise regression indicates that critical values obtained using correlated X variables are slightly lower than the corresponding ones for independent variables. The results for the best subset should be similar. To investigate this, all four real data sets were subjected to the following treatment. The values of the response variable were randomized[15]-(scrambled), thus breaking any association between the independent variables and the response. Best subset regression was then performed on the resulting data set and the maximum value of the F ratio recorded. This scrambling and best subset regression was performed 2000 times, and the 95th percentile of the ordered $F_{\max}$ values were noted, and these (critical $F_{\max}$ real) together with the corresponding

Table 2. Simulated Inflation Index for the 95th Percentile of the F_{max} Distribution.

k	n	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
5	20	1.83	1.67	1.46	1.24						
	50	1.73	1.52	1.32	1.15						
	100	1.69	1.48	1.28	1.13						
10	20	2.28	2.39	2.37	2.31	2.21	2.08	1.90	1.66	1.37	
	50	2.10	2.03	1.89	1.74	1.60	1.47	1.35	1.23	1.12	
	100	2.05	1.94	1.78	1.63	1.50	1.38	1.28	1.18	1.09	
20	20	2.78	3.26	3.63	3.99	4.37	4.80	5.28	5.83	6.45	7.14
	50	2.50	2.62	2.59	2.52	2.44	2.35	2.26	2.17	2.08	1.99
	100	2.42	2.45	2.37	2.26	2.16	2.06	1.97	1.88	1.80	1.71
50	20	3.47	4.63	5.86	7.41	9.52	12.53	17.00	23.93	35.25	54.87
	50	3.05	3.46	3.66	3.80	3.91	4.01	4.10	4.19	4.26	4.33
	100	2.92	3.17	3.24	3.26	3.26	3.26	3.25	3.24	3.23	3.21
75	20	3.79	5.29	7.03	9.41	12.88	18.24	26.95	41.91	69.33	123.76
	50	3.29	3.84	4.17	4.44	4.68	4.92	5.16	5.41	5.66	5.91
	100	3.13	3.49	3.64	3.73	3.81	3.87	3.94	4.00	4.06	4.12
100	20	4.01	5.77	7.93	11.00	15.73	23.41	36.65	60.99	109.27	214.26
	50	3.46	4.10	4.53	4.90	5.26	5.62	6.00	6.40	6.82	7.26
	100	3.28	3.71	3.92	4.07	4.20	4.33	4.46	4.59	4.72	4.86
150	20	4.33	6.46	9.26	13.49	20.42	32.50	55.04	100.43	200.55	447.17
	50	3.69	4.47	5.04	5.56	6.10	6.67	7.28	7.96	8.70	9.51
	100	3.49	4.01	4.30	4.54	4.76	4.99	5.23	5.48	5.75	6.03

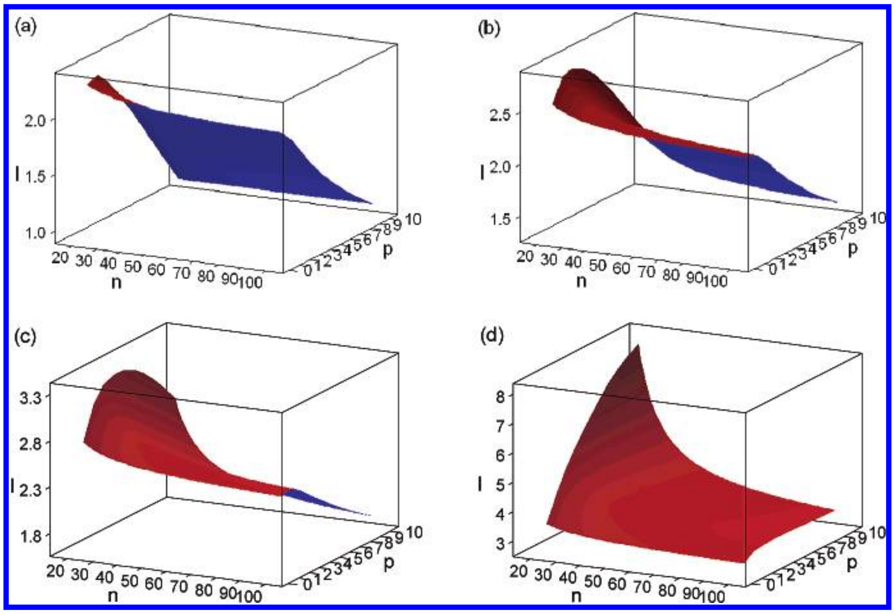

Figure 4. Inflation index $I(\upsilon 1, \upsilon 2, k)$ plotted against p and n for four values of k. (a) $k = 10$, (b) $k = 15$, (c) $k = 20$, and (d) $k = 50$.

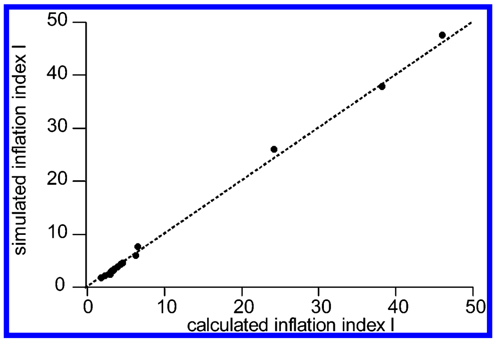

Figure 5. Plot of the simulated and predicted inflation indices calculated from eqs 6 and 7.

results using random data (critical F_{max} random) for the same size problem, that is, the same sample size n and number of available descriptors k, appear in Table 3 for three model sizes $p = 1$, 2, and 3. Also included are the observed F ratios from the actual regressions and, for comparison purposes, the usual tabulated critical F values.

Table 3. Comparison of Observed, Simulated Real Critical and Simulated Real F_{max} Values for the Three Real Data Sets

				critical	critical		usual
				F_{max}	F_{max}	observed	table
data set	n	k	p	random	real	F ratio	F value
Selwood	29	53	1	13.77	12.21	16.33	4.21
			2	13.42	10.88	24.64	3.37
			3	13.57	10.70	25.94	2.99
Kappa	35	28	1	11.52	11.03	24.69	4.14
			2	10.22	9.27	166.09	3.29
			3	9.44	8.52	182.22	2.91
Pyrethroid	19	34	1	14.32	12.68	5.09	4.45
			2	15.09	11.92	10.68	3.63
			3	16.75	13.35	9.91	3.18
Damborsky	15	9	1	10.94	11.58	12.40	4.67
			2	9.99	9.82	8.98	3.89
			3	9.61	9.32	8.26	3.59

An inspection of Table 3 reveals that, as suggested above, the simulated F_{max} values obtained by the y-scrambling in the real data sets are indeed less than the corresponding values for the random data. So, if an observed F ratio from a real regression analysis exceeds the corresponding one arrived at from the method developed in this paper using random data, then it indicates that the result is significant at the 5% level. The second-to-last column of Table 3 contains the values of the observed F ratio for the various model sizes, and as can be seen, they are all greater than the usual tabulated critical F values. However, further inspection of the table shows that, when the $F_{\max}$ values obtained by the y scrambling method are used, the pyrethroid results are in fact not significant and neither are the two or three variable models found for the Damborsky data set.

6. CONCLUSIONS

An examination of Table 2 clearly illustrates the problem of using standard F tables to assess the significance of a regression model found by subset selection. Although statisticians are aware of this inflation in the sample F, it is not generally known in the wider scientific community. Wilkinson[14] reports that out of 66 articles that he found in psychology which used variable selection only 19 were significant when compared to his Monte Carlo simulated critical values.

The research presented in this paper provides a method for assessing the true significance of best subset regression studies. Using the interpolation formula presented here for the inflation index together with standard tables of the F distribution, an analyst can obtain a realistic assessment of the true significance of their regression model. There are, however, two issues that need to be considered regarding the use of eqs 5−7 and Table 2. First, can the critical F values found here be used with other variable selection procedures? As the estimated $F_{\max}$ values provided in this paper have been obtained by an exhaustive search strategy, they can also be used with other variable selection routines. The reason for this is that variable selection methods which do not look at all possible combinations of variables may not have “seen” the optimal one before satisfying its stopping rule. Consequently, the critical maximum F ratios will be less than or equal to those obtained by the methods presented in this paper for the same combination of n, p, and k. Second, there is the issue of how many variables (p) should be included in a regression model. If p is not known before the selection procedure is applied, then p is a random variable and will need to be determined. This can be done by using a critical F to enter that is calculated using the formula given in eqs 5−7. For example, if the number of x variables available for inclusion in a regression is m, and q variables are already in the model, and another variable is being considered for inclusion in the model, then this critical F to enter can be obtained by substituting $p = 1$ and $k = m - \mathrm{q}$ in the eqs 5−7. The corresponding observed value of F, F_{obs}, is given by

$$F_{\mathrm{obs}} = \frac{{R_{q+1}}^2 - {R_q}^2}{(1 - {R_{q+1}}^2)/(n - q - 2)} \tag{8}$$

where ${R_{q+1}}^2$ and ${R_q}^2$ are the squared correlations of the models with and without the new variable. If this $F_{\mathrm{obs}} < F$ to enter, then the variable does not make a significant contribution to explaining the variation in the response and would not be entered into the model. We have experimented with this approach in our own laboratory and as expected have found that the number of variables entering a model is dramatically reduced. One of the consequences of this is that the “fitted” R^2 of the final model is also much reduced. But there is no point in having a model that fits your data if it cannot be relied upon and is the price to be paid for using a computer program to trawl through large amounts of data.

The situation can be improved to some extent by removing variables highly correlated with others prior to the supervised variable selection. Other ways of dealing with the problem of selection bias with large data sets exist. For example, Baumann has investigated the effect of using criteria other than maximizing R^2 for variable selection. He clearly demonstrates that the use of cross-validation (CV) as a means of variable selection, although reducing the risk of chance effects, with leave-multiple-out CV performing better than leave-one-out CV, does not eliminate the problem.

The results show that selection bias cannot be ignored because, if it is, then false claims about the usefulness of a group of variables in predicting a particular response will continue. We recommend that users of standard automated variable selection methods available in many statistical packages should ignore the reported significance levels and use eqs 5−7 or Table 2 to assess the significance of their final model.

ACKNOWLEDGMENT

The authors wish to express their thanks for the critical comments and suggestions from the reviewers which have helped to improve the research presented in this paper.

REFERENCES AND NOTES

(1) Livingstone, D. J. The Characterisation of Chemical Structures Using Molecular Properties − A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Mannheim, Germany, 2000.

(3) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D. J.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Makarenko, A. S.; Tanchuk, V. Yu.; Prokopenko, R. Virtual Computational Chemistry Laboratory − Design and Description. *J. Chem. Inf. Comput. Sci.* **2005**, *19*, 453−463.

(4) Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1996.

(5) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure−Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066−1078.

(6) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Strcuture−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(7) Copas, J. B. Regression, Prediction and Shrinkage. *J. R. Stat. Soc. B* **1983**, *45*, 311−354.

(8) Rencher, A. C.; Pun, F. C. Inflation of R^2 in Best Subset Regression. *Technometrics* **1980**, *22*, 49−53.

(9) Wilkinson, L.; Dallal, G. E. Tests of Significance in Forward Selection Regression with an *F*-to Enter Stopping Rule. *Technometrics* **1981**, *23*, 377−380.

(10) Miller, A. J. Selection of Subsets of Regression Variables. *J. R. Stat. Soc. A* **1984**, *147*, 389−425.

(11) Diehr, G.; Hoflin, D. R. Approximating the Distribution of the Sample R^2 in Best Subset Regressions. *Technometrics* **1974**, *16*, 317−320.

(12) Pope P. T.; Webster, J. T. The Use of an *F*-Statistic in Stepwise Regression Procedures. *Technometrics* **1972**, *14*, 327−340.

(13) McIntyre, S. H.; Montgomery, D. B.; Srinivasan, V.; Weitz, B. A. Evaluating the Statistical Significance of Models Developed by Stepwise Regression. *J. Market. Res.* **1983**, *20*, 1−11.

(14) Wilkinson, L. Tests of Significance in Stepwise Regression. *Psychol. Bull.* **1979**, *86*, 168−174.

(15) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QsiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(16) Baumann, K. Chance Correlation in Variable Aubset Regression: Influence of the Objective Function Mechanism and Ensemble Averaging. *QSAR Comb. Sci.* **2005**, *24*, 1033−1046.

(17) 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 Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.

(18) Livingstone, D. J.; Evans, D. A.; Saunders, M. R. Investigation of a Charge−Transfer Substituent Constant Using Computational Chemistry and Pattern Recognition Techniques. *J. Chem. Soc., Perkin Trans. 2* **1992**, 1545−1550.

(19) Ford, M. G.; Livingstone, D. J. Computational Chemistry and QSAR; Multivariate Techniques for Parameter Selection and Data Analysis Exemplified by a Study of Pyrethroid Neurotoxicity. *Quant. Struct.-Act. Relat.* **1990**, *9*, 107−114.

(20) Damborsky, J.; Schultz, T. W. Comparison of the QSAR Models for Toxicity and Biodegradability of Anilines and Phenols. *Chemosphere* **1997**, *34*, 429−446.

CI060113N