Return to a Note on Screening Regression Equations

LAURENCE S. FREEDMAN and DAVID PEE*

We revisit an article by D. A. Freedman on screening variables for regression models. After summarizing his asymptotic results we show that the theory does not entirely explain the results of computer simulations of his model. We demonstrate that this is due to the random correlation between simulated independent random variables. Finally, we explore some consequences of the asymptotic results. In the case of uncorrelated variables using the proposed two-stage screening procedure, it is possible to obtain significance tests for the final F statistic and t statistic that have the correct Type I error rates.

KEY WORDS: F statistic; Multiple correlation coefficient; Significance levels; Variable selection.

1. INTRODUCTION

In May 1983 *The American Statistician* published "A Note on Screening Regression Equations," by D. A. Freedman. In that article Freedman considered the problem of developing a regression model relating a variable Y to a subset of some possibly explanatory variables X_i ($i = 1, \ldots, p$). His particular concern was to investigate the effect of a procedure for screening out "unimportant" variables on the value of the multiple correlation coefficient, R^2, and on the value of the F statistic for the overall regression, calculated from the "final" regression model. He developed asymptotic theory for multiple linear regression with normal homoscedastic error for the case of a simplified two-stage selection procedure. In the first stage, the screen, all variables are entered into the regression equation. Those variables, whose partial regression coefficients are significant at a chosen probability level α, are then retained for the second, final regression model. Freedman reported the results of computer simulations to support the predictions of the asymptotic theory. The article warned that, when the ratio of the number of variables (p) to observations (n) is high, one will obtain, through screening, overall F statistics that are highly significant, even when none of the explanatory variables X_i are truly related to Y. Recent work (e.g., Freedman, Navidi, and Peters 1988) shows that bootstrap and jackknife methods do not appear to remedy the situation.

In this article we touch on two aspects of Freedman's work. First, we report on further simulations, note a discrepancy between the simulation results and the asymptotic theory, and explore the reasons for the discrepancy. Second, using the asymptotic theory, we investigate the effect of using different probability levels, α, for the screening, and the effect of varying the ratio of the number of variables to the number of observations upon the magnitude of the multiple correlation coefficient and the F statistic.

2. PREVIOUS RESULTS

We summarize here the results from Freedman (1983).

2.1 Asymptotic Theory

Let Y be the dependent variable, and let X_i ($i = 1, \ldots, p$) be the possibly explanatory variables. Let n be the number of observations, that is, the number of sets of ($Y, X_1, \ldots, X_p$) comprising the data. Let the ratio of the number of variables to the number of observations, p/n, be denoted by ρ.

Suppose that $Y, X_1, \ldots, X_p$ are independent normal variables, so that none of the X_i's relate to Y and none of the X_i's are intercorrelated. We now apply the screening procedure described in the previous section.

For the first stage of the procedure the following results hold asymptotically (i.e., as p and $n \to \infty$):

$$R_1^2 \to \rho \qquad (1)$$

and

$$F_1 \to 1.0, \qquad (2)$$

where R_1^2 is the multiple correlation coefficient and F_1 is the F statistic for the overall regression.

Moreover, the number of variables passing the screen, q_1, is given by

$$q_1 \to n\alpha_1\rho, \qquad (3)$$

where α_1 is the critical probability level of the screen.

For the second stage of the procedure the following results for the multiple correlation coefficient and F statistic hold:

$$R_2^2 \to g(\lambda)\rho, \qquad (4)$$

and

$$F_2 \to g(\lambda)(1 - \alpha_1\rho)/[\alpha_1(1 - g(\lambda)\rho)], \qquad (5)$$

where λ is the normal deviate corresponding to a two-tailed z test at level α_1 (e.g., if $\alpha_1 = .05$, then $\lambda = 1.96$) and

$$g(\lambda) = \alpha_1 + \sqrt{2/\pi}\,\lambda\exp(-\lambda^2/2).$$

The number of explanatory variables with partial regression coefficients significant at the 5% level at the second stage is q_2, and

$$q_2 \to np\Phi(1.96/\sqrt{(1 - \alpha_1\rho)/(1 - g(\lambda)\rho)}), \qquad (6)$$

where $\Phi(x) = 2\int_x^\infty (1/\sqrt{2\pi})\exp(-w^2/2)dw$.

*Laurence S. Freedman is Expert, Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD 20892. David Pee is Biostatistician, Information Management Services, Inc., Silver Spring, MD 20910. The authors thank David Byar and Charles Brown for helpful suggestions and Jennifer Gaegler for excellent secretarial assistance.

2.2 Simulation

Freedman chose the case in which $n = 100$, $p = 50$, and $\alpha_1 = .25$. Ten simulations of the screening procedure were reported. The six columns of Table 1 show the values of R_1^2, F_1, q_1, R_2^2, F_2, and q_2. In the first row we display their values expected by asymptotic theory. In the second row we show the means and standard errors of the values from Freedman's reported 10 simulations.

It may be seen from Table 1 that, whereas F_1 is about 1.0 and nonsignificant, the value of F_2 is inflated and is, on average, highly significant. Moreover, the number of variables q_2 found significant at the 5% level by the two-stage selection procedure is about twice that normally expected from applying a test of significance at the 5% level in a null situation. That is, the true chance of a Type I error is closer to 10% than to the nominal 5%. Thus Freedman advised caution in interpreting results from regressions derived from screening procedures.

3. NEW SIMULATIONS

Inspection of the first two rows of Table 1 reveals some apparent differences between Freedman's computer simulation results and those predicted by asymptotic theory. Although the values of R_1^2, F_1, and q_1 agree well with theory, those of R_2^2 and F_2 do not. The mean R_2^2 is .27 compared with the theoretical value of .36, whereas the mean F_2 is 2.96 compared with 3.97 predicted by theory. The simulated and theoretical values of q_2 appear to be close. We carried out a larger number of simulations of the same problem to check these apparent trends. The results are shown in the third row of Table 1. They confirm that R_1^2 and q_1 are consistent with the theory, although F_1 is slightly larger than predicted. They also make quite clear the discrepancies between simulation and theory for R_2^2 and F_2 suggested by the original simulations. In addition, the new simulations reveal that the mean value of q_2 (5.42) is larger than predicted by theory (4.75). Aside from the question of why the simulations do not agree with the theoretical values, these results are puzzling in another way. One would expect that if the final F statistic, F_2, were smaller than predicted, then the number of significant variables, q_2, would also be smaller. The opposite is found, however, namely that q_2 is larger than theory predicts.

The explanation for these disagreements is contained in the observation that, although $X_1, \ldots, X_{50}$ are generated as independent random variables, their realizations are in fact correlated and these correlations for $n = 100$ and $p = 50$ are nonnegligible. To verify this idea we simulated sets of X's that were constrained to be orthonormal. We did this by replacing each simulated set of X's by their principal components. The results are shown in the fourth row of Table 1, and the mean values of R_2^2, F_2, and q_2 are now consistent with the asymptotic theory.

Some insight into why correlation between the variables affects the values of R_2^2, F_2, and q_2 may be obtained by considering the simplest case in which two of the variables v_1, v_2 are correlated with coefficient r but are orthogonal to the remaining variables. These remaining variables are themselves mutually orthogonal. We can think of the two variables as being a linear transformation of two orthogonal variables, x_1 and x_2. For example, $v_1 = x_1$ and $v_2 = rx_1 + (1 - r^2)^{1/2}x_2$. The quantity R_1^2 is invariant under linear transformation and is, therefore, unaffected by the correlation. The average contribution from v_1 and v_2 to the quantity R_2^2, however, is not the same as that from uncorrelated variables x_1 and x_2. The contribution from x_1 and x_2 is .01434. (This value, being the contribution to R^2 from *two* uncorrelated variables, is 1/25 of the asymptotic expectation of R^2 for *50* uncorrelated variables, namely the .36 shown in the first row of Table 1.) When the correlation, r, between v_1 and v_2 is such that $|r| \leqq .15$, then their contribution to R_2^2 changes little from this value. For $|r| = .2$, however, the average contribution from v_1 and v_2 is lower at .01422, decreasing to .01398 when $|r| = .35$ and .01355 when $|r| = .5$. These results are obtained by numerical integration.

This case of bivariate correlation shows us qualitatively that correlations tend to reduce the value of R_2^2. The average $|r|$ between pairs of X's in our simulations, however, was .08 with a standard deviation of .05. Thus we would expect only a small reduction in R_2^2 in our simulations as a result of bivariate correlations, not enough to explain the difference between the theoretical value and the average value observed. Our earlier observation that the simulated values of R_2^2 are consistent with asymptotic theory when the X's are uncorrelated leads us to suppose that multiple correlations, where several variables are interrelated with one another, have a greater quantitative impact on R_2^2 than that

Table 1. Values of Multiple Correlation Coefficients, F Statistics, and Number of Significant Variables at First and Second Stage of Selection Procedure, According to Asymptotic Theory and the Means From Three Different Sets of Simulations

	First stage			Second stage		
Source	R_1^2	F_1	q_1[a]	R_2^2	F_2	q_2[b]
Asymptotic theory	.50	1.00	12.5	.36	3.97	4.75
Freedman's simulations	.48	.97	11.60	.27	2.96	4.90
(10 repetitions)	(.02)	(.10)	(1.4)	(.04)	(.28)	(.99)
New simulations	.500	1.040	12.51	.308	3.29	5.42
(500 repetitions)	(.003)	(.014)	(.20)	(.005)	(.05)	(.15)
Simulations of orthonormal variables	.495	1.025	12.15	.348	4.00	4.46
(100 repetitions)	(.007)	(.031)	(.38)	(.010)	(.08)	(.27)

NOTE: Standard errors are in parentheses.
[a] Number significant at the 25% level.
[b] Number significant at the 5% level.

produced by bivariate correlations, thus providing a possible explanation for the difference between the theoretical and simulated value of R_2^2.

Our general conclusion is that the results from the computer simulations do not meet those predicted by the asymptotic theory because of the correlations between the simulated X's. The asymptotic theory is valid for orthonormal X's, but not for stochastically independent X's. As n tends to infinity the correlations between the X's tend to 0, so for sufficiently large n we might expect the simulations to give results close to the asymptotic theory. The impracticality of inverting very large matrices, however, prevented us from verifying this, and there is some doubt as to the truth of this conjecture. The asymptotic theory nevertheless works well for a strictly orthogonal X matrix even when n is moderate, as was shown earlier.

4. DEDUCTIONS FROM THE ASYMPTOTIC THEORY

Equations (4)–(6) provide us with some simple relationships of the values of R_2^2, F_2, and q_2 in the final regression analysis with ρ, the ratio of variables to observations, and α_1, the critical level for screening. Although the regression problem and the screening procedure proposed here are too simplified to relate directly to most real-life problems, examining the effect of different values of ρ and α_1 on the values of R_2^2, F_2, and q_2 may nevertheless give us insights that are useful in more complicated situations.

Tables 2, 3, and 4 show different values of R_2^2, F_2, and $q_2/\rho n$, respectively, for a range of values of ρ and α_1. We have chosen four values of ρ from .1 to .75. The lower value (.1) was once cited by Cornfield (1976) as the minimum desirable ratio of variables to observations required for multiple regression problems. The upper value (.75) represents a study where there is an embarrassingly large number of variables relative to the observations available. The six values of α range from 1.0 to .01. The case α_1 = 1.0 corresponds to no screening, since all variables must pass the screen at this critical level. A common practice is to employ α_1 = .1 or .05.

Table 2 shows that the final multiple correlation coefficient decreases as the screening level α_1 becomes more stringent. More stringent screening allows fewer variables to be included in the final regression model, thus leading to a smaller value for R_2^2.

It is similarly clear that the lower is the number of variables per observation the smaller is R_2^2, remembering that none of the explanatory variables are truly related to the dependent variable.

Table 2. Effect of Changing Screening Level, α_1, and Variables/Observations Ratio, ρ, on the Final Multiple Correlation Coefficient, R_2^2

	ρ			
α_1	.75	.50	.25	.10
1.00	.75	.50	.25	.10
.50	.70	.46	.23	.093
.25	.54	.36	.18	.072
.10	.33	.22	.11	.044
.05	.21	.14	.070	.028
.01	.063	.042	.021	.0084

Table 3. Effect of Changing Screening Level, α_1, and Variables/Observations Ratio, ρ, on the Final F statistic, F_2

	ρ			
α_1	.75	.50	.25	.10
1.00	1.00	1.00	1.00	1.00
.50	3.83	2.60	2.12	1.95
.25	5.14	3.97	3.31	3.04
.10	6.06	5.35	4.81	4.55
.05	6.79	6.32	5.93	5.71
.01	8.95	8.78	8.61	8.51

NOTE: The significance of F_2 will depend on n. For n = 100 the degrees of freedom for F_2 will be, on average, 12.5 and 87.5. The critical values of the F distribution for these degrees of freedom are $P < .05$, $F = 1.86$; $P < .01$, $F = 2.40$; and $P < .001$, $F = 3.12$.

Table 3 shows that the final F statistic increases as the screening level α_1 becomes more stringent. The rate of increase with α_1 is less for lower values of ρ. The significance of the F statistic will depend on the number of observations, n. Even for moderate screening levels (α_1 = .25) and low ratios of variables to observations (ρ = .1), however, the average F statistic for the final overall regression will be highly significant ($\rho \simeq .001$) when n = 100 or more. In general, the level of α_1 is more influential than that of ρ in determining the value of F_2.

Table 4 shows values of $q_2/\rho n$. We have chosen to tabulate this quantity rather than q_2, since it is independent of n and it represents the proportion of the original variables that are found significant at the 5% level in the final regression. Ideally, one would like this proportion to be close to .05. If it is not, then the significance tests of regression coefficients of the retained variables do not have a true Type I error rate of 5%. From Table 4 we can see that the most serious departures from the value .05 are for large α_1 (.5 or .25) or large ρ (.75 or .5). For ρ = .25 or smaller and for α_1 = .05 or smaller the Type I error rate of the final significance test is not seriously different from .05.

5. DISCUSSION

The results from the asymptotic theory shown in Table 3 are disquieting, showing that with almost any screening level and a modest ratio of variables to observations, spuriously large F statistics are obtained. This is almost certainly true of other common screening procedures. The results in Table 4 are a little more comforting, since a stringent screening level and a reasonably low variable/observation ratio will preserve the Type I error rate of the final significance test on the regression coefficients of the variables. In particular, Table 4 tells us that if ρ is ¼ or less we should

Table 4. Effect of Changing Screening Level, α_1, and Variables/Observations Ratio, ρ, on the Proportion of Variables Finally Found Significant at the 5% Level, $q_2/\rho n$

	ρ			
α_1	.75	.50	.25	.10
1.00	.05	.05	.05	.05
.50	.17	.097	.066	.055
.25	.14	.095	.067	.056
.10	.095	.075	.061	.054
.05	.075	.066	.057	.052
.01	.057	.055	.052	.051

not be too concerned about the effect of screening on the Type I error. On the other hand, if ρ is ½ or greater we need to use quite stringent screening ($\alpha_1 \leqq .05$) to preserve the Type I error rate.

Even with the Type I error rate preserved at .05 we confront another problem. Suppose that there are 40 variables and 4 come out significant at the 5% level. These could easily consist of 2 genuinely predictive variables and 2 that are significant purely by chance. Without prior information, further experimentation, or other external knowledge, our only statistical device for distinguishing between the real and apparent predictors is to increase the stringency of the significance test, say to the 1% level. However, we thereby risk omitting useful variables from the regression model.

The message from Freedman's article was thus essentially cautionary and negative. Because of the difficulty of selecting variables for regression models, it is wise to advocate discretion. The theory developed by Freedman, however, does suggest the possibility that some of the statistical artifacts introduced by screening variables may be corrected for. For example, in the situation described in this article one could control the overall chance of finding a variable significant at the final regression to be 5% by adjusting the critical z level of the test to

$$1.96 \sqrt{(1 - \alpha_1 \rho)/(1 - g(\lambda)\rho)}.$$

Similarly, corrections to the test based on the final F statistic could easily be made on the basis of Equation (5). It is clear that to be useful this work would need to be extended to correlated variables and to other screening procedures, such as step-up and step-down variable selection.

[*Received November 1988. Revised February 1989.*]

REFERENCES

Cornfield, J. (1976), "Recent Methodological Contributions to Clinical Trials," *American Journal of Epidemiology*, 104, 408–420.

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

Freedman, D. A., Navidi, W., and Peters, S. C. (1988), "On the Impact of Variable Selection in Fitting Regression Equations," in *On Model Uncertainty and Its Statistical Implications* (Lecture Notes in Economics and Mathematical Systems, No. 307), ed. T. K. Dijlestra, Berlin: Springer-Verlag, pp. 1–16.