PREDICTION OF THE PROBABILITY OF SUCCESSFUL FIRST YEAR UNIVERSITY STUDIES IN TERMS OF HIGH SCHOOL BACKGROUND: WITH APPLICATION TO THE FACULTY OF ECONOMIC STUDIES AT THE UNIVERSITY OF NEW ENGLAND

D.M. Dancer and H.E. Doran

No. 48 - September 1990

ISSN 0 157-0188
ISBN 0 85834 893 4
General Background

This paper is motivated by informal observation over several years of first year students in the Faculty of Economic Studies at the University of New England. A very casual appraisal of students who have struggled has indicated that many of them had a poor background in Mathematics as well as a fairly low level of achievement in the New South Wales Higher School Certificate Examination.

A formal analysis of results could be a useful contribution to educational decision making. In particular, estimating the important determinants of performance could significantly improve the assessment of a student's chances of success. It was considered that factors which could influence the success rate of first year students would include the HSC aggregate score, preparation in Mathematics and the type and location of schooling.

The aim of this investigation is to relate the probability of a student passing a range of subjects to these 'background variables' and this is achieved by using the technique of probit analysis.

The Probit Model

The basic tool used in this paper is probit analysis. The probit model is a qualitative response (QR) model in which the dependent or endogenous variable may only take the discrete values 0 and 1. The particular form of the probit model used in this study is a univariate binary QR model where the normal distribution is specified as the function F.

It is convenient mathematically to define the dichotomous random variable \( y \) which takes the value of one if the event occurs and zero if it does not occur. It is assumed that the probability of an event depends on a vector of independent variables \( x \) and a vector of unknown parameters \( \theta \).

Amemiya (1981) writes the univariate dichotomous model as

\[
p_i = P(y_i = 1) = F[I(x_i, \theta)] \\
i = 1, 2, ..., n.
\]

where F is the cumulative distribution function of some random variable and \( I(x_i, \theta) \) is a scalar index and assumes the random variables \( y_i \) are independently distributed.

If a linear specification is chosen

\[I(x_i, \theta) = x_i' \beta\]

and the model can then be written as

\[P(y_i = 1) = F(x_i' \beta). \quad (1)\]

...
The probit model follows from the specification in Equation 1 by choosing F to be the cumulative distribution function (cdf) for a normal random variable. That is, \( F(.) = \Phi_z(.) \).

To simplify the explanation of the probit model, a specific example will be used. It could be hypothesized that a student’s probability of success is dependent on a number of factors. The first factor and probably the most important factor to consider is the HSC tertiary entrance score. This HSC score is used by the universities to determine a student’s eligibility for a certain degree.

Another factor that is considered to have a bearing on the level of success is the level of mathematics that a student undertakes at high school. The level of mathematics is particularly relevant if the student is considering a Science or an Economics degree.

In some respects, probit analysis is analogous to regression analysis. Let us suppose that we are interested in investigating the number of subjects passed. One way of approaching this problem would be to postulate that the number of subjects passed, \( Y \), could be regarded as a function of several variables. For example,

\[
Y = f(\text{HSC score}, \text{level of mathematics}, \text{place of school}).
\]

(2)

The linear regression model may then enable an equation in the following form to be estimated:

\[
Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3
\]

(3)

where

- \( X_1 = \text{HSC score} \)
- \( X_2 = \text{level of mathematics completed at high school} \)
- \( X_3 = \text{country or city school} \).

Note that if more than two levels of mathematics are being compared, then several dummy variables would be needed to capture the differences between the levels. We emphasise that such an approach could be used to calculate the average number of subjects passed for any combination of the three variables.

Another problem that is of interest is to estimate the proportion of students who pass two or more subjects. When analysing proportions or probabilities, probit analysis is an appropriate technique. An approach similar to the regression problem can be used. Again, a sample of students can be taken who have an HSC score, completed 2 Unit Mathematics and attended a city school.
These proportions could be interpreted as the probability that a student with certain qualifications will pass at least two or more subjects. When probabilities are to be estimated, the technique analogous to regression analysis is probit analysis. To illustrate this technique, an example will be given.

For the sake of simplicity, consider students at a university who have enrolled in an Economics degree. Let

\[ X_{1i} = \text{HSC tertiary entrance score.} \]
\[ X_{2i} = 1 \quad \text{if the student completed 2 Unit A Mathematics} \]
\[ = 0 \quad \text{if the student completed 2, 3 or 4 Unit Mathematics} \]
\[ X_{3i} = 1 \quad \text{if the student attends a city high school} \]
\[ = 0 \quad \text{if the student attends a country high school} \]

The variables \( X_{2i} \) and \( X_{3i} \) are dummy variables.

Define an index \( I_i' \), given by

\[ I_i' = \gamma_0 + \gamma_1 \cdot X_{1i} + \gamma_2 \cdot X_{2i} + \gamma_3 \cdot X_{3i}; \quad (4) \]

which is a single number summarising the values of \( X_1, X_2 \) and \( X_3 \) for the \( i \)th individual.

Such an index, which can take any real value, can be converted to a probability by reflecting it in any cumulative distribution function. The probit model uses the distribution function associated with a normal random variable. That is,

\[ p_i = F_X(I_i'), \]

where \( F_X(.) \) is the distribution function of a random variable \( X \) which is \( N(\mu, \sigma^2) \).

Now

\[ F_X(I_i') = P[X \leq \gamma_0 + \gamma_1 \cdot X_{1i} + \gamma_2 \cdot X_{2i} + \gamma_3 \cdot X_{3i}] \]
\[ = P[Z \leq \frac{(\gamma_0 + \gamma_1 \cdot X_{1i} + \gamma_2 \cdot X_{2i} + \gamma_3 \cdot X_{3i}) - \mu}{\sigma}] \]
\[ = P[Z \leq I_i] \quad (5) \]

where \( Z \sim N(0, 1) \) and

\[ I_i = \frac{(\gamma_0 - \mu)}{\sigma} + \frac{\gamma_1}{\sigma} \cdot X_{1i} + \frac{\gamma_2}{\sigma} \cdot X_{2i} + \frac{\gamma_3}{\sigma} \cdot X_{3i} \]

That is,

\[ I_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i}. \quad (6) \]
Because of the arbitrary nature of the coefficients $\gamma_i$ and $\beta_i$, it is always possible to choose

$$F_X(.) = \Phi_Z(.)$$

where $\Phi_Z(.)$ is the cumulative distribution function of the standardised normal random variable.

Consider some hypothetical values of $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$. Let $\beta_0 = -2.0$, $\beta_1 = 0.01$, $\beta_2 = -1.5$, $\beta_3 = -0.5$. Three individuals have been chosen to illustrate this example. The first student has an HSC score of 280, has completed 2 Unit A Mathematics and attended a city high school. The second student has an HSC score of 300, has completed 2 Unit Mathematics and attended a city high school. The third student has an HSC score of 350, has completed 3 Unit Mathematics and attended a country high school.

The index value for the first student is calculated as

$$I_1 = -2.0 + 0.01 \times 280 - 1.5 \times 0 - 0.5 \times 1 = -1.2.$$ 

Therefore,

$$P(Z \leq I_1) = P(Z \leq -1.2) = 0.12.$$ 

From Table 1, it can be seen that the probability that a student who has an HSC score of 280 with 2 Unit A Mathematics and has attended a city high school is quite low (0.12). The probability that a student who has an HSC score of 300 with 2 Unit Mathematics and has attended a city high school is 0.59. This indicates that this student has about a 50% chance of passing two or more subjects. The third student who has an HSC score of 350 with 3 Unit Mathematics and attended a country high school has a very high probability of passing two or more subjects at the end of first year at university. In the

<table>
<thead>
<tr>
<th>Student</th>
<th>HSC Score</th>
<th>Level of Mathematics</th>
<th>City or Country School</th>
<th>Index</th>
<th>Proby</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>280</td>
<td>2UA</td>
<td>City</td>
<td>-1.2</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>2U</td>
<td>City</td>
<td>0.5</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>350</td>
<td>3U</td>
<td>Country</td>
<td>1.5</td>
<td>0.93</td>
</tr>
</tbody>
</table>

above example, the coefficients, $b_0$, $b_1$, $b_2$ and $b_3$, were assigned arbitrary values. In reality, these numbers are always unknown, and must be estimated from the data.

The data for this project are in the form of a single observation on each decision maker. This means that maximum likelihood methods must be used for estimation purposes. For
a sample of $T$ observations, the likelihood function is

$$l = \prod_{i=1}^{T} f(y_i)$$

$$= \prod_{i=1}^{T} P_i^{y_i} (1 - P_i)^{1-y_i}$$

$$= \prod_{i=1}^{T} F(z_i' \beta)^{y_i} [1 - F(z_i' \beta)]^{1-y_i} \tag{7}$$

where $y_i = 1$ or 0. The logarithm of the likelihood function is maximised using numerical methods. The properties of the log likelihood function guarantee that the iterative procedure will converge to the global maximum based on any suitable set of starting values $\beta^{(0)}$. Also the maximum likelihood estimators of the $\beta$ parameters are consistent, asymptotically efficient and asymptotically normally distributed.

In this project, the two events which are considered separately are 'The student passes all his subjects' and 'The student passes all three main subjects'.

Data

The data for this study were obtained from two sources. The first source was the Administrative Data Processing section of the University of New England (UNE) which provided information on every student enrolled in the Faculty of Economic Studies in 1988. The following information was provided:

- the degree in which the student was enrolled
- the HSC aggregate mark
- each student's examination results for 1988.

For some students who live interstate, the UNE calculates an equivalent HSC aggregate mark which was then subsequently used in the analysis.

The second source of data was a questionnaire sent out to all students with a 1988 student number in the Faculty of Economic Studies. This survey was a postal survey with one follow up of those who had not replied the first time. Students were asked for the following information:

- State in which their education took place
- Highest level of Mathematics they attempted
- Highest level of English they attempted
• Education at a state or a private school

• Town or city in which they completed their education

The students were asked to answer these questions for the last two years at high school. On the basis of this information, the following variables are defined.

\[
D_1 = \begin{cases} 
1 & \text{for Bachelor of Agricultural Economic Students} \\
0 & \text{otherwise}
\end{cases}
\]

\[
D_2 = \begin{cases} 
1 & \text{for Bachelor of Financial Administration Students} \\
0 & \text{otherwise}
\end{cases}
\]

\[
M_1 = \begin{cases} 
1 & \text{for students who completed 2 Unit A Mathematics at high school} \\
0 & \text{otherwise}
\end{cases}
\]

\[
M_3 = \begin{cases} 
1 & \text{for students who completed 3 or 4 Unit Mathematics at high school} \\
0 & \text{otherwise}
\end{cases}
\]

\[
E_1 = \begin{cases} 
1 & \text{for students who completed 2 Unit English at high school} \\
0 & \text{otherwise}
\end{cases}
\]

\[
School = \begin{cases} 
1 & \text{for students who attended a state high school} \\
0 & \text{otherwise}
\end{cases}
\]

\[
City = \begin{cases} 
1 & \text{for students who attended a city high school} \\
0 & \text{otherwise}
\end{cases}
\]

\[
HSC = \text{the HSC tertiary entrance score for the student.}
\]

The basic model postulated is

\[ I = f(M_1, M_3, D_1, D_2, HSC, E_1, School, City) \quad (8) \]

**Results**

In this analysis, we are concentrating primarily on estimating the probability that a student will pass all subjects. To this end, a number of probit models are estimated including various combinations of the variables M1, M2, D1, D2, E1, HSC, School and City. The results of these estimations are presented in Appendix A.

The main features of the empirical results are:

- The HSC aggregate is highly significant statistically.
The level of Mathematics and whether the student attended a country or city school are nearly always significant.

With regard to Mathematics, once HSC aggregate is controlled, there appears to be no difference between 2 and 3 Unit Mathematics students. As we have chosen the 2 Unit group as base, we therefore omit the variable M3. Accordingly, for further interpretation, we have selected the model

\[ I = -2.4228 + 0.009725 HSC - 1.3452 M1 - 0.5471 City \]

\[ (0.887) \quad (0.00973) \quad (0.590) \quad (0.259) \]

The probabilities which result from this index are shown in Figures 1 and 2 below, in which probability is plotted against HSC aggregate for city educated students (Figure 1) and country educated students (Figure 2). Two points emerge very clearly from these figures.

- If a student enters the university with an HSC aggregate of 260 and 2 Unit Mathematics, then the probability of passing is approximately 0.52 for a student from a country school and approximately 0.32 for a student from a city school. However, if a student enters the university with an HSC aggregate of 260 and 2 Unit A Mathematics, then these probabilities drop dramatically to 0.1 and 0.03 respectively.

- The effect of 2 Unit Mathematics on a student's prospects is very marked. For example, for a student to have an even chance of passing all subjects with only 2 Unit A Mathematics, an HSC aggregate of about 380 if the student is from a country school and about 440 if the student is from a city school is required.

The implications of the model could also be interpreted from a different perspective. In Figures 3 and 4, we display the results of 2 Unit Mathematics students (Figure 3) and 2 Unit A Mathematics students (Figure 4). This enables us to see the difference between city and country educated students. It is apparent that country students have superior performance. Rather than inferring that a country education is superior, we suspect that we are observing a 'living away from home' effect. It would be interesting to see a similar analysis performed on students from a metropolitan university. Quite possibly the results would be reversed. We note in passing that a perusal of Appendix A shows little evidence that private or public school education has any effect on university results, at least in the Faculty of Economic Studies.

In addition to the above analysis, we carried out a similar analysis to assess the probability of students passing the core subjects common to all three degrees - namely,
Economics, Econometrics and Mathematics. The same broad conclusions emerge except that now, while the Mathematics effect is numerically very large, it is not statistically significant. The difference between country and city is no longer quite so marked.

Conclusion

This analysis has given rise to two main conclusions.

- Students who enter with an HSC aggregate of under 300 have a very low probability of passing all subjects. Such a conclusion has clear resource implications particularly in an environment in which Government pressure is towards increasing acceptance of poorly qualified students.

- Mathematical preparation appears to be very important. There is a clear differentiation between students who have at least 2 Unit Mathematics and those who do not.

References

Figure 1: Students from City Schools Passing All Subjects

Figure 2: Students from Country Schools Passing All Subjects
Figure 3: Students with 2 Unit Mathematics Passing All Subjects

Figure 4: Students with 2 Unit A Mathematics Passing All Subjects
Results of the Different Probit Analysis

Table 2: The Variable ‘Passing All Subjects’ vs Different Combinations

<table>
<thead>
<tr>
<th>Const</th>
<th>M1</th>
<th>M3</th>
<th>D1</th>
<th>D2</th>
<th>HSC</th>
<th>E1</th>
<th>School</th>
<th>City</th>
<th>$R^2$</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.074</td>
<td>0.661</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.068</td>
<td>0.637</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.025</td>
<td>0.637</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.075</td>
<td>0.653</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.034</td>
<td>0.637</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.203</td>
<td>0.710</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.163</td>
<td>0.645</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.159</td>
<td>0.669</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.146</td>
<td>0.661</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.132</td>
<td>0.669</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.213</td>
<td>0.685</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.198</td>
<td>0.710</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.188</td>
<td>0.694</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.169</td>
<td>0.669</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.242</td>
<td>0.718</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.219</td>
<td>0.702</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.214</td>
<td>0.710</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.255</td>
<td>0.726</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.219</td>
<td>0.710</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.247</td>
<td>0.726</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.272</td>
<td>0.702</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.283</td>
<td>0.726</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.227</td>
<td>0.669</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.281</td>
<td>0.726</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.251</td>
<td>0.742</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td>0.289</td>
<td>0.726</td>
</tr>
</tbody>
</table>


A Note on a Bayesian Estimator in an Autocorrelated Error Model. William Griffiths and Dan Dao, No. 3 - April 1979.


Bayesian Econometrics and How to Get Rid of Those Wrong Signs. William E. Griffiths, No. 31 - November 1987.


