RESEARCH REPORT No. 83

Determining Minimum Sample Sizes for Multiple Regression Grade Prediction Equations for Colleges

Richard Sawyer

February 1984

ACT

DETERMINING MINIMUM SAMPLE SIZES FOR MULTIPLE REGRESSION GRADE PREDICTION EQUATIONS FOR COLLEGES

Prepared by the Research and Development Division
The American College Testing Program

For additional copies write:
ACT Publications
P.O. Box 168
Iowa City, Iowa 52243

(Check or money order must accompany request.) Price: $1.00

ABSTRACT

The American College Testing Program offers research services through which colleges can predict the freshman grades of their future students. This paper describes research done to establish a minimum sample size requirement for calculating prediction equations for college freshman grade average. Results from all the studies suggest that eight-variable prediction equations based on representative samples of size 50 would have almost the same accuracy as prediction equations based on larger samples.

DETERMINING MINIMUM SAMPLE SIZES FOR MULTIPLE REGRESSION GRADE PREDICTION EQUATIONS FOR COLLEGES

Richard Sawyer

The American College Testing Program (ACT) offers research services through which colleges can predict the freshman grades of their future students (The American College Testing Program, 1983). The students' predicted grades are based on their ACT test scores in English, mathematics, social studies, and natural sciences, and on their self-reported high school grades in these four subject areas. The predicted grades are calculated by weighting the test scores and high school grades in least-squares regression equations that are specific to each college.[1]

The weights in a college's prediction equation are usually calculated from data on all students in a previous freshman class who took the ACT. Because these weights are estimates whose accuracy depends on the size of the base sample used to calculate them, and because error in estimating the weights propagates error in prediction, the freshman class size affects prediction error. It is possible, therefore, that weights calculated from very small freshman classes could be subject to large sampling errors, resulting in predictions of unacceptable accuracy.

One way to mitigate the effect of small sample sizes on prediction accuracy is to use information collaterally from several colleges in constructing prediction equations. Novick et al. (1972) further developed a Bayesian model due to Lindley (1970) in which this method was used. Novick et al. calculated for m = 22 junior colleges the standard least-squares and the Bayesian "m-group" prediction equations for freshman grade average, using the four ACT test scores as predictors. The mean number of students in the 22 colleges was approximately 246. Novick et al. then cross-validated the prediction equations against the following year's freshmen at these colleges. They obtained an average cross-validated Mean Absolute Error (MAE) of .58 grade units for both the least-squares and the Bayesian m-group prediction methods. When the prediction equations were developed from 25% samples of the base year freshman classes, the resulting mean cross-validated MAE was .61 grade units for the least-squares and .59 grade units for the Bayesian method. The results of Novick et al. suggest, therefore, that four-variable least-squares predictions for freshman classes with as few as 50 students would not be grossly inaccurate. The results further suggest that the Bayesian m-group method would yield more accurate predictions than least-squares when sample sizes are smaller than 50. Other centralized prediction methods, such as that due to Dempster, Rubin, and Tsutakawa (1981), also seem promising in this regard.

The focus of this paper is on standard least-squares predictions, since they are still the most extensively used predictions and are currently used by ACT. The purpose of the study is to determine for how small a college least-squares prediction equations can be developed without significant degradation in prediction accuracy. We shall consider of practical significance a 10% or larger increase in MAE over that which would occur at larger colleges.

One way to address this issue is to assume that the freshmen in a college are a random sample from a hypothetical population with postulated statistical characteristics. Under this assumption, determining the appropriate sample size for calculating prediction weights becomes a mathematical problem of relating measures of prediction accuracy to parameters of a statistical model. Sawyer (1982) took this approach; some of the results from that study are discussed later.

Students from colleges of different sizes may be samples from different populations of students, insofar as the predictability of their grades is concerned. Thus, a college's size, as an institutional characteristic that attracts certain kinds of students, could be related to the predictive validity of ACT test scores and high school grades. It is conceivable, for example, that the grades of students enrolled in very small colleges could be predicted more accurately than those of students enrolled in larger colleges. Sawyer and Maxey (1982) studied the sample size problem in this context; they found little relationship between prediction accuracy and college size for colleges with 90 or more freshmen. They also hypothesized that predictions of acceptable accuracy could be made for entire freshman classes with as few as 50 students.

[1]In practice, ACT averages the predictions from two four-variable multiple regression equations based on test scores separately and on high school grades separately. The accuracy of these predictions, though, is virtually the same as that of predictions based on a single eight-variable multiple regression equation (Sawyer and Maxey, 1979).

As a result of these two studies, ACT lowered the minimum sample size requirement for its predictive research services from 100 to 75 students, effective for 1979-80 freshmen. In this paper the accuracy of the grade predictions at colleges with 75-100 freshmen is summarized. The experience in predicting grades at these colleges is then discussed in the context of the previously cited studies. Finally, conclusions are drawn about the accuracy of predictions at colleges with fewer than 75 students.

Theoretical Considerations

Suppose the regression coefficients in a prediction equation are 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. (In the application described above, y_i is the college freshman grade average and $p = 8$.) Suppose $\mathbf{x}_i$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Therefore, the predictors $\mathbf{x}$ are assumed to be random rather than fixed; this aspect of the model reflects the inability of colleges to control precisely the test scores and high school grades of their entering freshmen.

The conditional distribution of y_i given $\mathbf{x}_i$ is assumed to be normal with mean $(1,\mathbf{x}_i')\,\boldsymbol{\beta}$ and variance σ^2. The regression coefficients are estimated by the usual least-squares estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \text{ where } \mathbf{X} = \begin{bmatrix} 1 & \mathbf{x}_1' \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & \mathbf{x}_n' \end{bmatrix},$$

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 $\hat{y} = (1,\mathbf{x}^{*\prime})\,\hat{\boldsymbol{\beta}}$.

Sawyer (1983) studied the moments of the distribution of the prediction error $\hat{y} - y^*$. The mean of $\hat{y} - y^*$ is, of course, 0; its standard deviation is

$$\text{RMSE} = \sigma \cdot K(n,p),$$

$$\text{where } K(n,p) = \sqrt{\frac{(n+1)(n-2)}{n(n-p-2)}}\,.$$

Sawyer found that when $K \leq 1.10$, the distribution of $\hat{y} - y^*$ is approximately normal. In this case, the mean absolute error of prediction $\text{MAE} = E(|\hat{y} - y^*|)$ is approximately

$$\text{MAE} = \sqrt{2/\pi} \cdot \text{RMSE}.$$

The function K(n,p) is an inflation factor due to estimating the regression coefficients; as $n \to \infty$, $K(n,p) \to 1$. For fixed values of K and p one can approximate the corresponding required base sample size n by

$$n = \frac{2K^2-1}{K^2-1} + \frac{K^2}{K^2-1}\,p\,. \qquad (1)$$

The coefficients in (1) are displayed in Table 1 for several values of K and p. They suggest that in predicting college freshmen grade average from an eight-variable multiple regression equation, a base sample size of approximately 53 would result in a 10% inflation in RMSE or MAE over that which would result if the population values of the coefficients were known.

TABLE 1

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 $\text{MAE} = K\sigma\sqrt{2/\pi}$ with $1 \leq p \leq 20$ predictors.

Sawyer and Maxey (1982) examined the accuracy of prediction equations at a random sample of 205 colleges that participated in the ACT Research Services in 1974-75 and in 1976-77. A separate prediction equation for each college was calculated from its 1974-75 data. Then, each resulting prediction equation was applied to data for the 1976-77 freshmen, and the predicted and actual grade averages were compared. (The two-year lag between base year and cross-validation year reflects the time lag encountered by colleges in developing and using prediction equations.)

The cross-validation statistics in Table 2 are summarized for five categories of colleges defined by their base sample size. The statistics P20, P50, and P100 refer to the proportion of students in a college whose predicted grade averages were within .20, .50, or 1.00 grade units, respectively, of their actual grade averages. The statistic CVR is the correlation between earned and predicted grade average in a college. The numbers in Table 2 are mean values of these cross-validation statistics among colleges in the sample.

Table 2 indicates that the predictive validity of ACT test scores and high school grades is only weakly related to freshman class size at colleges with 90 or more freshmen. For example, the average observed MAE ranged from .51 to .54 grade units over the five size categories. Similarly, the average cross-validated correlation ranged from .53 to .56 over the five size categories.

TABLE 2

Mean Cross-Validation Statistics, by Size of College Freshman Class
(Total Group Equation)

Size category	Number of colleges	Number of students (1976)	Cross-validation statistic				
			MAE	P20	P50	P100	CVR
90-100	15	2,544	.52	.25	.57	.87	.53
101-200	76	11,007	.51	.26	.59	.89	.55
201-500	50	15,951	.54	.24	.56	.87	.56
501-1000	35	29,603	.54	.24	.56	.87	.55
1001+	29	55,773	.53	.25	.57	.87	.56
All colleges	205	114,878	.53	.25	.57	.88	.55

Because of ACT's sample size requirements in effect at the time of the Sawyer and Maxey study, there were no colleges with total group sample sizes below 90. To obtain evidence about prediction accuracy for sample sizes below 90, albeit indirect, Sawyer and Maxey developed prediction equations from random subsamples of the 1974-75 freshman data from each college. The results, shown in Table 3, indicate that the MAEs associated with grade predictions based on random subsamples of size 50 are within 10% of the MAEs associated with predictions based on all records. Therefore, although direct evidence of the accuracy of grade predictions for colleges with fewer than 90 students was not available, it appeared that grade predictions of comparable accuracy could be made at colleges with as few as 50 freshmen.

TABLE 3

Mean College Cross-Validation Statistics for Prediction Equations Derived from Subsamples of Base Year Data

Size of subsample of base year data	Cross-validation statistics				
	MAE	P20	P50	P100	CVR
25	.65	.21	.48	.79	.41
50	.57	.23	.54	.85	.49
75	.55	.24	.55	.87	.52
100	.54	.24	.56	.88	.53
All records	.53	.25	.57	.88	.55

Follow-up Study

As a result of the two studies above, ACT lowered the minimum sample size requirement for its predictive research services from 100 to 75 students, effective for 1979-80 freshmen. Following is an examination of the accuracy of the grade predictions at the colleges whose sizes are in this range. Further evidence is also presented on the likely prediction accuracy at colleges with fewer than 75 students.

Prediction equations for freshman grade average were developed from the 1979-80 freshman grade data at all colleges with between 70 and 100 freshmen. (To accommodate small colleges with a few unexpectedly invalid records, ACT used an actual cut-off of five records less than the published cut-off of 75.) Separate subgroup equations were also developed for the males and females at each college. The prediction equations were then cross-validated against the grades of the 1981-82 freshmen at each college.

The results for the total group prediction equations, contained in Table 4a, confirm the expectation that predictions based on as few as 75 students would be about as accurate as predictions based on larger numbers of students. The mean MAE for colleges with 70-79 freshmen, for example, was .51 grade units; the same mean MAE was observed for colleges with 90-100 freshmen. In the Sawyer and Maxey study cited above, the mean MAE for colleges with 90-100 freshmen was .52 grade units, and the mean MAE for all colleges was .53 grade units.

It is interesting to note in Table 4a that the mean MAE for colleges with 80-89 freshmen (.55 grade units) is actually larger than the mean MAE for colleges with 70-79 freshmen (.51 grade units). This result might reflect differences in the predictive validity of the ACT at colleges in these two size categories. Given the estimated standard errors for these means, however, the differences could also be reasonably thought of as due to chance.

TABLE 4a

Mean Cross-Validation Statistics, by Size of College Freshman Class
(Total Group Equation)

Size category	Number of colleges	Number of students (1981)	Mean cross-validation statistics[a]				
			MAE	P20	P50	P100	CVR
70-79	33	2,643	.51(.01)	.24(.01)	.58(.01)	.89(.01)	.52(.02)
80-89	25	2,000	.55(.02)	.25(.02)	.56(.02)	.85(.02)	.46(.03)
90-100	10	849	.51(.03)	.28(.02)	.60(.02)	.88(.03)	.51(.03)
All colleges	68	5,492	.53(.01)	.25(.01)	.58(.01)	.88(.01)	.49(.02)

[a]Numbers in parentheses are estimated standard errors corresponding to the estimated means.

The results for the separate subgroup equations for males are contained in Table 4b. The mean MAE was .63 grade units for predictions based on 30-39 males, and .58 grade units for predictions based on 40-49 males. In the Sawyer and Maxey (1982) study, the mean MAE over all colleges was also .58 grade units. Therefore, it would appear that prediction equations based on as few as 40-49 males are about as accurate as predictions based on larger numbers of males.

The results for the separate subgroup equations for females are contained in Table 4c. The mean MAE was .47 grade units for predictions based on 50-59 females. In the Sawyer and Maxey (1982) study, the mean MAE over all colleges was .52 grade units. Therefore, prediction equations based on as few as 50-59 females are about as accurate as predictions based on larger numbers of females.

Because of the minimum sample size requirement now in effect, cross-validation statistics are not reported in Table 4a for colleges with fewer than 70 freshmen. The results in Tables 4b and 4c for the separate subgroup equations suggest, however, that total group equations developed from samples of as few as 40-50 freshmen would have nearly the same prediction accuracy as total group equations developed from larger samples.

TABLE 4b

Mean Cross-Validation Statistics, by Number of Males in Freshman Class
(Separate Subgroup Equation for Males)

Size category	Number of colleges	Number of students (1981)	Mean cross-validation statistics[a] MAE	P20	P50	P100	CVR
25-29	6	208	.74(.04)	.13(.03)	.35(.04)	.71(.03)	.47(.03)
30-39	8	293	.63(.04)	.21(.03)	.52(.03)	.81(.03)	.36(.07)
40-49	6	220	.58(.03)	.22(.03)	.52(.02)	.84(.03)	.39(.08)
50 and above	1	75	.41(—)	.36(—)	.71(—)	.93(—)	.54(—)
All colleges	21	796	.64(.03)	.20(.02)	.48(.03)	.80(.02)	.41(.04)

[a]Numbers in parentheses are estimated standard errors corresponding to the estimated means.

TABLE 4c

Mean Cross-Validation Statistics, by Number of Females in Freshman Class
(Separate Subgroup Equation for Females)

Size category	Number of colleges	Number of students (1981)	Mean cross-validation statistics[a] MAE	P20	P50	P100	CVR
25-29	5	147	.58(.12)	.27(.05)	.51(.08)	.84(.08)	.47(.09)
30-39	8	285	.55(.03)	.25(.02)	.56(.03)	.86(.03)	.38(.04)
40-49	20	858	.56(.03)	.25(.02)	.56(.03)	.86(.02)	.39(.04)
50-59	12	530	.47(.03)	.27(.02)	.63(.02)	.91(.03)	.57(.03)
60 and above[b]	3	133	.51(.06)	.24(.02)	.63(.05)	.88(.03)	.51(.09)
All colleges	48	1,953	.54(.02)	.26(.01)	.58(.02)	.87(.01)	.45(.03)

[a]Numbers in parentheses are estimated standard errors corresponding to the estimated means.

[b]Maximum sample size was 72.

Conclusions

Results from studies by Novick et al. (1972), Sawyer and Maxey (1982), and Sawyer (1982) suggest the likelihood that least-squares grade prediction equations based on data for as few as 50 students would be about as accurate as prediction equations based on much larger samples. The present study confirms that total group predictions based on 70 or more students have the same accuracy as predictions based on large samples. Moreover, the results from separate-sex prediction equations lend further support to the idea that a base sample size as low as 50 would be satisfactory.

One should keep in mind that these sample size recommendations pertain to entire freshman classes or to representative samples of freshman classes. Prediction equations based on greatly nonrepresentative samples may result in larger prediction errors when applied to more general student populations.

REFERENCES

The American College Testing Program (1983). ACT Research Services (December 1982 edition). Iowa City, Iowa: Author.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374), 341-353.

Lindley, D. V. (1970). *A Bayesian solution for some educational prediction problems, III* (Research Bulletin 70-33). Princeton, N.J.: Educational Testing Service.

Novick, M. R., Jackson, P. H., Thayer, D. T., and Cole, N. S. (1972). Estimating multiple regressions in m groups: A cross-validation study. *British Journal of Mathematical and Statistical Psychology*, 25, 33-50.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. *Journal of Educational Statistics*, 7(2), 91-104.

Sawyer, R., and Maxey, E. J. (1979). *The validity over time of college freshman grade prediction equations* (ACT Research Report No. 80). Iowa City, Iowa: The American College Testing Program.

Sawyer, R., and Maxey, E. J. (1982). *The relationship between college freshman class size and other institutional characteristics and the accuracy of freshman grade predictions* (ACT Research Report No. 82). Iowa City, Iowa: The American College Testing Program.