


Journal of Educational Statistics
Summer 1982, *Volume* 7, *Number* 2, *pp. 91–104*

SAMPLE SIZE AND THE ACCURACY OF PREDICTIONS MADE FROM MULTIPLE REGRESSION EQUATIONS

RICHARD SAWYER
The American College Testing Program

Key words: *Prediction*; *multiple regression*; *sample size*

ABSTRACT. Some rules of thumb are given for estimating the accuracy of predictions based on a multiple regression equation developed from a random sample of a multivariate normal population. The distribution of the prediction error in this case can be approximated usefully by a normal distribution. Formulas are given for the moments of the distribution and for other parameters such as the mean absolute error (MAE). The approximate inflation in MAE (over its asymptotic value) due to estimating the regression coefficients is a simple function of the base sample size and the number of predictors.

The coefficients in a multiple regression prediction equation must, in practice, be estimated from a base sample taken before the predictions are made. Because the accuracy in estimating these coefficients depends on the size of the base sample, and because error in estimating the coefficients propagates error in prediction, the base sample size affects prediction accuracy. In this paper, we consider the question, "How large should the base sample be to achieve a given level of accuracy in a future prediction?" We shall show that when a prediction equation is based on a sample from a multivariate normal population, the mean absolute error of prediction, relative to its minimum asymptotic value, can be closely approximated by a simple function of the number of predictor variables and the base sample size.

The accuracy of a prediction from a regression equation traditionally has been formulated in terms of the conditional error variance, given the predictor data in the base sample and the particular values of the predictor variables for which a prediction is to be made. This particular formulation of prediction accuracy is applicable to problems in which both of the following occur:

(a) either the base sample data have been collected or else one can control the values of the predictor data in the base sample; and
(b) one is interested in the accuracy of predictions at particular values of the predictor variables. (This could occur, for example, when one can select values of the predictor variables at which to make predictions.)

Derivations of this conditional variance can be found in the textbooks by Searle (1971, pp. 92–93), Draper and Smith (1966, pp. 58–61), and Graybill (1961, pp. 110–112).

In this paper we seek to answer the question posed above in the following context:

(a) one is interested in specifying prediction accuracy *before* collecting the base sample data;
(b) the predictors are not subject to experimental control;
(c) one is interested in specifying prediction accuracy averaged with respect to the distribution of the predictors. This would occur when instead of specifying prediction accuracy at the many possible individual values of the predictors, one wants a single, overall measure of prediction accuracy.

We therefore are working with a "random" regression model in which one considers the joint distribution of the predictors and the criterion variable in the base sample and for the future predictions. This sample size problem occurs in predicting college freshman grades from standardized test scores and high school grades. An example of this problem is discussed later in the paper.

Most results to date in this area pertain to the use of correlation coefficients as the overall measures of prediction accuracy mentioned in (c) above. Park and Dudycha (1974) related base sample size to $\text{Prob}[\rho^2 - \rho^2(\hat{\boldsymbol{\beta}}) \leqslant \varepsilon]$, where ρ is the population multiple correlation between the dependent variable and the predictors, and $\rho(\hat{\boldsymbol{\beta}})$ is the population correlation between the dependent variable and the predictions based on the coefficients $\hat{\boldsymbol{\beta}}$. Claudy (1972) generated multivariate normal populations with various correlation structures and compared the correlations in these populations of predictions based on samples of various sizes. Halinski and Feldt (1970) recommended a minimum subject-to-variable ratio of 10 to 1 on the basis of a simulation study. Miller and Kunce (1973), studying cross-validated correlations from vocational rehabilitation data, recommended the same subject-to-variable ratio.

Browne (1975) considered an alternative overall measure of prediction accuracy, the mean squared error (MSE) of prediction. Building on results obtained by Stein (1960), Kerridge (1967), and Darlington (1968), Browne formulated the precision of two methods for estimating MSE. He also presented a test of the hypothesis that using a prespecified subset of predictors would not increase MSE.

We consider an alternative measure of prediction accuracy, the mean absolute error prediction (MAE), i.e., the mean deviation of the distribution of prediction error. An advantage of MAE is that because of its simple definition its meaning is easily understood by people who use prediction equations, but who have received little statistical training. Several authors (e.g., Cramér, 1945, p. 181; Kendall & Stuart, 1969, pp. 42–43; Krutchkoff, 1970, p. 178; Mosteller, Rourke, & Thomas, 1970, pp. 206–207) acknowledge the intuitive appeal of a

measure like MAE, but cite its mathematical intractibility. We shall show that in the context of sampling from a multivariate normal population, the distribution of prediction error is approximately normal; hence, MAE is approximately $\sqrt{2/\pi}$ times the more commonly used root mean squared error. The results concerning MAE, relative to its minimum asymptotic value, therefore apply as well to the root mean squared error, relative to its minimum asymptotic value.

The Model

The regression coefficients are assumed to be estimated from a random sample $(y_i, \mathbf{x}_i')$, $(i = 1,\ldots,n)$, where y_i is the dependent variable and $\mathbf{x}_i$ is a vector of p predictor variables for the i-th case. Each $\mathbf{x}_i$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The conditional distribution of y_i given $\mathbf{x}_i$ is normal with mean $(1, \mathbf{x}_i')\boldsymbol{\beta}$ and variance σ^2. The regression coefficients $\boldsymbol{\beta}$ are estimated by the usual least squares estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}X'y \text{ where } X = \begin{bmatrix} 1 & \mathbf{x}_1' \\ \vdots & \vdots \\ 1 & \mathbf{x}_n' \end{bmatrix} \text{ and } \mathbf{y}' = (y_1,\ldots,y_n).$$

An additional independent observation $(y^*, \mathbf{x}^{*\prime})$ is to be taken and y^* is to be predicted by $\hat{y} = (1, \mathbf{x}^{*\prime}) \cdot \hat{\boldsymbol{\beta}}$.

The accuracy of the prediction $\hat{y}$ can be formulated in terms of the prediction error $\hat{y} - y^*$. Browne (1975) investigated $E[(\hat{y} - y^*)^2]$, the error variance or mean squared error (MSE). We shall focus on $E[|\hat{y} - y^*|]$, the mean absolute error of prediction (MAE) and on $P(t) = \text{Prob}[|\hat{y} - y^*| \leqslant t]$, expected proportion of future predictions within a distance t of the actual observations. All probabilities and expectations are with respect to the joint distribution of $(y^*, \mathbf{x}^{*\prime})$, and $(y_i, \mathbf{x}_i')$, $i = 1,\ldots,n$.

The development of a simple approximation to the relationship between prediction accuracy, n, and p depends on the following fact, proved by Kerridge (1967):

Lemma: The conditional distribution of $\hat{y} - y^*$, given $\mathbf{x}_1,\ldots,\mathbf{x}_n, \mathbf{x}^*$, is normal with mean 0 and variance $\sigma^2(1 + 1/n)[1 + (p/(n-p)) \cdot F(\mathbf{x}_1,\ldots,\mathbf{x}_n, \mathbf{x}^*)]$. As a random variable in $\mathbf{x}_1,\ldots,\mathbf{x}_n, \mathbf{x}^*$, F has an $F(p, n-p)$ distribution.

The unconditional distribution of $\hat{y} - y^*$ is, of course, asymptotically normal with mean 0 and variance σ^2. Although this distribution is much more complicated for finite n, we can approximate it through its moments.

Proposition: Let M be a positive integer. Then

$$E\left[(\hat{y} - y^*)^{2M-1}\right] = 0$$

when $2M \leqslant n - p$;

$$E\left[(\hat{y}-y^*)^{2M}\right]=\frac{\sigma^{2M}\dfrac{(2M)!}{M!}\left(\dfrac{n+1}{2n}\right)^M\displaystyle\prod_{j=1}^{M}(n-2j)}{\displaystyle\prod_{j=1}^{M}(n-p-2j)}$$

when $2M \leqslant n-p-1$.

Proof: The odd moments in the unconditional distribution of $\hat{y}-y^*$ are 0 because they are 0 in every conditional distribution, given $\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}^*$. As for the even moments,

$$\begin{aligned}E\left[(\hat{y}-y^*)^{2M}\right]&=E\left\{E\left[(\hat{y}-y^*)^{2M}\mid \mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}^*\right]\right\}\\&=\frac{\sigma^{2M}(2M)!}{M!}\left(\frac{n+1}{2n}\right)^M E\left\{\left[1+(p/(n-p))F\right]^m\right\}.\end{aligned}$$

Now the latter expected value is equal to

$$\sum_{j=0}^{M}\binom{M}{j}[p/(n-p)]^j E[F^j]=\sum_{j=0}^{M}\binom{M}{j}\frac{\Gamma[p/2+j]\Gamma[(n-p)/2-j]}{\Gamma[p/2]\Gamma[(n-p)/2]},$$

where $E[F^j]$ is as according to Kendall and Stuart (1969, p. 393). The summation can be simplified to

$$1+\sum_{j=1}^{M}\binom{M}{j}\frac{\prod_{k=1}^{j}[p+2(k-1)]}{\prod_{k=1}^{j}[n-p-2k]},$$

and this latter sum is equal to $\prod_{j=1}^{M}(n-2j)/\prod_{j=1}^{M}(n-p-2j)$. (This can be seen by expressing the identity $(1-t)^{-a}(1-t)^{-b}=(1-t)^{-a-b}$ in terms of power series. I wish to thank Professor Richard Askey for showing this.)

It follows immediately that $\mathrm{MSE}=\sigma^2(n+1)(n-2)/[n(n-p-2)]$. The kurtosis of $\hat{y}-y^*$ is $3\ (n-p-2)(n-4)/[(n-2)(n-p-4)]$, which is greater than 3 for every value of n and p for which $n-p\geqslant 5$.

In general, the $2M$-th moment of $\hat{y}-y^*$ is equal to the corresponding moment of a normal distribution multiplied by an inflation factor which depends on M, n, and p. The inflation factor is an increasing function of M and p, and a decreasing function of n.

The exact distribution of $\hat{y}-y^*$ is quite complicated but it can be approximated by a Gram-Charlier series (Cramér, 1945, pp. 227–229). For $n-p\geqslant 5$, the approximation using the first two nonzero terms is

$$\mathrm{Prob}[\hat{y}-y^*\leqslant t]\doteq\Phi(t/\sigma')+p/[4(n-2)(n-p-4)]\Phi^{(4)}(t/\sigma'),\quad(1)$$

where Φ is the standard normal distribution function, $\Phi^{(4)}$ is its fourth derivative, and $\sigma'=\sqrt{\mathrm{MSE}}$ is the root mean squared error. This is only an approximation; since for fixed n, $\hat{y}-y^*$ has only finitely many moments,

there is no possibility of making a convergent approximation by indefinitely adding more terms. Indeed, adding more terms may not even improve the approximation. The accuracy of these approximations is discussed in the Appendix.

By using the approximation (1), one has the following result:

Proposition: Under the assumptions stated previously,

$$\text{MAE} \doteq \sigma'\sqrt{2/\pi} \tag{2a}$$

$$\text{MAE} \doteq \sigma'\sqrt{2/\pi} - \sigma' \cdot \sqrt{2/\pi} \cdot p/[4(n-2)(n-p-4)] \tag{2b}$$

$$P(t) \doteq 2\Phi(t/\sigma') - 1 \tag{3a}$$

$$P(t) \doteq 2\Phi(t/\sigma') - 1 + \left[3(t/\sigma') - (t/\sigma')^3\right] \cdot \phi(t/\sigma') \cdot p/[2(n-2)(n-p-4)], \tag{3b}$$

where σ', as above, is the root mean squared error, and Φ and ϕ are the normal distribution and density functions, respectively. The approximations (2a) and (3a) are those resulting from using only the first term in (1), i.e., from using the normal approximation. Approximations (2b) and (3b) result from using both terms in (1).

Discussion

Under the normal approximation to the distribution of $\hat{y} - y^*$, MAE is equal to the product of $\sigma\sqrt{2/\pi}$ (its asymptotic value as the base sample size $n \to \infty$) and an inflation factor $K = \sqrt{(n+1)(n-2)/[n(n-p-2)]} > 1$ due to estimating the coefficients. The inflation factor is a function of n and the number of predictors p; on solving for n in terms of p and K, we find that

$$n = \frac{K^2(p+2) - 1 + \sqrt{\left[(p+2)K^2 - 1\right]^2 - 8(K^2-1)}}{2(K^2-1)}. \tag{4}$$

Although n is not a linear function of p, it can be well approximated by

$$n \doteq \frac{2K^2 - 1}{K^2 - 1} + \frac{K^2}{K^2 - 1}p. \tag{5}$$

(I wish to thank a referee for pointing out this approximation.) The coefficients in (5) are displayed for several values of K in Table I. The absolute differences between the sample sizes calculated from Table I and those calculated from (4) are less than 2 for $1 \leqslant p \leqslant 20$ and for every value of K listed in Table I. These aproximations can be used as rules-of-thumb like the subject-to-variable ratios cited earlier. The subject-to-variable ratio of 10 to 1 recommended by Halinski and Feldt (1970) and by Miller and Kunce (1973), for example, would result in an inflation in MAE of approximately 5 to 10 percent above that which would

TABLE I

Approximate Relationship Between Number of Predictors and Sample Size Required for Varying Degrees of Prediction Accuracy

Inflation Factor (K)	Approximate Required Sample Size[a]
1.01	$n = 50.8p + 51.8$
1.05	$n = 10.8p + 11.8$
1.10	$n = 5.8p + 6.8$
1.25	$n = 2.8p + 3.8$
1.50	$n = 1.8p + 2.8$

[a]Approximate base sample size (n) needed to achieve a MAE = $K \cdot \sqrt{2/\pi}$ with $1 \leq p \leq 20$ predictors.

occur if the regression coefficients were known. Because of the approximation (2a), the above interpretation of the relationship between K, n, and p applies as well to the root mean squared error σ'.

On replacing σ' in Formulas (2a) and (3a) with an appropriate estimator, one can estimate MAE and $P(t)$ for predictions based on a set of estimated regression coefficients. For example, an unbiased estimator of σ' is

$$K \cdot C(n, p) \cdot \text{SEE},$$

where $C(n, p) = \sqrt{(n - p - 1)/2} * \Gamma[(n - p - 1)/2]/\Gamma[(n - p)/2]$, and SEE is the standard error of estimate associated with the base sample regression equation. (The fact that $K \cdot C(n, p) \cdot \text{SEE}$ is an unbiased estimator of σ' can be seen by evaluating $E[\text{SEE}] = (\sigma/\sqrt{n - p - 1}) \cdot \int_0^\infty x^{1/2} f(x)\, dx$, where f is the density function for a chi-square random variable with $n - p - 1$ degrees of freedom.) Therefore, Formulas (2a) and (3a) permit one to interpret SEE in a manner analogous to Browne's use of the residual variance to estimate MSE.

The above approximation is based on the assumption of sampling from a multivariate normal population and on the normal approximation of the distribution of $\hat{y} - y^*$. Departures from these assumptions could result in misleading estimates of prediction accuracy. The following example suggests, however, that in at least one widespread use of multiple regression prediction equations in education, departures from the assumptions do not typically result in misleading estimates.

It should be noted that while the following example is concerned with measuring the accuracy of estimates of mean absolute error in a realistic setting, it does not seek to document the accuracy of the assumptions of the model. This model, like most statistical models, is only an approximation, not an exact description. The intent here is to document the model's usefulness rather than investigate the myriad ways in which its assumptions could be violated.

Example

The American College Testing Program (ACT) offers research services to colleges for measuring the local predictive validity of the ACT Assessment college entrance examination. These predictive research services summarize the relationships between ACT scores, high school grades, and grades of college freshmen. They also are used to generate regression coefficients for predicting the college grades of future applicants.

The sizes of the freshmen classes of the colleges that use ACT data vary from under 100 to over 5,000. A natural concern of the smallest colleges is that their size not severely affect the accuracy of their grade predictions. Moreover, although these colleges can set minimum standards for their students' test scores and high school grades, they cannot experimentally control the values of these predictor variables among their freshmen. Moreover, the values of the predictor variables can, depending on the college, be distributed over a fairly wide range. Finally, users of these prediction equations generally prefer a single measure summarizing the prediction accuracy for an entire freshman class. It is appropriate, therefore, to study this problem in the context of a random regression model.

The data base for this example consists of student records submitted by institutions through their participation in ACT's predictive research services. Each record contains a freshman overall grade average (reported on a 0- to 4-point grade scale), four ACT scores, and four student-reported high school grades in English, mathematics, social studies, and natural sciences. (For more detailed descriptive and technical information about ACT test scores, see the *Technical Report for the ACT Assessment Program*, 1973. Further information about the self-reported high school grades can be found in the *Technical Report* and in Maxey and Ormsby, 1971.)

The 605 colleges in the data base submitted about 250,000 student records for the years 1974–75 and 1976–77. From these 605 colleges, a stratified random sample of 205 colleges was selected. The prediction equations and cross-validation results below are based on the 1974–75 and 1976–77 student records submitted by these 205 colleges. (For further details on the sampling, refer to Sawyer & Maxey, 1979.)

A local prediction equation was calculated from the 1974–75 data submitted by each college in the sample. The predictors were the four ACT scores and the four high school grades, the criterion variable was college freshman overall grade average, and the form of the prediction equations was a standard eight-variable multiple linear regression. From SEE, the standard error of estimate associated with each college's prediction equation, a predicted MAE, P(.20), and P(.50) were computed by substituting an unbiased estimate of σ' in Formulas (2a) and (3a).

The prediction equation for each college was cross-validated on the 1976–77 data. Within each college, the average of the absolute errors (observed MAE) and the proportion of absolute errors less than .20 and .50 grade units (observed P(.20) and P(.50), respectively) were computed.

The following example illustrates these calculations for a particular college. Formula (5) indicates that if the expected mean absolute error of prediction is to be $K = 1.05$ times that associated with a prediction equation based on known population values of the eight regression coefficients, then the base sample size n should equal approximately 98. College X has approximately this number (101) of freshmen during 1974–75. At this college, the standard error of estimate associated with the regression of freshman grade average on the eight predictors was SEE = .68 grade units. The estimated mean absolute error associated with using this prediction equation for a future freshman class is therefore approximately $\sqrt{2/\pi} \cdot 1.05 \cdot \sqrt{45} \cdot (\Gamma(45)/\Gamma(45.5)) \cdot .68 = .56$ grade units. In fact, a mean absolute error of .56 grade units was observed in predicting 1976–77 freshman grades at this particular college. Of course, the predicted and observed measures of prediction accuracy are not always this close; the differences between predicted and observed measures of prediction accuracy among the 205 colleges in the sample are summarized at the end of this example.

The computations described above also were done separately for the males and females in each college. This was done to permit studying prediction accuracy with base samples that were smaller than those associated with the total group equations. The separate-sex equations were also calculated to determine whether the results would differ substantially for these subpopulations of students.

Because the records of some students did not have a sex indicator, the separate-sex prediction equations and their associated cross-validation statistics could not be computed from the records of all students of each sex. Moreover, separate sex equations were not computed from colleges with fewer than 25 records for a sex. Therefore, the results for the separate-sex predictions are based on fewer than all possible records in the entire sample. The distribution of base sample size among colleges is summarized in Table II.

Table III shows the quartiles of observed MAE, P(.20), and P(.50) among the colleges for the three sets of predictions. Note that according to all three measures, the separate-sex predictions for males were somewhat less accurate than the separate-sex predictions for females.

Table IV shows the quartiles of the absolute difference between predicted and observed MAE, P(.20), and P(.50) among the colleges in the sample. The median absolute differences for MAE, P(.20), and P(.50) were .04, .02, and .04, respectively. These values indicate that in most colleges prediction accuracy can be usefully estimated from Formulas (2a) and (3a).

TABLE II

Distribution of Base Sample Sizes for College Prediction Equations by Subgroup

Range in Base Sample Size	Subgroup		
	Total Group	Males	Females
25 – 50	.00	.17	.15
51 – 75	.00	.18	.16
76 – 100	.07	.07	.10
101 – 200	.37	.22	.24
201 – 500	.24	.21	.21
501 – 1000	.17	.11	.10
1000 +	.14	.03	.04
Minimum	99	26	25
Median	244	127	124
Maximum	3,804	1,430	1,691
Number of colleges	205	179	189
Number of student records	108,118	43,560	49,165

TABLE III

Quartiles of Observed Measures of Prediction Accuracy Among Colleges

Measure	Quartile	Subgroup		
		Total Group	Males	Females
MAE	Q_1	.46	.50	.44
	Q_2	.52	.58	.51
	Q_3	.58	.63	.57
P(.20)	Q_1	.22	.19	.22
	Q_2	.24	.23	.25
	Q_3	.28	.26	.29
P(.50)	Q_1	.52	.47	.52
	Q_2	.57	.53	.58
	Q_3	.63	.59	.64

TABLE IV

Quartiles of the Absolute Difference Between Predicted and Observed Measures of Prediction Accuracy Among Colleges

Measure	Quartile	Subgroup		
		Total Group	Males	Females
MAE	Q_1	.02	.02	.02
	Q_2	.04	.05	.05
	Q_3	.08	.10	.10
P(.20)	Q_1	.01	.01	.01
	Q_2	.02	.03	.03
	Q_3	.05	.06	.06
P(.50)	Q_1	.02	.02	.02
	Q_2	.04	.04	.05
	Q_3	.08	.09	.09

The estimates of MAE, P(.20), and P(.50) for the separate-sex predictions were slightly less accurate than the corresponding estimates for the total group predictions, reflecting the smaller base sample sizes. Further analysis of the data revealed that among colleges with fewer than 50 males, the median observed MAE for males was .60 grade units and the median absolute difference between predicted and observed MAE was .10 grade units. Among colleges with 50–75 males the median MAE for males was .59 grade units and the median absolute difference between predicted and observed MAE was .07 grade units. Among colleges with fewer than 50 females, the corresponding results were .59 and .10 grade units, respectively. Among colleges with 50–75 females, the corresponding results were .49 and .07 grade units, respectively. These results indicate that for most colleges the accuracy of separate-sex predictions can be usefully predicted from Formulas (2a) and (3a).

Appendix

Accuracy of the Gram-Charlier Approximations

A Monte-Carlo study was done to estimate the accuracy of the Gram-Charlier approximations to the distribution function of $\hat{y} - y^*$ and of the resulting MAE. Pseudo-random samples of 10,000 values of $\hat{y} - y^*$ were generated for each of several values of n and p. Statistics from the resulting samples were then computed and compared to the values predicted by Formulas (2a), (3a) and (2b), and (3b).

The values of $\hat{y} - y^*$ were generated in a manner suggested by the formula in the lemma: First, a variable W with an $F(p, n - p)$ distribution was generated from the International Mathematical and Statistical Libraries, Inc. (IMSL) routine GGAMR (IMSL, 1980). Then a standard normal variable X was generated from the McGill routine RNOR (Marsaglia, Ananthanarayanan,

& Paul, 1970). Finally, a value of

$$\hat{y} - y^* = X \cdot \sqrt{[1 + 1/n][1 + W \cdot p/(n - p)]}$$

was computed.

The results are displayed in Table A. In this table $\hat{F}$ denotes the empirical distribution function of $\hat{y} - y^*$ corresponding to the unknown distribution F; $\hat{F}$ was computed from the 10,000 observations in each category defined by n and p. F_1 and F_2 are the first and second Gram-Charlier approximations, respectively, to F. Similarly, $\hat{P}$ is the empirical distribution function of $|y - y^*|$ corresponding to the unknown distribution P, and P_1 and P_2 are the approximations to P derived from the approximations F_1 and F_2 (i.e., from Formulas (3a) and (3b), respectively). MAE_1 and MAE_2 are the mean absolute errors corresponding to F_1 and F_2 (i.e., from Formulas (2a) and (2b), respectively).

The normal approximation F_1 was within .014 of the empirical distribution $\hat{F}$ except in one case, where $(p, n) = (5, 10)$. It is difficult to construct a confidence interval for $\max_t |F(t) - F_1(t)|$ from $\max_t |\hat{F}(t) - F_1(t)|$; however, the 95th and 99th percentiles for $\max_t |\hat{F}(t) - F(t)|$ are approximately .013 and .016, respectively, when $\hat{F}$ is based on 10,000 observations.

In the simple linear regression of $\max_t |\hat{F}(t) - F_1(t)|$ on the inflation factor K, the constant was $-.0193$, the slope was .0262, the correlation was .87, and the standard error of estimate was .0023. This would suggest a strong linear relationship between the accuracy of the normal approximation and the inflation in MAE due to estimating the weights. For base samples for which $K = 1.10$, an approximate 95 percent confidence interval for the fitted value of $\max_t |\hat{F}_1(t) - F_1(t)|$ ranges from about .009 to .010; for $K = 1.25$, such an interval ranges from about .013 to .014.

In 14 of the 25 simulations, F_1 and F_2 were equally close to $\hat{F}$ (to within three decimal places), as measured by the maximum distance criterion. In the six simulations in which F_2 was closer to $\hat{F}$ than was F_1, the differences between the two distances was .003 or less. In the five simulations in which F_1 was closer, the difference between the two distances ranged up to .035 (for $(p, n) = (5, 10)$). The evidence would indicate that for values of n and p like those in this study, adding a second term to the Gram-Charlier approximation results in little or no improvement.

The results for the approximations P_1 and P_2 were similar. The approximation P_1 was within .014 of the empirical distribution function $\hat{P}$ except in three cases, namely $(p, n) = (3, 10)$, $(5, 10)$, and $(8, 25)$. There was a strong relationship between the accuracy of this approximation and K: In the simple linear regression of $\max_t |\hat{P}(t) - P(t)|$ on $K(n, p)$, the constant was $-.0452$, the slope was .0504, the correlation was .86, and the standard error of estimate was .0047. For $K = 1.10$, an approximate 95 percent confidence interval for the fitted value of $\max_t |\hat{P}(t) - P_1(t)|$ is (.009, .011); for $K = 1.25$, it is (.016, .019).

TABLE A

Results from Simulation Study of the Distribution of the Prediction Error

Number of Predictors (p)	Base Sample Size (n)	$\max_t \lvert \hat{F}(t) - F_1(t) \rvert$	$\max_t \lvert \hat{F}(t) - F_2(t) \rvert$	$\max_t \lvert \hat{P}(t) - P_1(t) \rvert$	$\max_t \lvert \hat{P}(t) - P_2(t) \rvert$	MAE_1 obs. MAE	MAE_2 obs. MAE
1	10	.008	.007	.009	.010	.007	.001
	25	.005	.004	.006	.006	.001	.001
	50	.008	.008	.008	.008	.000	.000
	75	.010	.010	.007	.007	.008	.008
2	10	.012	.014	.011	.017	.001	−.014
	25	.008	.008	.009	.010	.000	−.001
	50	.008	.008	.009	.009	−.007	−.007
	75	.010	.010	.009	.009	.006	.006
3	10	.014	.017	.019	.025	.009	−.025
	25	.008	.009	.009	.010	.000	−.002
	50	.008	.008	.010	.011	−.007	−.007
	75	.010	.010	.009	.009	.006	.006
5	10	.029	.064	.052	.123	.075	−.139
	25	.011	.012	.006	.006	.000	−.003
	50	.006	.005	.010	.010	.009	.009
	75	.013	.013	.008	.008	−.001	−.001
	100	.008	.008	.009	.009	−.004	−.004
8	25	.012	.009	.018	.012	.018	.011
	50	.012	.011	.011	.010	.008	.007
	75	.010	.010	.014	.014	.010	.010
	100	.008	.008	.012	.012	.006	.006
10	25	.013	.013	.009	.012	−.006	−.016
	50	.013	.012	.010	.009	.007	.005
	75	.006	.006	.009	.009	.007	.006
	100	.007	.007	.005	.005	.002	.002

The approximations P_1 and P_2 were equally close to $\hat{P}$ (to within .001) in 14 of the simulations. For $(p, n) = (5, 10)$, P_1 was considerably closer to $\hat{P}$ than was P_2.

The approximation MAE_1 was within .010 of the observed MAE except for $(p, n) = (5, 10)$ and $(8, 25)$. The estimated standard errors for the differences reported in the last two columns of Table A range from .006 to .008, except for $(p, n) = (5, 10)$, where the estimated standard error is .011.

References

American College Testing Program. *Technical report for the ACT Assessment Program*. Iowa City, Iowa: Author, 1973.

Browne, M. W. A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology*, 1975, *28*, 112–120.

Claudy, J. G. A comparison of five variable weighting procedures. *Educational and Psychological Measurement*, 1972, *32*, 311–322.

Cramér, H. *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press, 1945.

Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin*, 1968, *69*, 161–182.

Draper, N., & Smith, H. *Applied regression analysis*. New York: Wiley, 1966.

Graybill, F. *An introduction to linear statistical models* (Vol. 1). New York: McGraw-Hill, 1961.

Halinski, R. S., & Feldt, L. S. The selection of variables in multiple regression analyses. *Journal of Educational Measurement*, 1970, *7*(3), 151–158.

International Mathematical and Statistical Libraries, Inc. *IMSL Library 1 Reference Manual* (7th ed.). Houston, Tex.: Author, 1970.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics*, Vol. 1. New York: Hafner Publishing, 1969.

Kerridge, D. Errors of prediction in multiple regression with stochastic regressor variables. *Technometrics*, 1967, *9*, 309–311.

Krutchkoff, R. *Probability and statistical inference*. New York: Gordon and Breach, 1970.

Marsaglia, G., Ananthanarayanan, K., & Paul, N. McGill University Random Number Package "SUPER-DUPER". School of Computer Science, McGill University, 1970. (Documentation supplied by authors.)

Maxey, E. J., & Ormsby, V. J. *The accuracy of self-report information collected on the ACT test battery: High school grades and items of nonacademic achievement* (ACT Research Report No. 45). Iowa City, Iowa: American College Testing Program, 1971.

Miller, D. E., & Kunce, J. T. Prediction and statistical overkill revisited. *Measurement and Evaluation in Guidance*, 1973, *6*(3), 157–163.

Mosteller, F., Rourke, R., & Thomas, G. *Probability and statistical inference*. Reading, Mass.: Addison-Wesley, 1970.

Park, C. N., & Dudycha, A. L. A cross-validation approach to sample size determination for regression models. *Journal of the American Statistical Association*, 1974, *69*(35), 214–218.

Sawyer, R., & Maxey, E. J. The relationship between institutional size and other characteristics and the accuracy of college freshman grade predictions. *American Statistical Association* 1979 *Proceedings of the Social Statistics Section*. Washington, D.C.: American Statistical Association, 1979.

Searle, S. R. *Linear models*. New York: Wiley, 1971.

Stein, C. Multiple Regression. In I. Olkin, S. G. Churye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics*. Stanford, Calif.: Stanford University Press, 1960.

Author

SAWYER, RICHARD. Address: American College Testing Program, P.O. Box 168, Iowa City, IA 52243. Title: Statistician. Specialization: Statistics.