A Humane Tool for Aiding Computer Science Advisors, Computer Science Students, and Parents

Henry H. Walbesser
Professor of Computer Science
Baylor University
Computer Science Department
Waco, TX 76706
(254)710-4518
Henry_Walbesser@baylor.edu

Ning Hu
Graduate Research Assistant
Department of Computer Science
Baylor University
ningh@hotmail.com

Background: Over the past few years, the computer science department faculty at Baylor has observed that some students who perform adequately during the freshman and sophomore years have substantial difficulty during the junior and senior years of study. Baylor University is an institution committed to being caring of its students. The objective for this study grew out of these two realities.

Objectives: There are three objectives of this research. One objective is to identify students, no later than the sophomore year, who are less likely to succeed as computer science majors. A second objective is to accomplish this identification by using data from seniors majoring in computer science. A third objective is to begin to use this information at the end of their sophomore year when meeting with a computer science faculty advisor.

Design: A regression study is conducted on the data from all students classified as seniors, majoring in computer science in May 2001, showing grades in six freshman and sophomore courses, and showing grades for at least five junior or senior level computer science courses. These students and their course performance data constituted the study sample.

In computing a grade point average for graduation, Baylor counts only the grade received in a course taken during the most recent enrollment. For courses repeated, the grade received for any previous enrollment in that course is included on the student's transcript, but the grade is not included in computing the graduation grade point average. For the purposes of this study all grades received in a course are included in computing the student's grade point average in that course. For example, a student who fails the first introductory course in computer science and then receives a B on the repeat of that course would receive a (0 + 3)/2 or 1.5 for the purposes of this study.

Six courses taken during the freshman and sophomore years are identified as potential predictor variables. These include *Introduction to Computer Science I, Introduction to Computer Science II, Introduction to Computer Systems, Data Structures, Introduction to Discrete Mathematics, and Calculus I.* The grading schedule used at Baylor consists of A, B+, B, C+, C, D, and F with quality points 4, 3.5, 3, 2.5, 2, 1, and 0 respectively.

The criterion variable for the regression study consists of the grade point average of five or more junior or senior level computer science courses taken by the sample students.

An exhaustive linear and multiple regression analysis is conducted. A simple linear regression analysis is computed for each predictor variable and the criterion variable. All possible pairs of predictor variables and the criterion variable are used to compute multiple regressions. Similarly, all possible triples, quintuplets, and the sextuple of predictor variables taken together are used with the criterion variable to compute multiple regression equations.

A determinant algorithm is used to compute the regression coefficients. (Walbesser and Gonce-Winder, 1992) This algorithm results in standardized coefficients, b_i's.

Correlation or multiple correlation coefficients are also computed for each regression analysis. r^2 and R^2 are estimates of the percentage of variance in the criterion variable that can be explained by the predictor variable(s).

The largest of the correlation coefficients will determine the choice of the regression equation to use for predictive purposes. A raw score regression equation is computed from the standardized equation resulting from the determinant sion.nA standard error estimate of the multiple correlation is computed and used to construct a confidence interval for the predicted scores.

Sample Data: *Bear Access* is a computer-based student information system accessible by each professor in any campus academic department. The data shown are informal transcripts for each student majoring in that department. Faculty in a given department only have access to the student majoring in that department's degree program(s). [Note: Baylor University's mascot is a bear, if anyone was wondering about the system name.] 141 students are identified as satisfying the data requirements. Table One contains the demographic data of the sample.

Table One
Demographic Statistics for the Six Predictor Variables and the Criterion Variable

Descriptive	Predictor Variables					Criterion Variable	
Statistics	1	2	3	4	5	6	Y
N	140	141	141	141	127	137	141
Mean	3.34	3.17	2.86	2.89	3.03	2.78	2.73
SD	0.75	0.81	0.83	0.86	0.87	1.03	0.77
Median	3.50	3.50	3.00	3.00	3.00	3.00	2.82
SD_{median}	.007	.003	.002	.002	.002	.002	0.06
$Q_3 - Q_1$	0.08	0.04	0.02	0.02	0.02	0.02	0.70
Range	2.50	4.00	3.00	3.00	3.00	4.00	3.60

There are 141 students selected for the sample. The reason the N's show a variation from this number in the predictor variable N's has to do with waivers given to students on the basis of previous knowledge. The criterion variable mean value of 2.73 does suggest lower than expected performance among some of the students at the junior and senior levels of work.

The median values are each higher than the mean values for all variables, which suggests that the score distributions for the predictor variables and the criterion variable have a slight negative skewness.

$Q_3 - Q_1$ is the interquartile range. The small values for these ranges within the predictor variable suggest a large number of scores are concentrated around the medians of the predictor variables. The criterion variable distribution shows a larger interquartile range, and again adds some support for conducting the regression study. The standard deviation of the median is given by the relationship $Q_3 - Q_1$ / ÖN. (Walbesser and Gonce-Winder, 1991) There are very small standard deviations of the median for all of the variables. This is to be expected because the majority of grades cluster around the B to A range.

Findings: The findings are organized in the following manner. First the correlation coefficient matrix is presented. Next the R^2 values are presented for each of the regression runs. The standardized regression equation associated with the largest R^2 is then presented along with the raw score form of the regression equation, a standard error of estimate, and a confidence interval for the predicted scores.

Table Two presents the correlation coefficients between each pair of predictor variables and between each predictor variable and the criterion variable.

Table Two
Correlation Coefficient Matrix for the Predictor Variables and the Criterion Variable

	1	2	3	4	5	6	Y
1	1.00	0.54	0.30	0.32	0.54	0.45	0.41
2	0.54	1.00	0.44	0.40	0.50	0.52	0.53
3	0.30	0.44	1.00	0.69	0.35	0.47	0.64
4	0.32	0.40	0.69	1.00	0.35	0.44	0.61
5	0.54	0.50	0.35	0.35	1.00	0.58	0.40
6	0.45	0.52	0.47	0.44	0.58	1.00	0.58
Y	0.41	0.53	0.64	0.61	0.40	0.58	1.00

Table Three displays the amount of criterion variable variance that can be accounted for by using each predictor variable separately. In a regression analysis with one predictor variable, r^2 is the appropriate estimate of explained variance. (Walbesser and Gonce-Winder, 1991)

Table Three
Explained Variance Accounted for in the Criterion Variable Using Each Predictor variable Separately

Predictor Variables	r^2
1	0.16
2	0.30
3	0.42
4	0.42
5	0.18
6	0.33

Predictor variables show the largest amount of variance explained with each explaining 42%. Given these results, one would be led to guess that at least one of these two variables will be part of the final best regression solution.

Table Four displays the amount of the criterion variance that can be accounted for using each possible pair and each possible triple of predictor variables.

Table Four

Explained Variance Accounted for in the Criterion Variable by All Possible Pairs and All Possible Triples of Predictor Variables

Predictor Pairs	R^2	Predictor Triplets	R^2
1,2	0.31	1,2,3	0.49
1,3	0.45	1,2,4	0.47
1,4	0.42	1,2,5	0.31
1,5	0.14	1,2,6	0.38
1,6	0.26	1,3,4	0.48
2,3	0.48	1,3,5	0.45
2,4	0.47	1,3,6	0.51
2,5	0.29	1,4,5	0.44
2,6	0.36	1,4,6	0.48
3,4	0.46	1,5,6	0.37
3,5	0.42	2,3,4	0.52
3,6	0.48	2,3,5	0.48
4,5	0.42	2,3,6	0.53
4,6	0.44	2,4,5	0.48
5,6	0.32	2,4,6	0.52
		2,5,6	0.41
		3,4,5	0.47
		3,4,6	0.52
		3,5,6	0.51
		4,5,6	0.50

With each added predictor variable the R^2 estimates are increasing slightly in value.

Table Five completes the multiple regression computations using sets of four, five and six predictor variables. The largest R^2 value will determine which regression equation is selected.

Table 5

Explained Variance Accounted for in the Criterion Variable Using Quadruples, Quintuples, and A Sextuplet of Predictor Variables

Predictor Quadruples	R^2	Predictor Quintuples	R^2
1,2,3,4	0.52	1,2,3,4,5	0.53
1,2,3,5	0.49	1,2,3,4,6	0.57
1,2,3,6	0.54	1,2,4,5,6	0.55
1,2,4,5	0.48	1,3,4,5,6	0.56
1,2,4,6	0.52	2,3,4,5,6	0.57
1,3,4,5	0.42		
1,3,4,6	0.55	Predictor Sextuplet	R^2
1,3,5,6	0.53	1,2,3,4,5,6	0.57
1,4,5,6	0.51		
2,3,4,5	0.53		
2,3,4,6	0.56		
2,3,5,6	0.54		
2,4,5,6	0.53		
3,4,5,6	0.54		

An R^2 of 0.57 is the largest computed value for a multiple correlation coefficient. There are three combinations with this value (1,2,3,4,6), (2,3,4,5,6), and (1,2,3,4,5,6). Therefore, one of these combinations is used as the source of the regression equation. The sextuplet is selected because the other two are values that have been rounded up while the sextuplet value has been rounded down.

The standardized regression equation is
$Z_y = b_1z_1 + b_2z_2 + b_3z_3 + b_4z_4 + b_5z_5 + b_6z_6$.
Using the computed b_i values, the regression equation becomes
$Z_y = 0.088\ Z_1 + 0.169\ Z_2 + 0.272\ Z_3 + 0.237\ Z_4 - 0.088\ Z_5 + 0.279\ Z_6$

There is a procedure, suggested by Cohen and Cohen, for conducting a computational error check on the b_i's. (Cohen andCohen,1983) This check identifies the presence of an accumulated error as well as the presence of a computational blunder. The process is to examine the magnitude of the deviation from an identity for each b_i. The identity is illustrated for b_1.
$b_1 = r_{Y1} - (b_2 * r_{12}) - (b_3 * r_{13}) - (b_4 * r_{14}) - (b_5 * r_{15}) - (b_6 * r_{16})$. A similar identity is examined for b_2 - b_6. The observed error deviations are shown in Table 6.

Table 6
Error Estimates in the Computed b Values

b_i	Computational Error
b_1	0.0000001
b_2	0.0000004
b_3	0.0000007
b_4	0.0000008
b_5	0.0000008
b_6	0.0000003

The computational errors are trivial.

The raw score regression equation is computed from the standardized b_i's and the observed mean values of the predictor variables.
$Y_i' = 2.759 + [0.077 * (X_1 - 3.380)] + [0.146 * (X_2 - 3.209)]$
$+ [0.243 * (X_3 - 2.877)] + [0.208 * (X_4 - 2.926)]$
$- [0.060 * (X_5 - 3.083)] + [0.213 * (X_6 - 2.759)]$.

The standard error of multiple estimate is used to build a confidence interval around the predicted score Y_i'. $S_{error}Y_i' = SD_Y$ Ö $1 - R^2$.
In this example $S_{error}Y_i' = 0.7454$ * Ö $1 - 0.572 = 0.7454$ * Ö 0.428
$= 0.7454 * 0.6542 = 0.4876$.
The confidence interval is constructed from the relationship
$S_{error}Y_i'$ * standard normal Z score at desired confidence level. A confidence level of 85% was selected, and is arbitrary.
$S_{error}Y_i' = 0.4876 * 0.8023 = 0.3912$. Therefore, the raw score predicted value with an 85% confidence interval is $Y_i' \pm 0.3912$. This procedure was adapted from that shown in a statistics book by Hays. (William L. Hays, 1988)

Two Illustrative Examples: Assume the grade point system is based on a 4-point scale, which includes 2.5 for a C+ and 3.5 for a B+. Suppose a student has received all C's in the six-predictor variable courses. That student's predicted junior/senior level of performance is Y' = 1.974 with its 85% confidence interval of ± 0.3912. Thus, the predicted level of performance would be between 2.37 and 1.58. With 85% confidence, this student's faculty advisor can inform the student that the predicted junior/senior performance will range between a near middle C and a middle D. Some conversation about whether this suggests that computer science may not be a good career choice for this student is also appropriate.

As a second example, suppose that a student has repeated the systems course and the data structures course having received an F in both courses and a C in both courses when repeated. Using all grades received the student's profile in the six predictor variable courses is B, C, (F + C = D), (F + C = D), B, and C or 3, 2, 1, 1, 3, and 1. Applying the regression equation yields the result Y' = 1.2765 with a confidence interval of ± 0.3912. With 85% confidence, the student's faculty advisor can say the student's predicted level of performance would be between 1.67 and 0.89. With 85% confidence, this student's faculty advisor can say the student's predicted junior/senior performance will range between a high D and an F. The faculty advisor would also include in the advisement conversation that this predicted result does not foretell a promising career in computer science. A reference to that truth that everyone has different talents might be useful to include in the conversation. Reminding the student that finding a match between an individual's talents and selecting an appropriate career is something that an undergraduate university experience is meant to include is also important for such a student.

Recommendations for the Application of the Regression Analysis: Once the six-predictor variable courses have been completed at the level of C or better, the department computes that student's predicted junior/senior level of performance. A letter stating the results of the regression is prepared for the student and another copy for the student's parent(s) or guardian(s). At the scheduled advising appointment, the faculty advisor gives the letter to the student and explains its meaning for the student's future as a computer science major. A similar letter is sent to the parent(s) or guardian(s) with an explanation. It is important to note that this is not a cut-score procedure. The student is under no obligation to change majors as a result of the predicted future performance.

One desired result of this procedure is to help students who do not show academic promise in computer science to recognize this reality and to make a mature decision to find a more appropriate career path. A serendipitous result of this procedure may be to alert those students working at the C level that a greater effort may be necessary when taking the junior and senior level courses. It may serve as an intellectual wakeup call for some non-high risk students.

Each new class of seniors should be added to the data pool for the regression study. This will increase the accuracy of the regression analysis and multiple correlation estimates of the true population values for b_i's and R^2.

Some Concluding Remarks: A common departmental solution to the problem that prompted this study is to implement a cut-score. That is, some grade point average in some prescribed courses that a student must achieve before becoming an acknowledged computer science major. This is a perfectly acceptable solution, but it lacks humanity. Caring about the students leads to a desire to have the student make the informed decision.

We, as professors, appreciate that our responsibility is more than helping students to achieve a mastery of the content of our intellectual discipline and its problem solving heuristics. We also accept the responsibility for aiding each university student in the development of his or her adult maturity. Learning to use data to help with making both professional and personal decisions is an intimate part of this development. The regression solution contributes more effectively to this latter responsibility than does a cut-score solution.

Epilogue: At the faculty retreat on August 17,2001, the computer science department faculty decided to use the regression prediction as a part of faculty advising at the end of the sophomore year. The provision that the regression prediction be sent to parents was accepted by the department with the provision that it was legal to do so. The

university decided that parents could not be sent the information because of the confidentiality rules concerning student grades.

It was also agreed that subsequent years of data would be added to the regression analysis.

As a result of this decision, the raw score regression equation is available to each advisor through the department Web page. A faculty member inputs the values of the six-predictor variables for a student and the output is the predicted GPA at the junior senior years along with the confidence interval.

References

Jacob Cohen and Patricia Cohen. *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences, Second Edition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates,Publishers, 1983.

William L. Hays. *Statistics, Fourth Edition.* Fort Worth: Holt Rinehart and Winston, Inc., 1988.

Henry H. Walbesser and Cheryl Gonce-Winder. *Data Analysis: Volume One, Descriptive.* London: D. L. Walker Book Company, 1991.

Henry H. Walbesser and Cheryl Gonce-Winder. *Data Analysis: Volume Three, Factorial Designs.* London: D. L. Walker Book Company, 1992.