Judging the Significance of Multiple Linear Regression Models

David J. Livingstone*,†,‡ and David W. Salt§

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

Received November 4, 2004

Abstract: It is common practice to calculate large numbers of molecular descriptors, apply variable selection procedures to reduce the numbers, and then construct multiple linear regression (MLR) models with biological activity. The significance of these models is judged using the usual statistical tests. Unfortunately, these tests are not appropriate under these circumstances since the MLR models suffer from “selection bias”. Experiments with regression using random numbers have generated critical values ($F_{\max}$) with which to assess significance.

The motivation for this work came from the literature that was collected while preparing a review on variable selection.[1] Thirty years ago almost all QSAR studies were carried out using multiple linear regression (MLR) as the modeling “engine” and tabulated substituent constants to describe changes in chemical structure. There were, of course, exceptions to this. Some other statistical methods such as discriminant analysis were in use, and molecular connectivity descriptors provided an alternative to substituent constants. Dissatisfaction with the constraints of MLR and the deficiencies of substituent constants as molecular descriptors led to the increasing use of alternative analytical methods and computational chemistry software to characterize structure. The situation today is that there are dozens of mathematical and statistical methods used to create QSAR models, and the modeler has a choice of thousands of different molecular descriptors from which to create the models.[2] Nevertheless, MLR is still a popular method, as it does have advantages over other modeling techniques. One of its major advantages is that it is very easy to interpret the resulting regression equations. Another advantage is that it is possible to judge the quality of regression models by statistical tests. Unfortunately, these tests apply only if certain conditions are met.

It was recognized in this journal,[3] 25 years ago, that the consideration of large numbers of variables for inclusion in a supervised learning method[4] such as MLR increased the danger of chance correlations. Topliss demonstrated that the more descriptors that are considered for inclusion in a model, the greater the likelihood that a model may arise by chance. Although this was widely accepted, the guidelines were often misinterpreted and taken to mean that a certain number of data points were required for every descriptor in a model. This was not what the paper suggested but rather that a certain ratio of data points to descriptors *considered* should be maintained in order to reduce the possibility of chance correlations. Furthermore, it was assumed that the standard statistical tests, such as the F test, could still be used to assess the quality of MLR models derived from a subset taken from a larger number of variables. Unfortunately this is not true.

When an MLR model of a particular size has been constructed from a set of variables of the same size, then the F test can be used to judge its significance. If, however, the same size model has been constructed by taking a subset of a larger set of variables, then the model suffers from what is known as “selection bias”.[5] The effect of selection bias is to make the regression equation appear more significant than it really is. What this means in practice is that the critical F values which are used to judge significance need to be inflated but, the question is, by how much? We have investigated this by creating sets of random numbers and then fitting regression models of varying sizes to these sets of random data. Simple simulations of the regression of a single random y variable and three random x variables were carried out with a Minitab macro. Further regression simulations were carried out using C++ software written in-house running on PCs and Silicon Graphics workstations. This software computes random numbers based on a normal or uniform distribution and allows the user to choose the number of cases and the size of the pool of variables from which regression models are calculated. Regression models of a particular size may be computed, or the user may select to calculate all regression models from size one to a maximum model size. The number of simulations is also controlled by the user and in the results reported here was taken to be 50 000.

In each simulation, one of the variables is selected as the response (y variable) and the remaining variables are selected as descriptors (x set). In the case of one term regression models, the response is regressed against each one of the x variables in turn, and the best model, that is to say the model with the best fit, is recorded. When all the simulations have been run, the values of the F statistics ($F_{\max}$) are arranged in an ordered list. Values of the 10, 5, 2.5, and 1% confidence limits can be selected from this list by taking the $F_{\max}$ value that occurs at the top 10, 5, 2.5 and 1% of the simulations. For example, in 1000 simulations the 10% $F_{\max}$ value would occur at position 900 in the list (90th percentile). For two-term regression models, the response variable is regressed against all possible combinations of two variables of the x set and for higher order models all possible combinations of the descriptor variables are examined. The computer time required to run these experiments rises very rapidly as the size of the models examined rises and as the pool of variables is increased. A set of timed runs were shown to fit the model given in eq 1 where N is the number of possible regression models (for models of size p from k variables $N = k!/$

[*] To whom correspondence should be addressed: ChemQuest, Delamere House, 1 Royal Crescent, Sandown, IOW, PO36 8LZ. Tel +44 1983 401793, e-mail davel@chemquest.uk.com.

[†] ChemQuest.

[‡] Centre for Molecular Design, University of Portsmouth.

[§] Department of Mathematics, University of Portsmouth.

10.1021/jm049111p CCC: $30.25
Published on Web 01/14/2005

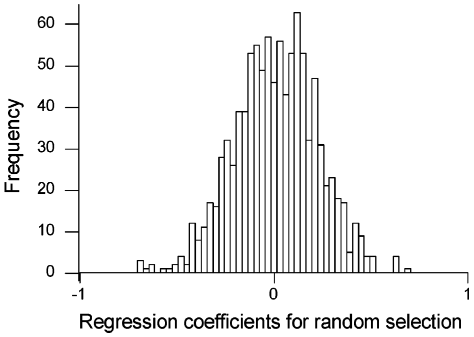

Figure 1. Distribution of slopes for random selection.

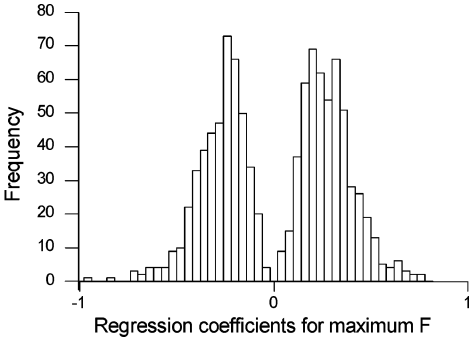

Figure 2. Distribution of slopes for maximum selection.

(p!(k-p)!), p is the size of the models, and n is the number of cases generated for the random variables

$$\text{time(s)} = 0.031N^{0.951}e^{0.382p + 0.008n} \qquad (1)$$

To illustrate the effect of selection bias, four sets of 25 random numbers were generated, one was chosen as the y variable, and this was regressed against one of the x variables chosen at random. The slope, R^2, and F ratio was recorded for this regression. The y variable was then regressed against all of the x variables and the slope, R^2 and F ratio of the best fit recorded. This process was repeated 1000 times. The results are shown in Figure 1 where the values of the slope for the random selection are distributed around a mean of zero. Since the variables are random numbers, there should be no correlation between them and this is the expected result.

The distribution of the values of the slope for the maximum selection, on the other hand, shows a bimodal distribution as illustrated in Figure 2 with the two means some distance from zero on the positive and negative sides.

The distribution of the F ratios (not shown) for the random choice shows a 95th percentile value of 4.18 which corresponds closely to the tabulated value of the F distribution. The 95th percentile value for the maximum fit regressions is equal to 6.65, quite considerably higher than the tabulated F distribution value.

Using our in-house software, simulations were performed for a range of the number of observations, n, the number of variables in a model, p, and the number of variables in the bucket to choose from, k (see Table 1). A range of replications were experimented with and it was found that 50 000 gave an acceptable compromise between time and precision. Fewer replications gave a faster completion time but at the expense of increased variation in the critical values between runs for the same parameter set up.

Table 1. Parameter Values Used in the Simulations

parameters varied	values used
sample size (n)	10, 20, 30, 40, 50, 75, 100
model size (p)	1, 2, 3, 4, 5, 6, 7, 8
bucket size (k)	5, 10, 20, 50, 100

Table 2. Comparison of Actual, Tabulated, and Simulated F Values for the Antimycin Data

terms	1	2	3
fit	13.55	18.0	17.5
table	4.62	3.85	3.54
$F_{\max}$ (k)	4.58	3.63	3.48
10 vars	11.06 (11.95)	10.03 (10.07)	9.80 (9.83)
23	13.88 (14.20)	14.85 (14.39)	17.39 (17.17)
53	17.22 (16.87)	21.11 (20.44)	29.73 (29.30)

The $F_{\max}$ critical values for 5% significance values were recorded and a power function response surface fitted to enable the critical values for other combinations of n, p, and k to be found. This function, shown in eq 2 where N, p, and k have the same meaning as for eq 1, has been tested using a number of combinations of parameters, and some of these are reported below for a typical problem.

$$\hat{F}_{\max} = \frac{29.96n^{3.18}N^{0.21}}{p^{0.82}}e^{[1.06(\ln(v_2))^2 - 0.97\ln(n) - 3.97]} \qquad (2)$$

Antimycin Example. This data set,[6] also known as the Selwood set, has been used by many other groups of workers to look at a whole range of different modeling tools. There are a number of problems with this dataset, but it serves to illustrate the importance of selection bias. In the original study 53 descriptors were calculated which were reduced to 23 based on the intercorrelation structure (unsupervised). These were further reduced by supervised selection to result in 10 variables. The training set consisted of only 16 compounds. Table 2 shows the F values obtained from the analysis in which one-, two-, and three-term models were fitted.

The second row of Table 2 gives the usual tabulated critical values of F for 5% significance, and certainly based on this evidence all three models are highly significant. The third row of the table shows the $F_{\max}$ values from the simulations when the pool size was the same as the model size ($k = p$); these, of course, should correspond to the tabulated values and, within the limits of precision of the number of simulations, do. However, when the number of variables in the pool is increased, the significance of these models is reduced. It can be seen that the $F_{\max}$ values are considerably bigger than the tabulated critical values for a pool size of 10, and for a pool size of 23, only one of the models ($p = 2$) might be considered significant. If the selection had been made using all 53 original variables, none of the three models would be significant. This result may be difficult to accept but, with 53 variables to select from, it is possible to obtain a three variable model using random numbers which can achieve an observed F as high as 29.73. The bracketed numbers in Table 2 are the $F_{\max}$ values predicted from the power function response surface and, as can be seen these results for the most part are reasonable, but there is still room for improvement.

Table 3. Upper 5% Points for F Statistic and $F_{\max}$ Values

cases	table F	$k = 5$	10	20	50	100
10	4.76	9.79	24.15	55.0	156.6	326.2
15	3.59	5.8	10.4	17.5	33.0	51.6
20	3.24	4.88	7.89	12.1	19.8	27.7
50	2.81	3.90	5.50	7.4	10.2	12.6
100	2.70	3.64	5.04	6.5	8.6	10.4

So, what does this mean in practice? If MLR models are constructed from a pool of variables which is the same size as the model then the standard F tables may be used to judge significance. If the pool is larger than the models, and the variables are chosen to maximize the fit, then use of the standard F tables will give an overoptimistic impression of the significance of the model. It is necessary to use critical values of the $F_{\max}$ distribution, a selection of which is reported here. It would be convenient to be able to simply replace standard F tables with a set of $F_{\max}$ tables, but unfortunately this is not possible because the $F_{\max}$ values are dependent on the size of the variable pool (k) as shown in Table 2. To produce "replacement" F tables would require four tables, for 1, 2.5, 5, and 10% critical values, for each size of the variable pool, k. To give an idea of what effect selection bias has on $F_{\max}$ values, Table 3 shows the value of the tabulated upper 5% critical values for F statistic and corresponding simulated $F_{\max}$ results for three-term regression models for different size variable pools (k) and numbers of cases (n).

Application of the power function response surface shown above may be used to obtain $F_{\max}$ values for any particular combination of n, p, and k. It is planned to make this function, or any improvement to it, available on the Centre for Molecular Design website (www.cmd.port.ac.uk). Work is in hand to extend the range of experimental values of $F_{\max}$ in order to investigate and perhaps improve the fit of this function.

Acknowledgment. This work was partially supported by INTAS grant 00-0363, Virtual Computational Chemistry Laboratory.

References

(1) Livingstone, D. J.; Salt, D. W. Variable Selection − Spoilt for Choice. In *Reviews in Computational Chemistry*; Lipkowitz, K., Ed.; Wiley−VCH: Hoboken, NJ, in press.

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

(3) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−44.

(4) Livingstone, D. J. Pattern Recognition Methods for use in Rational Drug Design. In *Molecular Design and Modeling: Concepts and Applications*; Langone, J. J., Ed.; Academic Press: New York, 1991; Vol. 203 of Methods in Enzymology, pp 613−638.

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

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

JM049111P