Southern Illinois University Carbondale
OpenSIUC

Research Papers

Graduate School

Fall 10-27-2017

Predicting Student Success: A Logistic Regression Analysis of Data From Multiple SIU-C Courses

Patrick Soule
psoule212@gmail.com

PREDICTING STUDENT SUCCESS:

A LOGISTIC REGRESSION ANALYSIS

OF DATA FROM MULTIPLE SIU-C COURSES

by

Patrick B. Soule

B.S., Southern Illinois University, 2015

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale

December, 2017

RESEARCH PAPER APPROVAL

PREDICTING STUDENT SUCCESS:

A LOGISTIC REGRESSION ANALYSIS

OF DATA FROM MULTIPLE SIU-C COURSES

By

Patrick B. Soule

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

Dr. B. Bhattacharya, Chair

Dr. M. Wright

Dr. R. Habib

Graduate School
Southern Illinois University Carbondale
July 18 2017

AN ABSTRACT OF THE RESEARCH PAPER OF

PATRICK B. SOULE, for the Master of Science degree in MATHEMATICS, presented on JULY 18 2017, at Southern Illinois University Carbondale.

TITLE: PREDICTING STUDENT SUCCESS: A LOGISTIC REGRESSION ANALYSIS OF DATA FROM MULTIPLE SIU-C COURSES

MAJOR PROFESSOR: Dr. B. Bhattacharya

The objective of this report is to improve prediction techniques regarding the future performance of students in select university courses through the utilization of multiple logistic regressions. This is achieved with the aid of statistical computing software which applies forward step-wise variable selection methods that identify influential variables sufficient to accurately predict student success. Once a logit model is constructed with the required parameters and predictors, the inverse logit function outputs a probability of student success. In all cases, logistic prediction models matched or exceeded the performance of current prediction methods while using an equal or lesser number of explanatory variables. These findings show that current prediction methods can improve by using a statistically justified procedure. It also suggests the inefficacy of some predictors used to currently estimate student performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Southern Illinois University Carbondale has recently begun using an early warning intervention program (EWIP). This program aims to detect struggling students in general education core courses early in the semester. Once a student is classified as struggling or at risk of not earning a grade of C or higher there are several ways the university initiates intervention. The student may be contacted by the instructor, an academic advisor or a residence hall academic peer associate who, after consultation, then best directs them to additional resources. This research paper investigates to detect such students, but does not address the subsequent intervention.

Each course participating in EWIP has its own predictive functions which are employed early in the semester, usually at the end of week three. The output of each function estimates a score between zero and one hundred percent which is then classified into one of four color codes. The following guidelines are used:

$$\text{Green} > .75$$
$$.75 \geq \text{Yellow} > .65$$
$$.65 \geq \text{Orange} > .55$$
$$.55 \geq \text{Red}$$

The precise cutoffs are tailored to each course but signify the same warning; the coding scheme labels students 'least at risk' to 'most at risk' by assigning the colors green, yellow, orange and red respectively. The goal of the university is to maximize the number of green coded students. Concern arises when students are misclassified, as is possible with an estimating function. The current models that determine academic risk are working well however were created ad hoc and are not, probably, generated statistically. Also, it does not use the final grade in consideration.

The current EWIP functions are linear in their predictors with weighted coefficients derived from specific knowledge of each course. It is not known if the individual predictors being used are best or if the coefficients are optimal. Using data to determine the weights of influential predictors provides a way to maximize the chance of correct predictions and therefore minimize misclassifications.

Optimization means that the university's limited available resources can be efficiently applied to the correct population with the greatest need. Successful interventions lead to students passing the course or avoiding a D, F or WF grade. Further, this leads to fewer students on academic probation or suspension which protects university retention rates.

By statistically analyzing previous student performance, we build multiple logistic regression models that predict student success. The binary nature of the equations best lend themselves to partitioning the response into two categories and not four. Whereas it is possible to assign color codes, particularly in an ordinal regression model, the end goal is to determine whether the student requires intervention. It is with this end goal in mind that the two methods will be compared.

CHAPTER 2

THE DATA

CLEANING

In this report the records of four Southern Illinois University Carbondale courses from 2015 fall semester are analyzed. These include: BIO200A, MATH101, MATH106 and MATH108. We combine all sections of a course for one semester and treat them as one sample. Records are collected in the third and eighth week.[1] The early week class sizes rarely equal the size of the samples at week eight: invariably students drop, withdraw or transfer from the course. Future records are subsequently left lacking; likewise, eighth week records occasionally contain students who transfer in after week three. In these events we use case-wise deletion.

We use the terms: variable, covariate, predictor, response variable, explanatory or independent variable. A variable is a quantity that is unknown, a covariate is a potential predictor of the response and explanatory or independent variable are used interchangeably. Descriptives, multicollinearity diagnostics, model fitting and model selection for BIO200A are covered in detail. The other courses follow similarly with figures to summarize results.

DESCRIPTIVE STATISTICS

BIO200A has records for students regarding Pretest, Homework, Quiz and Test scores as well as Attendance. Week three scores for Quiz are unfortunately incomplete until week eight.[2] Each of these explanatory variables is converted to a percent so they can be directly compared. The descriptive statistics for the potential predictors act as an additional check for errors incurred from data entry by illuminating any extreme outliers. Histograms, box-and-whisker plots and scatterplots help to show distribution and trends. From Figure 2.1 it is clearly observable that not all covariates are normally distributed. Homework and

[1] MATH101 collects in the fifth and eighth weeks due to course structure.

[2] Not all sections of BIO200A have a Quiz score recorded by week three.

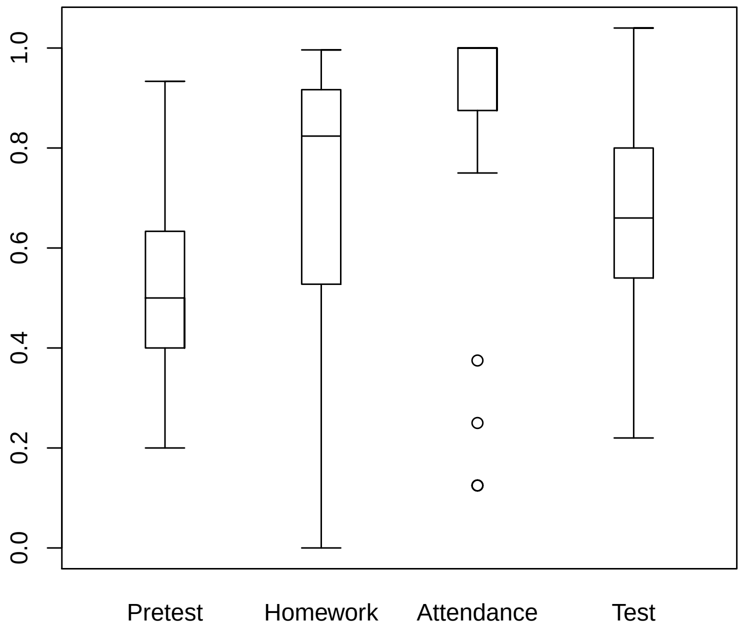

Figure 2.1. Boxplots of Covariates

Attendance are strongly skewed left. However, since logistic regression is used as opposed to linear, these variables don′t require a transformation.

CORRELATIONS

PerformanceAnalytics [2] is a convenient R [9] package that combines paired histograms, scatterplots, and correlation coefficients in a square matrix, see Figure 2.2. Displaying the data in this way helps to visualize how the explanatory variables relate to one another. In the lower triangular portion of the matrix the scatterplots have overlaid trend lines that attempt to describe the relationship. Note the effect caused by zeros in the data. For example, between Pretest and Test the trend line is strongly affected, creating the perception of a quadratic effect. This is considered an artifact of the smoothing technique. If the relationship were true it would imply, to a point, that higher pretest scores lead to decreased test scores.

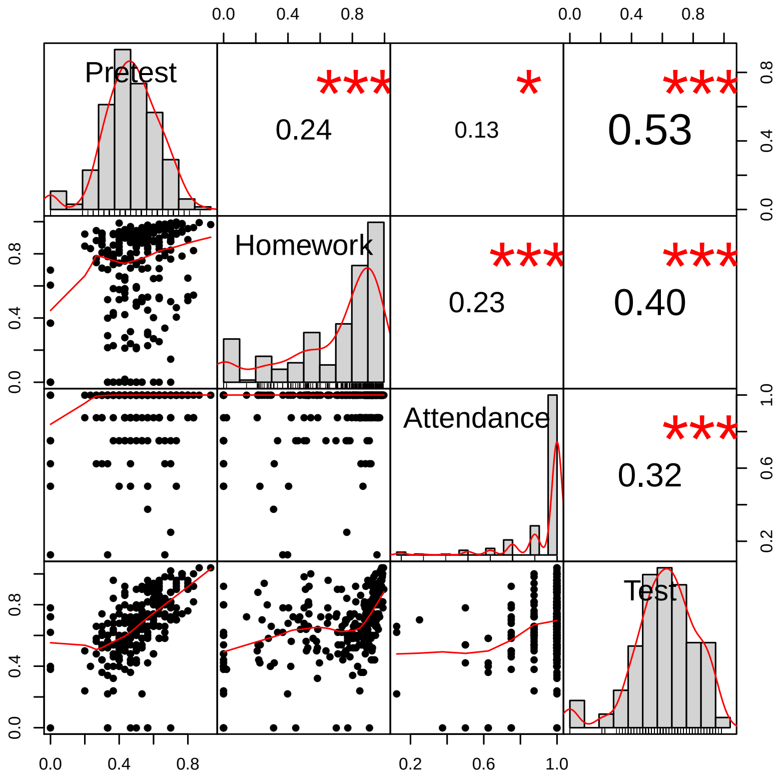

Figure 2.2. Descriptives and Correlations of Covariates

Pearson's R

Bivariate Pearson correlation coefficients are reported in the upper triangular portion of the matrix in Figure 2.2. Pearson′s r is a measure of the linear relation between two variables and is determined by calculating

$$\frac{\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2)}{\left[\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)^2\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2\right]^{1/2}}, \tag{2.1}$$

where x_1 and x_2 are distinct covariates and $\bar{x}$ is the arithmetic mean. Pearson′s r is bound between -1 and 1. A value of ±1 indicates a perfect positive or negative linear relationship, whereas 0 signifies no linear relationship between the variables.

Test and Pretest exhibit a moderate correlation with r = 0.53. Using Fisher′s z′

transformation, we find the 95% confidence interval about r to be (0.425, 0.612). The large sample size of 234 helps to keep the range relatively small. A Student′s t test is also calculated to determine if there is enough evidence to suggest a linear relationship in the population. Using the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{2.2}$$

and substituting our value of r, we have $t = 9.403$ on $n - 2 = 232$ degrees of freedom resulting in a p-value less than 0.0001. It should be noted that using both methods is redundant, since the above confidence interval does not contain zero.

Spearman′s rho & Kendall′s tau

Recall that not all covariates are normally distributed, in fact any other bivariate combination will require nonparametric methods. Spearman′s rho and Kendall′s tau are rank correlation coefficients that measure monotonic relationships. Monotonic functions cover a broader set of relationships than simply linear, specifically nonincreasing or non-decreasing functions. Using either method requires ranking the data and comparing pairs. We report both Spearman and Kendall coefficients together in Table 2.1. The Spearman rho values appear in the upper triangular portion of the matrix and Kendall's tau in the lower triangular portion.

Table 2.1. Matrix of Nonparametric Correlation Coefficients

Covariates	**Pretest**	**Homework**	**Attendance**	**Test**
Pretest	-	0.28^{***}	0.05	0.62^{***}
Homework	0.20^{***}	-	0.21^{**}	0.46^{***}
Attendance	0.04	0.17^{**}	-	0.25^{***}
Test	0.46^{***}	0.33^{***}	0.20^{***}	-

Asterisks denote level of statistical significance. *, **, and *** correspond to $p < 0.05$, < 0.01, and < 0.001 respectively.

Most values are low except when measuring Test versus Pretest and Test versus Homework. Spearman′s rho suggests a slightly increased correlation for each, this could be due to Spearman′s rank being a less robust statistic [3]. At any rate, they are still within the moderate range for the statistic.

MULTICOLLINEARITY

To build an informative and predictive logistic regression model, the explanatory variables must be independent of each other. Measuring correlations is one way to evaluate how much one variable explains another. If two variables essentially behave the same, including both for modeling purposes introduces redundancy as well as increases the size of parameter errors. Unfortunately, even if pairwise comparisons show low correlation coefficients there is still the possibility of collinearity. If this occurs, it becomes difficult to determine the distinct effects each covariate has on the response variable.

Although there is no definitive test for multicollinearity, there are diagnostic measures that are commonly used to assess the degree to which collinearity may be present. One such measure involves regressing each explanatory variable on all other covariates. If the coefficient of determination (R^2) is high, multicollinearity may be present. Of course high is a relative term, some [4] argue that values above 0.8 are concerning. The coefficient of determination is the portion of variation in the regressing variable that is explained by the other covariates and is calculated as follows.

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.3}$$

where $\hat{y}$ is the fitted value that the model predicts. This is a ratio of variances. The numerator of (2.3) is the sum of squared residuals (SSR), the squared difference between model predictions and actuality. The denominator represents total variance in the model. As the SSR increases, R^2 approaches 1. For our purposes we want R^2 to be low, otherwise the other covariates are explaining much of what is assumed to be an independent variable. In Table 2.2 we include two other measures that are related to the coefficient of determination: Tolerance $= 1 - R^2$, and Variance Inflation Factor (VIF) $= 1/(1 - R^2)$. Following the rule of thumb for $R^2 < 0.8$ this corresponds to Tolerance > 0.2 and a VIF < 5.

Table 2.2. Collinearity Measures

Statistic	**Regressor**			
	Pretest	Homework	Attendance	Test
R^2	0.279	0.172	0.117	0.393
Tolerance	0.721	0.828	0.883	0.607
VIF	1.386	1.208	1.133	1.647

VIF is interpreted [4] as the percentage of inflation in the variance of the coefficient due to collinearity. For Test, VIF = 1.647 which suggests that its variance is inflated by 65%. This may sound worse than it is. Consider the R^2 value of 0.393 which suggests that almost 40% of the variability of Test is explained by the other three covariates in the linear regression model. This is not enough to cause alarm. However, these results as well as the ones from the bivariate correlation coefficients suggest that Pretest and Test are likely the greatest cause of any potential collinearity.

CHAPTER 3

THE MODEL

BINOMIAL DISTRIBUTION

To make accurate statistical inferences we need to make certain assumptions about the distribution of the binary response variable. If we assume that each student is independent[1], that is one student′s success does not affect another, then we can assign a probability distribution to our response variable. In terms of our data on BIO200A, there were 234 students and 165 of those earned a grade of C or above. We let this proportion, 0.705 be a point estimate for the proportion of success of this class. Assume next semester BIO200A enrolls 250 students. We present the approximate normal distribution of Z in Figure 3.1. The distribution′s peak [2] represents the most likely outcome. In order to predict how many students will succeed in the next semester you may calculate the mean of the binomial distribution as $\mu = \pi n \approx 176$. This gives an estimate for the number of students to succeed, but which ones? To address this we consider the use of given covariates to construct a model.

LOGISTIC REGRESSION

Whether or not a student succeeds is the binary response variable we are trying to predict. Mathematically, we let Y represent the response as

$$Y = \begin{cases} 1 & \text{if student earns a grade of A, B or C} \\ 0 & \text{otherwise} \end{cases}$$

[1]. . . and identically distributed (a strong assumption for a student body). . .

[2]If π is left fixed, as n becomes large the binomial distribution approximates the normal curve.

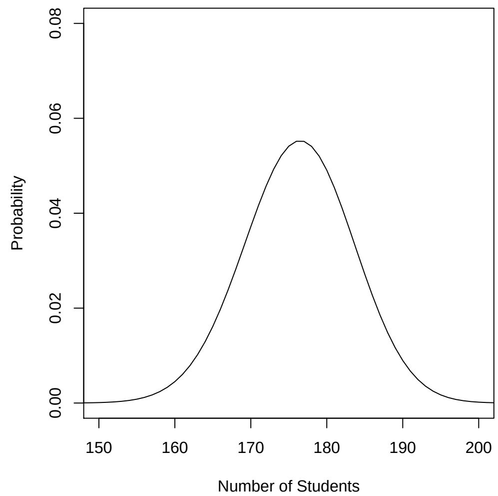

Figure 3.1. Approximate Normal Distribution of Z

The logistic regression function is in terms of success

$$\text{logit}[P(Y = 1)] = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \tag{3.1}$$

where $\pi(x)$ is equivalent to the probability of success, P(Y=1). Note that the right hand side (RHS) of (3.1) is linear in x and the LHS is a logarithmic scale of the odds of success. Although odds are commonly used and interpreted, the natural log function is not. With exponentiation and algebraic manipulation we yield

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{3.2}$$

to obtain a form that outputs only values between 0 and 1 which then are interpreted as probabilities. The equations (3.1) and (3.2) are simple in the fact that they only use one explanatory variable x. Using the method of maximum likelihood estimation we will build several different models of increasing complexity and choose one based on best fit criterion.

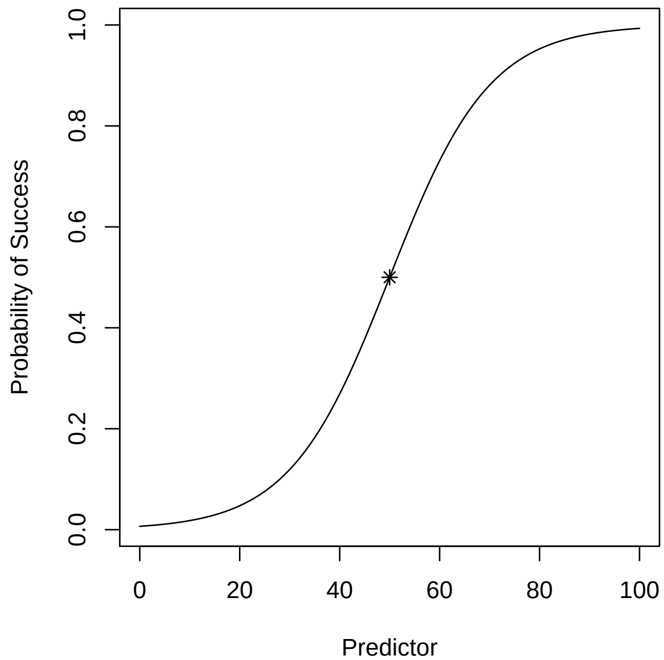

Figure 3.2. Logistic Regression Curve

Model Interpretation

Let us consider the following model.

$$\text{logit}[P(Y = 1)] = -5 + 0.1x$$

The coefficient of 0.1 is sometimes referred to as the weight of the predictor. Avoiding log interpretations, $e^{0.1} \approx 1.11$ translates to a 11% increase in the odds of success given one unit increase in x. We plot the function in Figure 3.2 and mark the point $\pi(50) = 0.5$. Note the nonlinear S-shaped curve that is monotonically increasing, which is typical in practice [1]. The slope of the curve is steepest at this point, which means volatility for a binary variable. There is a fifty percent chance of success which also means the same chance for failure.

METHOD OF MAXIMUM LIKELIHOOD

R uses the Newton-Raphson algorithm to find the maximum likelihood estimates (MLE) for each parameter in a model. It is an iterative method that involves an initial guess from a polynomial function, followed by successive iterates that rapidly converge to the maximum likelihood estimates [1]. This process determines the MLE′s for the weights in the model, but we need a way to determine what predictors should be in the model. We want a model that sufficiently describes the real life phenomenon. Any model is a simplification of reality but we must balance between too simple and over complex. Simple models are easier to interpret, but a more complex model may fit the data better. To assess this, we require that a model have statistical significance as well as conform to several goodness of fit tests.

STEPWISE VARIABLE SELECTION

We now build models with a single covariate first and check which produces the best fit. The covariate that provides the best fit to the data is selected and used in the next step. We then increase the complexity of the selected model by choosing one more predictor from the remaining and comparing fit. This process continues until increasingly complex models fail to produce a better fit. This is known as forward variable selection.

There are five measures of fitness that we will consider at each step: Akaike Information Criterion (AIC), deviance, concordance, McFadden′s pseudo R^2, and the significance of model parameters. AIC provides a measure of the information that a model provides. The calculation involves a penalty term which acts to preserve parsimony of parameters. We want to minimize

$$\mathrm{AIC}_k = -2[\text{log likelihood} - 2k] \tag{3.3}$$

where k is the number of parameters in the model. A similar statistic is deviance. Whereas AIC is measuring the tested model against the theoretically true model, deviance measures against the most complex model possible, a saturated model with an individual parameter for each observation. If we let L_T be the log likelihood of the tested model and L_S be the log likelihood of the saturated model, deviance is calculated as

$$\text{Deviance} = -2[L_T - L_S] \tag{3.4}$$

The deviance likelihood ratio statistic tests the hypothesis that all parameters not used in the tested model are zero. Deviance follows an approximately chi-squared distribution where the degrees of freedom are the number of observations minus the number of parameters used. Models can also be compared using the deviance statistic. If the parameters in a model are a subset of those in a more complex model, the difference in their deviances can be calculated and interpreted as a chi-square test for the hypothesis that the more complex model provides a better fit. A sufficiently high statistic would be evidence to suggest that the complex model is necessary.

Our third statistic measures predictive power. A receiver operating characteristic (ROC) curve is a plot of model sensitivity versus (1−specificity). The sensitivity of a model can be defined as the probability that the model predicts success given the student actually succeeded and specificity is the probability the model predicts failure given they failed.[3] Figure 3.3 displays the ROC curve for the model with 'Test' as the only explanatory variable.

The diagonal line represents the intercept model which is equivalent to guessing, essentially flipping a coin to decide. The desired shape is a high parabolic arch filling the upper left portion of the plot. Total area under the curve is equivalent to concordance. Pairing up the observed successes with observed failures, concordance checks that the model is assigning a higher probability to the successes and not the failures. A value of c = 0.5 is equivalent to guessing.[4]

[3]Failure here is defined as not succeeding in earning a grade of C or higher.

[4]It is possible to build a model that predicts worse than c = 0.5.

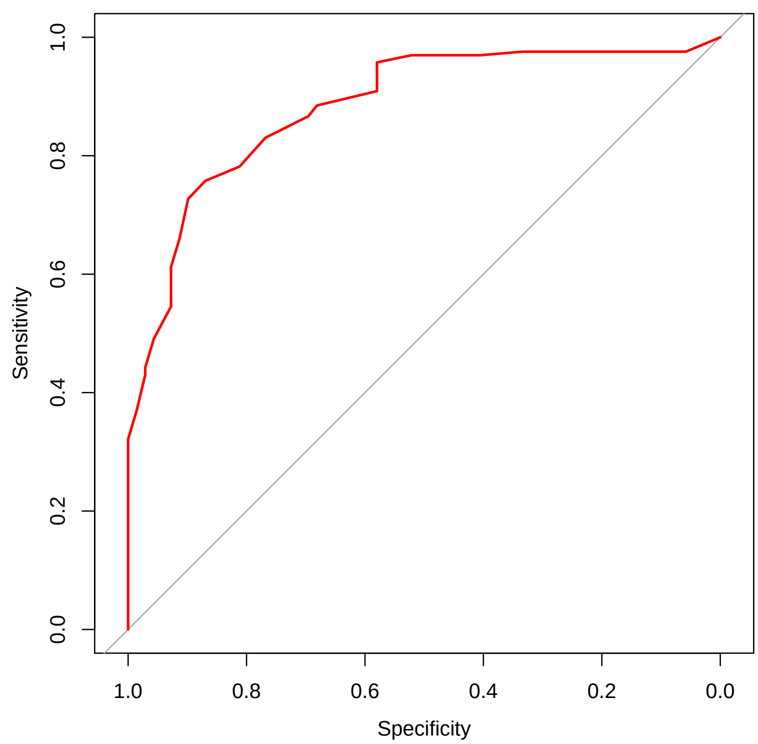

Figure 3.3. ROC Curve

McFadden′s pseudo R^2 is so-named because it does not compare variances as does R^2 but it does range over similar values: from 0 to 1. It is defined by subtracting the ratio of the log likelihood of the fitted model to the log likelihood of the intercept model from 1. General impression is that a higher value is more desirable with ′good′ fitting models in the $0.2 - 0.4$ range.[7]

Lastly, one can always test the statistical significance of the individual parameters used in the model. Along with the MLE for the parameters are standard errors (SE) which can be used to calculate confidence intervals about the estimates as well as test their significance, similar to what we saw in Section 2. The null hypothesis for logistic regression parameter significance states that the response is independent of the explanatory variable. To test this we calculate

$$z^2 = \left(\hat{\beta}/\text{SE}_{\hat{\beta}}\right)^2 \tag{3.5}$$

which has a chi-squared null distribution. Large values provide evidence that the predictor is affecting the response.

With these measures of model fit as a guide we proceed through a stepwise forward variable selection. We exhibit an instance of the process for week three BIO200A.

An Example

Table 3.1. Model Diagnostic Measures for Biology 200A Week 3

	Predictors				
	Intercept	**Pretest**	**Attendance**	**Homework**	**Test**
MLE	0.87	5.30	3.75	3.56	8.32
SE	0.14	1.06	1.02	0.56	1.27
χ^2	36.96	24.80	13.59	40.54	42.81
P value $<$	0.0001	0.0001	0.0001	0.0001	0.0001
AIC	285.82	256.04	271.22	237.40	207.06
Deviance	283.82	252.04	267.22	233.40	203.06
Concordance	0.5	0.74	0.62	0.80	0.88
McFadden's R^2	0	0.11	0.05	0.18	0.28

Statistical software calculates the estimates and measures of fit for BIO200A for week three which we summarize in Table 3.1. We omit the intercepts from one predictor models for the sake of clarity. The Test predictor is showing the best fit across multiple measures. Therefore 'Test' will be added to the logistic regression model and we proceed to the next step in variable selection. Table 4.5 shows all models considered. Adding 'Homework' substantially improves model fit compared to 'Pretest' and 'Attendance'. In the third step we see that 'Pretest' should be chosen over 'Attendance' however, both variables are not statistically significant. This could be a sign of multicollinearity. Notice that in the previous step, those predictors were also found to be not statistically significant. Perhaps 'Test' predictor is already explaining the contribution they could provide. With this and the principle of parsimony in mind, we choose 'Test' and 'Homework' as the predictors in the final model.

CHAPTER 4

RESULTS

MODEL CHOICE

Tables 4.5 through 4.12 contain model candidates for each course and are subdivided by week. Unsurprisingly, ‘Test’ is by far the most influential predictor in all instances. Models with just ‘Test’ as a predictor fit well but adding either ‘Homework’ or ‘Quiz’ predictors, when applicable, always increases the fit. Any models that have three terms are comprised only of the aforementioned predictors. ‘Pretest’ or ‘Attendance’ models only outperform an intercept model. As other terms are added the model fit marginally improves along with an undesired increase in parameter errors which lead to insignificant predictors. We conclude that ‘Pretest’ and ‘Attendance’ have little predictive power compared to the other covariates.

The final model choice for BIO200A at week three includes ‘Test’ and ‘Homework’ whereas the week eight model differs by the addition of ‘Quiz’. This suggests that if data were available in week three it could potentially improve model fit. MATH101 model uses all three predictors whereas MATH106 and MATH108 use ‘Test’ and ‘Homework’ only. In Table 4.1 we present the final models complete with parameter estimates, predictors are abbreviated to their first letter. The weight of each coefficient alone is not telling, rather the weight relative to others is what lends insight. Notice for example that Test in week three for BIO200A is weighted over twice as much as ‘Homework’. As the semester progresses, ‘Test’ in week eight is weighted over three times as much as ‘Homework’.

In Figure 4.1 we illustrate the influence of each predictor by plotting the logistic regression curve for final models of week three and week eight. Each curve holds the other covariates fixed at their medians. We see that in week eight the probability of success is very volatile inside the 0.4 to 0.5 percent range. Students earning above 50% as their average test percent have a very high probability of earning a C grade or higher, assuming they have

Table 4.1. Final Models

Course	Week	Model
BIO200A	3	$\text{logit}[\pi(x)] = -\ 5.74 + 7.49 * \text{T} + 3.01 * \text{H}$
	8	$\text{logit}[\pi(x)] = -\ 55.46 + 47.90 * \text{T} + 14.96 * \text{H} + 28.68 * \text{Q}$
MATH101	5	$\text{logit}[\pi(x)] = -\ 6.28 + 4.55 * \text{T} + 3.58 * \text{H} + 0.97 * \text{Q}$
	8	$\text{logit}[\pi(x)] = -\ 9.78 + 7.55 * \text{T} + 4.47 * \text{H} + 2.75 * \text{Q}$
MATH106	3	$\text{logit}[\pi(x)] = -\ 4.21 + 0.04 * \text{T} + 0.04 * \text{H}$
	8	$\text{logit}[\pi(x)] = -\ 13.98 + 0.13 * \text{T} + 0.08 * \text{H}$
MATH108	3	$\text{logit}[\pi(x)] = -\ 4.58 + 0.05 * \text{T} + 0.03 * \text{H}$
	8	$\text{logit}[\pi(x)] = -\ 9.57 + 0.10 * \text{T} + 0.05 * \text{H}$

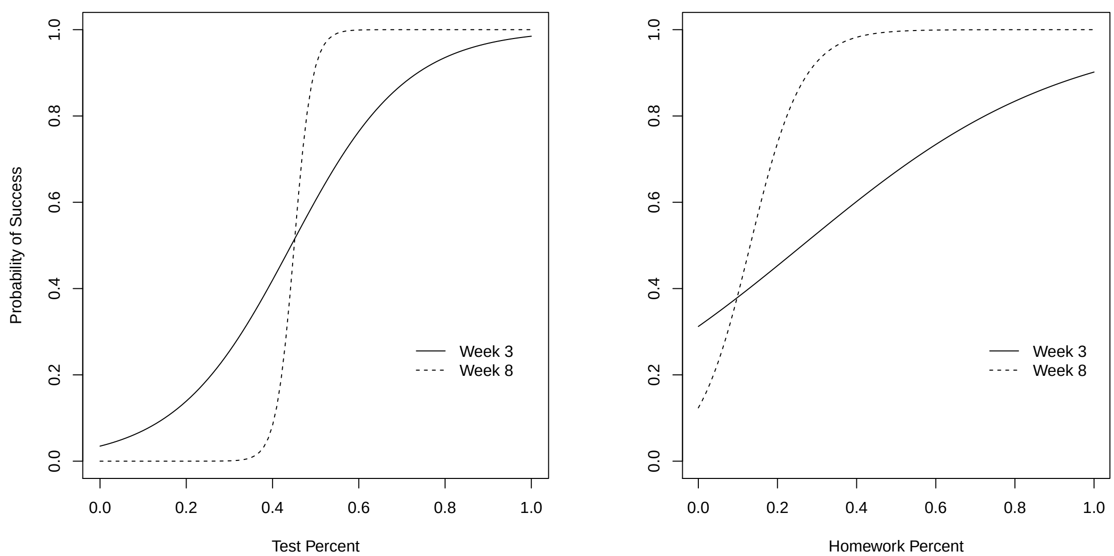

Figure 4.1. Between Week Comparisons for BIO200A Predictors

median homework and quiz scores. 'Homework' on the other hand has an almost linear relationship early semester whereas the mid-semester plot rapidly approaches a success probability of 1. This is due in part by the fixing of 'Quiz' and 'Test' at their median values of .76 and .66 respectively.

COMPARISON

Having selected the 'best' fitting model for each course we measure how well they perform compared to current predicting functions. Table 4.4 displays the current functions used for each course. Current functions predict a final course score out of one hundred. We

Table 4.2. Early Semester Classification by Function

Course	Predicting Function		Observed Success	Observed Failure	Correct Classifications
BIO200A	Linear	Success	139	15	82.48%
		Failure	26	54	
	Logistic	Success	146	16	85.04%
		Failure	19	53	
MATH101	Linear	Success	452	162	74.97%
		Failure	34	138	
	Logistic	Success	371	76	75.60%
		Failure	115	224	
MATH106	Linear	Success	60	32	70.63%
		Failure	5	29	
	Logistic	Success	74	18	76.98%
		Failure	11	23	
MATH108	Linear	Success	236	95	75.72%
		Failure	23	132	
	Logistic	Success	248	83	76.95%
		Failure	29	126	

Table 4.3. Mid-Semester Classification by Function

Course	Predicting Function		Observed Success	Observed Failure	Correct Classifications
BIO200A	Linear	Success	144	3	89.74%
		Failure	21	66	
	Logistic	Success	162	3	97.44%
		Failure	3	66	
MATH101	Linear	Success	440	91	82.47%
		Failure	46	209	
	Logistic	Success	390	48	81.58%
		Failure	96	252	
MATH106	Linear	Success	74	18	80.95%
		Failure	6	28	
	Logistic	Success	80	12	85.71%
		Failure	6	28	
MATH108	Linear	Success	262	69	82.72%
		Failure	15	140	
	Logistic	Success	270	61	83.33%
		Failure	20	135	

change this to a binary outcome by consulting the corresponding syllabus which determines the cutoff point for earning a C or higher. Then predicted outcomes are compared with actual observed data. There are four potential cases when predicting behavior. Correct classification is when the observed success or failure matches prediction. Tables 4.2 and 4.3 detail the results for all courses.

Table 4.4. Current Predictive Functions

Course	**Function**
BIO200A	0.15*P + 0.35*(A + H) + 0.50*T
MATH101	0.25*Q + 0.25*H + 0.50*T
MATH106/108	Max (0.25*P + 0.25 *H + 0.50*T, 0.25*H + 0.75*T)

Overall, with logistic regression the correct prediction rate improves for early week and mid-semester predictions except for MATH101. Improvements are more pronounced in mid-semester than in early week for BIO200A and MATH106. In MATH101, current functions and logistic models use the same predictors but with different weights. Since MATH101 is a modular course composed of four disjoint topics and a cumulative final, trends seen early semester may not be indicative of late semester performance. This is a possible explanation for the lack of improvement for the logistic model in week eight.

Table 4.5. Biology 200A Week 3 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	285.8	283.8	0.500	N/A	None
T	207.1	203.1	0.883	0.285	None
H	237.4	233.4	0.803	0.178	None
A	271.2	267.2	0.622	0.058	None
P	256.0	252.0	0.739	0.112	None
T + H	**187.0**	**181.0**	**0.909**	**0.362**	**None**
T + A	205.4	199.4	0.892	0.298	A
T + P	205.9	199.9	0.876	0.296	P
T + H + A	186.5	178.5	0.916	0.371	A
T + H + P	185.3	177.3	0.901	0.375	P
T + H + A + P	221.1	174.5	0.907	0.385	A

Table 4.6. Biology 200A Week 8 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	285.8	283.8	0.500	N/A	None
T	123.7	119.7	0.949	0.578	None
H	203.1	199.1	0.859	0.298	None
A	245.2	241.2	0.752	0.150	None
P	256.0	252.0	0.739	0.112	None
Q	133.2	129.2	0.940	0.545	None
T + H	75.6	69.6	0.983	0.755	None
T + A	105.9	99.9	0.964	0.648	None
T + P	123.8	117.8	0.951	0.585	P
T + Q	56.2	50.2	0.991	0.823	None
T + Q + H	**35.8**	**27.8**	**0.997**	**0.902**	**None**
T + Q + A	44.0	36.0	0.997	0.873	None
T + Q + P	58.0	50.0	0.992	0.824	P
T + Q + H + A	10.0	0.0	1.000	1.000	All
T + Q + H + P	36.5	26.5	0.998	0.907	P

Table 4.7. Math 101 Week 5 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	1046.2	1044.2	0.500	N/A	None
T	859.6	855.6	0.789	0.181	None
H	864.2	860.2	0.780	0.176	None
Q	930.7	926.7	0.725	0.113	None
T + H	796.5	790.5	0.829	0.243	None
T + Q	833.2	827.2	0.807	0.208	None
T + H + Q	**794.2**	**786.2**	**0.830**	**0.247**	**None**
T + H + T*H	782.9	774.9	0.828	0.258	T & H
T + H + Q + T*H	780.2	770.2	0.830	0.262	T & H

Table 4.8. Math 101 Week 8 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	1046.2	1044.2	0.500	N/A	None
T	745.1	741.1	0.849	0.290	None
H	778.3	774.3	0.830	0.258	None
Q	792.2	788.2	0.825	0.245	None
T + H	656.9	650.9	0.886	0.377	None
T + Q	692.1	686.1	0.872	0.343	None
T + H + Q	**644.1**	**636.1**	**0.891**	**0.391**	**None**
T + H + Q + T*Q	642.9	632.9	0.872	0.394	T & Q

Table 4.9. Math 106 Week 3 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	148.9	146.9	0.500	N/A	None
T	130.5	126.5	0.800	0.139	None
H	132.3	128.3	0.750	0.127	None
P	146.5	142.5	0.665	0.030	Intercept
T + H	**124.8**	**118.8**	**0.843**	**0.191**	**None**
T + P	131.7	125.7	0.792	0.144	P
T + H + P	124.5	116.5	0.822	0.207	P
T + H + T*H	114.0	106.0	0.861	0.279	Intercept & H
T + H + P + T*H	115.4	105.4	0.860	0.283	Intercept, T, H, P

Table 4.10. Math 106 Week 8 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	148.9	146.9	0.500	N/A	None
T	93.7	89.7	0.880	0.389	None
H	123.3	119.3	0.797	0.188	None
Q	104.4	100.4	0.879	0.317	None
P	146.5	142.5	0.665	0.030	Intercept
T + H	**76.2**	**70.2**	**0.930**	**0.522**	**None**
T + Q	83.5	77.5	0.913	0.473	None
T + H + Q	76.2	68.2	0.937	0.536	Q
T + H + T*H	78.2	70.2	0.930	0.522	All

Table 4.11. Math 108 Week 3 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	610.5	608.5	0.500	N/A	None
T	476.8	472.8	0.829	0.223	None
H	534.6	530.6	0.748	0.128	None
P	565.2	561.2	0.707	0.078	None
T + H	**453.9**	**447.9**	**0.855**	**0.264**	**None**
T + P	475.1	469.1	0.830	0.229	None
T + H + P	453.5	445.5	0.858	0.268	P
T + H + T*H	438.9	430.9	0.855	0.292	Intercept, T & H
T + H + P + T*H	437.5	427.5	0.856	0.298	Intercept, T & P

Table 4.12. Math 108 Week 8 Models

Predictors	**AIC**	**Deviance**	**Concordance**	**R^2**	**P val > 0.05**
None	610.5	608.5	0.500	N/A	None
T	381.4	377.4	0.899	0.380	None
H	492.2	488.2	0.815	0.198	None
Q	450.1	446.1	0.845	0.267	None
P	565.2	561.2	0.707	0.078	None
T + H	**348.2**	**342.2**	**0.926**	**0.438**	**None**
T + Q	366.6	360.6	0.912	0.407	None
T + P	383.4	377.4	0.900	0.380	P
T + H + Q	345.5	337.5	0.928	0.445	None
T + H + P	350.2	342.2	0.926	0.438	P
T + H + T*H	322.9	314.9	0.927	0.483	Intercept

REFERENCES

[1] Agresti, Alan, *An Introduction to Categorical Data Analysis*,Wiley, Inc., New Jerey, 2007.

[2] Brian G. Peterson and Peter Carl (2014). *PerformanceAnalytics: Econometric tools for performance and risk analysis.* R package version 1.4.3541. https://CRAN.R-project.org/package=PerformanceAnalytics

[3] Croux, C. and Dehon, C. (2010). *Influence Functions of the Spearman and Kendall Correlation Measures.* Statistical Methods and Applications, 19, 497-515.

[4] D. A. Belsley, E. Kuh and R. E. Welsch, *Regression Disgnostics: Identifying Influential Data and Sources of Collinearity*, Wiley,Inc., New York, 1980.

[5] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009.

[6] John Fox and Sanford Weisberg (2011). *An R Companion to Applied Regression*, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

[7] McFadden, D. (1974) *Conditional Logit Analysis of Qualitative Choice Behavior*. Pp. 105-142 in P.Zarembka (ed.), Frontiers in Econometrics. Academic Press.

[8] Patrick Breheny and Woodrow Burchett (2016). *Visreg: Visualization of Regression Models.* R package version 2.3-0. https://CRAN.R-project.org/package=visreg

[9] R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[10] Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

VITA

Graduate School
Southern Illinois University

Patrick B. Soule

Psoule212@gmail.com

Southern Illinois University at Carbondale

Bachelor of Science, Mathematics, May 2015

Research Paper Title:
Predicting Student Success: A Logistic Regression Analysis of Data from Multiple SIU-C Courses

Major Professor: Dr. B. Bhattacharya