The problem of underestimating the residual error variance in forward stepwise regression

L. S. FREEDMAN,[1] D. PEE[2] & D. N. MIDTHUNE[3]

[1] *Biometry Branch, Division of Cancer Prevention and Control,*
National Cancer Institute, Executive Plaza North, Suite 344,
Bethesda, MD 20892, USA
[2] *Information Management Services Inc., 6110 Executive Boulevard, Suite 310,*
Rockville, MD 20852, USA, and
[3] *Information Management Services Inc., 12501 Prosperity Drive, Suite 200,*
Silver Spring, MD 20904, USA

Abstract. Under the global null hypothesis that all covariates are unrelated to the outcome variables, forward stepwise regression procedures should have the property that the probability of selecting a given variable and finding it significant at the α level is equal to α. Because of the problem of underestimating the residual error variance the actual probability can be very different from α. This problem becomes of practical concern when the ratio of the number of variables to the number of observations becomes greater than 0·25, and is more serious for logistic than for linear regression.

1 Introduction

Regression analysis may be the most commonly used statistical method. In many applications the method is used to discover which of several covariates is related to an outcome variable of particular interest. In epidemiology the outcome variable may be continuous, e.g. blood pressure, in which case linear regression models are of interest. Often, however, the outcome is binary, e.g. the presence or absence of disease, and logistic regression methods are used. Because of limitations of time and money epidemiological investigations may involve only moderate numbers of subjects. However, the amount of information gathered on each subject may be substantial. Thus the ratio of the number of explanatory variables to the number of observations can be high.

Investigators often wish to identify a set of important covariates from amongst a large set of potential explanatory variables. To do so they employ variable selection procedures (see Hocking (1976) for a general review). Draper & Smith's (1966) text has been highly influential in its recommendation of stepwise procedures, particularly the forward stepwise method. This method has been implemented in all the major statistical software packages and is probably the most popular of variable selection methods used. Thus the statistical properties of the procedure are of considerable interest. Draper *et al.* (1971) noted that 'the conventional entry test for a new variable in stepwise regression is not theoretically correct, but is used because the necessary exact distributions are not known'. They derived some theoretical results for cases with two, three and four independent covariates. Pope & Webster (1972) provided numerical evidence that significance levels may be substantially inflated. Bendel & Afifi (1977) compared stopping rules in stepwise regression. Butler (1982) gave a method for calculating upper and lower bounds for the 'correct' p level of the best-fitting independent variable. Miller (1984) reviewed the selection of variables for regression at the Royal Statistical Society and his paper was followed by several contributions from the audience.

Significance tests carried out on the final model following a variable selection procedure are difficult to interpret for several reasons. First, the need to test each variable for inclusion in the model raises the problem of interpreting multiple significance tests. The papers by Draper *et al.* (1971) and Butler (1982) address this problem. Second, if important variables are omitted from the final model, then the estimates of the regression coefficients and the associated significance tests of the selected variables will probably be biased. This point is emphasized by Miller (1984). Third, the selection procedure affects the properties of the tests on the final model, as demonstrated by Freedman (1983) and elaborated upon by Freedman & Pee (1989).

Hosmer & Lemeshow (1989) therefore write: 'It is well known that the *p*-values calculated in stepwise selection procedures are not *p*-values in the traditional hypothesis testing context'. Although this is an accurate statement it does reduce the stepwise selection method to the level of exploratory data analysis. Our view is that the method could be used more reliably if some simple statistical properties of the overall procedure were better known.

In this paper we consider the properties of forward stepwise regression under the global null hypothesis that all covariates are unrelated to the outcome variable. This hypothesis is quite relevant where the investigator has no particular evidence for the importance of the covariates. For example, this occurs in nutritional epidemiology where large numbers of nutritional factors are recorded and their association with disease is then investigated. A simple requirement of the variable selection procedure is that, under the global null hypothesis, the probability of identifying a given covariate as significant should equal the nominal significance level. If this requirement holds, at least approximately, then the number of false leads resulting from the selection procedure can be controlled.

We demonstrate that, as mentioned by Miller (1984), forward stepwise selection leads to underestimation of the residual error variance which in turn leads to nominal *p* values for the regression coefficients which are too small. The novelty in our work is that we quantify this effect and thereby identify from our investigations the circumstances in which such an effect becomes of practical concern. We investigate both linear and logistic regression and present some recommendations for the use of forward stepwise regression methods with these models.

2 The computer simulation approach

Suppose we have N observations of an outcome variable Y and p covariates $X_1, \dots, X_p$. We will fit a regression model of Y against the Xs, selecting the covariates using a forward stepwise procedure with screening significance level α_1. Our interest is in how often a given variable will be selected and found significant at the α_2 level in the final model. If the probability of such an event is indeed α_2, then we view the significance level as correct.

A straightforward approach to answering this question is through computer simulation. We can take the X_is as fixed observations and generate a set of values of the outcome variable Y which are stochastically independent of the Xs. This is achieved by simply generating the Ys from pseudo-random numbers. For each new set of Y values we can then conduct a forward stepwise regression and record which of the X_is were selected for the final model and which were significant. Repeating this, we can observe how often a variable is selected for the final model or how often a variable is significant and compare this with the corresponding nominal significance level.

We have investigated forward stepwise selection in both linear and logistic regression. For linear regression we have studied the cases with known and unknown residual variance. To investigate linear regression, values of Y may be generated as standard normal deviates, without loss of generality. For logistic regression, Y is generated as a binary variable and each value of Y may be generated either from a Bernoulli distribution with probability π equal to the observed proportion of 1s in the data, or from a hypergeometric distribution with parameters $(N, N\pi)$. Using the Bernoulli distribution we do not constrain the proportion of 1s to be π *in each simulation*, whereas using the hypergeometric distribution we do make such a constraint. Either method will yield similar results for all but very small sample sizes.

To define the stepwise procedure the level of significance α_1 required for inclusion of variables in the model must be specified, as well as the level of significance α_2 in the final model. We do not allow removal of variables from the model once they have been included.

The method of screening variables depends upon the chosen model and is based on the z test for the partial regression coefficient in linear regression with known variance, the t test for the partial regression coefficient in linear regression with unknown variance, and the score test (Rao, 1965, p. 349) as described by Bartolucci & Fraser (1977) in logistic regression.

3 Results for uncorrelated variables

We used the method described in Section 2 to investigate the effect of the forward stepwise procedure on significance levels of regression coefficients when the X_is are orthogonal (i.e. uncorrelated). We examined cases where the number of observations, N, equalled 100, 200 and 400 and where the number of variables, p, was 25, 50 and 100. The screening level of significance α_1 was constant at 10%. We simulated each case 1000 times. Tables 1–3 show the average number of variables included in the final model and the average number found significant at the (α_2) 5% level (using a two-tailed test).

Table 1 shows the results for linear regression with known variance. In all cases the average proportions of variables included in the model are close to the nominal level of 10% and the average proportions significant are close to the nominal 5% level.

Table 2, which shows the results for linear regression with unknown variance, does not display the same pattern. All the average proportions are larger than the corresponding nominal levels. Furthermore, along the diagonals of the table it can be seen that the average proportions of variables included in the model are almost identical, as are the proportions found significant. This indicates that they depend upon the *ratio* of the number of variables to the number of observations ($\rho = p/N$), but not directly on the number of variables or the number of observations. The discrepancy between the tabulated proportions and their corresponding nominal levels is serious for a ratio ρ of 0·5. Other results not presented here show that the discrepancy grows larger as the ratio increases beyond 0·5.

Table 3 shows the results for logistic regression with the proportion π of 1s in the data equal to 0·5. These results are qualitatively similar to the results for linear regression with unknown variance. Proportions of variables included in the model or found significant are

Table 1. Linear regression with known variance: estimated proportions of variables included in the model at $P<0{\cdot}1$ and significant at $P<0{\cdot}05$ using forward stepwise selection.

Number of observations, N	$P<0{\cdot}1$			$P<0{\cdot}05$		
	$p=25$	$p=50$	$p=100$	$p=25$	$p=50$	$p=100$
100	0·104	0·101		0·051	0·051	
200	0·101	0·100	0·099	0·050	0·049	0·049
400	0·102	0·100	0·100	0·050	0·049	0·051

All standard errors 0·002 or less.

Table 2. Linear regression with unknown variance: estimated proportions of variables included in the model at $P<0{\cdot}1$ and significant at $P<0{\cdot}05$ using forward stepwise selection

Number of observations, N	$P<0{\cdot}1$			$P<0{\cdot}05$		
	$p=25$	$p=50$	$p=100$	$p=25$	$p=50$	$p=100$
100	0·116	0·141		0·064	0·084	
200	0·106	0·117	0·146	0·053	0·062	0·083
400	0·104	0·106	0·119	0·052	0·053	0·062

All standard errors 0·002 or less.

Table 3. Logistic regression: estimated proportions of variables included in the model at $P<0{\cdot}1$ and significant at $P<0{\cdot}05$ using forward stepwise selection

Number of observations, N	$P<0{\cdot}1$			$P<0{\cdot}05$		
	$p=25$	$p=50$	$p=100$	$p=25$	$p=50$	$p=100$
100	0·118	0·168		0·061	0·115	
200	0·110	0·120	0·174	0·054	0·063	0·120
400	0·103	0·107	0·122	0·052	0·053	0·065

All standard errors 0·003 or less.

Table 4. Logistic regression with 200 observations: estimated proportions of variables included in the model ($P<0{\cdot}1$) and significant ($P<0{\cdot}05$) using forward stepwise selection

Proportion π of $Y=1$	$p=50$		$p=100$	
	$P<0{\cdot}1$	$P<0{\cdot}05$	$P<0{\cdot}1$	$P<0{\cdot}05$
0·15	0·127	0·080	0·203	0·183
0·2	0·124	0·073	0·207	0·178
0·3	0·124	0·068	0·186	0·143
0·5	0·120	0·063	0·174	0·120

All standard errors 0·003 or less.

again higher than the nominal levels. In nearly all cases the proportions are higher than those in Table 2. The discrepancies between the proportions and their nominal levels are therefore more serious than for linear regression with unknown variance, but are still tolerable for ratios ρ of 0·25 and less.

We also investigated the effect of varying the proportion π of $Y=1$ on the proportion of variables selected or found significant. In Table 4 we show results for the case of 200 observations with either 50 or 100 covariates, with π equal to 0·15, 0·2, 0·3 or 0·5. The proportions increase as π departs from the value of 0·5. Thus discrepancies between the average proportion and their nominal values become even more serious for small values of π. The same results would hold for values of π greater than 0·5. For values of π below 0·3 or above 0·7 the nominal significance levels appear unreliable even when the ratio ρ of variables to observations is as low as 0·25.

4 Theory

Freedman (1983) presented an asymptotic theory for his two-stage selection procedure when outcome variable Y is normally distributed and covariates X_i $(i=1,\ldots,p)$ are uncorrelated. In this section we develop an asymptotic result for linear regression with unknown variance in which forward stepwise selection is used to select from a set of uncorrelated variables.

As in Freedman (1983) suppose, without loss of generality, that $\Sigma_{j=1}^{N} Y_j=0$, var $Y=1$, and that $X_{ij}=0$ except when $i=j$ in which case $X_{ij}=1$ $(i=1,\ldots,p;\ j=1,\ldots,N)$. Then the estimated regression coefficient $\hat{\beta}_i=Y_i$ $(i=1,\ldots,p)$.

For the forward stepwise procedure, re-order X_i according to the descending magnitude of $|Y_i|$. The first variable to enter is then X_1 and it is selected if $|Y_1|>t_{N-1,\alpha}s_{N,1}$ where $t_{N-1,\alpha}$ is the upper α percentile of the t distribution on $N-1$ degrees of freedom and

$$s_{N,1}{}^2=\frac{1}{N-1}\sum_{i=2}^{N} Y_i^2$$

Here α is the level of significance required at screening. (We have dropped the subscript 1 for convenience). In general X_q is selected if $|Y_q|>t_{N-q,\alpha}s_{N,q}$ where

$$s_{N,q}{}^2=\frac{1}{N-q}\sum_{i=q+1}^{N} Y_i^2$$

Now,

$$s_{N,q}{}^2=\frac{1}{N-q}\left(\sum_{i=1}^{N} Y_i^2-\sum_{i=1}^{q} Y_i^2\right)$$

and taking expectations

$$\begin{aligned}E(s_{N,q}{}^2)&=\frac{1}{N-q}[N-qE(Y^2||Y|>t_{N-q,\alpha}s_{N,q})]\\ &=\frac{1}{N-q}\left[\frac{N-qg(t_{N-q,\alpha}s_{N,q})}{\Phi(t_{N-q,\alpha}s_{N,q})}\right]\end{aligned} \qquad (1)$$

where

$$\Phi(y)=P(|Y|>y)=\left(\frac{2}{\pi}\right)^{1/2}\int_{y}^{\infty}\exp\left(-\frac{u^2}{2}\right)\mathrm{d}u$$

and

$$g(y)=\Phi(y)+\left(\frac{2}{\pi}\right)^{1/2} y\exp\left(-\frac{y^2}{2}\right)$$

Moreover, $E(q)=N\rho\Phi(t_{N-q,\alpha}s_{N,q})$ where $\rho=p/N$. Letting $N\to\infty$ we may substitute $E(q)$ for q in equation (1) and set $t_{N-q,\alpha}=z_\alpha$, the equivalent normal deviate, so that

$$s_{N,q}\to\left[\frac{1-\rho g(z_\alpha s_{N,q})}{1-\rho\Phi(z_\alpha s_{N,q})}\right]^{1/2} \qquad (2)$$

Equation (2) may be solved for $s_{N,q}$ by an iterative process. Starting with $s_{N,q}=1$, we evaluate the right-hand side of the equation to obtain a new value of $s_{N,q}$ and continue until convergence. The expected proportion of variables selected is then given by $\Phi(z_\alpha s_{N,q})$, and the proportion significant at the α_2 level is $\Phi(z_{\alpha_2}s_{N,q})$.

Table 5. Asymptotic expectations of the residual standard deviation $s_{N,q}$, and the proportions of variables selected ($P<0{\cdot}1$) or found significant ($P<0{\cdot}05$)

Ratio ρ of variables to observations	Residual standard deviation $s_{N,q}$	Proportion selected ($P<0{\cdot}1$)	Proportion significant ($P<0{\cdot}05$)
0·25	0·952	0·118	0·062
0·50	0·885	0·146	0·082

Table 6. Linear regression with 200 observations and 100 variables: estimated proportions of variables included in the model ($P<0{\cdot}1$) and significant ($P<0{\cdot}05$) using forward stepwise selection, for spherically correlated covariates

	Known variance		Unknown variance	
Correlation r	$P<0{\cdot}1$	$P<0{\cdot}05$	$P<0{\cdot}1$	$P<0{\cdot}05$
0·0	0·099	0·049	0·146	0·083
0·1	0·096	0·054	0·141	0·086
0·3	0·094	0·058	0·139	0·089
0·5	0·090	0·058	0·132	0·089
0·7	0·071	0·046	0·107	0·071
0·9	0·033	0·022	0·050	0·034

Table 5 shows the values of $s_{N,q}$ and the proportions of variables selected (at the 10% level) or found significant (at the 5% level) for $\rho=0{\cdot}25$ and $\rho=0{\cdot}5$. The proportions agree well with the values in Table 2 obtained by computer simulation.

The theory shows that the reason for discrepancy between the proportion of variables selected or found significant and their nominal significance levels is due to underestimation of the residual variance. In the case of known variance (Table 1) there is little or no discrepancy, because the residual variance is not estimated. However, with unknown variance, inclusion of covariates which are spuriously related to the outcome reduces the estimated residual variance, and thereby allows more covariates to enter the model (Table 2). A similar phenomenon occurs in logistic regression which requires the iteratively weighted least squares method for model fitting and inference.

5 Correlated variables

In practice covariates will not usually be orthogonal. We therefore investigated the influence of correlation between the covariates upon the results presented in Section 3. We chose the case of $N=200$ and $p=100$ for this investigation. Covariates were generated with correlation matrices of either the 'spherical' or 'autocorrelation' type. The spherical correlation matrix has elements $\{r_{ii}=1, r_{ij}=r;\ i,j=1,\ldots,100 \text{ and } i\neq j\}$. The autocorrelation matrix has elements $\{r_{ij}=r^{|i-j|};\ i,j=1,\ldots,100\}$. Both types of correlation structures were studied for $r=0{\cdot}1$, $0{\cdot}3$, $0{\cdot}5$, $0{\cdot}7$ and $0{\cdot}9$.

Table 6 shows the results of linear regression with known variance and unknown variance using the spherical correlation matrix. Results with the autocorrelation matrix were closely similar. In the case of known variance, the proportion of variables significant at the 5% level rises slightly above 0·05 for $r=0·1$, 0·3 and 0·5, decreases to just below 0·05 for $r=0·7$ and is only 0·02 when $r=0·9$. Hence for very highly correlated covariates the nominal significance levels are too high, but correlations of 0·7 or lower do not seriously affect the significance levels. In the case of unknown variance, the proportion of variables significant at the 5% level again rises, at $r=0·1$, 0·3 and 0·5, slightly above the 0·083 proportion found in the case when $r=0$, and decreases to 0·07 for $r=0·7$ and to 0·03 for $r=0·9$. Thus proportions of variables substantially above the nominal 5% are found significant for variables correlated at all levels below 0·9.

These results indicate that the results for correlated variables are not substantially different from those for variables which are uncorrelated unless very high correlations (above 0·7) exist.

6 Discussion

In this paper we have focused on a rather simple criterion for judging the significance levels resulting from a forward stepwise selection procedure. We invoke a global null hypothesis and ask what the probability is that a given variable will be found significant in the final model. If we do not know this probability then we have no control over the number of false leads which may be generated by our procedure. Epidemiologists are sometimes criticized (Feinstein, 1989) for reporting a plethora of risk factors for different diseases most of which are unconfirmed by further studies. In our view, uncritical use of significance testing following variable selection procedures has played its part in this phenomenon.

Our investigations of forward stepwise regression in the case of uncorrelated variables have revealed some serious discrepancies between the true probabilities of finding a given variable significant under the global null hypothesis and the nominal significance levels. These generally occur when the ratio of variables to observations is high, e.g. greater than 0·25. Rutter *et al.* (1991) have recently reported similar problems of bias in error rate estimates in stepwise discriminant analysis when the ratio of variables to observations was 0·40 or above. Although not shown here explicitly, the level of significance required for including variables in the model can also be influential, with more lenient inclusion rules magnifying the problem. Thus recommendations such as those given by Bendel & Afifi (1977) to increase α_1 to 0·15, 0·20 or 0·25 can exacerbate the discrepancy between nominal and actual significance.

The results in Section 4 suggest that, for linear regression with unknown variance, the problem with the significance levels may be rectified in the following way. First obtain an estimate of the residual variance by fitting the regression model with *all* the covariates included. Then conduct the forward stepwise procedure using that estimate of the residual variance. Our preliminary investigations of the properties of this procedure indicate that it will preserve the nominal significance levels.

In fact, under the null hypothesis, any consistent estimator of the residual variance should preserve nominal significance levels. The advantage of the estimate obtained from inclusion of all the covariates is that it is consistent even when some of the covariates *are* truly related to outcome Y. Using this estimate will more powerfully detect important covariates than using a variance estimate based on, for example, the inclusion of no covariates.

Alternatively, Copas & Long (1991), in a recent paper that is closely related to ours, suggest an elegant method of correcting the degrees of freedom of the residual variance for the variable selection process when the covariates are uncorrelated.

For this work we have used a global null hypothesis that none of the covariates considered in the forward selection procedure is related to the outcome variable. When covariates that are known to be related to outcome are available these should initially be forced into the model and the forward stepwise procedure could proceed from that model. In this case our recommendations would refer to the ratio of the number of covariates considered in the forward stepwise phase to the number of observations *minus the number of covariates forced into the model.*

In practice correlations will exist between the covariates. We have shown in Section 5 that unless covariates are very highly correlated ($r > 0{\cdot}7$) then the results we found for uncorrelated variables still apply qualitatively. If two very highly correlated variables are included in a stepwise selection procedure, however, then one can usually expect to include at most one of them in the model, since one competes with the other for inclusion. Hence the error rate of false inclusion is also considerably reduced. It is common to warn against inclusion of two highly correlated covariates in a regression model because of the problem of collinearity. There should be a similar warning against using highly correlated covariates in a stepwise procedure.

Our general recommendation is that when the ratio of variables to observations is 0·25 or greater, simulations of the type described in Section 6 should be conducted. These simulations are required to provide the investigator with adequate knowledge regarding the properties of the forward stepwise procedure employed, and to control the number of false leads arising from exercises which are essentially exploratory in nature.

Acknowledgements

The authors thank Charles Brown, David Byar, Mitchell Gail and John Gart for valuable discussions during the development of this work, and Jennifer Gaegler for excellent secretarial support.

References

Bartolucci, A. & Fraser, M. (1977) Comparative step-up and composite tests for selecting prognostic indicators associated with survival, *Biometrical Journal*, 19, pp. 437–448.

Bendel, R. B. & Afifi, A. A. (1977) Comparison of stopping rules in forward stepwise regression, *Journal of the American Statistical Association*, 72, pp. 46–53.

Butler, R. W. (1982) Bounds on the significance attained by the best-fitting regressor variables, *Applied Statistics*, 31 (3), pp. 290–292.

Copas, J. B. & Long, T. (1991) Estimating the residual variance in orthogonal regression with variable selection, *The Statistician*, 40, pp. 51–59.

Draper, N. & Smith, H. (1966) *Applied Regression Analysis* (New York, Wiley).

Draper, N., Guttman, I. & Kanemasu, H. (1971) The distribution of certain regression statistics, *Biometrika*, 58 (2), pp. 295–298.

Feinstein, A. R. (1989) Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials, *Journal of Clinical Epidemiology*, 42, pp. 481–489.

Freedman, D. (1983) A note on screening regression equations. *The American Statistician*, 37 (2), pp. 152–155.

Freedman, L. & Pee, D. (1989) Return to a note on screening regression equations, *The American Statistician*, 43 (4), pp. 279–282.

Hocking, R. R. (1976) The analysis and selection of variables in linear regression, *Biometrics*, 32, pp. 1–49.

Hosmer, D. W. & Lemeshow, S. (1989) *Applied Logistic Regression*, p. 111 (New York, Wiley).

Miller, A. (1984) Selection of subsets of regression variables (with discussion), *Journal of the Royal Statistical Society, Series A*, 147 (3), pp. 389–425.

Pope, P. T. and Webster, J. T. (1972) The use of an F-statistic in stepwise regression procedures, *Technometrics*, 14, pp. 327–340.

Rao, C. R. (1965) *Linear Statistical Inference and its Applications*, p. 349 (New York, Wiley).

Rutter, C., Flack, V. and Lachenbruch, P. (1991) Bias in error rate estimates in discriminant analysis when stepwise variable selection is employed, *Communications in Statistics—Simulations*, 20, pp. 1–22.