CHAPTER VI

USE OF AN EXTRA APTITUDE TEST SCORE FOR INCREASING THE ACCURACY OF PREDICTION

The problem to predict whether a student will pass or fail in a future test from his present achievement in a particular test, is a very important one. It is believed that there is always scope for improving the prediction, by adding an extra aptitude test score to the battery of tests being used as admission criteria for selection. Usually each extra test score is tested for significance by a suitable statistical test, and if the test variable is found to make a significant reduction in the variance of the criterion, it is considered worthwhile for addition. But we should also examine whether the addition of the extra test score materially improves the actual prediction also. We will find in this study, in the case considered, that even when the aptitude score reduces the residual variance to a significant extent, material improvement in the prediction is not obtained with the addition of the extra test score. The study also illustrates the uses of statistical methods in such an investigation.

Data

The data used in this paper refer to the Secondary School Certificate Examination marks (Bombay State of the year 1957 and the Preparatory Science Examination marks of the Maharaja Sayajirao University of Baroda of the year 1958, of 278 students, for whom these marks were available, X_1 denotes, marks in English at SSCE, X_2 denotes marks in Mathematics at SSCE, X_3 denotes marks in General Science at SSCE and Z denotes the total percentage marks at PScE. We shall first test the significance of the additional variable X_3 by the usual multiple regression analysis and then proceed to measure the extent of prediction by a discriminant analysis.

3. Multiple Regression Analysis

The corrected sums of squares and products are found to be:

	English	Mathematics	General Science	Criterion(PScE) Grand Total Percent
	X_1	X_2	X_3	Z
X_1	23936.90	10975.83	5976.63	13152.89
X_2		54899.32	13009.96	20600.92
X_3			20019.62	10611.46

The multiple regression equations computed from these are found to be:

$$Z = .3746X_1 + .2379X_2 + .2636X_3 + \text{const..} \quad ..(1)$$

$$Z = .4155X_1 + .2922X_2 + \text{const..} \quad ..(2)$$

The significance of the additional variable x_3 of General Science is tested by the following analysis of variance (Table 6.1):

Table 6.1 Testing the Significance of Gain Due to Addition of General Science Variable

ANALYSIS OF VARIANCE

Score of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Three-Variable Regression	12625.21			
Two-Variable Regression	11484.61			
Gain Due to Addition of General Science Variable	1140.60	1	1140.60	29.38*
Three-Variable Residual	10636.39	274	38.82	
Total	23261.60	277		

As the value of F is found to be highly significant from the F-table, we proceed to discriminant analysis.

4. Discriminant Analysis:

The analysis is done by the methods of Fisher's Discriminant Function and Wald's U-Statistic. The full particulars are reproduced to facilitate understanding and achieve completeness.

Let N_1 and N_2 be the number in the Pass Group and the Fail Group respectively

Let x_1 and y_1 denote marks in English in the Pass Group and the Fail Group respectively

Let x_2 and y_2 denote marks in Mathematics in the Pass Group and the Fail Group respectively

Let x_3 and y_3 denote marks in General Science in the Pass and the Fail Group respectively

The Statistical computations necessary for the analysis are shown below:

(1) Summations:

N_1	= 175	N_2	= 103
$\sum x_1$	= 10511	$\sum y_1$	= 5388
$\sum x_2$	= 12636	$\sum y_2$	= 6122
$\sum x_3$	= 11427	$\sum y_3$	= 6026
$\sum x_1^2$	= 643429	$\sum y_1^2$	= 289782
$\sum x^2_2$	= 942308	$\sum y^2_2$	= 378284

(1) Summations:

$\sum x^2_3 = 757117$ $\quad \sum y^2_3 = 358612$

$\sum x_1x_2 = 763577$ $\quad \sum y_1y_2 = 320181$

$\sum x_1x_3 = 688782$ $\quad \sum y_1y_3 = 315343$

$\sum x_2x_3 = 831806$ $\quad \sum y_2y_3 = 358842$

(2) Means and Mean-Differences:

$\bar{x}_1 = 60.0629$ $\quad \bar{y}_1 = 52.3107$ $\quad d_1 = 7.7522$

$\bar{x}_2 = 72.2057$ $\quad \bar{y}_2 = 59.4369$ $\quad d_2 = 12.7688$

$\bar{x}_3 = 65.2971$ $\quad \bar{y}_3 = 58.5048$ $\quad d_3 = 6.7923$

(3) Matrix of within-Group sums of squares and products S_{ij}:

$$S_{ij} = \begin{bmatrix} 20040.37 & 4557.76 & 2562.58 \\ 4557.76 & 44327.93 & 7386.59 \\ 2562.58 & 7386.59 & 17028.30 \end{bmatrix}$$

(4) Inverse Matrix S^{-1}_{ij}

$$S^{-1}_{ij} = \begin{bmatrix} .000051637044 & -.000004327171 & -.000005893779 \\ -.000004327171 & .000024679463 & -.000010054342 \\ -.000005893779 & -.000010054342 & -.000063974122 \end{bmatrix}$$

(5) D, Relative Weights and F:

$D = .0003050156x_1 + .0002132899x_2 + .0002604598x_3$

Relative Weight	34.48 Percent	39.72 Percent	25.80 Percent

$F_{3,274} = 40.61^{**}$ .. .. ..

(6) Classification equation and U-Statistic:

$U = .0841843x_1 + .0588680x_2 + .0718869x_3$

(7) Critical Region:

$A_1 = .0841843\bar{x}_1 + .0588680\bar{x}_2 + .0718869\bar{x}_3 = 14.00096$

$A_2 = .0841843\bar{y}_2 + .0588680\bar{y}_2 + .0718869\bar{y}_2 = 12.10840$

$\frac{1}{2}(A_1 + A_2) = 13.0547$

Therefore; For $U > 13.0547$ the individual is classified as coming from P_1 population of Pass

$U \leq 13.0547$ the individual is classffied as coming from P_2 population of Fail

(8) Error of classification and Effieiency of classification:

$$\delta^2 = S^{11}(\bar{x}_1 - \bar{y}_1)(\bar{x}_1 - \bar{y}_1) + S^{12}(\bar{x}_1 - \bar{y}_1)(\bar{x}_2 - \bar{y}_2)$$
$$+ S^{13}(\bar{x}_1 - \bar{y}_1)(\bar{x}_3 - \bar{y}_3) + S^{21}(\bar{x}_2 - \bar{y}_2)(\bar{x}_1 - \bar{y}_1)$$
$$+ S^{22}(\bar{x}_2 - \bar{y}_2)(\bar{x}_2 - \bar{y}_2) + S^{23}(\bar{x}_2 - \bar{y}_2)(\bar{x}_3 - \bar{y}_3)$$
$$+ S^{31}(\bar{x}_3 - \bar{y}_3)(\bar{x}_1 - \bar{y}_1) + S^{32}(\bar{x}_3 - \bar{y}_3)(\bar{x}_2 - \bar{y}_2)$$
$$+ S^{33}(\bar{x}_3 - \bar{y}_3)(\bar{x}_3 - \bar{y}_3).$$

$$\bar{\sigma}^2 = 1.89256$$

Hence $$\bar{\sigma} = 1.3757$$

$$\frac{A_2 - A_1}{2\bar{\sigma}} = .688$$

$$p_1 = 1-p_2 = \frac{1}{\sqrt{(2\pi)}} \int_{.688}^{\infty} e^{-t^2/2}\,dt = .2475 \text{ or } 24.75\%$$

where p_1 is the probability of making an error of Type I, that is, of classifying a student as one who will go successfully through the course when he actually does not, and $1-p_2$ is the probability of making an error of Type II, that is of classifying a student as one who will fail in the course in question while he actually passes.

5. Results

In using the above classification equation to classify 278 students used in this study, 22 errors i.e. 21.4%, of Type I were made while 48 errors i.e. 27.4%, of Type II were made. These percentages seem reasonably close to the expected 24.75 percent.

Table 6.2

Classification obtained by Three variable discriminant

	Actually Pass	Actually Fail	Total
Predicted Pass (by Three variable discriminant)	127	22	149
Predicted Fail (by Three variable discriminant)	48	81	129
Total	175	103	278

In using the classification equation from two variables, to classify the same 278 students, 23 errors of Type I i.e. 22.3% were made while 46 errors of Type II i.e. 26.3% were made. The following table shows the results obtained previously (Table 5.1):

Classification obtained by two-variable discriminant

	Actually Pass	Actually Pass	Total
Predicted Pass	129	23	152
Predicted Fail	46	80	126
Total:	175	103	278

On comparing these results, we observe that:

(i) the number of errors of both the kinds (I and II) remains almost the same in both the cases;

(ii) all the errors of one kinds as given by two variable discriminant are not the same as the errors of the same kind as given by three variable discriminant and vice versa.

This information is of great value. It suggests that

(i) it is better to start with the minimum essential variables, than with a number of variables together unless each extra variable to be added is sufficiently accurate to reduce the percentage error in prediction,

(ii) the straightforward application of extra score does not yield improvement in prediction of pass-fail on the whole, though the variable was found to account for a significant variance,

(iii) some special methods have to be sought to achieve further gain in prediction due to extra score.

6. A Method to Improve Prediction

When individuals were classified one by one, by two variable discriminant (1) and subsequently by three variable discriminant (2), the errors of Type I and II occurred as shown by serial numbers in the following tables:

Table 6.3 Serial Numbers Corresponding to Errors of Type I (i.e. Predicted Pass but Actually Fail) as Found by Two Variable Discriminant (1) and by Three Variable Discriminant (2)

Discriminant (1)	Discriminant (2)	Discriminant (1)	Discriminant (2)
10	10	181	181
45	45	183	183
48	48	187	187
119	119	206	x
123	123	x	215
133	133	218	x
x	138	227	227
141	141	228	228
146	146	241	241
151	151	254	x
156	156	255	255
x	171	260	260
179	x	277	277

From the above table, it can be seen that the cases corresponding to serial numbers 179, 206, 218 and 254 occurred as errors of Type I in two variable discriminant analysis but did not occur as errors of any kind in three variable discriminant analysis, that is, cases 179, 206, 218 and 254 were wrongly predicted pass by two variable discriminant but actually failed, while these cases were not wrongly predicted by three variable discriminant. On the other hand, cases corresponding to serial numbers 138, 171 and 215 occurred as errors of Type I in three variable discriminant but did not occur as errors of any kind in two-variable discriminant analysis.

Similar observations can be made for errors of Type II also. The following table shows the serial numbers of the cases which occurred as errors of Type II by two variable discriminant and three variable discriminant.

Table 6.4 Serial Numbers Corresponding to Errors of Type II (i.e. Predicted Fail but Actually Pass) as Found by Two-Variable Discriminant (1) and by Three-Variable Discriminant (2)

Dis.(1)	Dis.(2)	Dis.(1)	Dis.(2)	Dis.(1)	Dis.(2)
1	1	82	82	184	x
3	3	89	89	186	186
9	9	96	96	188	188
15	15	x	97	192	192
24	24	102	102	199	199
25	x	105	105	211	211
30	x	106	x	x	219
36	36	110	110	x	221
x	37	120	x	225	225
x	38	122	122	226	226
39	39	134	134	x	234
49	x	135	135	x	242
50	50	140	140	243	243
x	54	155	x	258	258
66	66	162	162	259	259
67	67	166	166	x	267
x	72	169	169	271	271
75	75	x	172	273	273
81	x	178	x	275	275

From the above table, it can be seen that the cases corresponding to serial numbers 25, 30, 49 ... etc., occurred as errors of Type II in two variable discriminant analysis but did not occur as errors of any kind in three variable discriminant analysis, that is, cases 25, 30, 49 ... etc. were wrongly predicted fail by two variable discriminant (1) while these cases were not wrongly predicted as such by three variable discriminant (2). On the other hand, cases corresponding to serial numbers 37, 38, 54 ... etc. occurred as errors of Type II in three variable discriminant but did not occur as errors of any kind in two variable discriminant analysis.

We further observe that percentage error of Type I is comparatively less than percentage error of Type II. From these observations, a simple suggestion to improve prediction could be given as follows:

(1) First analysis by two-variable discriminant (1) and take all predicted pass as pass. This will eliminate errors in 37, 38, 54 ... etc.

(2) Then analysis by three-variable discriminant (2) and take additional pass numbers given by discriminant (2) as 'pass' corresponding to 'fail'

by discriminant (1). This will eliminate errors in 25, 30, 49 etc. Thus the total number of errors of Type II will be reduced to 46-9 = 37 i.e. 21.1%. But this would increase three other errors of Type I and hence the total errors of Type I will be 23+3 = 26 i.e. 25.2% which does not exceed expected 26.3% obtained in case of two variables. The efficiency of correct classification is now (138 + 77) x100/278, that is, 77.3 percent.

How such gain in prediction could be obtained? What theoretical support can be given to these findings ?

Some considerations into the depth of testing of hypothesis will reveal that the above method that has been explored out of this analysis combines Neyman-Pearson test of hypothesis with the sequential test given by Abraham Wald and extension of this to multivariate analysis, developed by C.R. Rao. The gain in prediction is derived through the benefits of both the methods. The method of C.R. Rao will be described in the next chapter.