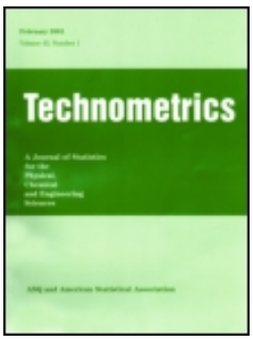

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Inflation of R^2 in Best Subset Regression

Alvin C. Rencher & Fu Ceayong Pun

To cite this article: Alvin C. Rencher & Fu Ceayong Pun (1980) Inflation of R^2 in Best Subset Regression, Technometrics, 22:1, 49-53

To link to this article: https://doi.org/10.1080/00401706.1980.10486100

Inflation of R^2 in Best Subset Regression

Alvin C. Rencher and Fu Ceayong Pun

Department of Statistics
Brigham Young University
Provo, UT 84602

When subset selection is used in regression the expected value of R^2 is substantially inflated above its value without selection, especially when the number of observations is less than the number of predictor variables. The extent of this increase was investigated by a Monte Carlo simulation. Tables are given with average values and percentage points of R^2 for the null case of independence between the response variable and the predictor variables. Approximation formulas are provided to supplement the coverage in the tables.

KEY WORDS

Regression
Subset selection
Multiple correlation coefficient
Stepwise regression
Monte Carlo

1. INTRODUCTION

Various procedures have been used in an attempt to find the "best" subset of a set of predictor variables in regression. Some of the methods which examine one variable at a time are forward selection, backward elimination and stepwise selection (see, for example, Draper and Smith, 1966, Chapter 6, or Gorman and Toman, 1966). Techniques which are equivalent to examining all possible subsets are discussed by Hocking and Leslie (1967) and Furnival and Wilson (1974). Berk (1978) compared the all subsets method with forward selection and backward elimination. Hocking (1976) reviewed several subset selection methods and discussed interpretation of results.

With any of these procedures, the usual F-statistics and R^2 are biased (see, for example, Berk, 1978, and Pope and Webster, 1972). Approximate percentage points for the distribution of the squared multiple correlation coefficient, R^2, under selection were given by Diehr and Hoflin (1974). They provided tables for $k = 5$ and $k = 10$ where k is the number of predictor variables. For other values of k a formula is given which requires an iterative solution.

In this paper, the results of Diehr and Hoflin are extended to include (1) average inflation of R^2 under selection as well as upper percentage points, (2) correlated predictor variables, and (3) the situation where the number of predictor variables exceeds the number of observations.

Consider the usual linear regression model

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times k$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an error vector. In this paper we consider only the null case, $\boldsymbol{\beta} = \mathbf{0}$. The matrix $\mathbf{X}$ can therefore consist of either fixed constants or random variables (satisfying certain assumptions) since R^2 has the same distribution either way in the null case.

This study was motivated by some rather unexpected results obtained from a set of data with $n = 26$ and $k = 54$. A stepwise regression program selected 8 predictor variables which yielded an R^2 value of .98. Considering this large R^2 value somewhat suspect because n was much smaller than k, we "scrambled" the response variables, i.e., randomly rematched the y values to different rows of the $\mathbf{X}$ matrix. The stepwise procedure again selected (a different subset of) 8 predictor variables with a resulting R^2 of .95. This gave little cause for confidence in the predictive power of the first set of 8 predictor variables selected and spurred us to find out if such large values of R^2 might be typical under selection when $n < k$ and $\boldsymbol{\beta} = \mathbf{0}$.

2. MONTE CARLO PROCEDURE

The mean and three percentage points of R^2 were obtained by Monte Carlo simulation. Data were gen-

Received February 1977; revised May 1978

erated according to (1) with $\beta = \mathbf{0}$ and $\varepsilon \sim N(\mathbf{0}, \mathbf{I})$. The McGill random number routine (Marsaglia, Ananthanarayanan and Paul, 1973) was used to obtain normally distributed random numbers. The procedure was checked by generating 3,000,000 random normal deviates. No significant departure from normality was found, with particular attention given to the tails.

For selection of a subset of predictor variables from the k available, we used the stepwise procedure described by Draper and Smith (1966, Sec. 6.4) or Efroymson (1960). The method is a modification of forward selection in which at each step the possibility of deleting a variable is considered. The variable which enters at each step is the one with the largest single degree of freedom F-ratio, provided this F-ratio exceeds the designated value for entry, F_{in}. After a variable has entered, the variables which entered previously are reexamined and the one with the smallest single degree of freedom F-ratio is deleted if this F-ratio is less than a specified value, F_{out}.

The stepwise procedure was altered slightly so that we could specify the number, p, of variables selected. If p variables have not entered before the F-value falls below F_{in} then both F_{in} and F_{out} are reduced as necessary to select p variables. The value of R^2 obtained by selection of p variables in this manner was denoted R_s^2. In each replication an R^2 value denoted R_o^2, was also obtained from p variables chosen randomly from the k available predictor variables.

After a predetermined number of replications, the average values of R_s^2 and R_o^2 were computed (denoted $\overline{R_s^2}$ and $\overline{R_o^2}$), along with the standard error of each and three percentage points of R_s^2. The number of replications varied from 500 to 2000. The standard error of $\overline{R_s^2}$ ranged from .00095 to .0089. We recognize that sometimes the stepwise technique will fail to find the subset with maximum R^2. Therefore, this procedure gives a conservative estimate of the average increase in R^2 due to selection.

Two covariance patterns were considered for the predictor variables. Most of the simulation involved the uncorrelated case, i.e., $\mathbf{X'X}$ diagonal. The correlated case was also examined where $\mathbf{X'X}$ was nondiagonal, ranging from a low to high degree of multicollinearity.

3. RESULTS OF THE SIMULATION

The expected value of R_o^2 when $\beta = \mathbf{0}$ is $p/(n-1)$ (see Kendall and Stewart, 1961, p. 341) and in all cases $\overline{R_o^2}$ was very close to that value. Accordingly we have tabulated $E(R_o^2) = p/(n-1)$ instead of $\overline{R_o^2}$ for comparison purposes. Values of $E(R_o^2)$, $\overline{R_s^2}$, and the 95th percentile point of R_s^2, denoted $R_{.95}^2$, are given in Table 1 for $n = 5,10(10)60$, $k = 5,10(10)40$ and $p = 2(2)10$. The 90th and 99th percentile points were also obtained in the simulation but are omitted from Table 1. Approximate values for other desired percentiles can be found using (12) in Section 4. The results in Table 1 are for independent predictor variables, i.e, $\mathbf{X'X}$ diagonal.

In Table 1 we see that $\overline{R_s^2}$ is always greater than $E(R_o^2)$. In some cases with large k, $\overline{R_s^2}$ is four or fives times as large as $E(R_o^2)$. Among the cases with $k > n$, 86% of the $R_{.95}^2$ values exceeded .75 and 66% exceeded .90. Even the $\overline{R_s^2}$ values tended to be large for $k > n$; 72% of these values exceeded .75 and 48% exceeded .90.

As expected, the inflation of R^2 is somewhat less with correlated predictors. We indicate the amount

TABLE 1—*Expected value of R^2 without selection compared with average and 95th percentile of R^2 with selection. The regressor variables are uncorrelated.*

		p = 2			p = 4			p = 6			p = 8			p = 10		
n	k	$E(R_o^2)$	$\bar{R}_s^2$	$R^2_{.95}$	$E(R_o^2)$	$\bar{R}_s^2$	$R^2_{.95}$	$E(R_o^2)$	$\bar{R}_s^2$	$R^2_{.95}$	$E(R_o^2)$	$\bar{R}_s^2$	$R^2_{.95}$	$E(R_o^2)$	$\bar{R}_s^2$	$R^2_{.95}$
5	5	.500	.784	.981												
	10	.500	.900	.991												
	20	.500	.952	.995												
10	5	.222	.421	.665	.444	.540	.851									
	10	.222	.567	.822	.444	.778	.955	.667	.894	.991						
	20	.222	.691	.877	.444	.912	.983	.667	.984	.999	.889	.997	1.000			
	30	.222	.751	.895	.444	.953	.989	.667	.994	.999	.889	.999	1.000			
	40	.222	.791	.907	.444	.965	.991	.667	.996	1.000	.889	.999	1.000			
20	10	.105	.299	.510	.211	.423	.658	.316	.488	.726						
	20	.105	.391	.562	.211	.585	.771	.316	.701	.858	.421	.786	.920			
	30	.105	.437	.605	.211	.654	.810	.316	.790	.905	.421	.876	.957	.526	.933	.984
	40	.105	.469	.618	.211	.700	.819	.316	.835	.920	.421	.916	.969	.526	.963	.989
30	10	.069	.202	.337	.138	.285	.451	.207	.326	.511						
	20	.069	.264	.388	.138	.405	.546	.207	.496	.648	.276	.561	.716			
	30	.069	.302	.439	.138	.472	.619	.207	.585	.730	.276	.669	.798	.345	.733	.850
	40	.069	.331	.456	.138	.514	.640	.207	.639	.759	.276	.733	.846	.345	.803	.901
40	10	.051	.147	.260	.103	.206	.339	.154	.233	.374						
	20	.051	.203	.319	.103	.309	.445	.154	.376	.527	.205	.424	.583			
	30	.051	.230	.336	.103	.361	.481	.154	.453	.592	.205	.523	.660	.256	.579	.713
	40	.051	.251	.349	.103	.397	.507	.154	.502	.626	.205	.584	.704	.256	.651	.775
50	30	.041	.184	.267	.082	.291	.397	.122	.367	.484	.163	.424	.559	.204	.469	.612
	40	.041	.201	.288	.082	.324	.432	.122	.413	.526	.163	.481	.598	.204	.537	.657
60	30	.034	.157	.229	.068	.247	.341	.102	.312	.425	.136	.360	.479	.169	.398	.523
	40	.034	.169	.249	.068	.272	.372	.102	.348	.457	.136	.406	.515	.169	.454	.569

of intercorrelation among the predictors by $(\Sigma 1/\lambda_i)/k$ where $\lambda_1, \lambda_2, \cdots, \lambda_k$ are the eigenvalues of the correlation matrix of the predictor variables. Four values of $(\Sigma 1/\lambda_i)/k$ were used for each of $k = 4, 8, 16$. The results for these four values are compared with the uncorrelated case in Table 2. Note that $(\Sigma 1/\lambda_i)/k = 1$ for the uncorrelated case.

4. APPROXIMATION FORMULAS

Zirphile (1975) used extreme value theory to obtain an asymptotic distribution of R_s^2. However, his results are obscured by several misprints, e.g., $\beta = -\ln[-\ln(\alpha)]$ should be $\beta = -\ln[-\ln(1-\alpha)]$. With appropriate corrections we found that only for large n did his formula give satisfactory results. In fact Zirphile suggested that n should be greater than 50. For some of the smaller values of n in (our) Table 1, his formula produced $R_{.95}^2$ values as high as 1.5.

The poor performance of Zirphile's formula for small n may be due in part to his use of the assumption that in the null case the distribution of R^2 is asymptotically normal. However, R^2 has a beta distribution in the null case (see Kendall and Stewart, 1961, p. 338) and following Gumbel (1958) we now obtain an asymptotic distribution, $G(R_s^2)$, based on the beta.

The distribution function of R^2 is given by the incomplete beta function

$$F(R^2) = \frac{1}{B(a,b)} \int_0^{R^2} (R^2)^{a-1}(1-R^2)^{b-1} dR^2 \qquad (2)$$

where $B(a, b)$ is the beta function with $a = p/2$ and $b = (n-p-1)/2$. Let $N = \binom{k}{p}$, the number of possible subsets of p predictor variables among the k available. The mode, u, of $dG(R_s^2)$ is approximately given by

$$F(u) = 1 - 1/N. \qquad (3)$$

If the N values of R^2 obtained from all possible subsets were independent the distribution function of the maximum value, R_s^2, would be $[F(R_s^2)]^N$. Using (3) this can be expressed as

$$[F(R_s^2)]^N = \left[1 - \frac{1 - F(R_s^2)}{N(1-F(u))}\right]^N. \qquad (4)$$

As N increases we obtain the extreme value distribution

$$G(R_s^2) = \exp\left[-\frac{1-F(R_s^2)}{(1-F(u))}\right]. \qquad (5)$$

To find percentile points, R_γ^2, where $P(R_s^2 \le R_\gamma^2) = \gamma$, we set $\gamma = G(R_\gamma^2)$ and solve for R_γ^2 to obtain

$$R_\gamma^2 = F^{-1}[1 + (1 - F(u)) \ln\gamma].$$

Using (3) this can be written

$$R_\gamma^2 = F^{-1}\left[1 + \frac{\ln\gamma}{N}\right], \qquad (6)$$

where F^{-1} is the upper percentage point of the beta distribution, i.e., the inverse of (2) which can be found using available tables or computer subroutines.

We did not succeed in obtaining an expression for $E(R_s^2)$ by integration using (5) and worked instead with Gumbel's third asymptote (1958, p. 275) for a limited distribution,

$$G(x) = \exp\left[-\left(\frac{1-x}{1-u}\right)^w\right], \qquad (7)$$

where u is defined as at (3) and w can be defined in our case as the order of the smallest derivative of $F(x)$ which does not vanish when evaluated at $x = 1$. It can easily be shown that for (2)

$$w = b = \frac{n-p-1}{2}. \qquad (8)$$

TABLE 2—*Expected value of R^2 without selection compared with average and 95th percentile of R^2 with selection for various levels of intercorrelation among the regressor variables.*

				$(\Sigma 1/\lambda_i)/k$									
				1.0 uncorrelated		1.25		12.5		125		1250	
n	k	p	$E(R_0^2)$	$\bar{R}_s^2$	$R_{.95}^2$	$\bar{R}_s^2$	$R_{.95}^2$	$\bar{R}_s^2$	$R_{.95}^2$	$\bar{R}_s^2$	$R_{.95}^2$	$\bar{R}_s^2$	$R_{.95}^2$
5	4	2	.500	.723	.969	.703	.967	.660	.969	.648	.963	.612	.950
10	8	2	.222	.532	.780	.492	.792	.471	.791	.429	.731	.401	.752
		4	.444	.720	.926	.697	.933	.677	.918	.620	.891	.540	.863
20	16	2	.105	.359	.528	.361	.552	.319	.538	.290	.507	.277	.489
		4	.211	.528	.706	.526	.716	.467	.689	.426	.663	.392	.632
		6	.316	.630	.803	.632	.830	.562	.772	.524	.751	.482	.716
		8	.421	.705	.880	.704	.888	.638	.837	.600	.814	.557	.790

To find $E(x)$ we have

$$E(x) = \int_0^1 x\, dG(x)$$

$$= \int_0^1 x\, d \exp\left[-\left(\frac{1-x}{1-u}\right)^w\right].$$

With the transformation $y = [(1 - x)/(1 - u)]^w$ this becomes

$$E(x) = (1-u)\int_0^{1/(1-u)^w} y^{1/w} d \exp(-y) - \int_0^{1/(1-u)^w} d \exp(-y).$$

As N increases $1/(1 - u)^w \to \infty$ and

$$E(x) = 1 - (1 - u)\Gamma(1 + 1/w). \tag{9}$$

By (3) $u = F^{-1}(1 - 1/N)$ and we have

$$1 - u \doteq F_{b,a}^{-1}(1/N) \tag{10}$$

where $F_{b,a}^{-1}$ denotes the inverse of (2) with the parameters a and b reversed. From (9) and (10) we obtain an approximate formula for $E(R_s^2)$:

$$E(R_s^2) \doteq 1 - F_{b,a}^{-1}(1/N)\Gamma(1 + 1/w). \tag{11}$$

We note that both (6) and (11) produce values constrained to lie between 0 and 1. However, both showed an upward bias when applied to the values in Table 1. The use of stepwise regression instead of all possible subsets may account for a small part of this bias but it is due mostly to the use of the assumption of $N = \binom{k}{p}$ independent observations from the distribution of R^2. In reality, of course, the N possible values of R^2 all arise from the same y vector and are not independent. The maximum tends to be lower and the "effective" value of N is less than $\binom{k}{p}$.

To correct this upward bias we found the adjusted value of N required for (6) and (11) to reproduce each value in Table 1 and then sought a function of N that would yield these adjusted values. The function $(\ln N)^{cN^d}$ was found to fit well. The values of c and d were chosen to allow a very small upward bias to compensate for stepwise regression. The final values used for (6) were $c = 1.8$ and $d = .04$. For (11) we used $c = 1.5$ and $d = .04$. The resulting approximation formulas were therefore

$$\hat{R}_\gamma^2 = F^{-1}[1 + \ln\gamma/(\ln N)^{1.8N^{.04}}] \tag{12}$$

and

$$\hat{E}(R_s^2) = 1 - F_{b,a}^{-1}(1/(\ln N)^{1.5N^{.04}})\,\Gamma(1 + 1/w), \tag{13}$$

assuming $w \neq 0$.

Approximations (12) and (13) were evaluated for each of the 91 combinations of n, k, p in Table 1. The average difference between $R_{.95}^2$ in Table 1 and $\hat{R}_{.95}^2$ from (12) was .0054. The standard deviation of $\hat{R}_{.95}^2 - R_{.95}^2$ was .025. For $\hat{E}(R_s^2) - \overline{R_s^2}$ the average of 91 values was .011 with a standard deviation of .046.

Based on this reasonably good fit to the values in Table 1 we can recommend use of (12) and (13) when interpolating for values bracketed by the n, k, and p in Table 1. Extrapolation beyond may be risky and needs further investigation.

5. EXAMPLE

Blumenthal, Trout and Chang (1976) used stepwise regression to relate four organoleptic properties of oils to gas chromatography peaks. For each response variable there were 6 observations and 24 predictor variables. With 2 variables selected, R^2 values of .9158, .9882, .9679 and .9898 were obtained for the 4 (correlated) response variables. The combination $n = 6$, $k = 24$, $p = 2$ is not covered in Table 1, but when these values were submitted to the simulation program with 500 replications, the result was $\overline{R_s^2} = .931$ with 90th, 95th, and 99th percentiles of .982, .986, and .996. Formulas (12) and (13) yielded $\hat{E}(R_s^2) = .896$ and $\hat{R}_{.95}^2 = .990$. These results would seem to support the authors' caution that "despite the observed good correlation, it is possible that the results could be accounted for by chance, because the number of samples and the number of possible combinations of data do not eliminate random or chance correlation."

6. SUMMARY

The shift in the distribution of R^2 under selection was examined by Monte Carlo methods. The stepwise selection procedure was used and investigation was limited to the null case where the response variable was independent of the predictor variables. A wide range of values was chosen for the number of observations and number of predictor variables with particular allowance made for the case where the number of predictor variables exceeds the number of observations.

For uncorrelated predictor variables (**X'X** diagonal) large increases were seen in $\overline{R_s^2}$, the average value of R^2 under selection. In cases with more predictor variables than observations, nearly half of the values of $\overline{R_s^2}$ were greater than .90. In such cases ($k \geq n$) the results of stepwise regression may be of little value unless substantiated by another sample or other information independent of the data.

Approximation formulas for the average value of R_s^2 and percentage points were provided to supplement the Monte Carlo values. The tables or formulas can be used as guidelines for assessing the significance of R^2 values obtained in best subset regression applications.

The case where the predictor variables were intercorrelated (**X'X** nondiagonal) was also investigated

and as expected the inflation in R^2 was somewhat less. The amount of reduction in average R^2 and percentage points was shown for various levels of multicollinearity.

7. ACKNOWLEDGEMENT

The authors wish to thank the referees for helpful suggestions which led to significant improvements in the paper.

REFERENCES

BERK, K. N. (1978). Comparing subset regression procedures. *Technometrics, 20,* 1–6.

BLUMENTHAL, M. M., TROUT, J. R., and CHANG, S. S. (1976). Correlation of gas chromatographic profiles and organoleptic scores of different fats and oils after simulated deep fat frying. *Journal of the American Oil Chemists' Society, 53,* 496–501.

DIEHR, G. and HOFLIN, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics, 16,* 317–321.

DRAPER, N. R. and SMITH, H. (1966). *Applied Regression Analysis.* New York: John Wiley & Sons.

EFROYMSON, M. A. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers,* ed. by A. Ralston and H. S. Wilf, pp. 191–203. New York: John Wiley & Sons.

FURNIVAL, G. M. and WILSON, R. W., JR. (1974). Regressions by leaps and bounds. *Technometrics, 8,* 27–51.

GORMAN, J. W. and TOMAN, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics, 16,* 499–511.

GUMBEL, E. J. (1958). *Statistics of Extremes.* New York: Columbia University Press.

HOCKING, R. R. and LESLIE, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics, 9,* 531–540.

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

KENDALL, M. G. and STEWART, A. (1961). *The Advanced Theory of Statistics, Vol. 2.* New York: Hafner Publishing Co.

MARSAGLIA, G., ANANTHANARAYANAN, K. and PAUL, N. (1973). Super-duper random number package. School of Computer Science, McGill University, Montreal, Quebec.

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

ZIRPHILE, J. (1975). Letter to the Editor. *Technometrics, 17,* 145.