Appendix A

A Note on Screening Regression Equations

DAVID A. FREEDMAN*

Consider developing a regression model in a context where substantive theory is weak. To focus on an extreme case, suppose that in fact there is no relationship between the dependent variable and the explanatory variables. Even so, if there are many explanatory variables, the R^2 will be high. If explanatory variables with small t statistics are dropped and the equation refitted, the R^2 will stay high and the overall F will become highly significant. This is demonstrated by simulation and by asymptotic calculation.

KEY WORDS: Regression; Screening; R^2; F; Multiple testing.

1. INTRODUCTION

When regression equations are used in empirical work, the ratio of data points to parameters is often low; furthermore, variables with small coefficients are often dropped and the equations refitted without them. Some examples are discussed in Freedman (1981) and Freedman, Rothenberg, and Sutch (1982, 1983). Such practices can distort the significance levels of conventional statistical tests. The existence of this effect is well known, but its magnitude may come as a surprise, even to a hardened statistician. The object of the present note is to quantify this effect, both through

* David A. Freedman is Professor, Statistics Department, University of California, Berkeley, CA 94720. This research developed from a project supported by Dr. George Lady, of the former Office of Analysis Oversight and Access, Energy Information Administration, Department of Energy, Washington, D.C. I would like to thank David Brillinger, Peter Guttorp, George Lady, Thomas Permutt, and Thomas Rothenberg for their help. Reprinted with permission by *The American Statistian*.

Common Errors in Statistics (and How to Avoid Them), by Phillip I. Good and James W. Hardin.
ISBN 0-471-46068-0

simulation (Section 2) and through asymptotic calculation (Section 3). For another discussion, see Rencher and Pun (1980).

To help draw the conclusion explicitly, suppose an investigator seeks to predict a variable Y in terms of some large and indefinite list of explanatory variables $X_1, X_2, \ldots$. If the number of variables is comparable to the number of data points, and if the variables are only imperfectly correlated among themselves, then a very modest search procedure will produce an equation with a relatively small number of explanatory variables, most of which come in with significant coefficients, and a high significant R^2. This will be so even if Y is totally unrelated to the X's.

To sum up, in a world with a large number of unrelated variables and no clear a priori specifications, uncritical use of standard methods will lead to models that appear to have a lot of explanatory power. That is the main—and negative—message of the present note. Therefore, only the null hypothesis is considered here, and only the case where the number of variables is of the same order as the number of data points.

The present note is in the same spirit as the pretest literature. An early reference is Olshen (1973). However, there is a real difference in implementation: Olshen conditions on an F test being significant; the present note screens out the insignificant variables and refits the equation. Thus, Olshen has only one equation to deal with; the present note has two. The results of this note can also be differentiated from the theory of pretest estimators described in, for example, Judge and Bock (1978). To use the latter estimators, the investigator must decide a priori which coefficients may be set to zero; here, this decision is made on the basis of the data.

2. A SIMULATION

A matrix was created with 100 rows (data points) and 51 columns (variables). All the entries in this matrix were independent observations drawn from the standard normal distribution. In short, this matrix was pure noise. The 51st column was taken as the dependent variable Y in a regression equation; the first 50 columns were taken as the independent variables $X_1, \ldots, X_{50}$. By construction, then, Y was independent of the X's. Ideally, R^2 should have been insignificant, by the standard F test. Likewise, the regression coefficients should have been insignificant, by the standard t test.

These data were analyzed in two successive multiple regressions. In the first pass, Y was run on all 50 of the X's, with the following results:

- **$R^2 = 0.50$, $P = 0.53$;**
- **15 coefficients out of 50 were significant at the 25 percent level;**
- **1 coefficient out of 50 was significant at the 5 percent level.**

Only the 21 variables whose coefficients were significant at the 25 percent level were allowed to enter the equation on the second pass. The results were as follows:

- **$R^2 = 0.36$, $P = 5 \times 10^{-4}$**
- **14 coefficients out of 15 were significant at the 25 percent level;**
- **6 coefficients out of 15 were significant at the 5 percent level.**

The results from the second pass are misleading indeed, for they appear to demonstrate a definite relationship between Y and the X's, that is, between noise and noise. Graphical methods cannot help here; in effect, Y and the selected X's follow a jointly normal distribution conditioned on having significant t statistics. The simulation was done 10 times; the results are shown in Table 1. The 25 percent level was selected to represent an "exploratory" analysis; 5 percent for "confirmatory." The simulation was done in SAS on the UC Berkeley IBM 4341 by Mr. Thomas Permutt, on April 16, 1982.

3. SOME ASYMPTOTICS

An asymptotic calculation is helpful to explain the results of the simulation experiment. The Y and the X's are independent; condition X to be constant. There is no reason to treat the intercept separately since the Y's and X's all have expectation zero. Finally, suppose X has orthonormal columns. The resulting model is

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is an $n \times 1$ random vector, X is a constant $n \times p$ matrix with orthonormal columns, where $p \geqq p$, while β is a $p \times 1$ vector of parameters, and ε is an $n \times 1$ vector of independent normals, having mean 0 and common variance σ^2. In particular, the rank of X is p. All probabilities are computed assuming the null hypothesis that $\beta \equiv 0$. Suppose

$$n \to \infty \text{ and } p \to \infty \text{ so that } p/n \to \rho, \text{ where } 0 < \rho < 1. \tag{2}$$

Let R_n^2 be the square of the conventional multiple correlation coefficient, and F_n the conventional F statistic for testing the null hypothesis $\beta \equiv 0$. Under these conditions, the next proposition shows that R_n^2 will be essentially the ratio of the number p of variables to the number n of data points: the proof is deferred.

Proposition. Assume (1) and (2). Then

$$R_n^2 \to \rho \text{ and } F_n \to 1 \text{ in probability.} \tag{3}$$

TABLE 1. Simulation Results

Repetition	First Pass					Second Pass					
	R^2	F	P[a]	#25%[b]	#5%[c]	R^2	F	$P \times 10^4$	#p[b]	#25%	#5%
1	0.50	0.98	0.53	15	1	0.36	3.13	5	15	14	6
2	0.46	0.84	0.73	9	0	0.15	1.85	700	9	6	2
3	0.52	1.07	0.40	16	4	0.36	2.93	7	16	16	9
4	0.45	0.83	0.75	7	1	0.14	2.13	500	7	5	4
5	0.57	1.35	0.15	17	2	0.44	3.82	0.2	17	17	9
6	0.46	0.84	0.73	12	1	0.22	2.06	300	12	11	2
7	0.41	0.70	0.89	4	0	0.12	3.33	100	4	3	1
8	0.42	0.72	0.88	12	1	0.27	2.66	40	12	11	3
9	0.39	0.64	0.94	8	0	0.20	2.90	60	8	8	4
10	0.63	1.69	0.03	16	4	0.48	4.80	0.008	16	16	9

[a] *P* is the significance level of the *F* test, scaled up by 10^4 in the second pass.
[b] #25% is the number of variables whose coefficients are significant at the 25% level; only such variables are entered at the second pass; #*p* is the number of such variables, that is, the number of variables in the second pass regression, repeated for ease of reference.
[c] #5% is the number of variables whose coefficients are significant at the 5% level.
Note: The regressions are run without intercepts.

Now consider redoing the regression after dropping the columns of X that fail to achieve significance at level α. Here, $0 < \alpha < 1$ is fixed. Let $q_{n,\alpha}$ be the number of remaining columns. Let $R^2_{n,\alpha}$ be the square of the conventional multiple correlation in this second regression, and let $F_{n,\alpha}$ be the F statistic. These are to be computed by the standard formulas, that is, without any adjustment for the preliminary screening.

To estimate $R^2_{n,\alpha}$ and $F_{n,\alpha}$, the following will be helpful. Let Z be standard normal and $\Phi(z) = P\{|Z| > z\}$. Analytically,

$$\Phi(z) = \sqrt{\frac{2}{\pi}} \int_z^\infty \exp\left(-\frac{1}{2}u^2\right) du.$$

Choose λ so that $\Phi(\lambda) = \alpha$. Thus, λ is the cutoff for a two-tailed z test at level α. Let

$$g(z) = \int_{\{|z|>z\}} Z^2 < 1.$$

For $0 \leqq z < \infty$, integration by parts shows

$$g(z) = \Phi(z) + \sqrt{\frac{2}{\pi}} z \exp\left(-\frac{1}{2}z^2\right). \tag{4}$$

Clearly,

$$E\{Z^2 \,|\, |Z| > z\} = g(z)/\Phi(z). \tag{5}$$

Then, as intuition demands,

$$E\{Z^2 \,|\, |Z| > z\} = 1 + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2}z^2\right) \Big/ \Phi(z) > 1. \tag{6}$$

Let Z_λ be Z conditional on $|Z| > \lambda$. Put $z = \lambda$ in (5) and recall that $\Phi(\lambda) = \alpha$:

$$g(\lambda)/\alpha = E\{Z^2 \,|\, |Z| > \lambda\} = E\{Z_\lambda^2\} > 1 \tag{7}$$

Using (6) and further integration by parts.

$$\text{var}\{Z^2 \,|\, |Z| > z\} = 2 + v(z), \tag{8}$$

where

$$v(z)^2(z) = \sqrt{\frac{2}{\pi}} w(z) \exp\left(-\frac{1}{2}z^2\right) \Big/ \Phi(z)^2 \tag{9}$$

and

$$w(z) = (z^3 + z)\Phi(z) - \sqrt{\frac{2}{\pi}} z^2 \exp\left(-\frac{1}{2}z^2\right).$$

In particular, v is continuous. Intuition suggests that v be positive. This fact will not be needed here, but it is true: see Diaconis and Freedman (1982, (3.15)–(3.16)).

Proposition. Assume (1) and (2). In probability: $q_{n,\alpha}/n \to \alpha\rho$ and $R^2_{n,\alpha} \to g(\lambda)$ and

$$F_{n,\alpha} \to \frac{g(\lambda)}{\alpha} \Big/ \frac{1 - g(\lambda)\rho}{1 - \alpha\rho}. \tag{10}$$

In the second regression, the t statistic for testing whether a coefficient vanishes is asymptotically distributed as

$$Z_\lambda \sqrt{\frac{1 - \alpha\rho}{1 - g(\lambda)\rho}}.$$

These results may be interpreted as follows. The number of variables in the first-pass regression is $p = \rho n + o(n)$; the number in the second pass is $q_{n,\alpha} = \alpha\rho n + o(n)$. That is, as may be expected, α of the variables are significant at level α. Since $g(\lambda) < 1$, the R^2 in the second-pass regression is essentially the fraction $g(\lambda)$ of R^2 in the first pass. Likewise, $g(\lambda) > \alpha$, so the asymptotic value of the F statistic exceeds 1. Since the number of degrees of freedom is growing, off-scale P values will result. Finally, the real level of the t test may differ appreciably from the nominal level.

Example. Suppose $N = 100$ and $p = 50$, so $\rho = \frac{1}{2}$; and $\alpha = 0.25$ so $\lambda \doteq 1.15$. Then $g(\lambda) \doteq 0.72$, and $E\{Z^2 \mid |Z| > \lambda\} \doteq 2.9$. In a regression with 50 explanatory variables and 100 data points, on the null hypothesis R^2 should be nearly $\frac{1}{2}$.

Next, run the regression again, keeping only the variables significant at the 25 percent level. The new R^2 should be around $g(\lambda) = 72$ percent of the original R^2. The new F statistic should be around

$$\frac{\mathcal{g}(\lambda)}{\alpha} \Big/ \frac{1-\mathcal{g}(\lambda)\rho}{1-\alpha\rho} \doteq 4.0.$$

The number of degrees of freedom should be around $\alpha\rho n \doteq 12$ in the numerator and $100 - 12 = 88$ in the denominator. (However, $q_{n,\alpha}$ is still quite variable, its standard deviation being about 3.) On this basis, a P value on the order of 10^{-4} may be anticipated.

What about the t tests? Take $\lambda' > \lambda$, corresponding to level $\alpha' < \alpha$. The nominal level for the test is α', but the real level is

$$\frac{1}{\alpha} P\left\{|Z| > \lambda' \sqrt{\frac{1-\alpha\rho}{1-\mathcal{g}(\lambda)\rho}}\right\}.$$

Since $\mathcal{g}(\lambda) > \alpha$, it follows that $1 - \alpha\rho > 1 - \mathcal{g}(\lambda)\rho$. Keep $\alpha = 0.25$, so $\lambda \doteq 1.15$; take $\alpha' = 5$ percent, so $\lambda' = 1.96$; keep $\rho = \frac{1}{2}$. Now

$$\lambda' \sqrt{\frac{1-\alpha\rho}{1-\mathcal{g}(\lambda)\rho}} = 2.3$$

and the real level is 9 percent. This concludes the example.

Turn now to the proofs. Without loss of generality, suppose the ith column of X has a 1 in the ith position and 0's everywhere else. Then

$$\hat{\beta}_i = Y_i \text{ for } i = 1, \ldots, p,$$

and the sum of squares for error in the first-pass regression corresponding to the model (1) is

$$\sum_{i=p+1}^{n} Y_i^2.$$

Thus

$$R_n^2 = \sum_{i=1}^{p} Y_i^2 \Big/ \sum_{i=1}^{n} Y_i^2$$

and

$$F_n = \frac{1}{p} \sum_{i=1}^{p} Y_i^2 \Big/ \frac{1}{n-p} \sum_{i=p+1}^{n} Y_i^2.$$

Now (3) follows from the weak law of large numbers. Of course, $E(R_n^2)$ and var R_n are known: see Kendall and Stuart (1969).

To prove (10), the t statistic for testing $\beta_i = 0$ is Y_i/s_n, where

$$s_n^2 = \frac{1}{n-p} \sum_{j=p+1}^{n} Y_j^2.$$

Thus, column i of X enters the second regression iff $|Y_i/s_n| > t_{\alpha,n-p}$, the cutoff for a two-tailed t test at level α, with $n - p$ degrees of freedom.

In what follows, suppose without loss of generality that $\sigma^2 = 1$. Given s_n, the events

$$A_i = \{|Y_i| > t_{\alpha,n-p} s_n\}$$

are conditionally independent, with common conditional probability $\Phi(t_{\alpha,n-p}s_n)$. Of course, $t_{\alpha,n-p} \to \lambda$ and $s_n \to 1$; so this conditional probability converges to $\Phi(\lambda) = \alpha$. The number $q_{n,\alpha}$ of the events A_i that occur is therefore

$$\alpha p + o(p) = \alpha \rho n + o(n)$$

by (2). This can be verified in detail by computing the conditional expectation and variance.

Next, condition on s_n and $q_{n,\alpha} = q$ and on the identity of the q columns going into the second regression. By symmetry, suppose that it is columns 1 through q of X that enter the second regression. Then

$$R_{n,\alpha}^2 = \sum_{i=1}^{q} Y_i^2 \Big/ \sum_{i=1}^{n} Y_i^2$$

and

$$F_{n,\alpha} = \frac{1}{q} \sum_{i=1}^{q} Y_i^2 \Big/ \frac{1}{n-q} \sum_{i=q+1}^{n} Y_i^2 .$$

Now $\Sigma_{i=1}^{n} Y_i^2 = n + o(n)$; and in the denominator of $F_{n,\alpha}$,

$$\sum_{i=q+1}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{q} Y_i^2 .$$

It remains only to estimate $\Sigma_{i=1}^{q} Y_i^2$, to within $o(n)$. However, these Y_i's are conditionally independent, with common conditional distribution: they are distributed as Z given $|Z| > z_n$, where Z is $N(0, 1)$ and $z_n = t_{\alpha,n-p} \cdot s_n$. In view of (5), the conditional expectation of $\Sigma_{i=1}^{q} Y_i^2$ is

$$q_{n,\alpha}\, g(z_n)/\Phi(z_n).$$

But $q_{n,\alpha} = \alpha\rho n + o(n)$ and $z_n \to \lambda$. So the last display is, up to $o(n)$,

$$\alpha\rho n g(\lambda)/\alpha = g(\lambda)\rho n.$$

Likewise, the conditional variance of $\Sigma_{i=1}^{q} Y_i^2$ is $q_{n,\alpha}\,\{2 + v(z_n)\} = O(n)$; the conditional standard deviation is $O(\sqrt{n})$. Thus

$$\sum_{i=1}^{q} Y_i^2 = g(\lambda)\rho n + o(n),$$

$$\frac{1}{q}\sum_{i=1}^{n} Y_i^2 = g(\lambda)/\alpha + o(1),$$

$$\frac{1}{n-q}\sum_{i=q+1}^{n} Y_i^2 = \frac{1 - g(\lambda)\rho}{1 - \alpha\rho} + o(1).$$

This completes the argument for the convergence in probability. The assertion about the t statistic is easy to check, using the last display.

REFERENCES

Diaconis P; Freedman D. "On the Maximum Difference Between the Empirical and Expected Histograms for Sums," *Pacific Journal of Mathematics*, 1982; **100**:287–327.

Freedman D. "Some Pitfalls in Large-Scale Econometric Models: A Case Study," *University of Chicago Journal of Business*, 1981; **54**:479–500.

Freedman D; Rothenberg T; Sutch R. "A Review of a Residential Energy End Use Model," Technical Report No. 14, University of California, Berkeley, Dept. of Statistics, 1982.

—— "On Energy Policy Models," *Journal of Business and Economic Statistics*, 1983; **1**:24–32.

Judge G; Bock M. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*, Amsterdam: North-Holland, 1978.

Kendall MG; Stuart A. *The Advanced Theory of Statistics*, London: Griffin, 1969.

Olshen RA. "The Conditional Level of the *F*-Test," *Journal of the American Statistical Association*, 1973; **68**, 692–698.

Rencher AC; Pun FC. "Inflation of R^2 in Best Subsets Regression," *Technometrics*, 1980; **22**:49–53.