Discriminant Analysis -

High School Student Mathematics Class Placement

By Simon King

November 20, 2010

Abstract

The purpose of this study was to explore high school student mathematics class placement through discriminant analysis. The teacher placements of the students into four populations were assessed using various standardized testing and student grades. Using logistic discriminant analysis, a miscalculation rate of 0.32 resulted. While standardized testing helps us paint a useful picture of a student, it is unlikely to paint a complete picture, the missing piece is a teacher knowing that student and other variables that effect a student's success and thus mathematics placement. These extra variables include but are not limited to, work ethic, commitment to learning and communication skills.

Introduction

The intention of this study is to explore student placement into math classes at an independent school. Student placements into mathematics classes are often contested by parents and students. Often, there is a lack of comparative qualitative evidence to support the teacher's recommendation. Therefore the challenge of this study is to use a discriminant analysis to help the assessment and placement of students into mathematic classes.

This study uses data from students from one grade (approximately 106 students) of an independent school. These students were placed into one of the following classes in 9th grade: Algebra I, Geometry, Geometry Honors or Algebra II Honors. A combination of teacher recommendation, grades and some standardized assessment is used to place students. Analysis of current placements will be used to create a classification rule to place future students of any grade who join the school into a math class based on different indicators.

Method

Data is collected for a single grade of 106 students for the following:

x_1 = Student PSAT percentile (based on college bound students)
x_2 = 8th grade mathematics class grade
x_3 = 2010 ERB quantitative test percentile (based on students attending Independent Schools)
x_4 = 2010 ERB Math 1 & 2 test percentile (based on students attending Independent Schools)
x_5 = 2009 ERB quantitative test percentile (based on students attending Independent Schools)
x_6 = 2009 ERB Math 1 & 2 test percentile (based on students attending Independent Schools)

Dataset: Discriminant Analysis – Simon King

The ERB (x_3 - x_6) is a standard assessment used by independent schools that ranks students according to their national and independent schools student percentile. For the purposes of this study we are using the independent schools student percentile as they provide a wider range of scores for the students being classified. For example, students ranked as 99th percentile on a national scale are ranked 77th – 99th percentile on an independent school scale. We can discriminate between the independent scores but not the national ones.

Additionally, their 9th grade mathematics class placement is recorded. This will be our response variable.Students are placed into the following populations of 9th grade mathematics classes:

- Algebra I (Y=1)
- Geometry (Y=2)
- Geometry Honors (Y=3)
- Algebra II Honors (Y=4)

We will explore classification through a discriminant analysis and a logistic discriminant model

Logistic regression model:

$$u(x) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$v(x) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$w(x) = \beta_{03} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Giving cumulative probabilities:

$$P(Y = 1) = \frac{1}{1 + e^{-u(x)}}$$

$$P(Y \leq 2) = \frac{1}{1 + e^{-v(x)}}$$

$$P(Y \leq 3) = \frac{1}{1 + e^{-w(x)}}$$

β_{01} = the average value of the logit for Y=1
β_{02} = the average value of the logit for Y=2
β_{03} = the average value of the logit for Y=3
β_1 = the difference between the logit means for student PSAT percentile and the average logit
β_2 = the difference between the logit means for 8th grade student math class grade and the average logit.
β_3 = the difference between the logit means for student 2010 ERB quantitative test percentile and the average logit.
β_4 = the difference between the logit means for student 2010 ERB Math 1 & 2 test percentile and the average logit
β_4 = the difference between the logit means for student 2010 ERB Math 1 & 2 test percentile and the average logit
β_5 = the difference between the logit means for student 2009 ERB quantitative test percentile and the average logit
β_6 = the difference between the logit means for student 2009 ERB Math 1 & 2 test percentile and the average logit

Different discrimination and classification methods will be explored, including multiple population discriminant analysis and logistic discriminant analysis

Holdout analysis and visuals will be used to determine which variates will be used for the classification for the initial discrimination analysis.

There is missing data as not all the students have taken all of the examinations available or the score is not on file. Therefore, after the variates have been determined, if a student has no value for that variate, the student will be ignored for the analysis. There are no distinct patterns of the students with missing data.

We will also remove a student who entered into 9th grade in AP Calculus. This student is an exception to any classification rule and two courses ahead of any other student. Similar exceptional students who might join the school are most like likely exceptions to classification rules but would not be difficult to identify.

Using the 100 students in the data with reported 9th grade mathematics placement, prior proportions will be used based on the proportions of students in the four groups as shown in *Figure 1*.

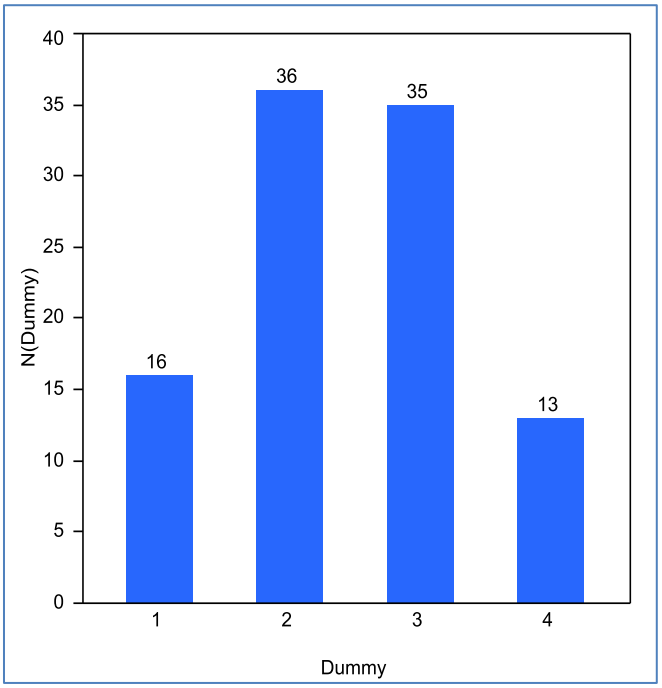


Figure 1 – prior proportions from sample

Results

For all of the variates, there is considerable overlap between groups. Using a single variate x_1 (PSAT) the boxplots in *figure 2* below show this overlap is considerable. The overlap is still present for two variates. In *figure 3,* x_2 (8^{th} mathematics class grade) and x_3(2010 ERB independent schools quantitative test percentile) demonstrate this. *Figure 2* and *figure 3* both indicate that the two populations that will have the highest error count will be dummy 2 (Geometry) and dummy 3 (Geometry Honors) as they both have two overlaps.

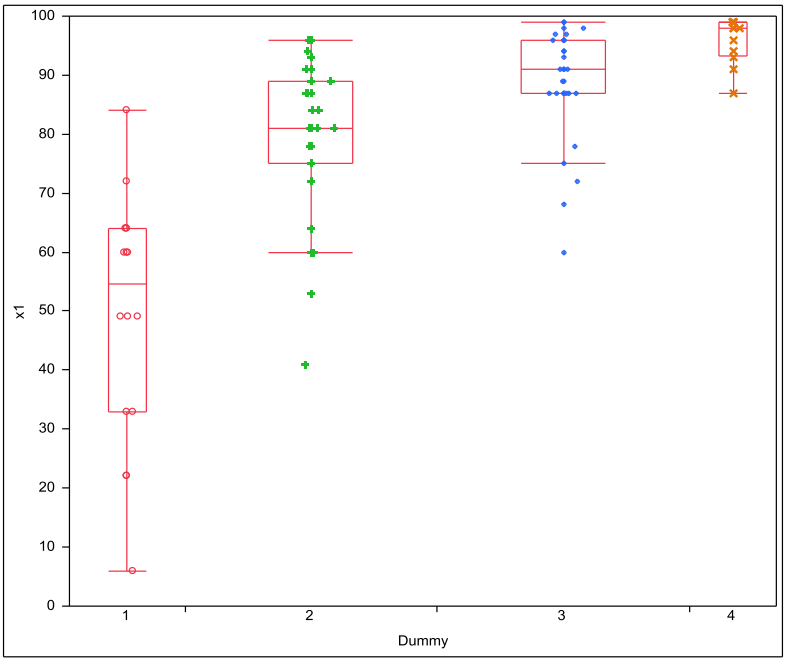


Figure 2 - One-way analysis of x_1 (PSAT) by Dummy Variables.

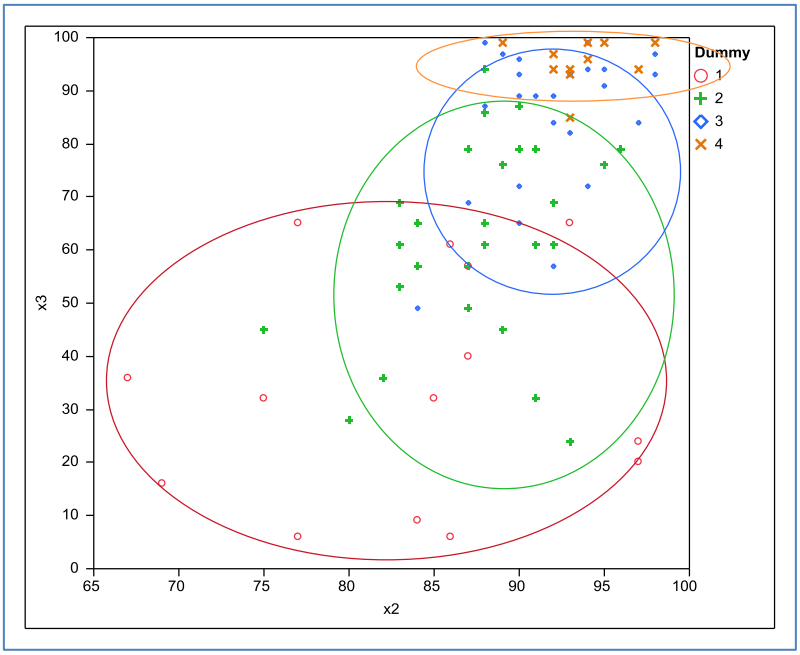


Figure 3 – Bivariate x_2(8^{th} Grade mathematics class grade) by x_3(2010 ERB independent schools quantitative test percentile)

The overlap is present for combinations of three variates. In *figure 4* below, variates x_1 (PSAT), x_5 (2009 ERB quantitative test percentile) and x_6 (2009 ERB Math 1 & 2 test percentile), normal contour ellipsoids (50% coverage) for the four groups have been used to demonstrate this. Note that the actual overlap will be greater.

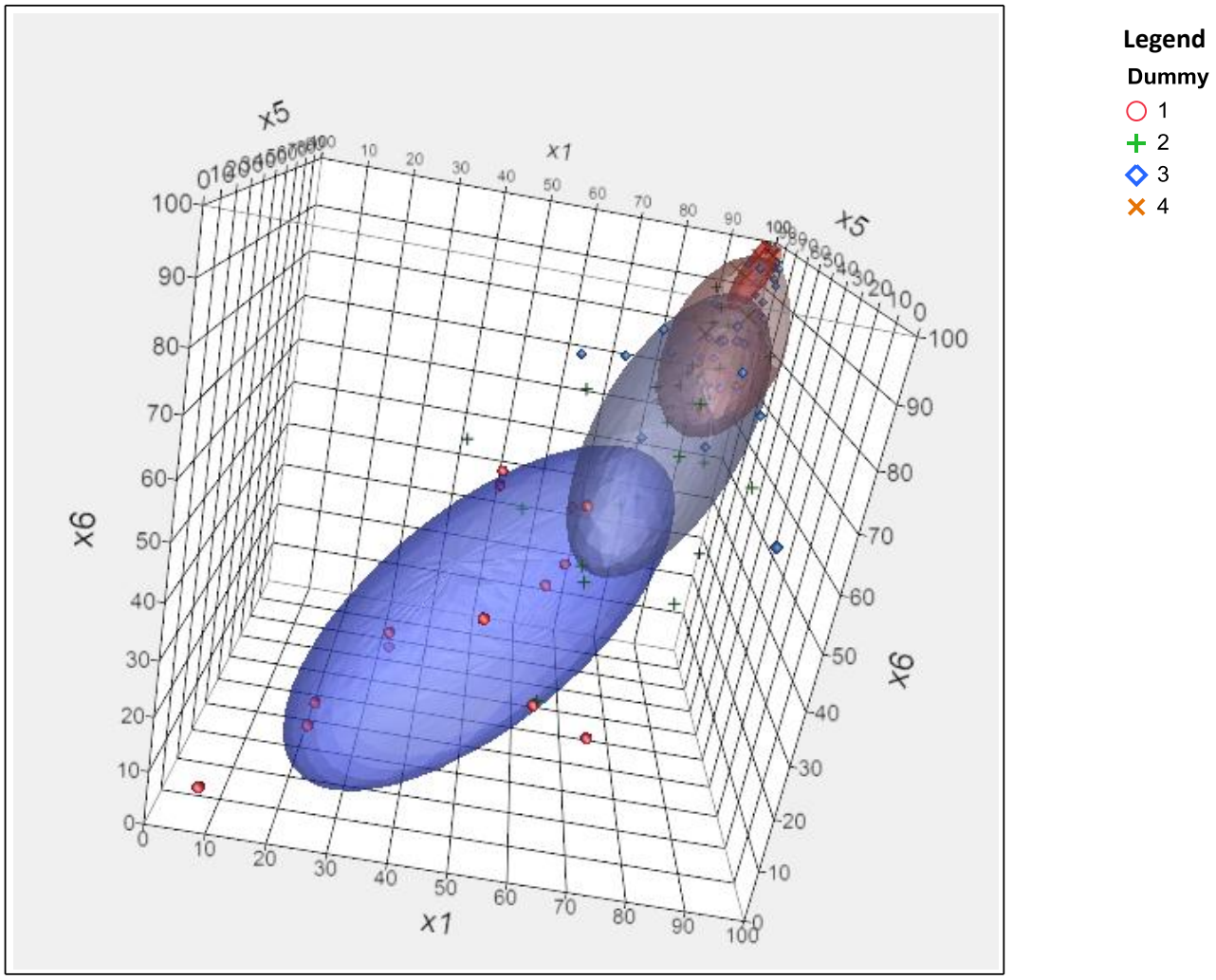


Figure 4 – 3D Scatterplot x_1, x_5 and x_6 with 50% Normal Contour Ellipsoids

Figure 5 shows the same as *figure 4*, but without the normal contour ellipsoids. It shows a particular intensity of group overlap where the variates have high values (80+).

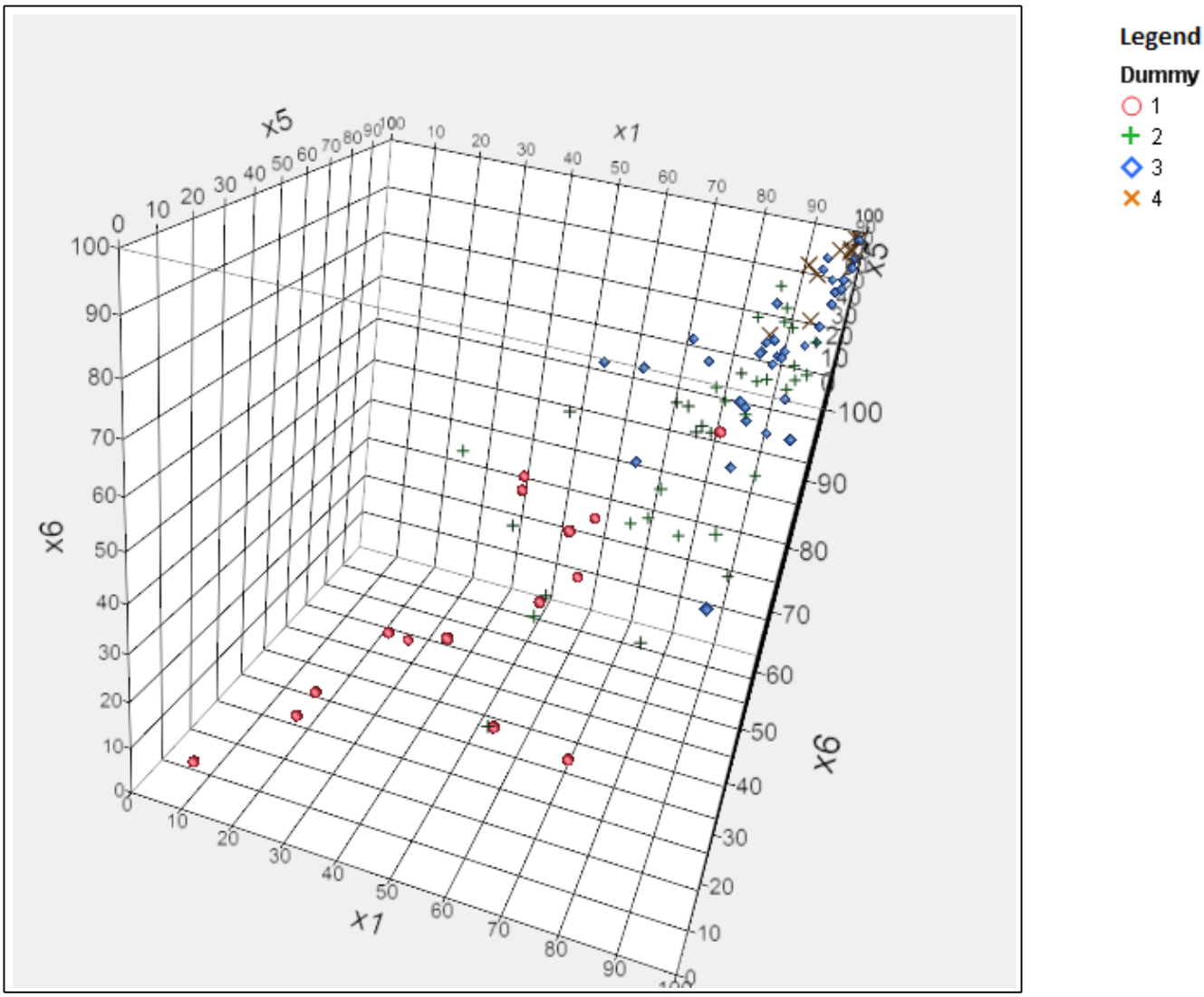


Figure 5 – 3D Scatterplot x_1, x_5 and x_6

There is a concern regarding the consistency of the ERB percentiles (x_3, x_4, x_5 and x_6). They are achievement tests and while year to year fluctuations are to be expected, the scores for an individual student should be roughly similar. Analyzing the difference in student scores between x_3 and x_5 (2010 and 2009 quantitative) and x_4 and x_6 (2010 and 2009 math 1 & 2) is shown in *figure 6*. While the overall means (3.27 and 1.24) are close to zero, the standard deviations are large (16.62 and 14.25) as are the ranges (106 and 102). This could indicate individual students with large fluctuations in percentiles year on year and as a result variates that might be poor indicators of groupings.

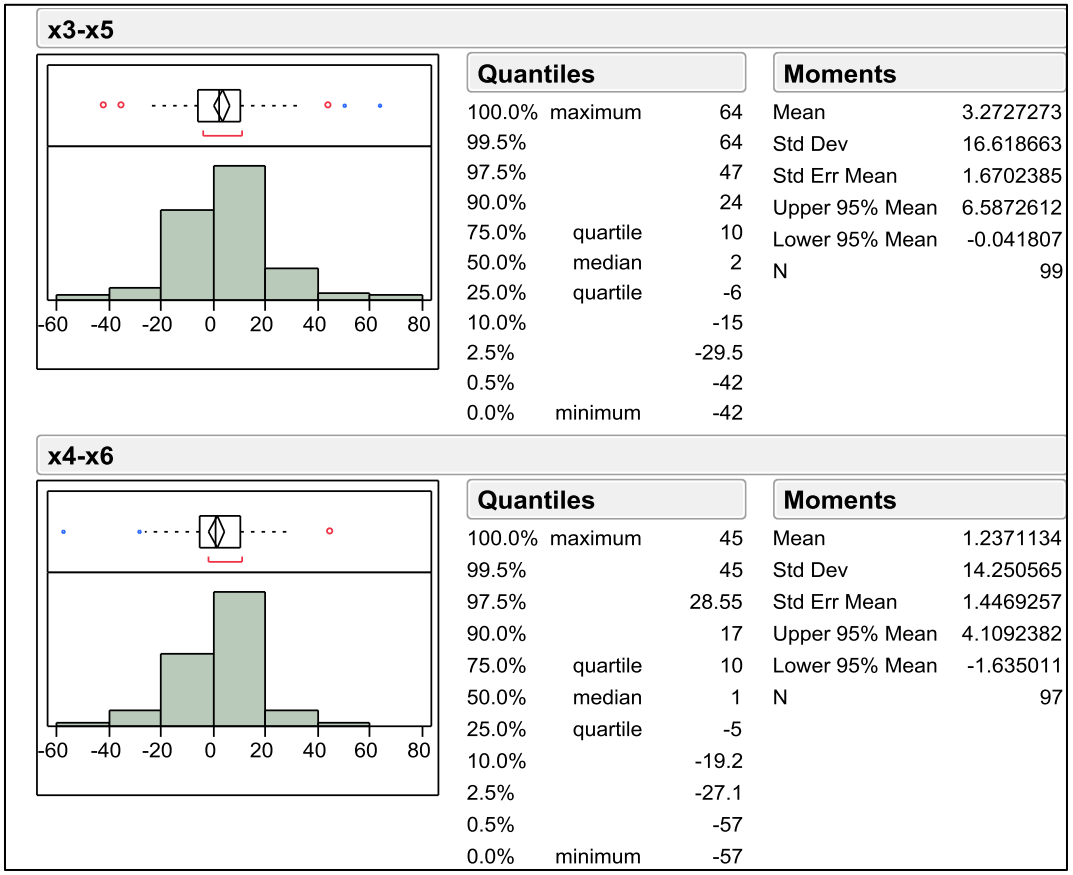


Figure 6 – Differences in year on year ERB scores (quantitative $x_3 - x_5$ and math 1 & 2 $x_4 - x_6$)

To determine which variates are good indicators, all were assessed using holdout of the other variates and all single variable misclassification rates being recorded (*figure 7*).

All the individual variates have relatively poor misclassification rates, and so none are removed. Pairs of variates were assessed for misclassification using holdout.

Variate	Misclassification Rate
x_1	0.4796
x_2	0.4875
x_3	0.5051
x_4	0.4536
x_5	0.4700
x_6	0.4200

Figure 7 – misclassification rate for single variates

The misclassification rates for pairs are better than for single variates, but still close (*figure 8*).

Variates	Misclassification Rate
x_1, x_2	0.4359
x_1, x_6	0.4286
x_3, x_4	0.4124
x_2, x_4	0.3766
x_5, x_6	0.4200

Figure 8 – misclassification rate – pairs of variates.

Variates	Misclassification Rate
x_2, x_5, x_6	0.3375
x_1, x_4, x_6	0.3474
x_3, x_4, x_5	0.3918

The misclassification rates for pairs are better than for single variates, but still close. Some variate triples were assessed but produced little to indicate useful differences in misclassification rates (*figure 9*).

Figure 9 – misclassification rate – triples.

No distinct economies of variates have presented themselves. Therefore, it is reasonable to classify using all six variates.

When running this analysis, 25 students are excluded due to missing scores. This probably does not create under coverage bias as the students are left out due to mostly incomplete files. This leaves us with 75 students which should be sufficient.

Class Level Information					
Dummy	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	13	13.0000	0.173333	0.173333
2	_2	28	28.0000	0.373333	0.373333
3	_3	24	24.0000	0.320000	0.320000
4	_4	10	10.0000	0.133333	0.133333

Prior probabilities are used proportionate to the classifications the teachers made of placing students into their respective mathematics classes (*Figure* 10) and dummy variables were assigned for the groups: Algebra I (Dummy 1), Geometry (Dummy 2), Geometry Honors (Dummy 3) and Algebra II Honors (Dummy 4).

Figure 10 – Frequency and Priors

However, we have a good reason to believe that we have a multivariate normal assumption violation (*figure 11*). All the variates demonstrate a lack of univariate normality (p-value<0.0001). Additionally, Mardia and Henze-Zirkler Tests reject multivariate normality (p-value<0.0001).

The explanation of this seems to be in the characteristics of the student body. Exploring the probability density function of variate x_1 (PSAT percentile – *figure 12*) indicates a left-skew. In other words, at this independent school the majority of students have high aptitude/attainment. All the variates follow this pattern.

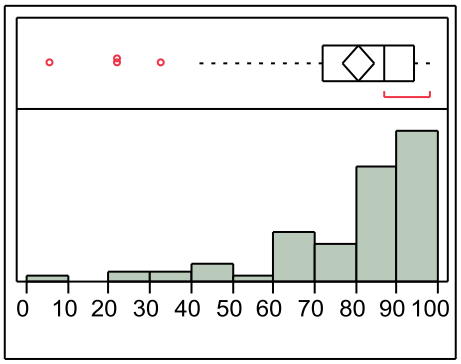


Figure 12 – distribution of variate x_1(PSAT percentiles)

Normality Test			
Equation	Test Statistic	Value	Prob
x1	Shapiro-Wilk W	0.80	<.0001
x2	Shapiro-Wilk W	0.91	<.0001
x3	Shapiro-Wilk W	0.90	<.0001
x4	Shapiro-Wilk W	0.90	<.0001
x5	Shapiro-Wilk W	0.91	<.0001
x6	Shapiro-Wilk W	0.89	<.0001
System	Mardia Skewness	213.5	<.0001
	Mardia Kurtosis	8.28	<.0001
	Henze-Zirkler T	19.58	<.0001

Figure 11 – testing multivariate normality

An alternative is the logistic discriminant analysis. The whole model test (*figure 13*) which tests the regressors as a whole, indicates significance (p-value<0.001).

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	46.032983	6	92.06597	<.0001*
Full	51.833414			
Reduced	97.866398			
RSquare (U)		0.4704		
Observations (or Sum Wgts)		75		
Converged by Objective				

figure 13 – whole model test

Exploring the analysis using variates $x_1 - x_6$, parameter estimates are given (*figure 14*). The outcome for our logit probabilities used 75 students and misplaced 24, giving a misclassification rate of 0.32 (*figure 15*). The logit probabilities and student group placements are given in appendix 1.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[1]	18.6436519	5.882222	10.05	0.0015*
Intercept[2]	24.3481319	6.482242	14.11	0.0002*
Intercept[3]	27.616954	6.7081383	16.95	<.0001*
x1	-0.0741904	0.0347207	4.57	0.0326*
x2	-0.1034826	0.0631324	2.69	0.1012
x3	-0.0172225	0.0268978	0.41	0.5220
x4	-0.0159573	0.0296062	0.29	0.5899
x5	0.02730834	0.0194721	1.97	0.1608
x6	-0.1042558	0.0347628	8.99	0.0027*

Figure 14 – Parameter estimates

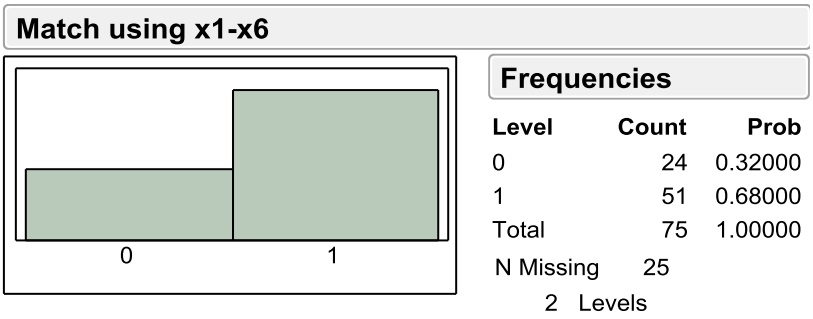


Figure 15 – misclassification (=0), correctly classified (=1)

From our holdout analysis, it is reasonable to assume that no single, pair or triple variates are good indicators of groupings, it is reasonable to use all six variates in our model. However, there is a serious concern of multicollinearity of the variates. It might be a better option to run the analysis using only one variate (or combine them in some method).

Exploring the misclassification by group, we find the following (*figure 16*). The table confirms what we might have suspect would happen from *figures 2 – 4*, that the groups bordered by two other groups (2 and 3) would have the highest misclassification rate.

Contingency Table

Most Likely Dummy x1-x6

Dummy Count	1	2	3	4	
1	10	3	0	0	13
2	3	19	6	0	28
3	0	7	16	1	24
4	0	0	4	6	10
	13	29	26	7	75

Figure 16 – Contingency table showing misclassifications by group

Exploring the residuals shows patterns of concern with groups 1 and 4 (*figure 17*). This is possibly explained by the rest of the data 'pulling' the expected group values for the data in the residuals in group 1 higher and lower for the data in the residuals in group 4.

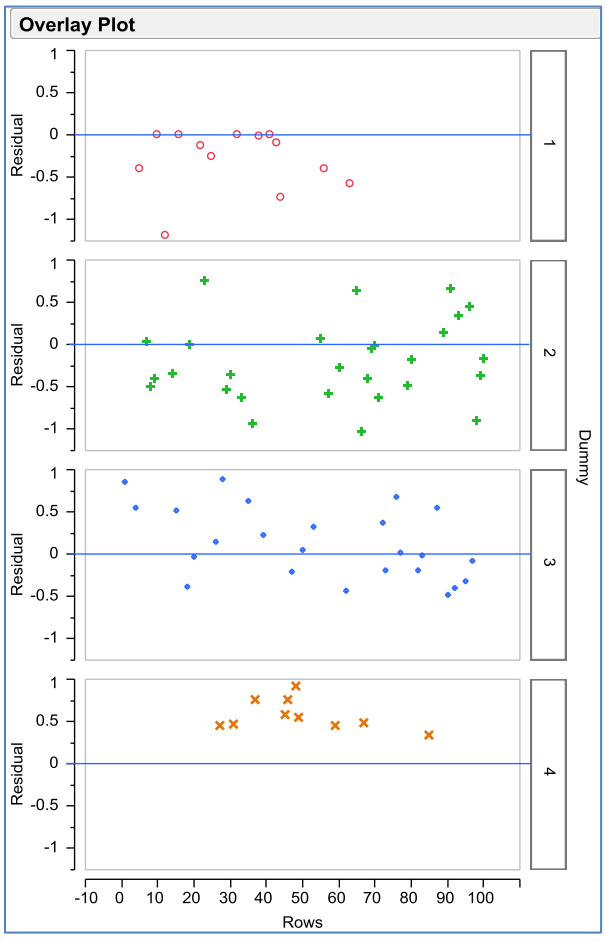
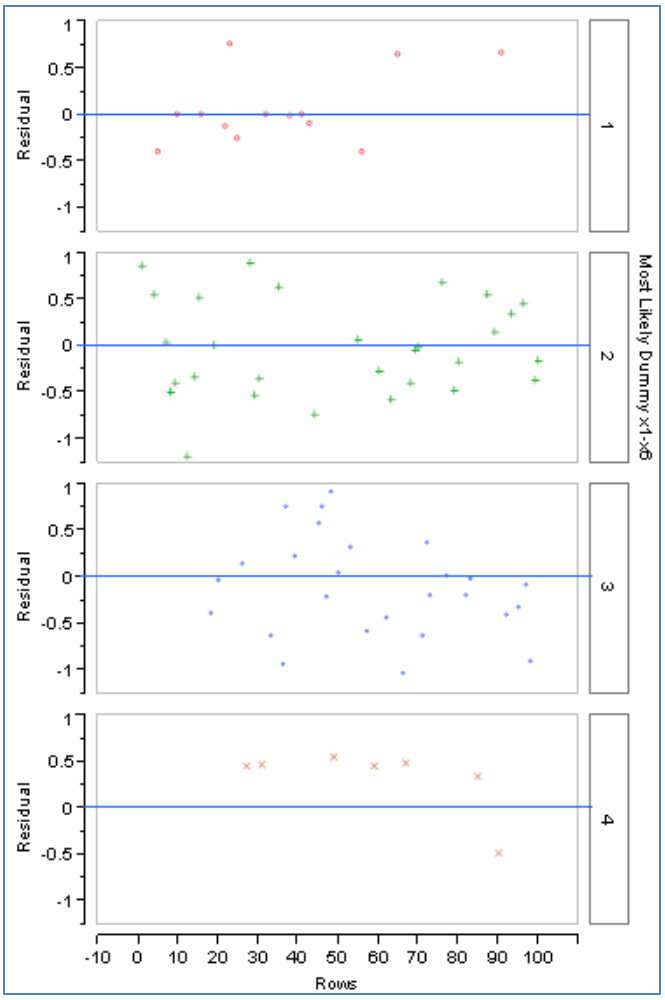


Figure 17 – residual plots by Dummary groups



If we now look at the residuals by their predicted groups for the data (*figure 18*), we find a more random pattern other than for group 4, which presents an unusual pattern.

Figure 18 – residual plots by predicted group classification

Exploring the new classifications with x_1, x_5 and x_6 (*Figure 19*) and using 50% normal contour ellipsoids to explore the new groupings, when comparing with *figure 4*, the ellipsoids are less overlapping. However, when looking at where the data values are placed, they are often in the wrong ellipsoid a long way from the ellipsoid they should be near to or contained within.

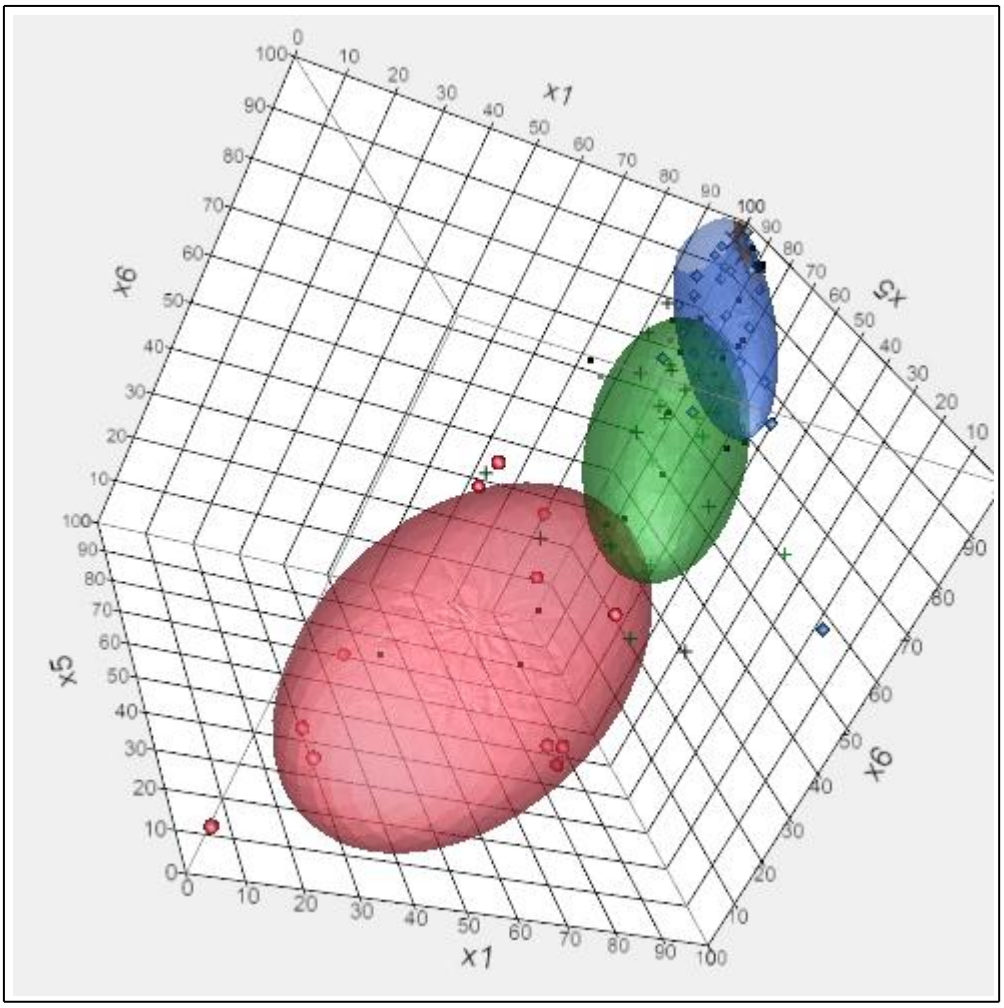


Figure 19 - 3D Scatterplot x_1, x_5 and x_6 with 50% Normal Contour Ellipsoids of predicted groups

Conclusions

The experiment produced an error rate of 0.32. This tells us that if we use our variates to place the students exclusively, the teachers who would usually place the students would possibly question one-third of the placements. Clearly, this is too high.

There are concerns with some of the variates, as there is inconsistency of the student percentiles in the standardized testing, particularly the ERBs ($x_3 - x_6$). This was demonstrated in *figure 6*. The students at this particular independent school take the ERBs examinations at the end of the year, but crucially the scores and percentiles do not affect their future mathematics placement and they are not included on the student transcript sent to colleges. The scores are used for school accreditation and a measurement of student achievement for school purposes only. Therefore, motivation is a concern in a 'zero stake' exam. A solution to this might have been to take the maximum ERB percentile (for quantative, say) over a series of years, but the exam is an assessment of year on year achievement rather than aptitude so this is difficult to justify. The PSAT percentiles taken in grade 10 were also not consistent indicators of groupings. Considering that the PSAT exam includes algebra II content and

some students have either taken algebra II, are taking algebra II or have not yet taken algebra II, this is not a level playing field. There are some students who have scored consistently well or consistently poorly in all the exams, and these students are relatively easy to place with or without a classification analysis. They are usually placed in dummy groups 1 or 4.

The classification analysis of a student can be used as a potential indicator if other variables are considered. Classification analysis places individual experimental units and so we are so much more aware of errors this creates compared with most other statistical inference. Therefore, using the classification as an initial guide for mathematics class placement would be beneficial for an individual student provided more information was then examined. This would include teacher recommendation as a comment on the drive, commitment and curiosity a student has with their learning. The classification analysis is also a useful information tool for parents and students. If a parent or student objects to a mathematics placement, looking at the probabilities of the logistic discriminant analysis (see appendix I) would help understand how much of a clear decision the placement was. For example, StudentID #10 is a simple decision for placement in Algebra I (logit probability 0.99973), while for StudentID #4 it is a closer decision between Geometry and Geometry Honors (logit probabilities 0.571 and 0.397 respectively).

The assumptions for the logistic regression remain questionable. There are patterns with the group 4 residuals and multicollinearity is a concern.

Appendix 1 – Results of Logistic Discriminant Analysis

StudentID	Course	Dummy	Most Likely Dummy	Match (=1)	Prob(1)	Prob(2)	Prob(3)	Prob(4)
1	Geometry Honors	3	2	0	0.017084	0.82209	0.153586	0.00724
2	Geometry	2	.	.	.	.	.	.
3	Geometry Honors	3	.	.	.	.	.	.
4	Geometry Honors	3	2	0	0.004502	0.571367	0.39687	0.027261
5	Algebra I	1	1	1	0.60934	0.388529	0.00205	8.13E-05
6	Algebra II Honors	4	.	.	.	.	.	.
7	Geometry	2	2	1	0.0721	0.886793	0.039478	0.001629
8	Geometry	2	2	1	0.003871	0.534604	0.429941	0.031583
9	Geometry	2	2	1	0.005259	0.608234	0.363095	0.023411
10	Algebra I	1	1	1	0.999733	0.000266	8.56E-07	3.39E-08
11	Geometry	2	.	.	.	.	.	.
12	Algebra I	1	2	0	0.013347	0.789072	0.188298	0.009282
13	Geometry Honors	3	.	.	.	.	.	.
14	Geometry	2	2	1	0.006824	0.666658	0.308404	0.018114
15	Geometry Honors	3	2	0	0.004077	0.547262	0.418627	0.030035
16	Algebra I	1	1	1	0.999169	0.000828	2.66E-06	1.05E-07
17	Geometry Honors	3	.	.	.	.	.	.
18	Geometry Honors	3	3	1	0.000165	0.047036	0.518376	0.434423
20	Geometry	2	2	1	0.058932	0.890563	0.048486	0.00202
21	Geometry Honors	3	3	1	0.000596	0.151181	0.672861	0.175363
22	Algebra I	1	.	.	.	.	.	.
24	Algebra I	1	1	1	0.88171	0.117844	0.00043	1.70E-05
25	Geometry	2	1	0	0.763844	0.235127	0.00099	3.92E-05
26	Geometry	2	.	.	.	.	.	.
27	Algebra I	1	1	1	0.750859	0.248037	0.001062	4.21E-05
28	Geometry Honors	3	3	1	0.001116	0.250021	0.646961	0.101902
29	Algebra II Honors	4	4	1	9.63E-05	0.028014	0.403746	0.568143
30	Geometry Honors	3	2	0	0.021813	0.848224	0.124312	0.005652
31	Geometry	2	2	1	0.003455	0.506556	0.454721	0.035268
32	Geometry	2	2	1	0.006542	0.657526	0.317047	0.018885
33	Algebra II Honors	4	4	1	0.000102	0.029747	0.417231	0.55292
34	Algebra I	1	1	1	0.999189	0.000809	2.60E-06	1.03E-07

35	Geometry	2	3	0	0.002401	0.417071	0.530501	0.050026
36	Geometry	2	.	.	.	.	.	.
37	Geometry Honors	3	2	0	0.006005	0.638582	0.334863	0.02055
38	Geometry	2	3	0	0.000857	0.203859	0.666499	0.128785
39	Algebra II Honors	4	3	0	0.000276	0.076172	0.60863	0.314922
40	Algebra I	1	1	1	0.989559	0.010406	3.38E-05	1.34E-06
41	Geometry Honors	3	3	1	0.001462	0.303819	0.615028	0.079692
42	Geometry Honors	3	.	.	.	.	.	.
43	Algebra I	1	1	1	0.999982	1.76E-05	5.64E-08	2.23E-09
44	Geometry	2	.	.	.	.	.	.
45	Algebra I	1	1	1	0.911208	0.088468	0.000312	1.24E-05
46	Algebra I	1	2	0	0.273737	0.717503	0.008424	0.000336
47	Algebra II Honors	4	3	0	0.000151	0.043177	0.5001	0.456573
48	Algebra II Honors	4	3	0	0.000277	0.076417	0.609133	0.314173
49	Geometry Honors	3	3	1	0.000311	0.085128	0.625136	0.289425
50	Algebra II Honors	4	3	0	0.000478	0.125141	0.664983	0.209398
51	Algebra II Honors	4	4	1	0.000135	0.038692	0.476112	0.485061
52	Geometry Honors	3	3	1	0.000765	0.186019	0.671094	0.142122
53	Geometry	2	.	.	.	.	.	.
54	Algebra II Honors	4	.	.	.	.	.	.
55	Geometry Honors	3	3	1	0.002061	0.380635	0.559475	0.057829
56	Geometry Honors	3	.	.	.	.	.	.
57	Geometry	2	2	1	0.096239	0.873429	0.029143	0.001189
58	Algebra I	1	1	1	0.605885	0.391953	0.00208	8.24E-05
59	Geometry	2	3	0	0.002839	0.45799	0.496549	0.042623
60	Geometry Honors	3	.	.	.	.	.	.
61	Algebra II Honors	4	4	1	9.54E-05	0.02776	0.401699	0.570446
62	Geometry	2	2	1	0.008838	0.719192	0.257955	0.014016
65	Geometry Honors	3	.	.	.	.	.	.
66	Geometry Honors	3	3	1	0.000143	0.040968	0.488683	0.470207
67	Algebra I	1	2	0	0.432745	0.562907	0.004181	0.000166
68	Geometry Honors	3	.	.	.	.	.	.
69	Geometry	2	1	0	0.640959	0.357178	0.001791	7.10E-05
70	Geometry	2	3	0	0.000594	0.150892	0.672825	0.175689
71	Algebra II Honors	4	4	1	0.000106	0.030672	0.424129	0.545093
72	Geometry	2	2	1	0.005285	0.60935	0.362064	0.023302

73	Geometry	2	2	1	0.036325	0.88248	0.077844	0.003351
74	Geometry	2	2	1	0.048645	0.890194	0.058688	0.002473
76	Geometry	2	3	0	0.0024	0.41691	0.530633	0.050058
77	Geometry Honors	3	3	1	0.00235	0.411907	0.534687	0.051056
78	Geometry Honors	3	3	1	0.00032	0.087339	0.628658	0.283683
79	Geometry Honors	3	.	.	.	.	.	.
80	Algebra I	1	.	.	.	.	.	.
81	Geometry Honors	3	2	0	0.007162	0.676954	0.298617	0.017266
82	Geometry Honors	3	3	1	0.000689	0.170702	0.67323	0.155379
83	Geometry	2	.	.	.	.	.	.
84	Geometry	2	2	1	0.003959	0.540087	0.42505	0.030904
86	Geometry	2	2	1	0.014768	0.803408	0.17344	0.008385
87	Geometry Honors	3	.	.	.	.	.	.
88	Geometry Honors	3	3	1	0.000334	0.09092	0.633946	0.274799
89	Geometry Honors	3	3	1	0.000622	0.156795	0.673375	0.169208
90	Geometry Honors	3	.	.	.	.	.	.
91	Algebra II Honors	4	4	1	5.95E-05	0.017496	0.301995	0.680449
92	Algebra I	1	.	.	.	.	.	.
93	Geometry Honors	3	2	0	0.004334	0.562167	0.405205	0.028294
94	Geometry	2	.	.	.	.	.	.
95	Geometry	2	2	1	0.159025	0.823665	0.016641	0.00067
96	Geometry Honors	3	4	0	0.000122	0.035241	0.455326	0.509311
97	Geometry	2	1	0	0.667794	0.330552	0.001591	6.31E-05
98	Geometry Honors	3	3	1	0.000156	0.044614	0.507146	0.448084
99	Geometry	2	2	1	0.349311	0.644522	0.005931	0.000236
100	Algebra II Honors	4	.	.	.	.	.	.
101	Geometry Honors	3	3	1	0.000215	0.060397	0.568425	0.370964
102	Geometry	2	2	1	0.454872	0.541152	0.003824	0.000152
103	Geometry Honors	3	3	1	0.000495	0.12897	0.666802	0.203733
104	Geometry	2	3	0	0.000945	0.220164	0.660693	0.118198
105	Geometry	2	2	1	0.006015	0.63894	0.334528	0.020517
106	Geometry	2	2	1	0.017062	0.821936	0.153753	0.007249

Appendix 2 - Sources

1. *ERB*. Web. 19 Nov. 2010. <http://erblearn.org/>.