ACT Research Report Series 87-4

Determining Minimum Sample Sizes for Estimating Prediction Equations for College Freshman Grade Average

Richard Sawyer

March 1987

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

DETERMINING MINIMUM SAMPLE SIZES FOR ESTIMATING PREDICTION EQUATIONS FOR COLLEGE FRESHMAN GRADE AVERAGE

Richard Sawyer

ABSTRACT

This report addresses the problem of sample size in developing prediction equations for college freshman grade average. Practical guidelines, based on theory and on analyses of data collected through the ACT predictive research services, are given.

DETERMINING MINIMUM SAMPLE SIZES FOR ESTIMATING PREDICTION EQUATIONS FOR COLLEGE FRESHMAN GRADE AVERAGE

The ACT Assessment Program is a system for collecting, processing, and reporting data to help students and educators involved in the transition from high school to college. A major component of the ACT Assessment is its predictive research services, through which colleges and universities conduct local predictive validity studies and develop prediction equations for guidance, selection, and placement (ACT, 1986). In this paper we focus on a practical statistical problem often encountered by institutions in developing their prediction equations, namely, the minimum sample size required to obtain accurate grade predictions.

The weights in a college grade prediction equation are typically estimated from the test scores, high school grades, and college grades of one freshman class, and are used to predict the grades of future freshmen. For the ACT Assessment, the prediction weights are estimated by standard least squares procedures. At small colleges, and at large colleges where a minority of students take the ACT, there may be few records from which to develop prediction equations. The question naturally arises, therefore, as to how small a sample can safely be used.

Because prediction weights are estimated regression coefficients whose accuracy depends on the size of the base sample used to estimate them, and because error in estimating the weights propagates error in prediction, sample size affects prediction accuracy. It is possible, therefore, that weights calculated from very small samples could be subject to large sampling errors, resulting in predictions of unacceptable accuracy.

Though affected by sampling error, prediction accuracy is primarily determined by the strength of the relationship between the predictor and criterion variables as measured, for example, by either the associated residual variance or the multiple correlation. (This, naturally, varies among colleges even of the same size.) Estimating regression coefficients from finite base samples inflates the prediction errors caused by the imperfect relationship between predictors and criterion. A useful way to study sample size in this context, therefore, is to determine the relationship between it and the resulting inflation in prediction error variance.

Theoretical Perspective

It is mathematically convenient to study predictions based on random samples from a multivariate normal population. If σ^2 is the conditional variance of the criterion variable y, given the predictor variables, and if the predicted criterion $\hat{y}$ is based on least squares estimates, then the root mean squared error of prediction, RMSE = $\{E(y-\hat{y})^2\}^{1/2}$, is RMSE = $\sigma\, K(n,p)$ where p is the number of predictor variables, n is the base sample size, and

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

Thus K is an inflation factor due to estimating the regression coefficients; note that for any fixed p, $K(n,p)\to 1$ as $n\to\infty$. Sawyer (1982) found that if $K \le 1.10$ then $y-\hat{y}$ is approximately normally distributed. For this case the mean absolute error of prediction, MAE = $E[\,|y-\hat{y}|\,]$, is approximately MAE $\doteq \sqrt{2/\pi}\cdot$ RMSE. Sawyer (1982) also found that for fixed values of K and p, one can approximate the corresponding required base sample size 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 freshman grade average from an eight-variable multiple regression equation, for example, 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. The corresponding required sample size for a two-variable prediction equation would be approximately 18.

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	50.8p + 51.8
1.05	10.8p + 11.8
1.10	5.8p + 6.8
1.25	2.8p + 3.8
1.50	1.8p + 2.8

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

Empirical Studies

In 1979, ACT lowered the minimum sample size requirement for its predictive research services from 100 to 75 students. In a study on the effects of this change, Sawyer (1984) found that there was no significant difference in the accuracy of grade predictions based on samples of size 70-99 and the accuracy of predictions based on larger samples. In 1983, ACT lowered its minimum sample size requirement still further, to 50 students. Following is an examination of the accuracy of grade predictions at those colleges, with base samples of 50-99 students, that have participated in the ACT predictive research services since 1983.

Prediction equations for freshman grade average were developed from the 1983-84 grade data at the 125 colleges with 50-99 cases. The predictor variables in these equations were the four ACT subtest scores (in English, mathematics, social studies, and natural sciences) and the four self-reported high school grades in the subject areas corresponding to the ACT subtests. To study the effect of the number of predictor variables on prediction accuracy, two-variable prediction equations, based on the ACT Composite (the average of the ACT subtest scores) and on HSA (the average of the self-reported high school grades), were also calculated. To determine the accuracy of prediction equations based on fewer than 50 cases, separate subgroup equations were also calculated for the females and males at each college.

All the prediction equations were then cross-validated against the grades of 1984-85 freshmen; that is, prediction equations developed from the 1983-84 freshmen were applied to the test scores and high school grades of the 1984-85 freshmen at each college, and the predicted and actual grades were compared. This procedure models the actual use of prediction equations by colleges, and it avoids the tendency of estimates of prediction accuracy derived from a single year's data to be overoptimistic.

The prediction equations developed from 1983-84 freshman data were used by colleges to predict the grades of 1985-86 freshmen; but, due to the time schedules colleges must follow in reporting data to ACT, these grades were not available when the analyses were done. Therefore, the prediction equations in this study were cross-validated against 1984-85 freshman grades, which were available. Sawyer and Maxey (1979) compared the accuracy of one- and two-year-old prediction equations and found negligible differences.

The predicted and actual grade averages of 1984-85 freshmen were compared in terms of observed mean absolute error (MAE), which is the average absolute difference between the predicted and actual grade averages at a college. The distributions of this cross-validation statistic over colleges are summarized in Tables 2, 3, and 4.

TABLE 2

Distribution of Cross-Validated Mean Absolute Error, by Base Sample Size and Number of Predictors (Total Group Equations)

Base sample size	Number of colleges	Number of predictors 2 Min.	Med.	Max.	Number of predictors 8 Min.	Med.	Max.
49-59	41	.36	.50	.74	.39	.53	.76
60-69	20	.41	.55	.67	.43	.56	.72
70-79	23	.38	.50	.70	.41	.53	.78
80-89	20	.37	.51	.77	.40	.55	.81
90-99	21	.33	.50	.65	.35	.53	.70

The results for the total group prediction equations, reported in Table 2, confirm the expectation that predictions based on as few as 50 students would be about as accurate as predictions based on larger numbers of students. The median MAE for colleges with 49-59 cases, for example, was .53 grade units for the eight-variable predictions; the same median MAE was observed for colleges with 90-99 cases. In a study by Sawyer and Maxey (1982), 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 that in Table 2 the median MAE for two-variable predictions at colleges with 60-69 cases (.55 grade units) is actually larger than the median MAE for colleges with 49-59 cases (.50 grade units). As the difference between these two medians is modestly statistically significant

($p < .05$), it might reflect differences in the predictive validity of the ACT at colleges in the two size categories.

The results for the separate subgroup equations for females, in Table 3, show the effect of the number of predictors on prediction accuracy. According to Sawyer and Maxey (1979), the mean MAE for eight-variable predictions for females, over all colleges with 100 or more students, is .50 grade units. The median MAEs for the two-variable predictions for females suggest that predictions based on samples with 20-29 cases are nearly as accurate, with a median MAE of about .52 grade units. The median MAEs for the eight-variable predictions suggest that sample sizes of 60 or more cases may be required to attain this level of accuracy.

TABLE 3

Distribution of Cross-Validated Mean Absolute Error, by Base Sample Size and Number of Predictors (Separate Subgroup Equations for Females)

Base sample size	Number of colleges	Number of predictors: 2 Min.	2 Med.	2 Max.	8 Min.	8 Med.	8 Max.
10-19	12	.36	.56	.93	.32	.64	1.23
20-29	26	.38	.52	.84	.39	.59	.91
30-39	30	.34	.53	.87	.35	.62	.93
40-49	30	.32	.48	.68	.36	.55	1.02
50-59	13	.31	.53	.76	.37	.59	.86
60 and over[a]	10	.33	.43	.66	.36	.46	.85

[a]Maximum sample size was 86.

The results for the separate subgroup predictions for males, in Table 4, show similar trends. According to Sawyer and Maxey (1979) the mean MAE for predictions for males over all colleges with 100 or more students is .56 grade units. The median MAEs for the two-variable predictions for males, in Table 4, suggest that predictions based on samples with 20-29 cases typically have MAEs of about .57 grade units. The median MAEs for the eight-variable predictions in the largest size category was .65 grade units.

TABLE 4

Distribution of Cross-Validated Mean Absolute Error, by Base Sample Size and Number of Predictors (Separate Subgroup Equations for Males)

		Number of predictors					
		2			8		
Base sample size	Number of colleges	Min.	Med.	Max.	Min.	Med.	Max.
10-19	20	.34	.62	1.34	.36	.72	2.91
20-29	37	.30	.57	.78	.45	.65	1.86
30-39	28	.39	.57	.90	.42	.65	1.93
40 and over[a]	11	.38	.54	.74	.42	.65	1.15

[a]Maximum sample size was 82.

A two-variable prediction equation based on ACT Composite score and HSA constrains the regression coefficients for the four ACT subtest scores to be the same; similarly, it constrains the regression coefficients for the four self-reported high school grades to be the same. These constraints should, other things being equal, result in larger prediction errors for the two-variable equation due to prediction bias. Because the four ACT subtest scores have the same scale and are moderately correlated with each other (and because the same is true of high school grades), one would expect the prediction bias to be minimal. Note that, in fact, the median MAEs in Tables 2, 3, and 4 for the two-variable equations are actually smaller than the corresponding median MAEs for the eight-variable equations. This suggests that any increase in bias caused by using two-variable equations is more than offset by decreased sampling error. Of course, this would not occur if predictor variables with dissimilar scales were averaged.

Conclusions

These results confirm the expectation that total group predictions based on 50 or more cases and eight or fewer predictor variables have nearly the same accuracy as predictions based on larger samples. Moreover, two-variable prediction equations based on as few as 20-29 cases would have essentially the same accuracy as prediction equations based on larger samples. On the other hand, the results from separate-sex prediction equations strongly suggest that eight-variable prediction equations based on much fewer than 50 cases would be noticeably less accurate.

REFERENCES

The American College Testing Program. (1986). ACT Research Services (December 1986 edition). Iowa City, Iowa: Author.

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. (1984). Determining minimum sample sizes for multiple regression grade prediction equations for colleges (ACT Research Report No. 83). Iowa City, Iowa: The American College Testing Program.

Sawyer, R., & 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.