Statistical Alternatives for Studying Success in Elementary Statistics: A Comparative Analysis of Logit and Linear Regression

Tricia M. Larson

Faculty Sponsor: Abdulaziz Elfessi, Department of Mathematics

ABSTRACT

Elementary Statistics (MTH 205) is one of the most difficult courses offered by this university, and many students struggle to pass. The purpose of our study was to determine some common characteristics of those who succeed, and of those who fail. We used stepwise linear regression and stepwise logistic regression, and compared their results. It was found by both techniques that a student's ACT score and high school rank percentile were the only significant determinants.

INTRODUCTION

Elementary Statistics (MTH 205) is a subject required for many majors, including those outside the Science and Allied Health discipline. Every semester, over 800 students enroll in an Elementary Statistics course at UW-L. Of these students, 10% to 15% drop the course, and 22% earn a D or an F. Is there a way to improve these numbers? What can be done to better allocate available resources in order to provide assistance to those who need it the most? In order to answer these questions, we must first determine some common characteristics of students who struggle and of those who succeed, and provide a method of predicting success rates based upon these characteristics.

Several studies have indicated that the most significant determinants of success in college are high school grade point average (GPA), and percentile ranking on the American College Test (ACT). These studies include Park and Kerr (1990), who used a multinomial logit approach to determining academic success, and Beecher and Fischer (1995) who used stepwise linear regression analysis. Park and Kerr focused on introductory Economics courses, while Beecher and Fischer focused on college in general. Are there more determinants, such as enrollment in earlier math courses, which apply to success in a statistics course specifically? The primary purpose of our study is to add more understanding of the various factors that influence the academic performance of students in Elementary Statistics. In doing so, we will not only provide an invaluable service to students who put their time and money at risk when they choose to enroll in the course; also, faculty and the University at large will benefit because more time and funds will be spent on the most effective resources. Our expectations at the onset of this study were that more determinants of success would be identified.

A secondary goal of this study is to provide a basis for the comparative analysis of linear and logistic regression. Theory does not provide a basis for comparison, because the two techniques are very different theoretically. We can only compare their practical implications. In this study we can observe whether linear and logistic regression will produce similar results, and we can also compare the errors associated with the two techniques.

METHOD

The study was conducted by administering a self-completion survey to the students in the third week of the Fall 2002 semester, and by obtaining additional information through the registrar's office. The students' grades at the end of the semester were used to model the grade (dependent variable) and characteristics of the students (independent variables, or indicator variables).

Independent Variables

Many independent variables were considered in this study. The survey that was administered to the students in the beginning of the semester addressed their major (**school**), their classification (**classif**), their gender (**gender**), their marital status (**marital**); how far they live from school in miles (**miles**), how many hours in a week that they work (**work**), how many hours per week that they are involved in extracurricular activities (**extrac**); whether they had taken previous math classes at the university level (**mathuniv**), high school statistics (**staths**), and high school calculus (**calchs**); how many hours per week that they study for statistics (**study**), and their expected final grade in statistics (**expect**). In addition, the information that we collected from the registrar's office include the students' ACT scores (**ACT**) and their high school rank percentile (**rank**).

Dependent Variable

The dependent variable for this study was the students' final grades (**grade**) in the course, as obtained by the registrar's office.

Procedure

Two statistical procedures, Stepwise Linear Regression and Stepwise Logistic Regression, were used to form the models; then the results were compared. Regression analysis is a statistical tool that utilizes the relation between two or more variables so that one variable can be predicted from the other (Neter, 23). Linear regression treats the grade as a continuous variable, meaning it can take on any number between 0.0 and 4.0. A 0.0 would represent an F, and a 4.0 would represent an A. Linear regression results in an equation with the following format:

$$Y = \alpha_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots. + \varepsilon$$

where Y = the value of the grade (the dependent variable),
α_0 is a constant,
β_i's are regression weights ($i = 1\ldots p$ where p is the number of independent variables),
$X_1, X_2, \ldots$ are the independent variables (predictors),
ε is the error term that is assumed to sum to zero.

Logistic regression works differently. In this case, the observed grade is a dichotomous variable, meaning it can either be a 1 or a 0, success or failure. We defined a grade of C or above to represent a success, and anything below to represent a failure. The logistic model, rather than predicting a student's grade as does linear regression, predicts the probability of a success. Logistic regression results in the following equation:

$$P(\text{event} \mid X_1, \ldots) = \frac{\exp[\alpha_{AD} + \beta_1(X_1) + \beta_2(X_2) + \ldots.]}{1 + \exp[\alpha_{AD} + \beta_1(X_1) + \beta_2(X_2) + \ldots.]}$$

where event = success or failure (1 or 0),
α_{AD} is a constant,
β_i's are regression weights ($i = 1\ldots p$ where p is the number of independent variables),
$X_1, X_2, \ldots$ are the independent variables (predictors),
Exp = e = 2.71828 is the base of the natural logarithm.

RESULTS

The sample size for this study was 272 students who had taken the survey, completed the course, and for whom the registrar's office had complete information. For the first step of this study, SPSS® (Statistical Software for the Social Sciences) was used to compute basic descriptive statistics to note frequencies, and to observe whether the mean grade differs among levels of the categorical independent variables. Table 1 lists the frequency of each observed grade and their percentages. Table 2 lists the mean grade, sample size, and standard deviation for the independent variables. The following observations may be worth noting:

- Those who had taken calculus in high school performed slightly better than those who had not.

- The mean grades did not significantly differ between those who had and had not taken statistics in high school, those who had and had not taken earlier math courses at the university level, or among the three levels of time spent studying.

- Those who expected higher grades earned higher grades, and those who expected lower grades earned lower grades.

- Females performed slightly better than males.

- The more quantitative disciplines (Science and Allied Health, Business) performed better than the qualitative (Liberal Studies, Art and Communication).

Regression

The next step in this study was to perform stepwise linear regression. The resulting independent variables were the students' ACT scores and high school rank percentile. The regression equation is as follows, where rank represents the percentage of students who ranked below that student:

$$\text{expected grade} = -2.3624 + .03695(\text{rank}) + .07371(\text{ACT})$$

The standard error of the rank coefficient is .00515, while the standard error of the ACT coefficient is .0191. The coefficient of multiple determination is .247, indicating that the fitted model explains about 25% of the variability in the students' grades. The model was tested, and was found to be significant. The analysis of variance results were the following:

$$F(2,269) = 44.227 \qquad \text{P-value} < .001$$

The next step was to produce a stepwise logistic regression model. The resulting independent variables were, again, ACT and rank. The regression equation is as follows:

$$\text{probability of success} = \frac{\exp[-6.04 + .107(\text{ACT}) + .063(\text{rank})]}{1 + \exp[-6.04 + .107(\text{ACT}) + .063(\text{rank})]}$$

The standard error of the rank coefficient is .013, while the standard error of the ACT coefficient is .052. The odds ratio for rank is 1.065, while the odds ratio for ACT is 1.113. Thus, with an increase in a student's high school rank percentile of one percentage point, that student's probability of success should increase by about 6.5%. Likewise, with an increase in the student's ACT score of one point, his or her probability of success in statistics should increase by about 11.3%. When the regression equation was tested, 81.6% of the students' grades were correctly predicted.

Table 1. Frequencies

Observed Grade	Frequency	Percent
A	26	9.6
AB	31	11.4
B	70	25.7
BC	24	8.8
C	65	23.9
D	15	5.5
F	41	15.1
Total	272	100.0

Table 2. Descriptive Statistics

Factor	Level	Mean Grade	N	Std. Dev.
CALCHS	Yes	2.587 (BC)	115	1.1067
	No	2.197 (C)	157	1.1832
STATHS	Yes	2.223 (C)	56	1.0951
	No	2.398 (BC)	216	1.1827
MATHUNIV	Yes	2.389 (BC)	117	1.2476
	No	2.342 (BC)	155	1.1031
STUDY	0-2 hours	2.346 (BC)	39	1.0076
	3-5	2.335 (BC)	164	1.2311
	6 or more	2.435 (BC)	69	1.0978
EXPECT	A	3.239 (AB)	46	.8079
	AB	2.580 (BC)	75	1.0875
	B	2.197 (C)	99	1.0852
	BC	1.855 (C)	31	1.0969
	C	1.190 (D)	21	1.0305
GENDER	Female	2.517 (BC)	177	1.1230
	Male	2.074 (C)	95	1.1939
SCHOOL	Science and Allied Health	2.855 (B)	62	1.0689
	Education	2.436 (BC)	39	1.1708
	Business Administration	2.328 (BC)	67	1.1983
	Health, Physical Education, Recreation, and Therapy	2.328 (BC)	41	1.0908
	Liberal Studies	2.109 (C)	46	1.0950
	Art and Communication	1.750 (C)	14	1.1393

CONCLUSIONS

Both linear and logistic regression found only ACT score and high school rank percentile significant enough to include in the models. This is consistent with the earlier studies mentioned previously. However, common sense seems to suggest that knowledge of math courses prior to statistics should be a very important determinant of one's performance in statistics. Also, one would think that such behavioral characteristics as amount of studying would be important. Perhaps with a larger sample size, more variables such as these may have been found significant. Nonetheless, there are noteworthy implications associated with our results. While these preparation factors (ACT score and high school rank) by no means predestine statistics students for success or failure, they can serve as warning signals, indicating a possible need for further preparation and a reassessment of the level of effort that will be required to succeed. Another suggestion that could be made from these results would be for the university, while assigning freshmen to course sections, to group similar preparation levels (low ACT and class rank versus high ACT and class rank) into classes together. By doing so, instructors will be aware that this group of students may potentially struggle, and teach accordingly. Unfortunately, freshmen are the only students who would benefit from this, as all other students choose their own schedules.

Regarding the comparison of linear and logistic regression, the results were similar. Both techniques found the same indicator variables to be significant. The standard errors of the class rank and ACT coefficients were also similar. Logistic regression was more accurate; however, with the dichotomous variable there are fewer opportunities for mistakes. These results demonstrate that one technique is not favored over the other per se, and that the choice between the two depends by and large on the type of study and the preferred nature of the outcome.

LIMITATIONS

For future projects, there are several improvements that could be made. First of all, the sample size of D's and F's was rather small. To increase the sample size, it may have been helpful to include withdrawals as failures, as these students will also have to retake the course. Time, and possibly money have been wasted by these students as well. Also, rather than using the ACT score as an independent variable, the math portion of the ACT score may have been more indicative. Another alternative to the ACT score could be the math portion of the placement test that all students are required to take before beginning college.

REFERENCES

John Neter, William Wasserman, Michael H. Kutner, *Applied Linear Statistical Models, Second Edition.* Richard D. Irwin, Inc, 1985. p23.

Kang H. Park and Peter M. Kerr (Spring 1990), "Determinants of Academic Performance: A Multinomial Logit Approach." *Journal of Economic Education.*

Mark Beecher and Lane Fischer (1995), "Determinants of College Success." *The Journal of College Admission.*