Int. J. Engng. Ed. Vol. **10,** No.3, pp. 268-273.1994
Printed in Great Britain.

0742-0269/91 $3.00+0.00

A Simple Model for Predicting Success in an Engineering Programme*

G. H. GEORGE
E. MOORE
M.C. PATEY
Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's,
Newfoundland, Canada AIB 3X5

This paper presents a general approach for developing a test of the promotion criteria in a typical engineering programme. As an illustration we describe some mathematical models which could be used to predict a person's probability of succeeding in Memorial University's engineering programme based on performance in an early term in the programme. Logistic regression is employed due to the binary nature of the response variable (whether or not graduation occurs after 5 *years). Multiple logistic regression on the grades obtained in various courses provides little additional predictive power over a simple logistic regression on the early term's average grade. The results appear to support the faculty's choice of minimum average term grade for promotion to the next term in the programme.*

INTRODUCTION

IN ANY academic programme a measure of a student's ability is required. If possible the student's aptness for a specific discipline should be determined. The idea is a general approach to the quantification of a student's ability through the development of probability models. With such a model you can determine the likelihood of a person succeeding in his or her chosen discipline. This article deals with the application of this approach to the engineering programme at Memorial University.

The engineering programme at Memorial University of Newfoundland consists of eight academic terms, the first two of which are common to all disciplines of engineering and six work terms during which the students are placed with an employer to gain greater experience in the field of their choice.

In 1992, one of us employed an engineering work term student to investigate the relationship between success in graduating from the engineering programme on time (no more than 5 years after starting in the programme) and performance in the university courses required of Newfoundland students prior to their entry to the engineering programme [1].

Subsequently, the authors set out to develop some sort of means for predicting the chances of graduating from Memorial University's engineering programme in 4, 5 and 6 years after completion of academic term 2 for the first time. Our ultimate task was to provide the Faculty of Engineering with information to assist in determining what modifications to the promotion criteria in the engineering programme may be needed.

In 1990 the promotion criteria from academic terms 1 and 2 were modified. Until this time the two terms were viewed as a unit with promotion to term 3 requiring an overall average of 60% between the two terms. Now the first two terms are viewed individually with promotion from each term requiring a 60% average and a grade of at least 50% in each course. One will therefore be able to use the information contained within the original report [2] in the future to compare success rates before and after the change in the programme's structure.

THE DATA

Considering the nature of the problem, the most obvious variables to include are the grades of the person under consideration. It would seem logical to assume that the higher a person's grades the better the chance of succeeding in their chosen discipline. Other factors such as age, marital status, financial resources, life-style, number of dependents, etc., may or may not have a measurable influence on a person's chance of success. While it would be worthwhile to consider these variables, the task presented to us was specifically to be based on grades and nothing else. As such the analysis deals only with the purely academic side of the problem. The variables used were the marks from the seven courses completed in term 2 and the average of these seven marks.

* Paper accepted 12 April 1994.

The important thing to remember in this kind of study is to use the most recent data available. The models desired by the associate dean were to describe a person's probability of graduating in 4, 5 or 6 years after completing the first two academic terms of the engineering programme. Four years is the shortest time in which the remainder of the programme can be completed. Because of this the models were developed using the marks of the students in term 2 of the engineering programme in 1985 and tested using the term 2 marks of the students completing this term in 1986. Unfortunately, a consequence is that the resulting model applies only to those students who completed term 2 before the change in 1990. These were the most recent data available.

Several criteria were used in determining whom to include in the design and test data sets. The models were intended to predict the chance of success for a person completing term 2 for the first time. Anyone repeating the term in either of the two sets of data was excluded. Due to the set-up of the university's computer system, any marks obtained by a person prior to 1980 were not available. Since it would be impossible to determine whether they were repeating the term, those who registered as students in the university prior to 1980 were excluded from the analysis. There were a few cases for which no grades were available - they were also left out of the study. Finally, anyone who dropped out of the programme with clear standing was also left out. Many of them did not fail any term on their path to graduation and cannot be counted as failures. The main consideration here is to acquire as much data as possible without biasing the model or treating students unfairly.

LOGISTIC REGRESSION AND THE BMDP PROGRAM

Logistic regression models were used in the analysis. If we use an ordinary linear regression equation with several variables the dependent (response) variable could have any value. We want this value to lie between 0 and 1, because we want the *probability of graduating* for a given set of grades. In addition, models designed using regression analysis are based on data collected for the independent and dependent variables. In our case, the collected data for the dependent variable has two possible values-0 for failure and 1 for success. The logistic regression function (LRF), given by

$$P(S) = \frac{1}{1+\exp(-\mathbf{X\beta})} = \frac{1}{1+\exp\left(-\left(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n\right)\right)}$$

accomplishes both these tasks. It limits the range of $P(S)$ to (0, 1) for all values of the X_i s and it also requires the response variable to be binary. (Note: for people familiar with matrix notation, in the equation above, $\boldsymbol{\beta}$ is the vector of coefficients and $\mathbf{X}$ is the row matrix of regressors ('independent' variables).) While the values of the various X_i s have been collected and are known, the values of the β_i s are generally unknown. The problem is to determine the values of the coefficients such that the differences between the resultant model or LRF and the actual values for the dependent variables are minimized.

The amount of mathematical computation required to determine the β_i s is considerable. This is why the BMDP statistical software package was used in the analysis. There are several features of this program which are relevant to the discussion at this point. These include the stepwise regression procedure, the *p*-values-to-enter/remove and the C.C. Brown and Hosmer-Lemeshow goodness-of-fit tests. The features to be discussed at this point are located at the beginning of each step in the program output.

The BMDP program develops models in a stepwise fashion. This means it enters the independent variables into the model according to their significance, that is, the most significant variable is entered into the model first, then the next most significant, etc. At each step the program calculates the coefficients of the variables in the model at that point and various other relevant statistics. It then recalculates the significance of the terms not in the model. The significance is measured by the variable's *p*-value and its chi-squared value, given in the table of 'Statistics to Enter and Remove Terms' at the beginning of each step in the program. A lower p-value indicates greater significance for that variable compared to the other variables. If two terms have the same *p*-value the term with the higher chi-squared value is the more significant.

The *p*-values-to-enter/remove determine what terms are allowed into the model. At a given step a variable is entered into the model if its *p*-value is lower than the *p*-value-to-enter. In a similar manner a term is removed from the model if its *p*-value at a given step is higher than the *p*-valueto-remove. The *p*-values-to-enter/remove are specified by the programmer in the BMDP instruction language program. Note that these *p*-values do not affect the output at anyone step - they only limit the number of steps in the program.

The next features of importance are the two goodness-of-fit tests mentioned above. The 'Hosmer-Lemeshow' goodness-of-fit test determines whether the predicted values based on the model at this stage in the development fit the data. A small *p*-value here indicates that the predicted values do not fit the data. The 'C.C. Brown' goodness-of-fit test compares the developed logistic model to a more general family of models of which the logistic model is a member. A small *p*-value for this test means that the logistic model is inappropriate for the data under consideration.

More complete descriptions of the procedure and related issues can be found in the references [2]-[13].

DEVELOPING THE MODEL

Preliminary analysis of the grades from 1985 revealed several interesting features. First of all, there tends to be strong multicollinearity among regressor variables. In regression analysis multicollinearity or intercorrelation among the regressor variables tends to have a detrimental effect on the regression coefficients and the inferences that can be made about them. A correlation value between two variables (denoted r_{ij}) near –l or 1 means that there is a high probability that one variable can be expressed as a linear function of the other. The coefficient of multiple determination, R^2, denotes the proportional reduction in the variability of one term through the introduction of a set of variables into the regression analysis. R^2 values near 1 indicate a very strong linear relationship between one variable and the others in the model. Table 1 shows the R^2 values between each variable used in this analysis, except for 'average' and the other variables.

'Average' was excluded because a linear relationship is known to exist between it and the other seven variables. This is shown by its R^2 value in Table 1. (These figures were obtained using BMD P Program '6R Partial Correlation and Multivariate Regression' to regress average on the seven other variables.) Note the high correlation between average and the seven courses (the slight deviation from a value of 1 is caused by rounding error in the recorded value for average in the data file). This is expected since average is computed from the other variables.

The value of R ranges in value from –1 to 1. Values of R >0.7 or <–0.7 indicate that approximately half of the variability in the dependent variable is explained by the independent variables. This is indicative of a possible linear relationship between the dependent and independent variables. Table 1 shows high R values for mechanics, circuits, calculus and vectors. These large values indicate strong correlations among the independent variables in all of the graduation models. This means that multicollinearity exists among the regressor variables (i.e. average, materials, mechanics, etc.).

Table 1. R^2 and R values for the covariates

Variable	R^2 value	R value
Materials	0.44131	0.66431
Mechanics	0.51816	0.71984
Circuits	0.61393	0.78353
Calculus	0.61655	0.78520
Vectors	0.68405	0.82707
Design	0.26326	0.51309
Elective	0.39611	0.62937
Average	0.99725	0.99862

As further evidence of this is the failure of the tolerance test in latter steps of the logistic regression models. This test eliminates variables from the model that are too highly correlated with the other independent variables or covariates. The BMDP Logistic Regression program uses a default tolerance value of 0.0001. If $(1-R^2)$ among one variable and the other regressor variables is less than the tolerance value then the variable is excluded from the model because of its high multiple correlation. When the *p*-values-to-enter/remove were relaxed (i.e. increased) more variables (thus more steps) were allowed into the model. The nature of the R calculation is such that entering more variables always increases the multiple correlation of a variable in the model. The ultimate result of this was that the majority of terms allowed into the model at earlier steps failed the tolerance test in later steps. This is more evidence of multicollinearity among the covariates. There are also several informal tests for the presence of intercorrelation. Large changes in the estimated regression coefficients from one step to the next, regression coefficients with algebraic signs the opposite of what is logically expected and large coefficients of simple correlation between the independent variables are all indicative of serious multicollinearity. This would account for the high multiple correlations among the covariates and the high simple correlations between 'average' and some of the other terms. The simple correlations are shown in the correlation matrix of the regressor variables in Table 2.

Because of the high correlation between 'average' and the rest of the independent variables, this term was excluded from further multivariate analysis. New models 'were developed using only the seven course variables. In the first three steps of

Table 2. Correlation matrix of the regression variables

Variable	Materials	Mechanics	Circuits	Calculus	Vectors	Design	Elective	Average
Materials	1.000							
Mechanics	0.530	1.000						
Circuits	0.610	0.641	1.000					
Calculus	0.406	0.495	0.534	1.000				
Vectors	0.445	0.568	0.605	0.776	1.000			
Design	0.406	0.396	0.351	0.086	0.147	1.000		
Elective	0.406	0.400	0.578	0.425	0.527	0.274	1.000	
Average	0.714	0.820	0.862	0.746	0.808	0.453	0.682	1.000

each of the new models all coefficients are positive and there is a reasonable degree of fit between the predicted and observed values. However, the standard errors of the regression coefficients increase as more terms are entered into the models. This means the probability that the estimated regression coefficients are close to the true regression coefficients is reduced. Additional problems, such as the failure of the tolerance tests, lack of fit and high correlations among regression coefficients were encountered despite the exclusion of the average term. All of these factors are indicative of strong multicollinearity among the covariates. Because of this only the first three steps of each model were used in the testing procedure described below. This makes the testing somewhat simpler while eliminating a lot of the problems listed above. Multicollinearity among grades will generally be impossible to avoid.

THE TESTING PROCEDURE

The models were tested using the term 2 marks of the students at this stage of the programme in the winter of 1986. For each year of graduation four models were tested-one containing only the variable 'average' and one for each of the first three steps of the new models. The β_i s for each of the models are shown in Table 3. In this figure, β_o is the constant in the logistic regression function and the variable entered at each step is indicated in parentheses at that step. Each β_i corresponds to the *i*th term entered into the model and is the coefficient of that term. For example, in step 1 of the 4 year graduation model 'circuits' was entered with a coefficient of β_1 = 0.1727. In step 2 'materials' was entered with a coefficient of β_2 = 0.1378 while 'circuits' now has a coefficient of β_3 = 0.1634.

The probability of each person graduating within the specified time was computed using the corresponding models. These probabilities were then summed for the entire class to give the expected number of graduates for each model. This follows from

$$E(X) = \sum_{i=0}^{n} p(x_i)$$

for binary responses, where *E* (*X*) is the expected value of *X, n* is the number of cases and *P*(X_i) is the probability of success for X_i.

The relative error for each model can then be calculated according to the formula

$$\text{Relative error} = \frac{(\text{actual \# grads}) - (\text{expected \# grads})}{(\text{actual \# grads})}$$

and multiplied by 100 to be expressed as a percentage. The model with the smallest relative error would be the most accurate predictor of the probability of succeeding in the engineering program. It is advisable in research of this nature to continuously test models with new data. This may reveal that the test data or the data used to design the model was an atypical group of people for this kind of analysis.

One could also work backwards through your models to find a required grade given a certain probability of graduating. When this was done for a probability of 50% (when you have just as much chance as failing as passing) the required averages were all approximately 60%. This lends support to the current policy at Memorial University of requiring a 60% average for promotion to term 2.

The model developed using data from the class which completed term 2 in 1985 was tested on data from the class which completed term 2 in 1986. The results for all 12 models are displayed in Table 4.

Table 3. The logistic regression models tested

Model	β_0	β_1	β_2	β_3
Four year: average	-23.69	0.3638		
Step 1 (circuits)	-9.072	0.1727		
Step 2 (materials)	-16.58	0.1634	0.1378	
Step 3 (vectors)	-29.13	0.1426	0.1886	0.14260
Five year: average	-21.59	0.3450		
Step 1 (circuits)	-8.202	0.1713		
Step 2 (calculus)	-16.32	0.1256	0.1563	
Step 3 (materials)	-20.95	0.1093	0.1550	0.09532
Six year: average	-18.20	0.3076		
Step 1 (circuits)	-8.145	0.1896		
Step 2 (design)	-17.63	0.1982	0.1258	
Step 3 (calculus)	-12.77	0.1425	0.1464	0.09615

β_0 is the constant. Subscripts of the β_i s indicate the step at which the corresponding term was entered.

Table 4. Actual and predicted numbers of graduates for each model (students completing term 2 in 1986)

Model	Actual number of graduates	Predicted number of graduates	Relative error (%)
Four year model: average	67	70.4	5
Step 1	67	92.7	38
Step 2	67	90.6	35
Step 3	67	67.9	1
Five year model: average	87	78.8	9
Step 1	87	96.8	11
Step 2	87	84.4	3
Step 3	87	82,4	5
Six year model: average	91	87.5	4
Step l	91	100.0	10
Step 2	91	101.4	11
Step 3	91	97.4	7

CONCLUSIONS

One logistic regression model for the probability of graduating on time from the BEng programme at Memorial University given a term 2 average of x is

$$P(S) = \frac{1}{1+\exp(+23.69 - 0.3638x)}$$

This leads to a prediction of 70.4 graduates (on time) for the class completing term 2 in 1986. The actual number graduating on time for that class is 67, an error of 5%.

Similar expressions for the probability of graduating no more than 5 years and no more than 6 years after completion of term 2 can be obtained from Table 3. The comparison between predicted and actual numbers of graduates for the 1986 term 2 class can be found in Table 4. With 'average' as the sole regressor, the error remained below 10%. Replacing average by the set of individual course grades produced errors ranging from 1 to 38% across the various models.

We also recommend looking at the influence the other factors mentioned above (i.e. finances, children, etc.) may have on a person's chances of graduating.

It appears that logistic regression serves well for the task of modelling success in Memorial University's engineering programme.

Acknowledgements - We would like to thank the following people without whose assistance this project would not have been completed: Janet Kearney, Yvonne King, Darryl Keating and the staff at Computing and Communications at Memorial University.

REFERENCES

1. D. Keating, *Memorial University's Engineering Programme: Modelling Success Rates,* EAST Report 93-002, Faculty of Engineering and Applied Science, Memorial University of Newfoundland (1992).
2. M. C. Patey, *Modelling Success in Memorial University's Engineering Programme,* EAST Report 93-001, Faculty of Engineering and Applied Science, Memorial University of Newfoundland (1993).
3. W. J. Dison (ed.), *BMDP Statistical Software,* University of California Press, Berkeley, CA (1983).
4. C. C. Brown, On a goodness-of-fit test for the logistic model based on score statistics, *Commun. Statistics: Theory Methods,* **11** (10), 1087-1105 (1982).
5. J. L. Devore, *Probability and Statistics for Engineering and the Sciences,* 3rd edn, Brooks/Coles Publishing Company, Pacific Grove, CA (1991). .
6. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression,* John Wiley, New York (1989).
7. D. W. Hosmer and S. Lemeshow, Goodness-of-fit for the multiple logistic regression model, *Commun. Statistics: Theory Methods,* A9 (10) (1986).
8. R. H. Myers, *Classical and Modern Regression with Applications,* Prindle Weber & Schmidt, Boston, MA (1986).
9. J. Neter, W. Wasserman and M. H. Kutner, *Applied Linear Regression Models,* 2nd edn, Richard D. Irwin, Homewood, IL (1985).
10. J. Neter, W. Wasserman and M. H. Kutner, *Applied Linear Statistical Models,* 2nd edn, Richard D. Irwin, Homewood, IL (1985).
11. R. L. Prentice, A Generalization of the Probit and Logit Methods for Dose Response Curves, *Biometrics,* 32 (4),761 -768 (1976).
12. S. Weisberg, *Applied Linear Regression,* 2nd edn, John Wiley, New York (1985).
13. D. R. Wittink, *The Application of Regression Analysis,* Allyn & Bacon, Needhan Heights, MA (1988).

Dr G. H. George obtained a bachelor of science degree in mathematics at the University of Southampton in 1980 followed by a PhD in astronomy at University College Cardiff (Wales) in 1983. He served as a lecturer in North East Surrey College of Technology in England during 1983-1986 and as an assistant professor of mathematics at the University of Bahrain (Arabian Gulf) until 1988 when he took up his present appointment in the Faculty of Engineering and Applied Science at the Memorial University of Newfoundland.

Dr E. Moore has undergraduate degrees from the Universities of Edinburgh and St Andrews in engineering and mathematics, respectively. His doctorate in engineering mathematics is from the University of Waterloo. He has taught at Teesside University, UK and the Technical University of Nova Scotia, Canada. Currently, he is associate dean of undergraduate studies at the Memorial University of Newfoundland.

Matthew Patey is currently an undergraduate student in the Memorial University of Newfoundland's co-operative engineering programme. The study discussed in this paper was carried out as part of his second work term. He is currently interested in the area of ocean engineering and the development of subsea engineering systems.