Identifying Students at Risk

Dr. Wolfgang Rolke
University of Puerto Rico - Mayaguez

November 10, 2014

Abstract

We use admissions data to calculate estimates of the probability for a student to return for the second year, for a student to graduate within a reasonable time span as well as their GPA after the freshman year. This information can be used to very early identify students at risk of failure at UPRM.

Contents

1 Introduction **1**

2 Using IGS **1**

3 Other Variables **2**

4 School GPA **2**

5 Pre-Scaling **2**

6 IES (Índice Estatístico de Solicitud) **3**

7 Least Squares Regression **5**

8 IES vs IGS **5**

9 Simplifying IES **6**

10 Identifying Students at Risk **7**

11 Logistic Regression **7**

12 Additional Variables **9**

13 Statistical Model for Logistic Regression **10**

14 The Coefficients in a Logistic Regression Model **10**

15 Performance of these Models **11**

16 More Detailed Studies **12**

17 Information on Freshman Class **15**

18 Implementation **17**

19 Conclusions **17**

20 Appendix: Detailed Information on Regression Fits **18**

List of Figures

1	Artificial example of negative regression coefficients despite positive correlations.	4
2	Students with a low IGS but not IES (in red), and vice versa (in blue).	6
3	Artificial examples of logistic regression curves.	7
4	Logistic regression curves for predicting return for the second year using IES and IGS, respectively.	8
5	Logistic regression curves for predicting graduation at 150% using IES and IGS, respectively.	9
6	Boxplots of estimated return percentages and of rate of Graduation by their percentiles. Blue lines are true percentages	12
7	Residual vs Fits plot for IES	19

List of Tables

1	Coefficients for logistic regression	10
2	Logistic Regression Coefficients for Predicting Return for Second Year. Overall for all students together as well as broken down by Faculty	13
3	Logistic Regression Coefficients for Predicting Return for Second Year. Overall for all students together as well as broken down by Orientation	13
4	Logistic Regression Coefficients for Predicting Graduation. Overall for all students together as well as broken down by Faculty	14
5	Logistic Regression Coefficients for Predicting Graduation. Overall for all students together as well as broken down by Orientation	14
6A	Percentiles for New Students	15
6B	Percentiles for New Students by Faculty	15
6C	Percentiles for New Students by Orientation	15
7A	Percentiles for New Students	16
7B	Percentiles for New Students by Faculty	16
7C	Percentiles for New Students by Orientation	16
A 1	Information on fit of Return for Second Year	18
A 2	Information on fit for Graduating at 150%	19
A 3	Correlations Between Predictors	20
A 3	Correlations Between Predictors	20
A 3	Correlations Between Predictors	21
A 3	Correlations Between Predictors	21

1 Introduction

UPR, like many other Universities and Colleges, has a sizable dropout rate, that is, students do not return for the second year or later do not graduate. While there are many reasons for a student to leave the University before graduating, at least some of them might be addressed by a properly designed intervention system. Offering special counseling, tutoring or other services would hopefully improve the situation. Crucial for this effort is the ability to identify those students with the highest risk of failure. In this report we will investigate ways to do this at the earliest possible moment, namely the first day of classes, and based on the data collected from the students during the admissions process.

This work originated in the meetings of a small group during the Spring semester of 2013−2014. The group was brought together by Dr. Héctor Jiménez, then Director of the Office of Investigación Institucional y Planificación. The other members of the group were Dra. Damaris Santana, Dr. Raul Macchiavelli and Dra. Bernadette Delgado Acosta. The goal was originally to study how well the IGS score predicts success at UPRM with the aim to eventually replace it with a superior measure. Eventually the focus shifted to the question on how admissions data could be used to identify students at risk of dropping out before graduating.

It should be noted that this analysis is based on admissions data for UPR Mayaguez only. The methodology described here, though, is a general one and could be used at the other Recintos as well to the University system as a whole.The exact numbers such as the coefficients will then of course change.

2 Using IGS

The current admissions system is solely based on the IGS ("Indice General de Solicitud") score of a student calculated as

$$IGS = 0.5\cdot(GPA*100)+0.25(AptVerbal-200)*\frac{2}{3}+0.25(AptMatem-200)*\frac{2}{3}$$

There are a number of issues with this formula. First of all, it uses only three of the variables available at the time of admission. Secondly it is not clear why this weighting scheme is used. In fact, we can write the formula also as

$$IGS = 50\cdot GPA + \frac{1}{6}(AptVerbal + AptMatem - 400)$$

because Aptverbal and AptMatem are between 200 and 800, $AptVerbal + AptMatem - 400$ is between 0 and 1200, and so $\frac{1}{6}(AptVerbal+AptMatem-400)$ is between 0 and 200, as is $50GPA$, so in the formula GPA accounts for 50% and the other two variables for 25% each. It is not clear why this should be so.

Thirdly, it does not distinguish between a student who went to an academically rigorous High School and whose high GPA therefore indicates a likely good student, and another one who might have the same or an even higher GPA but who attended a school with much lower standards.

3 Other Variables

On the admissions form there are a number of other variables that could be used for predicting performance as well. They are AprovEspanol, AprovIngles, AprovMatem, which have the same scale as AptVerbal and AptMatem ($200-800$), as well as Niv_Avanzado_Espanol, Niv_Avanzado_Ingles, Niv_Avanzado_Mate_I and Niv_Avanzado__Mate_II. These variables indicate whether a student has taken some advanced exams. If so the score is from $1-5$. If no exam was taken the score is 0.

4 School GPA

As pointed out before, just using the High School GPA is problematic because different schools have quite different academic standards. Even worse, the formula for IGS essentially penalizes students who have gone to tough High Schools, were a high GPA is more difficult to achieve. In our analysis we will therefore introduce a new variable designed to measure the quality of a High School, and therefore improve our understanding of the true meaning of a students High School GPA. This is done as follows:

1) Take all the students from the same High School who have finished their first year at UPRM and calculate their mean Freshman GPA.

2) For those same students calculate their mean High School GPA.

3) Divide the first by the second

As examples consider the two most extreme cases: Students from High School #3943 had a mean High School GPA of 3.8 and a mean Freshman GPA of 1.3, so their School GPA is $1.3/3.8 = 0.34$. On the other end of the scale, students from High School #2973 had a mean High School GPA of 3.2 and a mean Freshman GPA of 3.0, so their School GPA is $3.0/3.2 = 0.94$. This clearly illustrates the usefulness of this variable: According to IGS students from #3943 should be better than those from #2973 (3.8 vs 3.2 GPA) but in reality once they get to UPRM they are doing much worse (1.3 vs 3.0 Freshman GPA). Our new variable shows this very clearly (0.34 vs 0.94).

This procedure is used for all those High Schools with at least 20 students, for all others a student is assigned the overall average (0.745).

5 Pre-Scaling

The numerical values of the predictors differ by several orders of magnitude, so in order to allow a direct comparison of the coefficients we will standardize each of them by subtracting their mean and dividing by their standard deviation. This does have the undesirable effect to make the formula somewhat more complicated and we might ultimately decide to eliminate this step. The results discussed in this report do not depend on this scaling.

6 IES (Ińdice Estatístico de Solicitud)

We now use these variables to derive an equation useful for predicting the standard measure for (early) success in College, namely the Freshman GPA. Based on data for the 25495 students accepted from 2002 to 2013 and using the standard method of least squares we find the equation

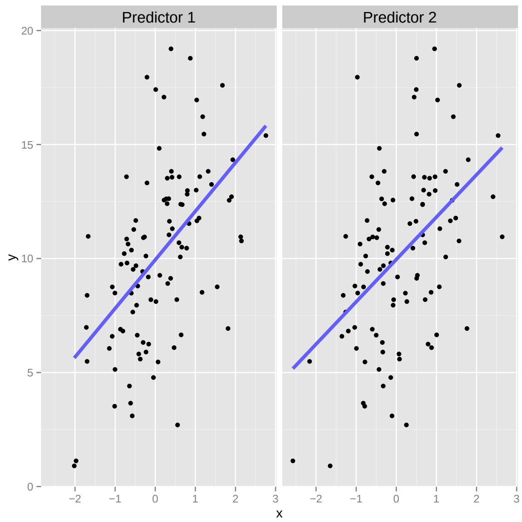
IES =	+	0.352	GPA.Escuela.Superior
	+	0.210	SchoolGPA
	+	0.056	Aprov.Espanol
	+	0.041	Aptitud.Verbal
	-	0.036	Aptitud.Matem
	+	0.036	Niv_Avanzado_Mate_II
	+	0.028	Aprov.Ingles
	+	0.022	Niv_Avanzado_Espa
	+	0.021	Aprov.Matem
	-	0.015	Niv_Avanzado_Ingles
	-	0.0005	Niv_Avanzado_Mate_I

First off, clearly the High School GPA is the single best predictor for success, followed closely by the School GPA. Notice that the other two variables used in the calculation of IGS, Aptitud.Ingles and Apitud.Matem, have weights considerably smaller than High School GPA.

One oddity of this equation is that some of the coefficients are negative. This seems to contradict the fact that individually all of the variables are positively correlated with the response. It seems to suggest that a student with a lower Aptitud.Matem score would be expected to have a higher Freshman GPA!

The problem here is one quite common in multiple regression called multicolinearity. What we need to consider are not just the correlations of the predictors and the response but also the correlations between the predictors, some of which are sizable. For example (not surprisingly) the correlation between Aptitud.Matem and Aprov.Matem is 0.82. Why this matters is illustrated by the following artificial example. Here we have a data set with a response and two predictors. Figure 1 shows the scatter-plots of the predictors vs. the response, together with the least squares regression lines:

Figure 1: Artificial example of negative regression coefficients despite positive correlations.



Clearly both predictors have a strong positive relationship with the response, and correspondingly the slopes of the least squares regression lines are positive as well. But now calculating the multiple regression equation we find

$$\text{Response} = 9.95 + 2.34 \text{ Predictor 1} - 0.23 \text{ Predictor 2}$$

and again we have a negative coefficient! Here the correlation between the predictors is 0.91.

This phenomena has been observed many times in real life cases. In Statistics it is a version of the well known Simpson's paradox. In the behavioral sciences it is often referred to as Positive Net Suppression. Its major consequence is that interpreting the coefficients in a multiple regression is a difficult and dangerous thing to do. Thankfully in our case the ultimate goal is not understanding the model but simply predicting success. Moreover, avoiding any negative coefficients would come at a steep price: the best possible model with no negative coefficients is considerably worse than the full model.

Some of the coefficients in the model are very small, so one might consider simplifying it by eliminating some variables. It can be shown, though, that in fact all variables are statistically significant. Moreover, because all the data is already available in electronic form and because prediction rather than interpretation is our goal there is not really any reason to simplify the model.

7 Least Squares Regression

The statistical method used in the calculation of the formula for IES is called least squares regression. This is one of the best known and most widely used methods in Statistics, dating back all the way to the famous rediscovery of the asteroid Ceres by Carl Fridrich Gauss in 1801. It has been in used in practically every area of science since.

The basic idea of this method is to find the straight line that most closely follows a cloud of dots. "most closely" here means that it minimizes the sum of squares of the distances between the observed and predicted response values.

The main issue when applying this method is to insure that the assumptions of homoscatasticity and linearity are satisfied. The main tool for this is a plot of the residuals vs the fitted values, as shown in the appendix. In our case these assumptions are quite correct.

8 IES vs IGS

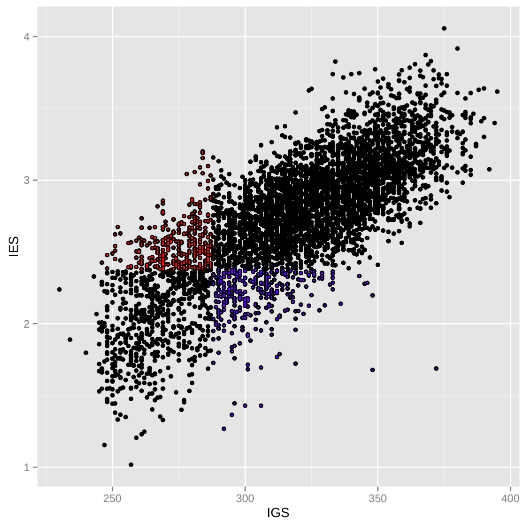
What can we gain by using IES instead of IGS? A general measure of the quality of the fit in a regression is R^2, the coefficient of determination. R^2 has the interpretation of the variation in the response explained by the predictors, so a value close to 0% means that none of the predictors is useful whereas a value close to 100% would mean a model capable of predicting the response almost perfectly. For our data we find

$$\begin{array}{ll} \text{IGS} & R^2 = 18.7\% \\ \text{IES} & R^2 = 30.4\% \end{array}$$

so IES has a well over 60% higher explanatory power than IES. The adjusted R^2 are equal to the standard R^2.

Let's consider another outcome measure, namely graduation. Here we will consider students who have been at the University long enough to graduate within 150% of the official study time of their major, so students enrolled in a 4 year program are counted as graduated if they have done so within 6 years. In figure 2 we have the plot of IES vs IGS and we focus on two groups: those in the bottom 20^{th} percentile of IES but not in the bottom 20^{th} percentile of IGS (drawn in blue) and vice versa (in red).

Figure 2: Students with a low IGS but not IES (in red), and vice versa (in blue).



Now of those with a low IES score but a not so low IGS score (blue) 20% graduated anyway, but of those with a low IGS score but a not so low IES score (red) 43% graduated. So IES is also a better predictor of graduation than IGS.

9 Simplifying IES

In the next section we will change our focus to the problem of predicting success in College. At some point in the future, though, the University might consider using IES for the purpose of admissions. In that case it might be desirable to have a simpler formula than the one in section 6, maybe even a formula that one can calculate by hand. For this we will do the following: first we will no longer pre-scale the data, because this was done mostly to make the coefficients size comparable. Next we multiply the SchoolGPA by 4 so it has the usual scale of GPAs. We also remove a number of the variables that are only marginally useful for predicting the Freshman GPA, which will also have the nice effect of removing any negative correlations. This then leads to the following formula

IES* =		1.05	GPA.Escuela.Superior	*	100
	+	0.87	SchoolGPA	*	100
	+	0.75	(Aprov.Espanol	-	200)
	+	0.57	(Aptitud.Verbal	-	200)
	+	0.2	(Aprov.Ingles	-	200)
	+	0.9	(Aprov.Matem	-	200)

This model has an $R^2 = 30.1\%$ and is therefore just about as good as the one in section 6.

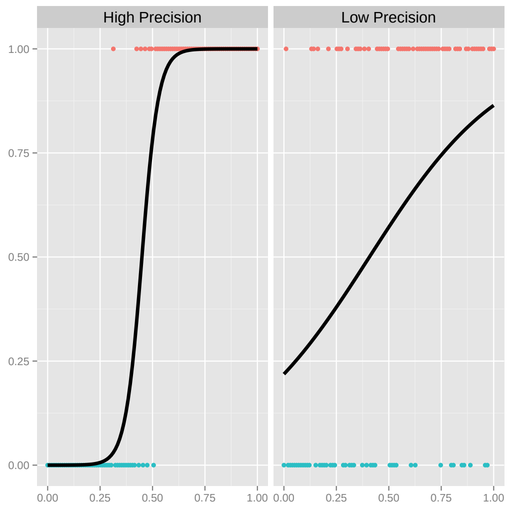
10 Identifying Students at Risk

In the last section we already considered the graduation rate as an outcome measure. Now we will focus on this, as well as a second one, namely the return rate for the second year. Again we will derive a formula for predicting these outcomes. Of course now they are binary (Yes-No), so what our formula is going to yield is the probability that a student returns for the second year or that the student graduates at 150% of the official time. The statistical technique for this type of problem is known as logistic regression.

11 Logistic Regression

In this type of problem we have a binary outcome measure (coded as 0 and 1) and one or more predictors. Two examples of logistic regression curves are shown in figure 3.

Figure 3: Artificial examples of logistic regression curves.



In the left panel the dots at the bottom and the top are fairly well separated, with almost all dots corresponding to small values of the predictor equal to 0 (in blue) and almost all dots with a high value of the predictor equal to 1 (in red). This results in a curve that stays close to 0 (and therefore a very small probability for a "1"), then rises sharply up and staying there until the right side of the graph (and therefore predicting a "1" with a high probability).

In contrast in the panel on the right the dots on the bottom and the top both go from almost the left to the right, indicating that a ”1” is quite likely even if the predictor is small, and vice versa. This results in a curve that already starts out on the right with a probability well above 0 and then gently rises, though never reaching 1.

Clearly in a situation as shown on the left the model has a much higher predictive power, and so this is what one would hope for in practice. Figure 4 shows the logistic regression fits for both IGS (in red) and IES (in blue) where the outcome measure is return for the second year and figure 5 does the same for whether or not a student graduates:

Figure 4: Logistic regression curves for predicting return for the second year using IES and IGS, respectively.

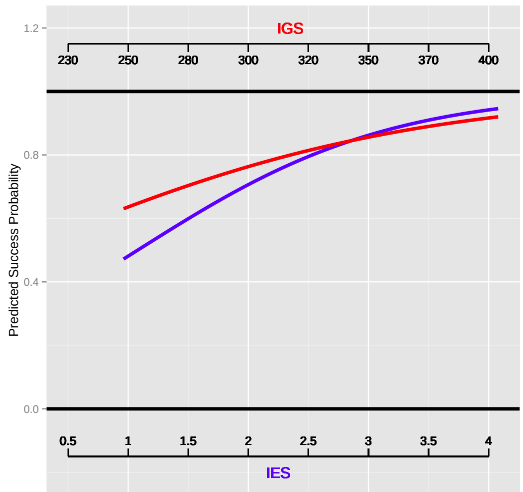
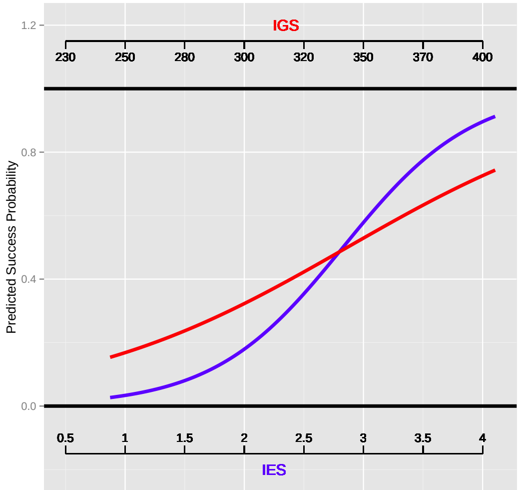


Figure 5: Logistic regression curves for predicting graduation at 150% using IES and IGS, respectively.



Clearly IES yields a much better predictor than IGS

12 Additional Variables

In the regression fit of IES shown above we have actually used three more variables:

- Gender of the Student, coded as 0=Male and 1=Female
- Educational achievement of the father.
- Educational achievement of the mother.

These last two have values
1 : None
2 : Grados 1 al 9
3 : Grados 10 al 12
4 : Completó Escuela Superior
5 : Asistió a la Universidad, pero no Terminó
6 : Grado Asociado
7 : Bachillerato
8 : Maestría
9 : Doctorado

If the information is missing the students is assigned the mean value (5.7)

If the goal were to use the IES in the admissions process use of these variables would clearly not be acceptable, both for legal and ethical reasons. However, for the purpose of identifying students at risk these variables should be acceptable and will increase the predictive power of the model.

13 Statistical Model for Logistic Regression

The resulting models for predicting return for a second year and for graduating at 150% of the time are shown in table 1. As before the predictors (except Gender) were standardized so that the coefficients are in principle size-comparable. The model for the return for a second year is based on 23881 students because at the time o the writing of this report this information for the class of 2013 is not yet available. Similarly the calculations for graduating are based on the 15766 students for whom graduation at 150% can be determined at this time.

Table 1: Coefficients for logistic regression

Second	Year	Graduated	at 150%
GPA.Escuela.Superior	0.325	GPA.Escuela.Superior	0.559
Gender	0.236	Gender	0.503
SchoolGPA	0.203	SchoolGPA	0.446
Aptitud.Verbal	0.150	Aprov.Matem	0.272
Aprov.Espanol	-0.129	Niv_Avanzado_Mate_II	0.117
Aprov.Ingles	-0.127	Aptitud.Matem	-0.106
Aprov.Matem	0.125	Niv_Avanzado_Espa	0.104
Niv_Avanzado_Espa	0.115	Aprov.Ingles	-0.096
Niv_Avanzado_Mate_I	0.056	Aprov.Espanol	-0.053
Niv_Avanzado_Ingles	-0.056	Father	0.039
Aptitud.Matem	0.053	Niv_Avanzado_Ingles	0.027
Niv_Avanzado_Mate_II	-0.034	Mother	0.021
Mother	-0.012	Aptitud.Verbal	0.018
Father	0.011	Niv_Avanzado_Mate_I	-0.005

It is interesting to note that Gender is the second most important predictor in either case.

14 The Coefficients in a Logistic Regression Model

What is the correct interpretation of these coefficients? In an ordinary least squares problem with just one predictor this is very straight forward. Say we have the equation

$$y = \beta_0 + \beta_1 x$$

then β_1 is the increase in y due to a 1 unit increase in x. If we have k predictors and the equation

$$y = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$$

this is still true, although as previously discussed this interpretation can already become suspect because of the correlations between predictors.

What is the situation in logistic regression? If, as we have done in this work, we are using the logit link function we actually are fitting a model of the form

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$$

that is the log-odds of success are modeled as a linear function of the predictors.

Let's consider student #xxxEF77 in the data set. He is a male (Gender=0) and the method predicts a probability of 0.842 for him to return for the second year. Now what would be the probability if this student were female (Gender=1) and all the other values of the predictors were the same? Here the coefficient of Gender is 0.236 so one might guess the probability to be $0.842+0.236*1 = 1.078$, which of course is nonsense.

What is happening here? If we invert the link function we find

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + .. + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + .. + \beta_k x_k)}$$

and for our student if $x_1 = 0$ we get $p = 0.842$ but if $x_1 = 1$ we get $p = 0.871$, which is the correct probability.

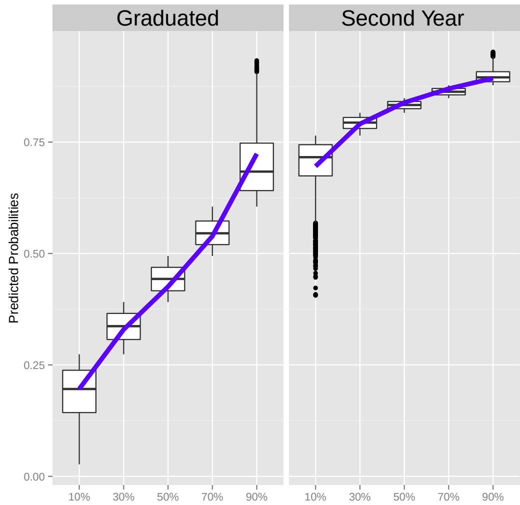
So the problem is that the coefficients effect the probability via the inverse of the link function, not directly. This function, though, is monotonically increasing, so the one feature preserved is that a larger coefficient leads to a larger change in probability. Therefore what matters in the coefficients is not their actual value but their relative sizes. It is therefore valid to conclude that GPA.Escuela.Superior is the most important predictor because it has the largest coefficient, but the actual value of 0.325 is essentially meaningless.

15 Performance of these Models

How well do these models predict the actual percentages? To study this consider the following exercise: let's concentrate for the moment on the return for the second semester. Let's focus on those students that fall into the bottom 20^{th} percentile of the predicted return probabilities. For this cohort the mean predicted return probability is 69.5%. But we have the actual data, so we can check, and indeed 70.1% of these students did return for a second year!

The boxplots of the predicted success probabilities if we repeat this exercise for the other percentiles and also for graduation at 150% are shown in figure 6. Boxplots are drawn at the midpoints of the percentile ranges.

Figure 6: Boxplots of estimated return percentages and of rate of Graduation by their percentiles. Blue lines are true percentages



We have an excellent match between the predicted and the actual percentages.

16 More Detailed Studies

It seems reasonable that whatever indicates a student to be at risk should also depend on what the student is studying at UPR. For example, for a student in Engineering a strong math background is likely more important than for a student in the Humanities. Using this methodology it is possible to tailor the models to varies groups of students. We have so far considered two stratifications:

1) by Faculty, broken down by ADEM, ARTES, CIAG, CIENCIAS and INGE

2) by Orientation, broken down by Analysis Oriented, Information Oriented, Mathematical and Service Oriented.

Tables 2 to 5 have have the coefficients for all combinations.

Table 2: Logistic Regression Coefficients for Predicting Return for Second Year. Overall for all students together as well as broken down by Faculty

	Overall	ADEM	ARTES	CIAG	CIENCIAS	INGE
Intercept	1.454	1.353	1.441	1.110	1.469	1.236
GPA.Escuela.Superior	0.325	0.383	0.275	0.228	0.248	0.488
Gender	0.236	0.353	-0.004	0.388	0.248	0.331
SchoolGPA	0.203	0.248	0.229	0.236	0.175	0.188
Aptitud.Verbal	0.150	0.018	0.135	0.114	0.058	0.319
Aprov.Espanol	-0.129	-0.038	-0.070	-0.127	-0.104	-0.265
Aprov.Ingles	-0.127	-0.215	-0.139	-0.158	-0.016	-0.184
Aprov.Matem	0.125	0.219	0.114	-0.094	0.185	0.140
Niv_Avanzado_Espa	0.115	0.057	0.057	0.071	0.166	0.122
Niv_Avanzado_Mate_I	0.056	-0.0004	0.016	0.164	0.028	0.071
Niv_Avanzado_Ingles	-0.056	0.059	0.038	-0.030	-0.106	-0.084
Aptitud.Matem	0.053	0.036	-0.100	0.195	-0.160	0.357
Niv_Avanzado_Mate_II	-0.034	-0.253	-0.003	0.034	-0.062	-0.035
Mother	-0.012	0.015	-0.029	0.071	0.016	-0.063
Father	0.011	0.053	-0.004	0.049	-0.036	0.039

Table 3: Logistic Regression Coefficients for Predicting Return for Second Year. Overall for all students together as well as broken down by Orientation

	Overall	Analysis	Information	Mathematical	Service
Intercept	1.454	1.296	1.457	1.383	1.436
GPA.Escuela.Superior	0.325	0.271	0.415	0.438	0.292
Gender	0.236	0.299	-0.008	0.320	-0.045
SchoolGPA	0.203	0.199	0.273	0.220	0.237
Aptitud.Verbal	0.150	0.116	0.052	0.239	0.050
Aprov.Espanol	-0.129	-0.070	-0.147	-0.231	-0.067
Aprov.Ingles	-0.127	-0.129	-0.053	-0.120	-0.161
Aprov.Matem	0.125	0.136	-0.205	0.170	0.072
Niv_Avanzado_Espa	0.115	0.103	0.044	0.133	0.001
Niv_Avanzado_Mate_I	0.056	0.040	-0.176	0.062	0.434
Niv_Avanzado_Ingles	-0.056	-0.064	0.039	-0.073	0.118
Aptitud.Matem	0.053	-0.008	0.178	0.159	-0.084
Niv_Avanzado_Mate_II	-0.034	-0.063	-0.059	-0.045	-0.262
Mother	-0.012	0.023	-0.118	-0.044	-0.035

Table 4: Logistic Regression Coefficients for Predicting Graduation. Overall for all students together as well as broken down by Faculty

	Overall	ADEM	ARTES	CIAG	CIENCIAS	INGE
Intercept	-0.498	0.001	0.090	-0.451	-0.521	-1.029
GPA.Escuela.Superior	0.559	0.538	0.678	0.621	0.668	0.873
Gender	0.503	0.484	0.206	0.090	0.544	0.385
SchoolGPA	0.446	0.435	0.510	0.544	0.440	0.388
Aprov.Matem	0.272	0.234	0.037	0.286	0.304	0.477
Niv_Avanzado_Mate_II	0.117	0.145	0.129	0.078	0.153	0.121
Aptitud.Matem	-0.106	0.066	0.036	0.072	-0.171	0.091
Niv_Avanzado_Espa	0.104	0.060	0.118	0.212	0.134	0.085
Aprov.Ingles	-0.096	-0.086	-0.153	-0.240	-0.030	-0.146
Aprov.Espanol	-0.053	-0.061	-0.014	-0.061	-0.014	-0.152
Father	0.039	-0.016	-0.078	0.135	0.053	0.078
Niv_Avanzado_Ingles	0.027	0.171	0.071	0.097	0.008	-0.036
Mother	0.021	0.077	0.068	-0.028	0.043	0.0003
Aptitud.Verbal	0.018	0.086	0.188	0.078	0.042	0.060
Niv_Avanzado_Mate_I	-0.005	0.072	0.109	-0.037	-0.016	-0.011

Table 5: Logistic Regression Coefficients for Predicting Graduation. Overall for all students together as well as broken down by Orientation

	Overall	Analysis	Information	Mathematical	Service
Intercept	-0.498	-0.096	0.022	-1.033	-0.267
GPA.Escuela.Superior	0.559	0.663	0.760	0.886	0.549
Gender	0.503	0.288	0.570	0.435	0.446
SchoolGPA	0.446	0.511	0.342	0.393	0.353
Aprov.Matem	0.272	0.282	-0.204	0.430	0.261
Niv_Avanzado_Mate_II	0.117	0.214	-0.346	0.120	-0.092
Aptitud.Matem	-0.106	0.009	0.378	0.092	-0.218
Niv_Avanzado_Espa	0.104	0.123	0.197	0.083	0.129
Aprov.Ingles	-0.096	-0.115	0.037	-0.092	-0.211
Aprov.Espanol	-0.053	0.010	-0.255	-0.155	0.032
Father	0.039	0.031	0.044	0.075	-0.154
Niv_Avanzado_Ingles	0.027	0.039	0.134	-0.032	0.006
Mother	0.021	0.069	-0.001	-0.004	0.018
Aptitud.Verbal	0.018	0.085	0.346	0.100	0.073
Niv_Avanzado_Mate_I	-0.005	0.027	0.037	-0.011	0.093

17 Information on Freshman Class

Using the methods described above we can now provide the following information for each student in the Freshman class, shown in tables 6A, 6B and 6C.

Table 6A: Percentiles for New Students

	IGS	IES	Return	Graduate
xxxx743B	322	3.2	84.3	68.6
xxxxAB1B	305	3.1	83.3	68.6
xxxxC677	297	2.7	78.9	37.3
xxxx7A55	319	2.8	84.2	42.7
xxxx52D9	333	3.0	89.9	64.1

Table 6B: Percentiles for New Students by Faculty

	Faculty	IES	Return	Graduate
xxxx743B	INGE	2.9	77.7	45.6
xxxxAB1B	CIAG	3.1	76.0	68.9
xxxxC677	INGE	2.3	68.6	19.2
xxxx7A55	INGE	2.6	83.1	29.8
xxxx52D9	CIENCIAS	3.0	88.5	66.0

Table 6C: Percentiles for New Students by Orientation

	Orientation	IES	Return	Graduate
xxxx743B	Mathematical	2.8	82.8	49.0
xxxxAB1B	Analysis Oriented	2.8	79.9	79.6
xxxxC677	Mathematical	2.5	75.9	19.5
xxxx7A55	Mathematical	2.1	85.8	31.0
xxxx52D9	Mathematical	3.2	90.2	53.0

In order to identify the students most at risk it might be more informative to consider their respective rankings within the freshman class, expressed as their percentiles. These are shown in tables 7A, 7B and 7C.

Table 7A: Percentiles for New Students

	IGS	IES	Return	Graduate
xxxx743B	47.9	90.5	57.0	90.0
xxxxAB1B	30.8	86.2	50.8	90.0
xxxxC677	23.8	38.3	26.9	34.9
xxxx7A55	44.9	51.9	56.6	46.2
xxxx52D9	62.4	79.1	92.5	85.3

Table 7B: Percentiles for New Students by Faculty

	Faculty	IES	Return	Graduate
xxxx743B	INGE	63.0	28.6	54.5
xxxxAB1B	CIAG	83.3	22.7	87.7
xxxxC677	INGE	19.5	7.6	12.0
xxxx7A55	INGE	36.8	52.5	26.4
xxxx52D9	CIENCIAS	71.4	83.2	84.8

Table 7C: Percentiles for New Students by Orientation

	Orientation	IES	Return	Graduate
xxxx743B	Mathematical	55.7	50.8	58.3
xxxxAB1B	Analysis Oriented	46.1	33.6	94.3
xxxxC677	Mathematical	22.7	18.6	12.8
xxxx7A55	Mathematical	8.3	68.3	27.9
xxxx52D9	Mathematical	83.5	90.7	64.2

18 Implementation

Clearly the calculations needed to carry out this analysis have to be done by computer. I have used the statistical analysis program R for this purpose. R is the de facto standard in Statistics today, with the benefit of being freeware. Moreover all of its methods have been thoroughly tested by the leading Statisticians in the world.

Unlike the formula for IGS, which has been unchanged for at least 15 years, the formulas for IES as well as those for predicting success should be updated regularly, possibly every year. This will insure that they are always the best for the current generation of students.

Unfortunately R is not a simple program to use, and if the University decides to employ these ideas we will need to develop a system simple enough to be used by a non-expert. This is possible but would require a considerable effort.

19 Conclusions

We have shown that advanced statistical methods can be used to predict the probabilities of students returning for the second year as well as for graduating at 150%. Using this information we can identify those students at the highest risk of failure at UPRM. Hopefully a well designed intervention program aimed at these students can then be used to lower the failure rates. In this study we have focused solely on the data available from the students admissions information. It might be worthwhile to consider collecting additional information, maybe via an email survey. Also, additional information becomes available as the school year progresses, for example the students grades after the first semester. Such information could then also be used to update our models.

20 Appendix: Detailed Information on Regression Fits

Table A 1: Information on fit of Return for Second Year

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.454	0.024	60.384	0
Gender	0.237	0.037	6.351	0
Father	0.011	0.020	0.561	0.575
Mother	-0.012	0.019	-0.607	0.544
SchoolGPA	0.203	0.019	10.768	0
GPA.Escuela.Superior	0.325	0.018	18.459	0
Aptitud.Verbal	0.150	0.023	6.631	0
Aptitud.Matem	0.053	0.029	1.812	0.070
Aprov.Ingles	-0.127	0.023	-5.655	0.00000
Aprov.Matem	0.125	0.030	4.194	0.00003
Aprov.Espanol	-0.129	0.022	-5.866	0
Niv_Avanzado_Espa	0.115	0.026	4.499	0.00001
Niv_Avanzado_Ingles	-0.056	0.026	-2.132	0.033
Niv_Avanzado_Mate_I	0.056	0.019	2.893	0.004
Niv_Avanzado_Mate_II	-0.034	0.022	-1.554	0.120

Null deviance: 24220 on 25494 degrees of freedom
Residual deviance: 23361 on 25480 degrees of freedom
AIC: 23391.5

Table A 2: Information on fit for Graduating at 150%

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.498	0.027	-18.584	0
Gender	0.504	0.039	12.961	0
Father	0.039	0.018	2.122	0.034
Mother	0.021	0.018	1.143	0.253
SchoolGPA	0.446	0.022	20.657	0
GPA.Escuela.Superior	0.559	0.022	25.287	0
Aptitud.Verbal	0.018	0.025	0.709	0.478
Aptitud.Matem	-0.106	0.031	-3.388	0.001
Aprov.Ingles	-0.096	0.024	-4.060	0.00005
Aprov.Matem	0.272	0.032	8.546	0
Aprov.Espanol	-0.053	0.025	-2.169	0.030
Niv_Avanzado_Espa	0.105	0.025	4.265	0.00002
Niv_Avanzado_Ingles	0.027	0.027	1.013	0.311
Niv_Avanzado_Mate_I	-0.005	0.017	-0.328	0.743
Niv_Avanzado_Mate_II	0.117	0.021	5.578	0.00000

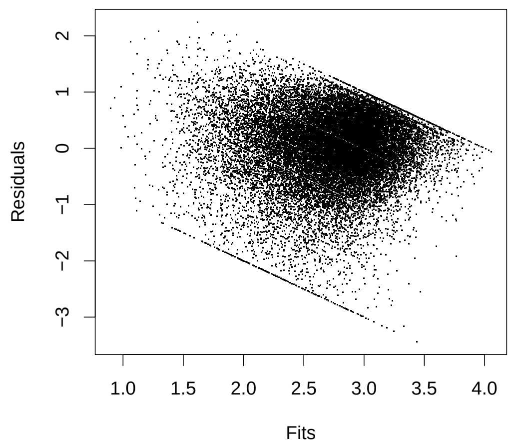
Null deviance: 21649 on 15765 degrees of freedom

Residual deviance: 19389 on 15751 degrees of freedom

AIC: 19418.9

In both cases the deviance is in line with the degrees of freedom, so there is no reason to suspect a problem with the fits.

Figure 7: Residual vs Fits plot for IES



The residual vs fits plot shows no problem with the assumptions of least squares regression. The diagonal appearance of the graph is due to the fact that the response variable GPA after the freshman year is bounded by $0.0 - 4.0$, and does not indicate a problem with the fit.

Table A 3: Correlations Between Predictors

	Gender	Father	Mother	SchoolGPA	GPA.Escuela.Superior
Gender	1	-0.06	-0.06	0.02	0.15
Father	-0.06	1	0.50	0.26	-0.03
Mother	-0.06	0.50	1	0.22	-0.01
SchoolGPA	0.02	0.26	0.22	1	-0.22
GPA.Escuela.Superior	0.15	-0.03	-0.01	-0.22	1
Aptitud.Verbal	-0.04	0.15	0.16	0.21	0.18
Aptitud.Matem	-0.27	0.17	0.17	0.25	0.16
Aprov.Ingles	-0.10	0.29	0.27	0.32	0.05
Aprov.Matem	-0.23	0.18	0.18	0.25	0.22
Aprov.Espanol	0.16	0.12	0.13	0.20	0.25
Niv_Avanzado_Espa	0.06	0.11	0.11	0.15	0.30
Niv_Avanzado_Ingles	-0.01	0.21	0.18	0.24	0.20
Niv_Avanzado_Mate_I	-0.02	0.02	0.02	0	0.09
Niv_Avanzado_Mate_II	-0.06	0.08	0.08	0.09	0.25

Table A 3: Correlations Between Predictors

	Aptitud.Verbal	Aptitud.Matem	Aprov.Ingles
Gender	-0.04	-0.27	-0.10
Father	0.15	0.17	0.29
Mother	0.16	0.17	0.27
SchoolGPA	0.21	0.25	0.32
GPA.Escuela.Superior	0.18	0.16	0.05
Aptitud.Verbal	1	0.46	0.51
Aptitud.Matem	0.46	1	0.45
Aprov.Ingles	0.51	0.45	1
Aprov.Matem	0.47	0.82	0.48
Aprov.Espanol	0.60	0.39	0.43
Niv_Avanzado_Espa	0.37	0.33	0.28
Niv_Avanzado_Ingles	0.36	0.37	0.50
Niv_Avanzado_Mate_I	0.08	0.15	0.09
Niv_Avanzado_Mate_II	0.23	0.38	0.19

Table A 3: Correlations Between Predictors

	Aprov.Matem	Aprov.Espanol	Niv_Avanzado_Espa
Gender	-0.23	0.16	0.06
Father	0.18	0.12	0.11
Mother	0.18	0.13	0.11
SchoolGPA	0.25	0.20	0.15
GPA.Escuela.Superior	0.22	0.25	0.30
Aptitud.Verbal	0.47	0.60	0.37
Aptitud.Matem	0.82	0.39	0.33
Aprov.Ingles	0.48	0.43	0.28
Aprov.Matem	1	0.40	0.35
Aprov.Espanol	0.40	1	0.36
Niv_Avanzado_Espa	0.35	0.36	1
Niv_Avanzado_Ingles	0.38	0.32	0.67
Niv_Avanzado_Mate_I	0.16	0.07	0.20
Niv_Avanzado_Mate_II	0.41	0.21	0.47

Table A 3: Correlations Between Predictors

	Niv_Avanzado_Ingles	Niv_Avanzado_Mate_I	Niv_Avanzado_Mate_II
Gender	-0.01	-0.02	-0.06
Father	0.21	0.02	0.08
Mother	0.18	0.02	0.08
SchoolGPA	0.24	0	0.09
GPA.Escuela.Superior	0.20	0.09	0.25
Aptitud.Verbal	0.36	0.08	0.23
Aptitud.Matem	0.37	0.15	0.38
Aprov.Ingles	0.50	0.09	0.19
Aprov.Matem	0.38	0.16	0.41
Aprov.Espanol	0.32	0.07	0.21
Niv_Avanzado_Espa	0.67	0.20	0.47
Niv_Avanzado_Ingles	1	0.23	0.43
Niv_Avanzado_Mate_I	0.23	1	0.10
Niv_Avanzado_Mate_II	0.43	0.10	1