

PREDICTING LONG TERM SUCCESS IN A PARTICULAR PROGRAM
BY IDENTIFYING GATEWAY COURSES

A THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE GRADUATE SCHOOL OF THE
TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
COLLEGE OF ARTS AND SCIENCES

BY
KATHRYN E. JOHANON B.S.

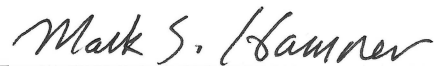
DENTON, TEXAS
AUGUST 2015

TEXAS WOMAN'S UNIVERSITY
DENTON, TEXAS

May 12, 2015

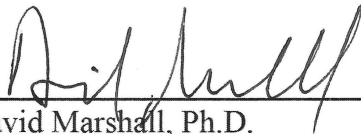
To the Dean of the Graduate School:

I am submitting herewith a thesis written by Kathryn E. Johanon entitled "Predicting Long Term Success in a Particular Program by Identifying Gateway Courses." I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Mathematics.



Mark Hamner, Ph.D., Major Professor

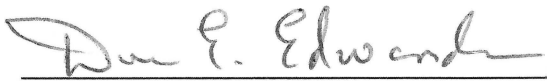
We have read this thesis and recommend its acceptance:



David Marshall, Ph.D.



Don Edwards, Ph.D.



Department Chair

Accepted:



Dean of the Graduate School

ABSTRACT

KATHRYN E. JOHANON

PREDICTING LONG TERM SUCCESS IN A PARTICULAR PROGRAM BY IDENTIFYING GATEWAY COURSES

AUGUST 2015

The purpose of this study is to see if students' grades in a math class they complete in the beginning of their college careers are indicative of future success in their major. Using pre-existing historical data from Texas Woman's University, we chose Math 1013 as a test gateway course and noted participants' grades and if they graduated in their original major or in another major within five years of taking the course. With this data, we developed nineteen possible models using Log-Linear Analysis. After exploring the development of our models, we then tested the models to find the model that most closely fits our data.

TABLE OF CONTENTS

	Page
ABSTRACT.....	iii
LIST OF TABLES.....	v
LIST OF FIGURES	vi
 Chapter	
I. INTRODUCTION	1
II. LITERATURE REVIEW	3
III. DATA EXPLORATION	6
Chi-Square Test of Association	11
IV. DEVELOPMENT OF LOG-LINEAR MODELS	14
Zero Cells.....	18
Saturated Model.....	18
Non-Comprehensive Models	21
Mutual Independence Model	28
Partial Independence Models.....	29
Conditional Independence Models	30
Homogenous Association Model.....	32
V. ASSESSMENT OF LOG-LINEAR MODELS	35
Likelihood-Ratio	35
Dissimilarity Index.....	39
Final Model with Coefficients	41
VI. CONCLUSION.....	43
REFERENCES	44

LIST OF TABLES

Table	Page
1.1 Rates for Graduation within Five Years	1
3.1 Frequencies for Data Sample	7
3.2 Rates for Graduation by Grade for Data Sample	10
3.3 GPA Frequency for Graduating Students	10
3.4 Expected Values for Chi-Square Test of Association.....	11
4.1 Equations to Find Coefficients for Saturated Model	19
4.2 Expected Values for Non-Comprehensive Models, Part I.....	22
4.3 Expected Values for Non-Comprehensive Models, Part II	25
4.4 Expected Values for Non-Comprehensive Models, Part III	27
4.5 Expected Values for Mutual and Partial Independence Models	28
4.6 Expected Values for Conditional Independence Model	31
4.7 Expected Values for Homogeneous Association Model	33
5.1 Statistics for Non-Comprehensive Models	37
5.2 Statistics for Comprehensive Models	38
5.3 Equations to Find Coefficients for Homogeneous Association Model	41

LIST OF FIGURES

Figure	Page
3.1 Proportions of Grade, Class, and Major.....	8
3.2 Interaction between Grade and Class.....	9
3.3 Interaction between Grade and Major.....	9
3.4 Interaction between Class and Major.....	9

CHAPTER I

INTRODUCTION

Success can be defined in a variety of ways for different people, but for a college student, graduation is usually an eventual objective. In 2014, the National Center for Educational Studies showed that the average five-year graduation rate for first-time-in-college freshmen students attending public universities are just over fifty percent, with slight variations depending on if the students are women, men, or both (seen in Table 1.1). The TWU Fact Book for 2015 shows that the five-year graduation rates for first-time-in-college freshmen students at Texas Woman's University, also found in Table 1.1, are much lower than the national average, especially considering that the population of TWU students between 2005 and 2014 was an average of 91.4% female.

Table 1.1

Rates for Graduation within Five Years

	2005-2010	2006-2011	2007-2012	2008-2013	2009-2014
National Female/Male	51.1	51.9	52.3	--	--
National Female	53.9	54.9	55.5	--	--
National Male	47.8	48.3	48.6	--	--
TWU Female/Male	35.7	41.1	37.5	38.4	35.6

Are there academic factors that we can look at early in students' college careers that indicate if they will graduate within 5 years? Can a student's grade in a course they take as a freshman or sophomore indicate future success of graduation, especially in their

original major? In this study, we will be looking at this connection between course grades, classification, major, and graduation.

CHAPTER II

LITERATURE REVIEW

College retention and graduation rates are important and studied often. In this chapter, we will review some of the literature regarding graduation and retention rates and how to best predict success in a student's college career.

Harackiewicz, Barron, Tauer, and Elliot (2002) followed a group of students' entire college career, starting when they took an introductory psychology course in their first semester through graduation. They studied the interaction between variables from high school (SAT/ACT scores, overall performance, motivation, and goals) and both short term and long term college (performance, motivation, goals, courses, major, and GPA) using multiple regression analysis. Although their study does not use graduation as a measure of success, it does discuss the link between classes taken in a student's first semester and future success in GPA as well as major.

A 2006 study by Dougherty, Mellor, and Jian calculated the connection between a student's participation in Advanced Placement in high school and college graduation. They found that 64% of the students who passed AP exams graduated college within five years.

Much of the research uses either students' high school grades or their SAT/ACT scores or a combination of both to predict graduation rates. While some suggests that high school grades are not as reliable to use for college admissions as SAT/ACT scores,

Geiser and Santelices found different results in their 2007 research. Their results showed that high school GPA is an excellent predictor of both college freshman grades and graduation.

Camara and Ehternacht (2000) also look at the many studies that have been done measuring the usefulness of SAT scores and high school grades in predicting college success. Some studies include different types of success criteria such as awards, honors, and teacher ratings; however the more common criteria are grades, GPA, persistence, and graduation. The results in these studies are consistent: college successes can be predicted jointly by SAT scores and high school GPA. While graduation is almost always the objective, it can be harder to predict since it can be affected by financial and familial concerns. Predictions for persistence and graduation are still consistent though despite these outside influences.

In a 2001 College Board Research Report, Burton and Ramist also find that the most accurate forecast of different kinds of college success comes from the grouping of high school grades and SAT scores. When predicting graduation alone however SAT scores give a slightly more accurate prediction. Again, they stress that graduation predictions can be less accurate because of outside influences.

An ACT Policy Report in 2004 also recognizes that there are many influences, both academic and non-academic, that prevents students from continuing their education to the point of graduation. The authors encourage universities to develop an early alert system based on these elements that would recognize students who are unlikely to persist. One of factors that could be monitored is students' grades at the beginning of their

college career. Universities could then offer assistance to at-risk students to encourage them to further their education and graduate (Lotkowski, Robbins, & Noeth).

In 2008, Texas Woman's University implemented an Early Warning System for Retention. This early alert system sends emails to students in the 5th week of the semester warning them that they are at risk of failing their course. It then encourages them to contact their instructor or advisor, seek assistance from a variety of tutoring options, and directs them to health, counseling, and disability services if needed. This early intervention can be a great help to students who are struggling in their courses.

It is this idea of an early alert system that our research expands upon to help at-risk students recognize that they may be in danger of not realizing their goal of graduation. We will look into the relationship between students' course grades and graduation to determine what the classes that students typically take as freshman and sophomore students indicate about their future success.

CHAPTER III

DATA EXPLORATION

In this Chapter, we will discuss our data in detail and do a preliminary test to confirm that there is a relationship between our variables. To begin, we had to choose a course to be our test Gateway Course. While it could be a beginning-level course of any topic that satisfies the core requirements for that subject, we chose an introductory mathematics course, Math 1013. During the five fall semesters between 2005 and 2009, a total of 2,063 Texas Woman's University students took Math 1013. Since we wanted to limit our participants to students who were more towards the beginning of their college careers, we used only students who were of the freshman or sophomore class at the time they took the course. We further reduced our sample by only using students who completed the course, earning a grade between A and F, and excluding students who withdrew from the course, receiving a grade of W or WF. After making these exclusions, we were left with a total sample size of 1,406. While we also noted that of our 1,406 students, we had 1,332 women and only 74 men, we did not make an exclusion based on gender. In order to maintain a higher level of student privacy, we did not assign any kind of identification number to the students, so we did not exclude students who might have repeated the course between those years.

Of the data available, we are mostly interested in three variables we call Grade, Class, and Major. The first, Grade (G), is the grade the student received in the course.

There are five outcomes ($g = 5$): A, B, C, D, or F. The second, Class (C), is the classification of the student at the time they took the course. We will be limiting Class to two options ($c = 2$): Freshman (FR) or Sophomore (SO). The last variable, Major (M), is really a combination of three things: the declared major at the time of taking the course, the graduation status of the student five years after taking the course, and the major at the time of graduation. The three possible outcomes ($m = 3$) for Major are Original, Other, and None. “Original” means the student graduated within five years of taking the course in the same major declared at the time the course was taken. “Other” means the student did graduate within five years but in a major other than the one originally declared at the time of taking the course. “None” means that the student did not graduate in any major within five years. In Table 3.1, we have the Frequency Table for our sample of data.

Table 3.1

Frequencies for Data Sample

Grade	Class	Major			Total
		Original	Other	None	
A	FR	123	40	132	295
A	SO	125	27	51	203
B	FR	94	29	181	304
B	SO	61	18	61	140
C	FR	37	19	135	191
C	SO	31	10	44	85
D	FR	10	1	43	54
D	SO	6	3	22	31
F	FR	2	0	55	57
F	SO	3	4	39	46
Total		492	151	763	1406

(Continued)

Grade	Class	Major			Total
		Original	Other	None	
A		248	67	183	498
B		155	47	242	444
C		68	29	179	276
D		16	4	65	85
F		5	4	94	103
Total		492	151	763	1406
	FR	266	89	546	901
	SO	226	62	217	505
Total		492	151	763	1406

Figure 3.1 illustrates the proportions of students in the different variables of Grade, Class, and Major, and Figures 3.2, 3.3, and 3.4 illustrate the interactions between the variables.

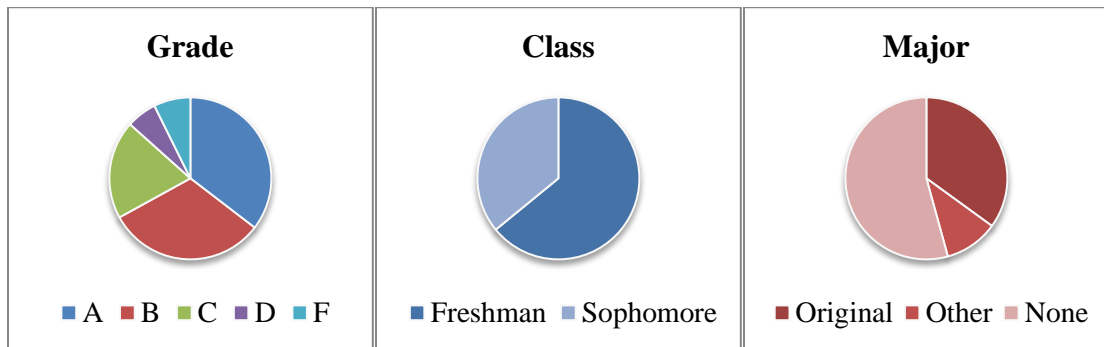


Figure 3.1: Proportions of grade, class, and major

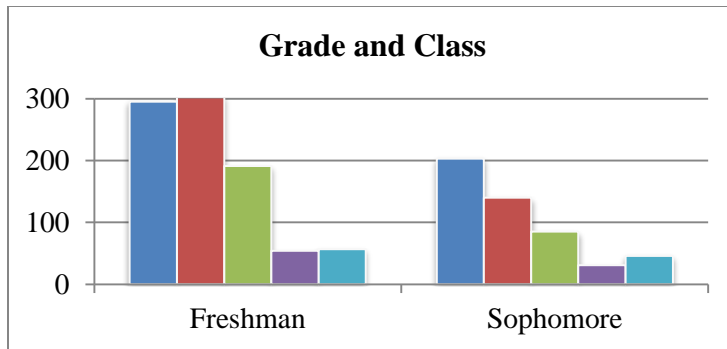


Figure 3.2: Interaction between grade and class

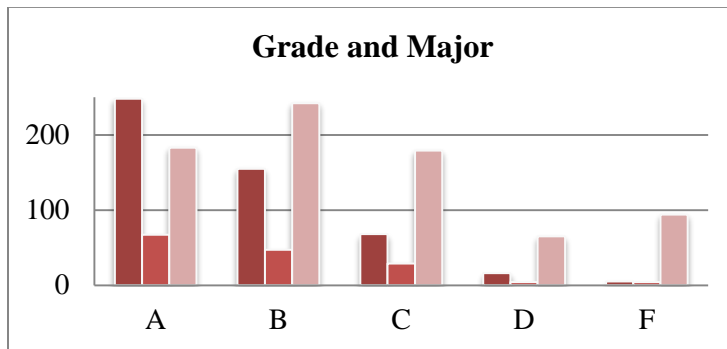


Figure 3.3: Interaction between grade and major

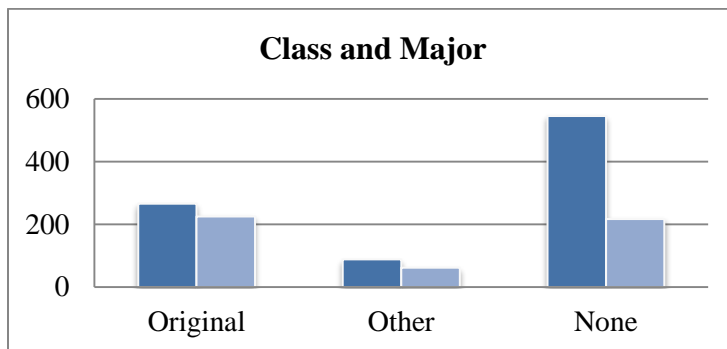


Figure 3.4: Interaction between class and major

The Graduation Rates for our sample is shown in Table 3.2. Compared to the rates we discussed in Chapter I, the average for our sample is higher than the TWU average, but it is still lower than the national average. Students who make an A in the

course however, have a much higher graduation rate. While on the other end of the spectrum, less than ten percent of students who made an F graduated within five years.

Table 3.2

Rates for Graduation by Grade for Data Sample

Grade	Graduating/Total	Graduation Rate
A	315/498	63.25
B	202/444	45.50
C	97/276	35.14
D	20/85	23.53
F	9/103	8.74
All	643/1406	45.73

Another piece of data that we are interested in is the grade point average (GPA) of the 643 students who graduated in either an original or other major. Because students cannot graduate with a GPA below 2.0, our range of GPA values, seen in Table 3.3, is between 2.0 and 4.0.

Table 3.3

GPA Frequency for Graduating Students

Grade	GPA				Total
	2.00-2.49	2.50-2.99	3.00-3.49	3.50-4.00	
A	6	23	131	155	315
B	4	42	112	44	202
C	9	38	48	2	97
D	0	9	11	0	20
F	3	3	3	0	9
Total	22	115	305	201	643

Chi-Square Test of Association

We perform a Chi-Square Test of Association to confirm that there is a relationship between a student's Grade, Class, and Major. We begin by defining the null and alternate hypotheses:

H_0 = Grade, Class, and Major are Independent

H_A = They are Dependent

We then calculate Expected Values.

$$exp_{\chi^2} = obs_{GC} * obs_M / n$$

To find the Expected Value for Freshmen who made an A in the course and Graduate in the Original Major, take the Total for Freshmen who made an A and multiply it by the Total for students who Graduate in Original Major, then divide that by the Total for all students. The rest of the Expected Values are found in the same way and can be seen in Table 3.4.

Table 3.4

Expected Values for Chi-Square Test of Association

Grade	Class	Major	Observed	Expected	Chi-Square
A	FR	Original	123	103.229	3.787
A	FR	Other	40	31.682	2.184
A	FR	None	132	160.089	4.928
A	SO	Original	125	71.036	40.996
A	SO	Other	27	21.802	1.240
A	SO	None	51	110.163	31.773

(Continued)

Grade	Class	Major	Observed	Expected	Chi-Square
B	FR	Original	94	106.378	1.440
B	FR	Other	29	32.649	0.408
B	FR	None	181	164.973	1.557
B	SO	Original	61	48.990	2.944
B	SO	Other	18	15.036	0.584
B	SO	None	61	75.974	2.951
C	FR	Original	37	66.836	13.319
C	FR	Other	19	20.513	0.112
C	FR	None	135	103.651	9.482
C	SO	Original	31	29.744	0.053
C	SO	Other	10	9.129	0.083
C	SO	None	44	46.127	0.098
D	FR	Original	10	18.896	4.188
D	FR	Other	1	5.799	3.972
D	FR	None	43	29.304	6.401
D	SO	Original	6	10.848	2.166
D	SO	Other	3	3.329	0.033
D	SO	None	22	16.823	1.593
F	FR	Original	2	19.946	16.146
F	FR	Other	0	6.122	6.122
F	FR	None	55	30.932	18.726
F	SO	Original	3	16.097	10.656
F	SO	Other	4	4.940	0.179
F	SO	None	39	24.963	7.893
Total			1406	1406	196.015

Now to calculate the Chi-Square statistic, the following equation is used:

$$\chi^2 = \sum (obs - exp)^2 / exp$$

We show the calculation for Freshmen who made an A Grade in the course and Graduate in the Original Major.

$$(obs - exp)^2/exp = (123 - 103.229)^2/103.229 = 3.787$$

This is done for the other 23 groups, as also seen in Table 3.4, and these are then summed to find the Chi-Square Statistic:

$$\chi^2 = 196.015$$

The degrees of freedom are found:

$$df = (c - 1)(r - 1) = (3 - 1)(10 - 1) = 18$$

With an alpha of 0.05, we find the critical value to be 28.86 and the p-value to be 6.27×10^{-32} . Based on these values, we reject the null hypothesis that Grade, Class, and Major are Independent. We know then that there exists a relationship between these three variables, but what exactly is that relationship? We do further testing to determine this.

CHAPTER IV

DEVELOPMENT OF LOG-LINEAR MODELS

There are several ways to evaluate the relationship between these variables. We choose to use Log-Linear Analysis. Log-linear analysis treats all the categorical variables as response variables, but is suitable even for clear explanatory/response cases when there are more than two responses such as in our case with three outcomes for Major. This type of analysis is beneficial because it creates models that look at every possible relationship between the variables by examining expected cell frequencies.

To show how all the models will be built, we will demonstrate with the Saturated Model {GCM}. This model includes every variable and interaction possible and is a perfect fit for our data since the observed cell frequencies equal the expected cell frequencies. Here we have the multiplicative form of the general equation for the Saturated Model.

$$F_{ijk} = \eta \tau_i^G \tau_j^C \tau_k^M \tau_{ij}^{GC} \tau_{ik}^{GM} \tau_{jk}^{CM} \tau_{ijk}^{GCM}$$

We start by defining f_{ijk} as the observed frequency of cell ijk and F_{ijk} as the expected frequency for that cell. The first term in the equation, η , is similar to an intercept. Unless it is the only term in the equation, it does not have much meaning except as the foundation from which each additional term will deviate. The rest of the terms, τ , are the effects that each variable and interaction have on the expected value. The number of

effect terms included depends on the model. For the multiplicative form of the equation, effects that are equal to one have no impact on the expected cell frequencies, whereas values greater than one will increase the frequencies and values less than one will decrease them.

For the individual variable terms, each variable X with x number of outcomes adds $x - 1$ terms to the Saturated Model. Since Grade has five outcomes, there will be four terms: τ_i^G , where

$$i = \begin{cases} 1 & \text{if a student makes a Grade of A} \\ 2 & \text{if a student makes a Grade of B} \\ 3 & \text{if a student makes a Grade of C} \\ 4 & \text{if a student makes a Grade of D} \\ - & \text{if a student makes a Grade of F} \end{cases}$$

No term is included if a student makes a Grade of F. Class, with two outcomes, adds one term: τ_j^C , where

$$j = \begin{cases} 1 & \text{if a student is in the Freshman Class} \\ - & \text{if a student is in the Sophomore Class} \end{cases}$$

Again no term is included if a student is in the Sophomore Class. And the last of the individual variables with three outcomes and two terms, we have Major: τ_k^M , where

$$k = \begin{cases} 1 & \text{if a student graduates in their Original Major} \\ 2 & \text{if a student graduates in a Different Major} \\ - & \text{if a student does not graduate in any Major (None)} \end{cases}$$

Like the other two categories, there is no term if a student does not graduate in any Major. These individual terms represent the unequal distribution of the variables across the other groups.

When two or three variables are associated, we add interaction terms to the model showing the degree to which they are connected. The two-way interaction between Grade and Class, τ_{ij}^{GC} , will have $(g-1)(c-1)$, or four terms. Eight terms, $(g-1)(m-1)$, are added by the interactions between Grade and Major: τ_{ik}^{GM} . And the interaction between Class and Major, τ_{jk}^{CM} , adds $(c-1)(m-1)$ or two terms. The three-way interaction, τ_{ijk}^{GCM} , adds $(g-1)(c-1)(m-1)$ or eight terms to the model.

We now take the natural log of both sides of the equation,

$$\ln(F_{ijk}) = \ln(\eta \tau_i^G \tau_j^C \tau_k^M \tau_{ij}^{GC} \tau_{ik}^{GM} \tau_{jk}^{CM} \tau_{ijk}^{GCM})$$

and rewrite the equation in its additive form.

$$\begin{aligned} \ln(F_{ijk}) &= \ln(\eta) + \ln(\tau_i^G) + \ln(\tau_j^C) + \ln(\tau_k^M) + \ln(\tau_{ij}^{GC}) + \ln(\tau_{ik}^{GM}) + \ln(\tau_{jk}^{CM}) \\ &\quad + \ln(\tau_{ijk}^{GCM}) \end{aligned}$$

If we let $\ln(F_{ijk}) = N_{ijk}$, $\ln(\eta) = \theta$, and $\ln(\tau) = \lambda$, we have the following general equation for the Saturated Model.

$$N_{ijk} = \theta + \lambda_i^G + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{GC} + \lambda_{ik}^{GM} + \lambda_{jk}^{CM} + \lambda_{ijk}^{GCM}$$

It is in this additive form that the effects of each individual variable and interaction can be easily seen. Here, the effects and interactions equal to zero have no effect on the expected cell frequencies, while positive values increase the frequencies and negative values decrease them. Degrees of freedom for all models are equal to the number of parameters equal to zero.

Using different combinations of these effects and interactions, we will build a total of nineteen different models divided into several categories. All of our models will have a hierarchical structure. As we have seen, our Saturated Model includes not only the three-way interaction but also all possible two-variable interactions and single-variable effects. The rest of the models are unsaturated but still hierarchical. Models that have two-variable interactions will include the single-variable effects that make up those interactions.

Using at most two variables, we will build ten Non-Comprehensive Models, some that only include individual effects and others that consist of one interaction. We will also build eight other models that include all three of our variables: one Mutual Independence Model, three Partial Independence Models, three Conditional Independence Models, and one Homogenous Association Model. With all of our models however, even though they are hierarchical and include all of the lower-order terms, when interpreting the final models with their coefficients, we will only look at the highest-order interaction terms and will ignore the lower-order terms.

Zero Cells

Before we start our analysis, it is important to address the issue of zero cells. Since Log-Linear Analysis involves taking the natural log of each observed value, we cannot have any values equal to zero, causing the calculation to become undefined. Of the 1,406 students in our sample, we have no freshmen with a grade of F graduating in an “Other” major. This is only a Sampling Zero since it is possible to have students in this category if the sample size is large enough.

Without having to increase the sample size, we have several options. First, we could define zero divided by zero to be zero, but in this situation all expected values for this category would also be zero. This outcome would not be very likely in larger populations. The other two options are similar to each other: add a small amount to either each observed value or to the zero cell only. Smith and Cicchetti (1974) show in their research that adding only to the zero cells has the least impact on the final outcome. We follow this path and add the commonly recommended amount of 0.5 to our zero cell (Knoke & Burke, 1980). All of our models will show a total population of 1,406.5. We are now ready to start discussing our different models.

Saturated Model

We begin with the Saturated Model whose general equation we have already found.

$$N_{\{GCM\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GC} + \lambda^{GM} + \lambda^{CM} + \lambda^{GCM}$$

Again, this model is a perfect fit for our data, with the expected values being equal to the observed values.

$$exp_{\{E-P\}} = obs_{GCM}$$

To build the full equation for the Saturated Model, we look to the equations formed by our thirty categories, shown in Table 4.1, and solve for the coefficients.

Table 4.1

Equations to Find Coefficients for Saturated Model

$\ln 132 = b_0 + b_1^G + b_1^C + b_{11}^{GC}$	$\ln 31 = b_0 + b_3^G + b_1^M + b_{31}^{GM}$
$\ln 125 = b_0 + b_1^G + b_1^M + b_{11}^{GM}$	$\ln 10 = b_0 + b_3^G + b_2^M + b_{32}^{GM}$
$\ln 27 = b_0 + b_1^G + b_2^M + b_{12}^{GM}$	$\ln 43 = b_0 + b_4^G + b_1^C + b_{41}^{GC}$
$\ln 181 = b_0 + b_2^G + b_1^C + b_{21}^{GC}$	$\ln 6 = b_0 + b_4^G + b_1^M + b_{41}^{GM}$
$\ln 61 = b_0 + b_2^G + b_1^M + b_{21}^{GM}$	$\ln 3 = b_0 + b_4^G + b_2^M + b_{42}^{GM}$
$\ln 18 = b_0 + b_2^G + b_2^M + b_{22}^{GM}$	$\ln 2 = b_0 + b_1^C + b_1^M + b_{11}^{CM}$
$\ln 135 = b_0 + b_3^G + b_1^C + b_{31}^{GC}$	$\ln 0.5 = b_0 + b_1^C + b_2^M + b_{12}^{CM}$
$\ln 39 = b_0$	$\ln 123 = b_0 + b_1^G + b_1^C + b_1^M + b_{11}^{GC} + b_{11}^{GM} + b_{11}^{CM} + b_{111}^{GCM}$
$\ln 51 = b_0 + b_1^G$	$\ln 40 = b_0 + b_1^G + b_1^C + b_2^M + b_{11}^{GC} + b_{12}^{GM} + b_{12}^{CM} + b_{112}^{GCM}$
$\ln 61 = b_0 + b_2^G$	$\ln 94 = b_0 + b_2^G + b_1^C + b_1^M + b_{21}^{GC} + b_{21}^{GM} + b_{11}^{CM} + b_{211}^{GCM}$
$\ln 44 = b_0 + b_3^G$	$\ln 29 = b_0 + b_2^G + b_1^C + b_2^M + b_{21}^{GC} + b_{22}^{GM} + b_{12}^{CM} + b_{212}^{GCM}$
$\ln 22 = b_0 + b_4^G$	$\ln 37 = b_0 + b_3^G + b_1^C + b_1^M + b_{31}^{GC} + b_{31}^{GM} + b_{11}^{CM} + b_{311}^{GCM}$
$\ln 55 = b_0 + b_1^C$	$\ln 19 = b_0 + b_3^G + b_1^C + b_2^M + b_{31}^{GC} + b_{32}^{GM} + b_{12}^{CM} + b_{312}^{GCM}$
$\ln 3 = b_0 + b_1^M$	$\ln 10 = b_0 + b_4^G + b_1^C + b_1^M + b_{41}^{GC} + b_{41}^{GM} + b_{11}^{CM} + b_{411}^{GCM}$
$\ln 4 = b_0 + b_2^M$	$\ln 1 = b_0 + b_4^G + b_1^C + b_2^M + b_{41}^{GC} + b_{42}^{GM} + b_{12}^{CM} + b_{412}^{GCM}$

Here is the full equation for our Saturated Model:

$$\begin{aligned}
\ln(F_{ijk}) = & 3.663562 + 0.268264 t_1^G + 0.447312 t_2^G + 0.120628 t_3^G - 0.572519 t_4^G + \\
& 0.343772 t_1^C - 2.564949 t_1^M - 2.277267 t_2^M + 0.607205 t_{11}^{GC} + \\
& 3.461437 t_{11}^{GM} + 1.641279 t_{12}^{GM} + 0.743852 t_{21}^{GC} + 2.564949 t_{21}^{GM} + \\
& 1.056765 t_{22}^{GM} + 0.777314 t_{31}^{GC} + 2.214747 t_{31}^{GM} + 0.795663 t_{32}^{GM} + \\
& 0.326386 t_{41}^{GC} + 1.265666 t_{41}^{GM} + 0.284837 t_{42}^{GM} - 0.749237 t_{11}^{CM} - \\
& 2.423213 t_{12}^{CM} - 0.217869 t_{111}^{GCM} + 1.865279 t_{112}^{GCM} + 0.094034 t_{211}^{GCM} + \\
& 1.812514 t_{212}^{GCM} - 0.194918 t_{311}^{GCM} + 1.943982 t_{312}^{GCM} + 0.589905 t_{411}^{GCM} + \\
& 0.654443 t_{412}^{GCM}
\end{aligned}$$

While this model is a perfect fit, it is complex and can be difficult to interpret.

Our goal would be to find a simpler model that would still produce equivalent expected values and provide an accurate description of our data while being easier to interpret. To do this we will begin with the simplest models and work our way up. First we will build each model, showing its general equation and expected values. Then, we will calculate the likelihood-ratio statistics, and using our established alpha and each model's degrees of freedom, we will see the impact of each variable and association. We will also see which model fits our data the best and if any of them can be used in place of the Saturated Model. Once we find a model that is a good fit for our data, we will calculate the coefficients and construct the final equation.

Non-Comprehensive Models

There are ten non-comprehensive models. These simple models only include one or two variables. Although we are interested in the association between all three of our variables, it will be worthwhile to study these models and their statistics. By going through every possible model, we can gain valuable information about the importance of the individual variable and of the small associations between two variables. By recognizing the associations that affect our numbers, we will know the vital components and be able to follow the step-by-step evolution of our final model. With that in mind, we will begin discussing the Non-Comprehensive Models.

The first and simplest model is the Equal-Probably Model {E-P}. With this model, each outcome has an equal probability of occurring. The Equation is simply a constant, not including any of our variables:

$$N_{\{E-P\}} = \theta$$

The Expected Values (Table 4.2) are all the same: the total number of students divided by our 30 different outcomes.

$$exp_{\{E-P\}} = n/(g * c * m)$$

The next three models, represented by {G}, {C}, and {M}, only look at one variable. For example, the {G} Model only shows variation in expected values based on Grade with no regard to Class or Major. The {C} and {M} Models are similar, only

focusing on Class or Major and ignoring the other two variables. Their Equations only include the one variable:

$$N_{\{G\}} = \theta + \lambda^G$$

$$N_{\{C\}} = \theta + \lambda^C$$

$$N_{\{M\}} = \theta + \lambda^M$$

The Expected Values (Table 4.2) are found by dividing the Observed Values of the one variable by the product of the number of outcomes of the ignored variables.

$$exp_{\{G\}} = obs_G / (c * m)$$

$$exp_{\{C\}} = obs_C / (g * m)$$

$$exp_{\{M\}} = obs_M / (g * c)$$

For example with the {G} Model, the Expected Value for students making a Grade of B would be the same for all of the six possible combinations of Class and Major.

Table 4.2

Expected Values for Non-Comprehensive Models, Part I

Grade	Class	Major	Obs	{E-P}	{G}	{C}	{M}
A	FR	Original	123	46.88	83.00	60.10	49.20
A	FR	Other	40	46.88	83.00	60.10	15.15
A	FR	None	132	46.88	83.00	60.10	76.30
A	SO	Original	125	46.88	83.00	33.67	49.20
A	SO	Other	27	46.88	83.00	33.67	15.15
A	SO	None	51	46.88	83.00	33.67	76.30

(Continued)

Grade	Class	Major	Obs	{E-P}	{G}	{C}	{M}
B	FR	Original	94	46.88	74.00	60.10	49.20
B	FR	Other	29	46.88	74.00	60.10	15.15
B	FR	None	181	46.88	74.00	60.10	76.30
B	SO	Original	61	46.88	74.00	33.67	49.20
B	SO	Other	18	46.88	74.00	33.67	15.15
B	SO	None	61	46.88	74.00	33.67	76.30
C	FR	Original	37	46.88	46.00	60.10	49.20
C	FR	Other	19	46.88	46.00	60.10	15.15
C	FR	None	135	46.88	46.00	60.10	76.30
C	SO	Original	31	46.88	46.00	33.67	49.20
C	SO	Other	10	46.88	46.00	33.67	15.15
C	SO	None	44	46.88	46.00	33.67	76.30
D	FR	Original	10	46.88	14.17	60.10	49.20
D	FR	Other	1	46.88	14.17	60.10	15.15
D	FR	None	43	46.88	14.17	60.10	76.30
D	SO	Original	6	46.88	14.17	33.67	49.20
D	SO	Other	3	46.88	14.17	33.67	15.15
D	SO	None	22	46.88	14.17	33.67	76.30
F	FR	Original	2	46.88	17.25	60.10	49.20
F	FR	Other	0.5	46.88	17.25	60.10	15.15
F	FR	None	55	46.88	17.25	60.10	76.30
F	SO	Original	3	46.88	17.25	33.67	49.20
F	SO	Other	4	46.88	17.25	33.67	15.15
F	SO	None	39	46.88	17.25	33.67	76.30

The next set of models we will look at each have two variables, but they are independent of each other. They are represented by the following: {G, C}, {G, M}, and {C, M}. The {G, M} Model states that the data is governed by the variables of Grade and Major, but they are not associated and Class has no bearing on any outcome. The

Equations for these models include terms for both of the variables considered, with the {G, M} Model including the G term and the M term.

$$N_{\{G,C\}} = \theta + \lambda^G + \lambda^C$$

$$N_{\{G,M\}} = \theta + \lambda^G + \lambda^M$$

$$N_{\{C,M\}} = \theta + \lambda^C + \lambda^M$$

The Expected Values (Table 4.3) are found by dividing the product of the Observed Values of the first variable and the Observed Values of the second variable by the product of the total number of students and the number of outcomes for the variable ignored.

$$exp_{\{G,C\}} = obs_G * obs_C / (n * m)$$

$$exp_{\{G,M\}} = obs_G * obs_M / (n * c)$$

$$exp_{\{C,M\}} = obs_C * obs_M / (n * g)$$

For example, the Expected Values from the {G, M} Model for students making a Grade of B and graduating with a Major of Other, would be the same whether they are Freshmen or Sophomores.

Table 4.3

Expected Values for Non-Comprehensive Models, Part II

Grade	Class	Major	Obs	{G, C}	{G, M}	{C, M}
A	FR	Original	123	106.40	87.10	63.07
A	FR	Other	40	106.40	26.82	19.42
A	FR	None	132	106.40	135.08	97.81
A	SO	Original	125	59.60	87.10	35.33
A	SO	Other	27	59.60	26.82	10.88
A	SO	None	51	59.60	135.08	54.79
B	FR	Original	94	94.86	77.66	63.07
B	FR	Other	29	94.86	23.91	19.42
B	FR	None	181	94.86	120.43	97.81
B	SO	Original	61	53.14	77.66	35.33
B	SO	Other	18	53.14	23.91	10.88
B	SO	None	61	53.14	120.43	54.79
C	FR	Original	37	58.97	48.27	63.07
C	FR	Other	19	58.97	14.86	19.42
C	FR	None	135	58.97	74.86	97.81
C	SO	Original	31	33.03	48.27	35.33
C	SO	Other	10	33.03	14.86	10.88
C	SO	None	44	33.03	74.86	54.79
D	FR	Original	10	18.16	14.87	63.07
D	FR	Other	1	18.16	4.58	19.42
D	FR	None	43	18.16	23.06	97.81
D	SO	Original	6	10.17	14.87	35.33
D	SO	Other	3	10.17	4.58	10.88
D	SO	None	22	10.17	23.06	54.79
F	FR	Original	2	22.11	18.10	63.07
F	FR	Other	0.5	22.11	5.57	19.42
F	FR	None	55	22.11	28.07	97.81
F	SO	Original	3	12.39	18.10	35.33
F	SO	Other	4	12.39	5.57	10.88
F	SO	None	39	12.39	28.07	54.79

The rest of the Non-Comprehensive Models, represented by {GC}, {GM}, and {CM}, are similar to the ones we just discussed with two variables, but now the models are stating that the variables are dependent on each other. For the Model {CM}, this means that Class and Major are dependent on each other, while Grade has no impact on the data. The Equation, like the other models, will include separate terms for each variable, but this time also includes the interaction term of the two variables.

$$N_{\{GC\}} = \theta + \lambda^G + \lambda^C + \lambda^{GC}$$

$$N_{\{GM\}} = \theta + \lambda^G + \lambda^M + \lambda^{GM}$$

$$N_{\{CM\}} = \theta + \lambda^C + \lambda^M + \lambda^{CM}$$

The Expected Values (Table 4.4) are found by dividing the Observed Values for the intersection of the two variables by the number of outcomes in the variable not considered.

$$exp_{\{GC\}} = obs_{GC}/m$$

$$exp_{\{GM\}} = obs_{GM}/c$$

$$exp_{\{CM\}} = obs_{CM}/g$$

For the {CM} Model, the Expected Values for the Class/Major category of Sophomore students graduating in their Original Major would be the same for all five Grades.

Table 4.4

Expected Values for Non-Comprehensive Models, Part III

Grade	Class	Major	Obs	{GC}	{GM}	{CM}
A	FR	Original	123	98.33	124.00	53.20
A	FR	Other	40	98.33	33.50	17.90
A	FR	None	132	98.33	91.50	109.20
A	SO	Original	125	67.67	124.00	45.20
A	SO	Other	27	67.67	33.50	12.40
A	SO	None	51	67.67	91.50	43.40
B	FR	Original	94	101.33	77.50	53.20
B	FR	Other	29	101.33	23.50	17.90
B	FR	None	181	101.33	121.00	109.20
B	SO	Original	61	46.67	77.50	45.20
B	SO	Other	18	46.67	23.50	12.40
B	SO	None	61	46.67	121.00	43.40
C	FR	Original	37	63.67	34.00	53.20
C	FR	Other	19	63.67	14.50	17.90
C	FR	None	135	63.67	89.50	109.20
C	SO	Original	31	28.33	34.00	45.20
C	SO	Other	10	28.33	14.50	12.40
C	SO	None	44	28.33	89.50	43.40
D	FR	Original	10	18.00	8.00	53.20
D	FR	Other	1	18.00	2.00	17.90
D	FR	None	43	18.00	32.50	109.20
D	SO	Original	6	10.33	8.00	45.20
D	SO	Other	3	10.33	2.00	12.40
D	SO	None	22	10.33	32.50	43.40
F	FR	Original	2	19.17	2.50	53.20
F	FR	Other	0.5	19.17	2.25	17.90
F	FR	None	55	19.17	47.00	109.20
F	SO	Original	3	15.33	2.50	45.20
F	SO	Other	4	15.33	2.25	12.40
F	SO	None	39	15.33	47.00	43.40

Having built all ten of the Non-Comprehensive Models, we now move on to discuss the models that include all three variables.

Mutual Independence Model

The Mutual Independence Model is the first model we have looked at that includes all three of our variables, although there are no association terms. This model, $\{G, C, M\}$, states that the variables of Grade, Class, and Major all govern the data but are mutually independent of each other. The Equation includes all three of the individual terms but no interaction terms.

$$N_{\{G,C,M\}} = \theta + \lambda^G + \lambda^C + \lambda^M$$

The Expected Values, shown in Table 4.5, are the product of the three separate Observed Values for Grade, Class, and Major, divided by the squared value of all the students.

$$exp_{\{G,C,M\}} = obs_G * obs_C * obs_M / n^2$$

Table 4.5

Expected Values for Mutual and Partial Independence Models

Grade	Class	Major	Obs	{G, C, M}	{GC, M}	{GM, C}	{CM, G}
A	FR	Original	123	111.66	103.19	158.96	94.18
A	FR	Other	40	34.38	31.78	42.94	31.69
A	FR	None	132	173.16	160.03	117.29	193.32
A	SO	Original	125	62.55	71.01	89.04	80.02
A	SO	Other	27	19.26	21.87	24.06	21.95
A	SO	None	51	97.00	110.12	65.71	76.83
B	FR	Original	94	99.55	106.34	99.35	83.97
B	FR	Other	29	30.65	32.75	30.12	28.25
B	FR	None	181	154.38	164.91	155.11	172.36

(Continued)

Grade	Class	Major	Obs	{G, C, M}	{GC, M}	{GM, C}	{CM, G}
B	SO	Original	61	55.76	48.97	55.65	71.34
B	SO	Other	18	17.17	15.08	16.88	19.57
B	SO	None	61	86.48	75.95	86.89	68.50
C	FR	Original	37	61.88	66.81	43.58	52.20
C	FR	Other	19	19.05	20.57	18.59	17.56
C	FR	None	135	95.97	103.61	114.73	107.14
C	SO	Original	31	34.66	29.73	24.42	44.35
C	SO	Other	10	10.67	9.16	10.41	12.17
C	SO	None	44	53.76	46.11	64.27	42.58
D	FR	Original	10	19.06	18.89	10.26	16.08
D	FR	Other	1	5.87	5.82	2.56	5.41
D	FR	None	43	29.55	29.29	41.66	33.00
D	SO	Original	6	10.68	10.84	5.74	13.66
D	SO	Other	3	3.29	3.34	1.44	3.75
D	SO	None	22	16.56	16.82	23.34	13.11
F	FR	Original	2	23.21	20.11	3.20	19.57
F	FR	Other	0.5	7.15	6.19	2.88	6.59
F	FR	None	55	35.99	31.19	60.25	40.18
F	SO	Original	3	13.00	16.09	1.80	16.63
F	SO	Other	4	4.00	4.95	1.62	4.56
F	SO	None	39	20.16	24.95	33.75	15.97

Partial Independence Models

Now we will move into discussing models that not only include all three of our variables but also include interaction terms between two of the variables. The first models fitting this description are the Partial Independence Models. These models are represented by the following: {GC, M}, {GM, C}, and {CM, G}. With these models, two of the variables are associated with each other while being jointly independent of the third variable.

Each Equation has all three of the individual terms G, C, and M, and also the interaction term that is unique to the model.

$$N_{\{GC,M\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GC}$$

$$N_{\{GM,C\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GM}$$

$$N_{\{CM,G\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{CM}$$

The Expected Values, shown in Table 4.5, are found by multiplying the Observed Values of the independent variable by the Observed Values of the Interaction term, and then divided that product by the total number of students.

$$exp_{\{GC,M\}} = obs_{GC} * obs_M / n$$

$$exp_{\{GM,C\}} = obs_{GM} * obs_C / n$$

$$exp_{\{CM,G\}} = obs_{CM} * obs_G / n$$

The Equation used to find the Expected Values of the model {GC, M} may look familiar because we have already used it once before in our study. It is identical to the equation used to find the Expected Values for our Chi-square calculations. The Expected Values for the {GC, M} Model differ slightly however, since we had to marginally increase our sample to take care of the existing zero cell.

Conditional Independence Models

Now we introduce the Conditional Independence Models. These models consist of two interaction terms. The two variables that appear only once in the different terms are said to be independent from each other while being conditional on the third variable

that appears in both terms. For example, with the Model {GC, GM}, we would say that Class and Major are conditionally independent given Grade. The Equations include all three individual variable terms and the two interaction terms.

$$N_{\{GC,GM\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GC} + \lambda^{GM}$$

$$N_{\{GC,CM\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GC} + \lambda^{CM}$$

$$N_{\{GM,CM\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GM} + \lambda^{CM}$$

The Expected Values, found in Table 4.6, are the product of the Observed Values of the two interaction terms divided by the Observed Values of the individual variable present in both of the interactions.

$$exp_{\{GC,GM\}} = obs_{GC} * obs_{GM} / obs_G$$

$$exp_{\{GC,CM\}} = obs_{GC} * obs_{CM} / obs_C$$

$$exp_{\{GM,CM\}} = obs_{GM} * obs_{CM} / obs_M$$

Table 4.6

Expected Values for Conditional Independence Models

Grade	Class	Major	Obs	{GC, GM}	{GC, CM}	{GM, CM}
A	FR	Original	123	146.91	87.04	134.08
A	FR	Other	40	39.69	29.29	39.58
A	FR	None	132	108.40	178.67	130.95
A	SO	Original	125	101.09	90.85	113.92
A	SO	Other	27	27.31	24.92	27.42
A	SO	None	51	74.60	87.23	52.05

(Continued)

Grade	Class	Major	Obs	{GC, GM}	{GC, CM}	{GM, CM}
B	FR	Original	94	106.13	89.70	83.80
B	FR	Other	29	32.18	30.18	27.77
B	FR	None	181	165.69	184.12	173.17
B	SO	Original	61	48.87	62.65	71.20
B	SO	Other	18	14.82	17.19	19.23
B	SO	None	61	76.31	60.16	68.83
C	FR	Original	37	47.06	56.36	36.76
C	FR	Other	19	20.07	18.96	17.13
C	FR	None	135	123.87	115.68	128.09
C	SO	Original	31	20.94	38.04	31.24
C	SO	Other	10	8.93	10.44	11.87
C	SO	None	44	55.13	36.52	50.91
D	FR	Original	10	10.16	15.93	8.65
D	FR	Other	1	2.54	5.36	2.36
D	FR	None	43	41.29	32.71	46.51
D	SO	Original	6	5.84	13.87	7.35
D	SO	Other	3	1.46	3.81	1.64
D	SO	None	22	23.71	13.32	18.49
F	FR	Original	2	2.78	16.97	2.70
F	FR	Other	0.5	2.50	5.71	2.66
F	FR	None	55	52.22	34.83	67.27
F	SO	Original	3	2.22	20.59	2.30
F	SO	Other	4	2.00	5.65	1.84
F	SO	None	39	41.78	19.77	26.73

Homogeneous Association Model

We now introduce the last model, the homogeneous association model

{GC, GM, CM} that includes all three interactions GC, GM, and CM. This is the most complicated model, with only the full interaction term of GCM missing from the

Saturated Model. The Equation for this model has the three variable terms plus the three interaction terms:

$$N_{\{GC,GM,CM\}} = \theta + \lambda^G + \lambda^C + \lambda^M + \lambda^{GC} + \lambda^{GM} + \lambda^{CM}$$

The Expected Values for this model, seen in Table 4.7, are not found with a simple equation but through a series of calculations called Iterative Proportional Fitting Procedure (IPFP). This is a process of small proportional adjustments. Using the margin totals of the Observed Values table and starting values of one for the Expected Values, a series of adjustments begin. First the adjustments are made so the row and column totals are correct, then the row and layer totals are corrected, and finally the column and layer totals are corrected. This threefold adjustment is repeated over and over until the margin totals for the expected values equal the margin totals for the observed values. In our case, we repeated the process thirteen times until we found our new expected values. The values generated by the IPFP are Maximum Likelihood Estimates (MLE) of the expected cell frequencies.

Table 4.7

Expected Values for Homogeneous Association Model

Grade	Class	Major	Obs	{GC, GM, CM}
A	FR	Original	123	127.58
A	FR	Other	40	37.89
A	FR	None	132	129.52
A	SO	Original	125	120.42
A	SO	Other	27	29.11
A	SO	None	51	53.48

(Continued)

Grade	Class	Major	Obs	{GC, GM, CM}
B	FR	Original	94	90.20
B	FR	Other	29	29.66
B	FR	None	181	184.14
B	SO	Original	61	64.80
B	SO	Other	18	17.34
B	SO	None	61	57.86
C	FR	Original	37	38.67
C	FR	Other	19	17.93
C	FR	None	135	134.40
C	SO	Original	31	29.33
C	SO	Other	10	11.07
C	SO	None	44	44.60
D	FR	Original	10	7.70
D	FR	Other	1	2.13
D	FR	None	43	44.17
D	SO	Original	6	8.30
D	SO	Other	3	1.87
D	SO	None	22	20.83
F	FR	Original	2	1.85
F	FR	Other	0.5	1.88
F	FR	None	55	53.77
F	SO	Original	3	3.15
F	SO	Other	4	2.62
F	SO	None	39	40.23

Having built all of our Log-linear Models and generated their Expected Values, it is now time to see if one of these Models will be able to take the place of the Saturated Model.

CHAPTER V
ASSESSMENT OF LOG-LINEAR MODELS

Likelihood-Ratio

To find the model that is the best fit for our data, we will use the Likelihood-Ratio Statistic (L^2). L^2 statistics are calculated using the table of expected values for each model. The natural logs of both the expected values and the observed values are taken. This step of the process is why we had to find a way to work around our zero cell. The natural log of the expected value is subtracted from the natural log of the observed value. The difference is then multiplied by twice the observed value. This process is repeated for each expected value in our table. All the values are then summed to give us the L^2 statistic for that model.

$$L^2 = \sum 2 * obs * (\ln obs - \ln exp)$$

Our null and alternative hypotheses for each of these models will be the same:

H_0 =The data is governed by the model and the expected frequencies do not significantly differ from the observed data

H_A =The data is not governed by the model and the expected frequencies do significantly differ from the observed data

To find the best fitting model, we are looking for a low L^2 statistic relative to the degrees of freedom. Degrees of freedom for all models are equal to the number of parameters equal to zero. We will compare the corresponding p-value to our alpha to see if it is significant and we need to reject our null hypothesis or if the data is indeed governed by the model and the expected frequencies do not significantly differ from the observed data. In this case, that model can replace the more complex Saturated Model.

There is the question of which alpha to use. Since we are interested finding a model that describes all the relationships between our three variables in the best way, we do not want to eliminate any relationship that our sample seems to be indicating. Therefore, we want to reduce the possibility of Type II errors. In order to do this, some suggest that the alpha should be between 0.10 and 0.35 to reduce the risk of accepting a model that is not a true representative of our data, even at the increased risk of not detecting a real match (Knoke & Burke, 1980). We, however, will keep our alpha at 0.05, while keeping in mind that values below 0.35 could lead to a less than ideal model.

Since our Saturated Model's Expected Values are equal to its Observed Values, making it a perfect fit, its L^2 statistic, degrees of freedom, and p-value are all equal to zero. We accept the null hypothesis; the data is governed by the model. As we have said though, our goal is to find a simpler model.

We will now go through the eighteen other models that we previously discussed to show their L^2 statistics, degrees of freedom, p-values, and significance. We begin with the simplest models, the Non-Comprehensive Models. Again, although we are interested in a final model that shows the relationship between all three of our variables, here is

where we will be able to clearly see the effect of each variable and interaction term as it is added. We will then move on to the more complex models in order to find the best fitting one for our data.

Table 5.1

Statistics for Non-Comprehensive Models

Model	L^2	df	p-value	sig
{E-P}	1344.43	29	7.55E-265	yes
{C}	1231.12	28	1.38E-241	yes
{M}	896.08	27	2.19E-171	yes
{G}	791.17	25	8.81E-151	yes
{C, M}	782.77	26	2.94E-148	yes
{CM}	741.18	24	5.31E-141	yes
{G, C}	677.86	24	1.12E-127	yes
{GC}	662.68	20	1.72E-127	yes
{G, M}	342.82	23	9.28E-59	yes
{GM}	178.45	15	4.89E-30	yes

The goodness-of-fit statistics for the Non-Comprehensive Models are listed in Table 5.1 and are sorted from highest to lowest L^2 statistic. Starting with the highest value belonging to the Equal Probability Model, we then see the three individual variable Models. The rest of the Non-Comprehensive Models came out in pairs. We see first see the Model with two variables independent of each other followed by the Model with the two variables dependent on each other. While the pairs with Class/Major and Grade/Class have extremely high L^2 statistics with only a slight improvement with the association term added, the pair with Grade/Major is different. Even as independent variables, the L^2 statistic is half the value of the previous and when the association term is added, the value is cut in half again down to 178. This shows that the association of

Grade and Major is noteworthy. As we have already said though, all of these L^2 statistic values are too high and based on their accompanying p-values, we reject the null hypotheses that these models fit the data. We will now discuss the models that include all three of our variables.

Table 5.2

Statistics for Comprehensive Models

Model	L^2	df	p-value	sig
{G, C, M}	229.51	22	1.74E-36	yes
{GC, M}	214.33	18	1.34E-35	yes
{CM, G}	187.92	20	2.72E-29	yes
{GC, CM}	172.74	16	2.39E-28	yes
{GM, C}	65.14	14	1.44E-08	yes
{GC, GM}	49.96	10	2.72E-07	yes
{GM, CM}	23.55	12	0.02341	yes
{GC, CM, GM}	6.67	8	0.57259	no
{GCM}	0	0	0	no

Table 5.2 has the goodness-of-fit statistics for the Models that include all three variables, again sorted from highest to lowest L^2 statistic. The first three Models, the Mutual Independence Model and the two Partial Independence Models without the Grade/Major association term, have L^2 statistics that are even higher than the {GM} Model's 178. The Conditional Independence Model {GC, CM}, also missing the GM association term, is close at 173. All of these Models' high L^2 statistics and significant p-values force us to reject the null hypotheses.

When we move on to the Models that include the Grade/Major interaction term, we start to see an improvement in the L^2 statistics. The Partial Independence Model

{GM, C} has an L^2 statistic of 65, while the Conditional Independence Model {GC, GM} has an L^2 statistic of 50. These are still too high however, and based on their p-values, we reject these models as being a good fit.

The last Conditional Independence Model {GM, CM} has an L^2 statistic that is on 24 and its p-value is the lowest we have seen at 0.02. This is the first model that has a p-value greater than 0.01. Although a commonly used alpha, we recall the discussion of not wanting to accept a model that excludes relationships that do exist. Therefore, we recognize that this is still significant, and the model does not offer the best fit.

Finally we have the Homogeneous Association Model {GC, GM, CM}. This model has an L^2 value of only 6.67, with a p-value of 0.57. This model is an excellent fit for our data, and the difference between it and the saturated model is not significant. It is interesting to note that even if we raise our alpha to somewhere in the suggested range of 0.10 to 0.35, the difference is still not significant. This model is a solid representation of our data. With eight fewer terms than the saturated model, this uniform association model is much simpler while still maintaining enough terms to describe the interactions of our variables. While all three variables may still be associated like our saturated model shows, the homogenous association model tells us that the relationship between any two variables seems to remain stable regardless of the third variable's value.

Dissimilarity Index

We will now calculate one more statistic to see how well the Homogenous Association Model fits our data. We choose the Dissimilarity Index. This statistic will tell us what proportion of Observed Values would need to be relocated in order for the

model in question to fit perfectly. It starts with the absolute values of the differences between each Observed Value and the Expected Value. Those values are then summed and divided by twice the sample size.

$$D = \sum |obs - exp| / 2n$$

Possible values for D range between 0 and 1; however, to show that the proposed model is a good fit for our data, we are looking for a small value for D (preferably less than 0.03). This would indicate that the variances between the Observed Values and Expected Values are minimal. For the {GC, GM, CM} Model, the Dissimilarity Index is found as follows:

$$D_{\{GC,GM,CM\}} = \sum |obs - exp| / 2n = 54.92197 / (2 * 1406.5) = 0.0195243$$

This value shows us that 1.95% of our observations would need to shift into another category in order for this model to fit the observed data of the sample perfectly. This is within the guidelines of non-important lack-of-fit ($D < 0.03$), confirming again that this Model is indeed a good fit for our data.

Compare this to the Dissimilarity Index of Model {GM, CM}:

$$D_{\{GM,CM\}} = \sum |obs - exp| / 2n = 124.34329 / (2 * 1406.5) = 0.0442031$$

This time, we see that 4.42% of observations would need to be moved to achieve a perfect fit. This is too high of a percentage to trust the model, as we had previously concluded by the high Likelihood-Ratio Statistic.

Final Model with Coefficients

Now that we have confirmed that the Homogenous Model {GC, GM, CM} is the best-fitting of our non-saturated models, we will build the full equation. We again use the equations formed by our categories to solve for our model's coefficients. Although we once again have thirty equations, we will only be using the simplest twenty-two shown in Table 5.3, since we only have twenty-two coefficients.

Table 5.3

Equations to Find Coefficients for Homogenous Association Model

$\ln 40.2 = b_0$	$\ln 20.8 = b_0 + b_4^G$
$\ln 53.5 = b_0 + b_1^G$	$\ln 53.8 = b_0 + b_1^C$
$\ln 57.9 = b_0 + b_2^G$	$\ln 3.2 = b_0 + b_1^M$
$\ln 44.6 = b_0 + b_3^G$	$\ln 2.6 = b_0 + b_2^M$
$\ln 129.5 = b_0 + b_1^G + b_1^C + b_{11}^{GC}$	$\ln 29.3 = b_0 + b_3^G + b_1^M + b_{31}^{GM}$
$\ln 120.4 = b_0 + b_1^G + b_1^M + b_{11}^{GM}$	$\ln 11.1 = b_0 + b_3^G + b_2^M + b_{32}^{GM}$
$\ln 29.1 = b_0 + b_1^G + b_2^M + b_{12}^{GM}$	$\ln 44.2 = b_0 + b_4^G + b_1^C + b_{41}^{GC}$
$\ln 184.1 = b_0 + b_2^G + b_1^C + b_{21}^{GC}$	$\ln 8.3 = b_0 + b_4^G + b_1^M + b_{41}^{GM}$
$\ln 64.8 = b_0 + b_2^G + b_1^M + b_{21}^{GM}$	$\ln 1.9 = b_0 + b_4^G + b_2^M + b_{42}^{GM}$
$\ln 17.3 = b_0 + b_2^G + b_2^M + b_{22}^{GM}$	$\ln 1.8 = b_0 + b_1^C + b_1^M + b_{11}^{CM}$
$\ln 134.4 = b_0 + b_3^G + b_1^C + b_{31}^{GC}$	$\ln 1.9 = b_0 + b_1^C + b_2^M + b_{12}^{CM}$

We now have our final equation:

$$\begin{aligned}
N_{ijk} = & 3.694530 + 0.284752 t_1^G + 0.363578 t_2^G + 0.103196 t_3^G - 0.658115 t_4^G + \\
& 0.290249 t_1^C - 2.545545 t_1^M - 2.731978 t_2^M + 0.594316 t_{11}^{GC} + \\
& 3.357219 t_{11}^{GM} + 2.123610 t_{12}^{GM} + 0.867314 t_{21}^{GC} + 2.658689 t_{21}^{GM} + \\
& 1.526811 t_{22}^{GM} + 0.812849 t_{31}^{GC} + 2.1264889 t_{31}^{GM} + 1.338367 t_{32}^{GM} + \\
& 0.461374 t_{41}^{GC} + 1.625430 t_{41}^{GM} + 0.321088 t_{42}^{GM} - 0.826748 t_{11}^{CM} - \\
& 0.620661 t_{12}^{CM}
\end{aligned}$$

With eight fewer terms, this Homogenous Association Model {GC, GM, CM} is an excellent fit for our sample data.

CHAPTER VI

CONCLUSION

In this study, we have looked at the connection between course grades, classification, major, and graduation. Using Log-Linear Analysis, we developed a Homogenous Association Model that accurately describes the relationship between our variables that we saw in our sample data. The equation shows the connection between students' course grades, classification, major, and graduation.

For future research, we need to repeat this study with more samples, both from this course and other freshman level courses in all subjects. This would help us to discover the courses that best act as gateway courses, both for specific majors and overall. We could then expand our research to look at other factors that prevent students from reaching their goal of graduation. By identifying these Gateway Courses and educating students about graduation rate predictions, we can help more students achieve long term success in their program.

REFERENCES

- Agresti, A. (2002). *Categorical data analysis*. (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Burton, N.W. & Ramist, L. (2001). Predicting success in college: SAT studies of classes graduating since 1980. College Board Research Report No 2001-2. New York: College Entrance Examination Board.
- Camara, W.J. & Echternacht, G. (2000). The SAT I and high school grades: Utility in predicting success in college. College Board Research Notes No RN-10. New York: College Entrance Examination Board.
- Dougherty, C., Mellor, L., and Jian, S. (2006). The relationship between advanced placement and college graduation. National Center for Educational Accountability.
- Early Warning System for Retention. (2008). Academic Affairs. Texas Woman's University. <http://www.twu.edu/academic-affairs/early-warning-process.asp>
- Geiser, S., & Santelices, M.V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. Center for Studies in Higher Education, Research & Occasional Paper Series. Berkeley, CA.

- Harackiewicz, J.M., Barron, K.E, Tauer, J.M., & Elliot, A.J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562-575.
- Knoke, D. & Burke, P. (1980). Log-linear models. Sage University Paper series on quantitative applications in the social sciences, 07-020. Beverly Hills, CA: Sage Publications.
- Log-linear models. Lecture notes on analysis of discrete data, lesson 10. Pennsylvania State University. <http://onlinecourses.science.psu.edu/stat504/node/117>
- Lotkowski, V.A., Robbins, S.B., & Noeth, R.J. (2004). The role of academic and non-academic factors in improving college retention. Act Policy Report. Iowa City, IA: ACT, Inc.
- http://www.act.org/research/policymakers/pdf/college_retention.pdf
- National Center for Education Statistics. (2014). Graduation rate from first institution attended for first-time, full-time bachelor's degree-seeking students at 4-year postsecondary institutions, by race/ethnicity, time to completion, sex, control of institution, and acceptance rate: Selected cohort entry years, 1996 through 2007. Table 326.10. https://nces.ed.gov/programs/digest/d14/tables/dt14_326.10.asp
- Rodriguez, G. (2007). Log-linear models for contingency tables. Lecture notes on generalized linear models, chapter 5. Princeton University.
- <http://data.princeton.edu/wws509/notes/c5.pdf>

Smith, V.K. & Cicchetti, C.J. (1974). A note on fitting log-linear regression with some zero observations for the regressand. *Metroeconomica*, 26 1-3: 282-284.

TWU Fact Book. (2015). Office of Institutional Research and Data Management. Texas Woman's University. <http://www.twu.edu/institutional-research/fact-book.asp>

Zaiontz, C. (2015). Real statistics using Excel. <http://real-statistics.com>