2008-07-07

Ordinal Regression to Evaluate Student Ratings Data

Emily Brooke Bell
Brigham Young University - Provo

ORDINAL REGRESSION TO EVALUATE STUDENT RATINGS DATA

by

Emily B. Bell

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics
Brigham Young University
August 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Emily B. Bell

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	H. Dennis Tolley, Chair
Date	Scott D. Grimshaw
Date	David A. Engler

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the project of Emily B. Bell in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

H. Dennis Tolley
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

ORDINAL REGRESSION TO EVALUATE STUDENT RATINGS DATA

Emily B. Bell

Department of Statistics

Master of Science

Student evaluations are the most common and often the only method used to evaluate teachers. In these evaluations, which typically occur at the end of every term, students rate their instructors on criteria accepted as constituting exceptional instruction in addition to an overall assessment. This presentation explores factors that influence student evaluations using the teacher ratings data of Brigham Young University from Fall 2001 to Fall 2006. This project uses ordinal regression to model the probability of an instructor receiving a good, average, or poor rating. Student grade, instructor status, class level, student gender, total enrollment, term, GE class status, and college are used as explanatory variables.

ACKNOWLEDGEMENTS

Thanks to Dr. Tolley and Dr. Grimshaw for much help from beginning to end.

To my mom for a lifetime of encouragement and support.

To Joshua for believing that I could finish.

CONTENTS

CHAPTER

APPENDIX

TABLES

Table

FIGURES

Figure

1. ORDINAL REGRESSION

Ordinal logistic regression is a statistical tool used when the response is categorical with ordered outcomes. The outcome of the model provides predicted probabilities for each level of the response. Ordinal data is found in areas such as psychometrics, customer satisfaction, and taste testing where the outcomes are related to a latent variable but we can only observe relative levels.

Traditional tools of linear regression do not provide an adequate framework to model the probabilities of each outcome. For example, if linear models are used to analyze ordinal data, predictions can occur outside the probability space. Also, linear models assume constant variance, which is violated in the ordinal case because the variance depends on the probabilities. For these reasons additional tools are needed to analyze ordinal data.

McCullagh (1980) developed these tools to deal with situations in which the response variable is not measured on a continuous scale. His method used the successive dichotomizations of the data forming cumulative splits to the data. This allows for the ordinal nature of the response to be utilized and results in an analysis similar to logistic regression. This cumulative odds model simultaneously considers predictors across all cumulative splits to model the probability of success (O'Connell, 2006).

Consider a special case where three levels are observed: "poor," "average," and "good." Let $\theta_1 = \pi_1$ and $\theta_2 = \pi_1 + \pi_2$ where π_1 denotes the probability of "good," π_2 denotes the probability of "average," and π_3 denotes the probability of "poor." Thus, θ_1 is the probability of "good" and θ_2 is the probability of "good" or "average."

The logit function uses the log of the odds ratio to model the probabilities of each level. The odds ratio uses the probability of the event occurring to the probability of the event not occurring. The logit function is then set equal to the linear combination of explanatory variables being considered. This function for θ_1

becomes

$$logit(\theta_1) = \log\left\{\frac{\pi_1}{\pi_2+\pi_3}\right\} = \alpha_1 + x'\beta.$$

This represents the log of the probability of a good rating to the probability of an average or poor rating. Exponentiating each side,

$$\theta_1 = \pi_1 = P[\text{good}|x] = \frac{\exp\left(\alpha_1 + x'\beta\right)}{1+\exp\left(\alpha_1 + x'\beta\right)}.$$

The logit function used to model θ_2 is

$$logit(\theta_2) = \log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \alpha_2 + x'\beta.$$

This represents the log of the probability of a good or average rating to the probability of a poor rating. Exponentiating each side gives

$$\theta_2 = \pi_1 + \pi_2 = P[\text{good or average}|x] = \frac{\exp\left(\alpha_2 + x'\beta\right)}{1+\exp\left(\alpha_2 + x'\beta\right)}.$$

For a specific x, predictions for π_1, π_2, and π_3 are computed from θ_1 and θ_2 using $\theta_1 = \pi_1$, $\theta_2 = \pi_1 + \pi_2$, and $\pi_3 = 1 - \pi_1 - \pi_2$.

One common assumption in ordinal logistic regression is the proportional odds assumption. This model assumes that the true β-values are the same for each θ, and the only difference exists in the intercept. The test used by many statistical software packages to verify this assumption is not a powerful test and will nearly always result in small p-values when "the sample size is large or when continuous variables are used in the model" (O'Connell, 2006). The model for this analysis did assume proportional odds.

2. DATA

The data for this project comes from the teacher evaluations at Brigham Young University from Fall 2001 to Fall 2006. The original data set contains approximately 970,000 observations. Students who withdrew from a class, received an incomplete, were enrolled in a pass/fail class, or received no grade were excluded from the analysis. This resulted in a final data set of 943,320 student ratings. Because these data contain sensitive information, all specific information (i.e. college name, instructor name, student ID, etc.) were blinded. This blinding limits the interpretation of interesting effects but was required as a condition for providing the data. An example of the limited interpretation is that differences between colleges could be explored but how those differences were impacted by field of study could not be determined.

These data differ from other data collected in teacher evaluation analyses in several significant ways. First, the data contain all the evaluations a given student made over multiple courses and all the evaluations from a given instructor over multiple courses and students. In this analysis, this obvious correlation structure will be ignored since the purpose of the project is to investigate regression on ordinal data. Second, the data contains the actual grade the student received in the course. Most research uses the grade the student reports as their expected grade in the course. Table 2.1 shows summary statistics for student and course characteristics. These summary statistics are comparable to the BYU student and course demographics.

Student ratings of teachers are the most commonly used and often the only used tool to evaluate teaching (Seldin, 1999). Assessing teaching is important because it can lead to improvement; however, few schools are satisfied with the measures they have in place to assess teaching and use the information gained from the assessment (Seldin, 2006). Typically, schools will perform assessments at the end of each semester or term and allow students to give feedback on the course and teacher. The feedback

Table 2.1: Summary statistics for the BYU teacher evaluation data

Student Characteristics		Course Characteristics	
Gender		Class Level	
Female	51.9%	100-200	58.2%
Class Level		300-400	36.2%
F/S	31.7 %	Graduate	5.8%
J/S	62.8 %	Term	
Graduate	5.5 %	Fall	51.3 %
Nationality		Winter	40.3 %
US	92.0%	Spring	4.9 %
Age		Summer	3.8 %
Minimum	13	Instructor Status	
Q1	19	Full	65.7 %
Median	21	Part	26.9 %
Q3	23	Student	7.4 %
Maximum	80	Total Enrolled	
ACT Composite		Minimum	1
Minimum	10	Q1	22
Q1	24	Median	36
Median	27	Q3	66
Q3	29	Maximum	1684
Maximum	36		
GPA			
Minimum	0.08		
Q1	3.15		
Median	3.48		
Q3	3.72		
Maximum	4.00		

often includes specific questions related to learning objectives, teaching style, use of classroom time, and teacher-student interaction.

Table 2.2 shows the questions used in the student evaluations at Brigham Young University. Students have access to the survey beginning two weeks before the end of the semester. During this interval they can rate their teachers online by answering questions and providing comments they feel will benefit the teacher. Students are asked specific questions regarding their experience in the course related to grading procedures, course materials, and course objectives and goals. Additionally, they give the course an overall rating. Students are also asked questions regarding their experience with the instructor. The questions cover topics such as student-teacher interaction and effectiveness in explaining concepts. Students also give their teacher an overall rating.

The response of the ordinal logistic regression analysis presented here used the overall teacher rating. Students give their teacher a rating of one through eight. A rating of one represents “very poor,” and a rating of eight represents “exceptional.” Ratings of one through four were classified as “poor,” ratings of five and six were classified as “average,” and ratings of seven and eight were classified as “good.” The ratings were grouped for two reasons. First, the labels assigned to each rank show different ability levels. For example, the labels associated with a rating of one through three all include the word “poor,” which suggests that these levels can be grouped together. Additionally, research suggests that students rarely take advantage of lower ratings which results in a positive evaluation bias (Seldin, 2006). For this reason, a four rating was placed in the poor category, as students are unlikely to give a lower rating. The labels for a rating of seven or eight indicated that the performance of the instructor was above average, so these ratings were grouped together in the “good” category. A rating of five or six were then considered as average.

Response rate is important to consider when examining teacher evaluation data.

Table 2.2: Teacher evaluation questions asked to students attending Brigham Young University

	Course Questions
1.	Comparing this course with other university courses you have taken, please indicate an OVERALL rating for the course.
2.	I learned a great deal in this course.
3.	Course materials and learning activities were effective in helping students learn.
4.	This course was well organized.
5.	Evaluations of students' work (e.g., exams, graded assignments and activities) were good measures of what students learned in the course.
6.	Course grading procedures were fair.
7.	This course helped me develop intellectual skills (such as critical thinking, analytical reasoning, integration of knowledge).
8.	This course provided knowledge and experiences that helped strengthen my testimony of the Gospel of Jesus Christ.
9.	For this course, about how many hours per week did you spend in class?
10.	What percentage of the time you spent in class was valuable to your learning?
11.	For this course, about how many hours per week did you spend out of class (doing assignments, readings, etc.)?
12.	What percentage of the time you spent out of class was valuable to your learning (as opposed to just busy work)?
	Instructor Questions
1.	Comparing this course with other university courses you have taken, please indicate an OVERALL rating for the instructor.
2.	The instructor showed genuine interest in students and their learning.
3.	The instructor provided adequate opportunities for students to get help when they needed it.
4.	The instructor provided opportunities for students to become actively involved in the learning process.
5.	The instructor gave students prompt feedback on their work.
6.	The instructor provided students useful feedback on their work.
7.	The instructor responded respectfully to students' questions and viewpoints.
8.	The instructor was effective in explaining difficult concepts and ideas.
9.	The instructor appropriately brought Gospel insights and values into secular subjects.
10.	The instructor was spiritually inspiring insofar as the subject matter permitted.
11.	This instructor and course contributed to the Mission and Aims of a BYU Education (i.e., Spiritually Strengthening, Intellectually Enlarging, Character Building, Leading to Lifelong Learning and Service).

Table 2.3: Results of the logistic regression analysis to model response rate

Parameter	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	0.0556	0.00362	236.5134	< 0.0001
Year 1	-0.6028	0.00419	20707.4402	< 0.0001
Year 2	0.0601	0.00299	403.4598	< 0.0001
Year 3	0.3245	0.00302	11569.9527	< 0.0001
Year 4	-0.2946	0.0030	9642.4963	< 0.0001
Fall	0.2561	0.00311	6775.6795	< 0.0001
Spring	-0.1818	0.00528	1183.0523	< 0.0001
Summer	-0.1700	0.00586	840.9977	< 0.0001
Level 100-200	0.0115	0.00286	16.2131	< 0.0001
Level 300-400	-0.1419	0.00284	2494.4168	< 0.0001
GE Class	0.1433	0.00200	5132.0504	< 0.0001

To have representative information, the response rate needs to be above 65% (Seldin, 2006). Presently at BYU there is concern because the shift from in-class student evaluations to web-based student evaluations has resulted in a plummeting response rate. In the Statistics Department, the response rate has dropped from over 95% to between 50 and 65%. The following analysis is provided for those interested in factors that affect response rate. Table 2.3 shows the results of a simple logistic regression performed on the response rate.

When compared to year 5, the response rate in years 1 and 4 are significantly lower, and years 2 and 3 have a higher response rate. Because of the blinding of the data, it is unknown what school years these values correspond to. Fall has a higher response rate than Winter semester, and both Spring and Summer terms have lower response rates than Winter semester. Classes numbered 100 through 200 have about the same response rate as graduate level classes, but 300 to 400 level classes have a lower response rate. GE classes have a higher response rate than non-GE classes. The odds ratios for these variables are presented in Table 2.4. Those odds ratios that are greater than one indicate that the probability of a response for that variable is greater than the probability of a response for the reference group. For example, a

Table 2.4: Odds ratios with 95% confidence intervals for the response rate analysis

Effect	Point Estimate	95% Confidence Interval
Year 1 vs 5	0.328	0.324, 0.332
Year 2 vs 5	0.636	0.630, 0.642
Year 3 vs 5	0.828	0.821, 0.836
Year 4 vs 5	0.446	0.442, 0.450
Fall vs Winter	1.174	1.166, 1.182
Spring vs Winter	0.758	0.748, 0.768
Summer vs Winter	0.767	0.755, 0.779
100-200 vs Graduate	0.888	0.876, 0.900
300-400 vs Graduate	0.762	0.751, 0.772
GE vs Non-GE	1.332	1.322, 1.342

student is 33% more likely to complete a student rating in a GE class than a non-GE class. One factor that could not be measured in this study and likely has a large effect on this outcome involves courses that offer extra credit.

3. ORDINAL REGRESSION MODEL

3.1 College Effect

Studies have confirmed that ratings vary across academic fields. The results suggest that scientific and quantitative classes, such as math classes, receive the lowest ratings on average. Humanities classes are found to receive the highest rating, with the social science classes in between. Although the reason for this difference is not clear, one possible explanation offered by Cashin (1990) is that the classes that receive the lowest ratings may be more difficult to teach, and therefore receive the lowest ratings. Academic discipline must be taken into account when analyzing student ratings data. A surrogate used for academic discipline in this analysis is college.

At BYU, the data suggests that there are differences in the ordinal regression models between college. Consider seven different college clusters, defined in Table 3.1. Because the college name was blinded in the data set, comparison between colleges cannot be made. For example, while college 6 and 11 have similar characteristics in the analysis, the blinding makes it impossible to determine if these colleges have a common field of study.

Table 3.1: College clusters

Cluster	Colleges
A	2
B	12
C	8, 9, and 13
D	4 and 14
E	6 and 11
F	1,5, and 7
G	10 and 15

The college clusters were created by first letting each college receive its own model initially. This first model also included all interactions with college. This model allowed for each college to have its own additive effect and a possibly different coefficient for each explanatory variable. The goal was to identify colleges with similar coefficients that could be clustered together. Plots were constructed for each interaction, which plotted the slope and college effect. The first college identified as different from the others was College 2. The effect of College 2 was significantly different from the other colleges, and the slopes for College 2 explanatory variables often differed in magnitude and sign. A two-cluster model was considered, which separated College 2 and clustered all remaining colleges together. The likelihood ratio test comparing this two-cluster model to the full model with all colleges considered separately indicated that this model lost too much information and other clusters needed to be considered.

College 12 also seemed to differ from the other colleges in many explanatory variables. A three-cluster model considered College 2 and College 12 and clustered the remaining colleges together. The likelihood ratio test indicated that this model was an improvement, but more clusters still needed to be considered.

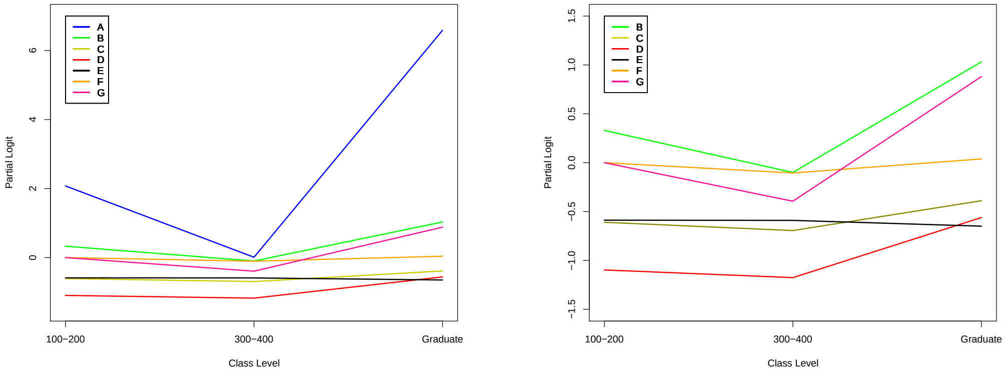
Identification of the five remaining clusters was based on comparing the signs of the slopes across the different explanatory variables. Those colleges with consistently positive slopes were clustered, as well as those with negative slopes. The effects related to College 10 were consistently insignificant which suggested that College 10 was closely related to College 15, the reference college used in the model. A series of models were considered that used various clusterings until a likelihood ratio test indicated similar performance with the model where each college had a different model.

The explanatory variables in the model consist of those that are common to student ratings data, along with the college interactions with class level and total enrollment. The significance of the interactions with college indicate that the results with regards to class level and total enrollment need to be discussed in context of a particular college cluster.

3.1.1 Class Level

Research suggests that higher-level courses receive higher ratings. This finding appears to be related to the idea that students find courses more enjoyable if they have a prior interest in the subject matter. As students progress in their education, they have more control over the courses in which they enroll. That is, students choose a major based on their interests, and these will be higher-level courses. Studies confirm that prior student interest impacts student responses (Johnson, 2003).

The data suggests the effect of class level differs with college cluster. Figure 3.1 shows the partial logit effect of class level for each college cluster, which conveys the change in the likelihood of a teacher receiving a good rating when all else is held constant. Figure 3.1(a) includes all college clusters to show the distinction between Cluster A and the other clusters, and the Figure 3.1(b) shows all colleges except Cluster A to show the differences between these clusters which are obscured in Figure 3.1(a) due to the vertical axis scale.



(a) All Colleges

(b) College 2 Omitted

Figure 3.1: Partial logit for each college group based on class level

Notice that Cluster A is positioned significantly higher than the other colleges. This suggests that teachers of all classes in Cluster A get higher ratings than teachers in other colleges. Also, the increase in ratings that teachers of graduate-level classes can expect in Cluster A is more dramatic than any other level in the other colleges. This suggests that teachers of graduate-level classes in Cluster A can almost always expect a good rating.

Figure 3.1(b) omits Cluster A and reveals the three main trends for the remaining colleges. First, two college clusters have a V-shaped relationship between ratings and class level. The V shape indicates that teachers of 300–400-level classes can expect lower ratings when compared to both 100–200 and graduate-level classes when all else is held constant. In these clusters, the increase in ratings when moving to graduate-level classes is greater than 100–200-level classes. Another pattern shows no great difference between 100–200- and 300–400-level classes, but shows an increase in graduate-level classes. Finally, two clusters show no meaningful difference across class level. As class level changes in these clusters, a slight difference exists in the effect of class level on student ratings, but this effect seems minimal when compared to the other clusters.

3.1.2 Total Enrollment

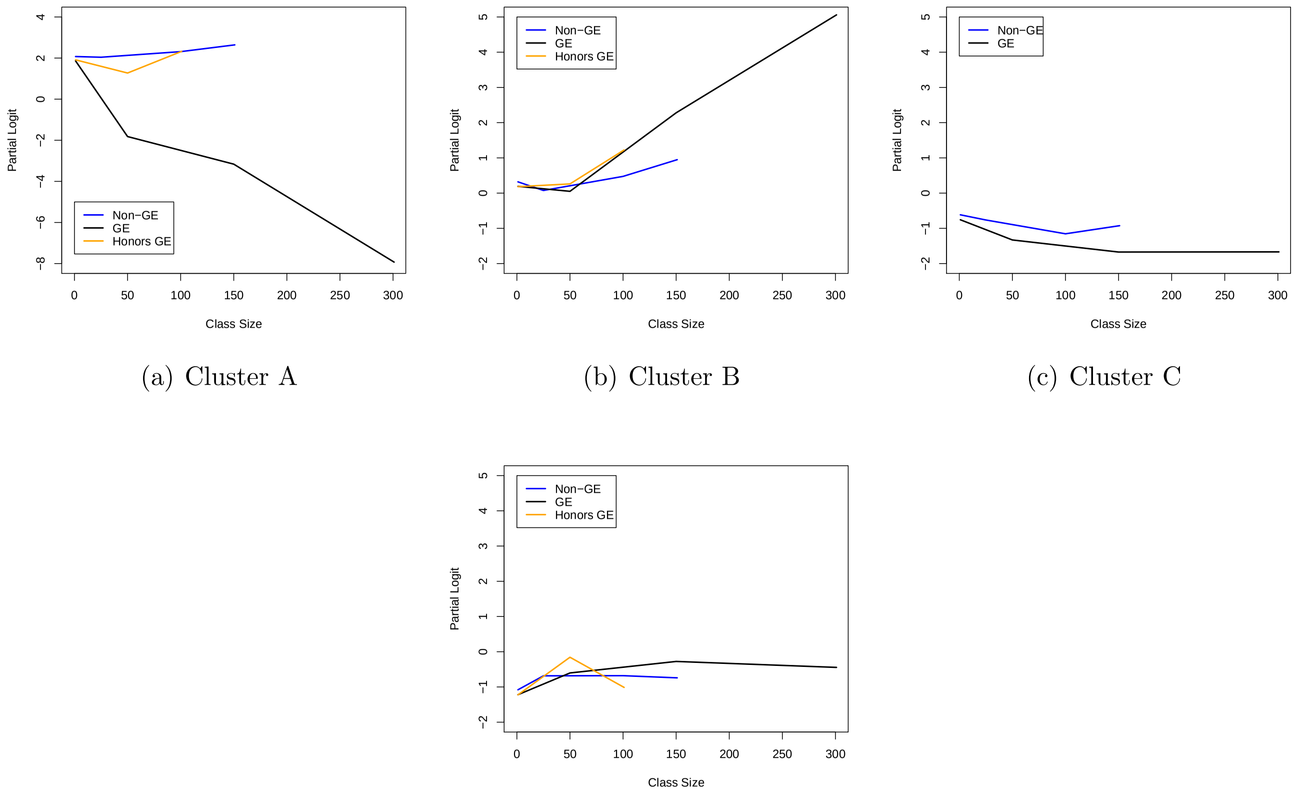
Total enrollment also impacts the student rating. The general association between total enrollment and student ratings follows a U-shaped pattern. When enrollment is small, evaluations tend to be high, but as class size increases ratings tend to decrease. At some point, ratings begin to increase again for very large classes (Fernandez, 1998).

Total enrollment appears to be confounded with GE class status. Non-GE classes typically have lower enrollment when compared to GE classes, and GE classes can be divided into regular and honors status. Honors classes typically have smaller class sizes. Because not every college offers honors classes, this effect is not estimated in all college clusters. The total enrollment effect was estimated using a piecewise linear spline for non-GE, GE, and GE honors. Each spline was fitted to only those total enrollments typical for each type.

The spline was originally considered with knots at enrollment of twenty-five, fifty, and, in increments of fifty up to three hundred. Classes with enrollment greater than three hundred were "clumped" at three hundred. Partial regression plots were used to determine which knots were necessary by evaluating the wiggliness of the curve. The spline was then refitted with only necessary knots.

Figures 3.2 and 3.3 present the piecewise linear approximations of total enrollment for each college cluster. As with class level, Cluster A has a different effect than the other college clusters. Non-GE classes show a slight increase in ratings as enrollment increases. Notice that the graph for Cluster A is on a different scale than the others to accommodate the dramatic difference seen in ratings for GE classes. For classes with enrollment lower than fifty, the decline in student ratings is quite drastic. The rate of decline slows a little for classes with between fifty and one hundred fifty students, and beyond one hundred fifty, the rate of decline increases.

Cluster B shows a different effect for GE classes. At first, ratings slightly decrease, but once class size surpasses fifty the increase in teacher ratings is dramatic, far exceeding both honors and non-GE classes. Two other clusters, D and E, show a slight increase in the effect of total enrollment. In Cluster D, the increase occurs until total enrollment reaches one hundred fifty, and then total enrollment tends to have no effect. For cluster E, the estimated effect of class size is too wiggly because the knots were selected based on the entire data set. The remaining colleges show that for GE classes, as total enrollment increases, teacher ratings decrease.



(a) Cluster A

(b) Cluster B

(c) Cluster C

(d) Cluster D

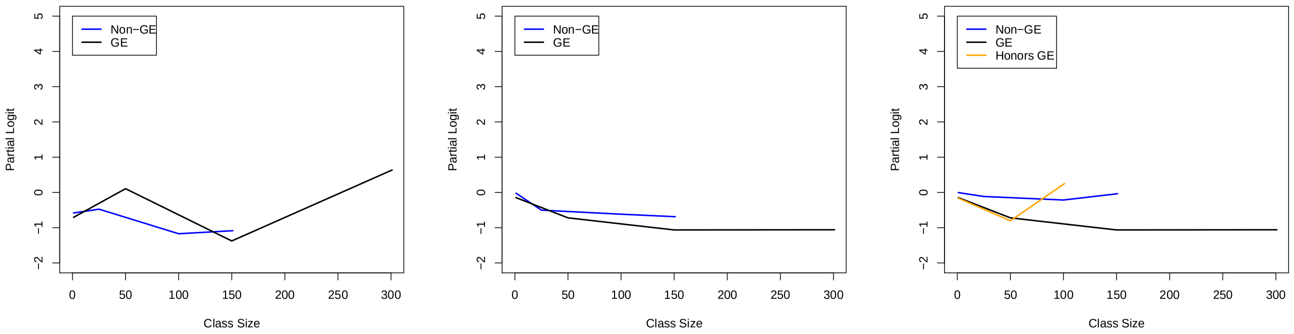
Figure 3.2: Partial logit plots for each college cluster based on total enrollment

For non-GE classes, most college clusters show the expected U-shaped effect. For most of these clusters, the result for large classes does not reach the same height as small classes. The one exception to this is Cluster B, in which the partial logit effect for large classes exceeds that for small classes. Clusters E and F show a general downward trend in student ratings as enrollment increases. Both college clusters show little influence on ratings when total enrollment surpasses one hundred. Cluster D shows a slight increase in ratings as total enrollment increases, but once enrollment surpasses fifty the effect on teacher rating is minimal.

For the three college clusters that offer honors classes, two show the expected U-shaped pattern for enrollment. The effect on teacher ratings decreases until enrollment reaches fifty, at which point increasing enrollment leads to increasing teacher ratings. In both cases, the logits for large classes exceed those for small classes. Cluster D shows the opposite pattern. An increase in total enrollment leads to an increase in teacher rating until enrollment reaches fifty, at which point the effect of enrollment on teacher rating is negative.

3.2 Main Effects Common to all College Clusters

Some effects are common across all college clusters. Student grade was found to be an important predictor of the rating a teacher receives. Other explanatory variables such as instructor status, term, class standing, and gender showed the same effects across clusters. Table 3.2 shows the odds ratios for these variables. Those odds ratios that are less than one indicate a small probability of receiving a good rating when compared to the reference category, and those that are greater than one indicate an increase in probability of a good rating. Each variable will be considered in the sections that follow.



(a) Cluster E

(b) Cluster F

(c) Cluster G

Figure 3.3: Partial logit plots for each college cluster based on total enrollment

Table 3.2: Odds ratios for covariates included in every college cluster

Effect	Point Estimate	95% Confidence Limits
Instructor Status Part vs. Full	0.912	0.903, 0.922
Instructor Status Student vs. Full	0.930	0.904, 0.957
Term Spring vs. Fall	1.246	1.220, 1.272
Term Summer vs. Fall	1.206	1.178, 1.235
Term Winter vs. Fall	1.080	1.071, 1.090
Student Grade	1.284	1.277, 1.291
Class Standing Upper vs. Lower	0.857	0.848, 0.866
Class Standing Graduate vs. Lower	0.810	0.783, 0.837
Female vs. Male	1.027	1.018, 1.036

3.2.1 Student Grade

Student grade was included in the model because it is highly contested in determining what factors influence student ratings. Nearly all studies conclude that there exists a positive association between student grade and teacher evaluation, but the degree of this association and the underlying cause remain contested (Johnson, 2003).

Those wishing to explain the effect of student grade without wanting to suggest that it represents a "biasing effect" often look for other explanations. The teacher effectiveness theory is the most commonly accepted theory used to explain the positive correlation between student grade and teacher evaluations. This theory bases its core idea on the assumption that students learn more from effective teachers. Because they learn more, they earn higher grades. Thus, the positive relationship between student grade and teacher evaluation is "not only not the result of an underlying bias, but is a desirable feature of student evaluations of teaching" (Johnson, 2003).

The other prominent theory regarding student grades, which represents the opposite end of the spectrum, assumes that students reward teachers who reward them. This view, often called the grade satisfaction theory, proposes that the effect of student grade on teacher evaluations is a biasing effect because grades are not related to effective teaching or student learning (Johnson, 2003).

Regardless of the true cause of the positive correlation between student grade and teacher evaluations, the fact that a correlation exists suggests that ratings must be adjusted for the student grade prior to their use by administration (Johnson, 2003).

At BYU, student grade has a positive effect on teacher ratings ($\chi^2 = 8506.06$, df=1, p-value < 0.0001). In fact, student grade was the most important predictor in determining the rating a student would give a teacher. Figure 3.4 shows the partial logit curves for student rating for each college cluster. The steady increase in the partial logit results in a significant increase in the likelihood that a teacher will get a good rating. To investigate different grade effects between college clusters, the hypothesis test for a student grade by college interaction had χ^2=219.2779, df=6, p-value < 0.0001. The interaction was not warranted, as the effect of student grade did not vary across college clusters.

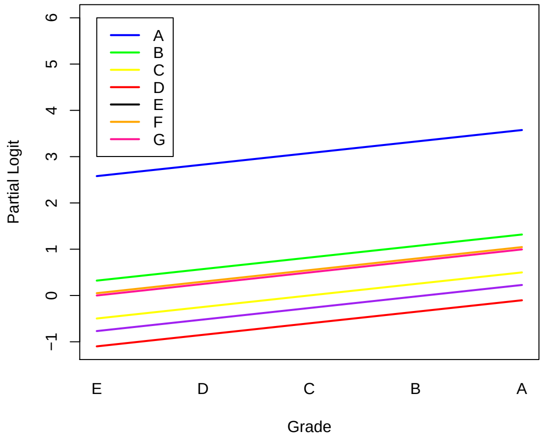


Figure 3.4: Partial logit for student grade

3.2.2 Instructor Status

Instructor status impacts how students will rate their teacher ($\chi^2 = 273.63$, df=2, p-value < 0.0001), but this statistically significant result appears to imply little practical difference between the three groups after adjusting for the other explanatory variables. Figure 3.5 shows the partial logit effect of instructor status on teacher rating, and this plot indicates that the confidence intervals overlap.

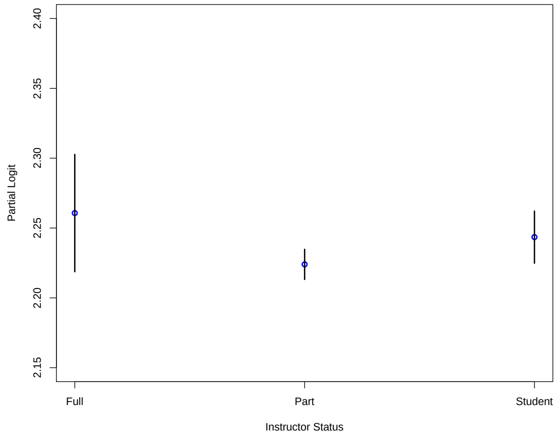


Figure 3.5: Partial logit for instructor status

3.2.3 Semester

The semester a course was offered affects the rating that a teacher will receive ($\chi^2 = 742.7$, df=3, p-value < 0.0001). The partial logit effect of term on teacher ratings is shown in Figure 3.6. Fall and winter semesters are very close to each other, and spring and summer terms also appear quite close to each other. However, this difference, while statistically significant, seems to have little practical meaning after adjusting for other explanatory variables.

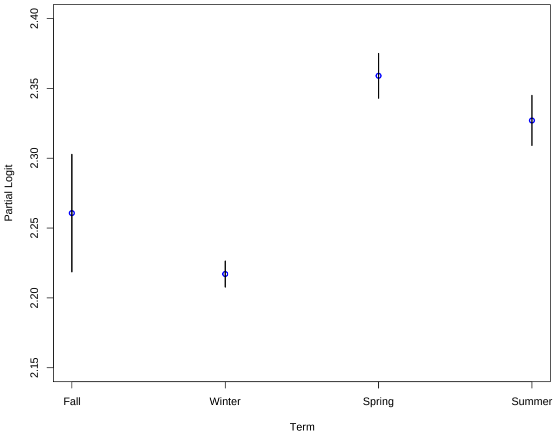


Figure 3.6: Partial logit for term

3.2.4 Class Standing

A characteristic of the student that influences ratings is their class standing (freshman, sophomore, junior, senior, and graduate).

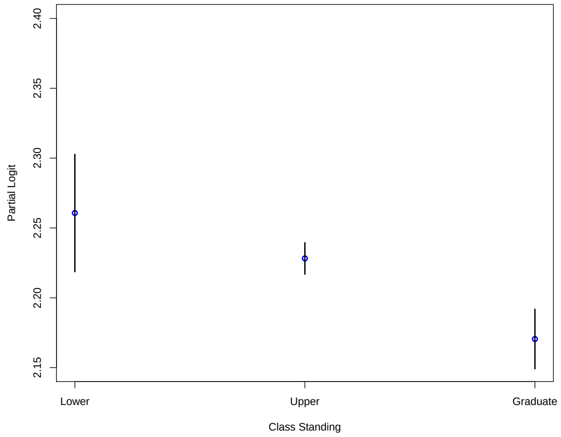


Figure 3.7: Partial logit for class standing

This finding is confirmed in the BYU data, which indicates that class standing of the student does influence the rating a teacher receives ($\chi^2 = 845.74$, df=3, p-value $<$ 0.0001). The effect at BYU differs from other research that indicates that as students progress in school, they give higher ratings on average. A common explanation is that upper-level and graduate students are enrolled in courses that interest them because they are taking more elective classes in their major interest of study. At BYU, lowerclassmen, referring to Freshmen and Sophomores, give higher ratings on average than upperclassmen and graduate students. The largest difference occurs between upperclassmen and graduate students. Figure 3.7 helps to explain the effect of class standing on teacher rating by showing the partial logit effect. There is some overlapping of the confidence intervals so the difference between BYU's results and those of other researchers may lack practical significance.

3.2.5 Gender

Gender has been shown in the literature to influence the teacher ratings. The most common association relates the gender of the teacher in relation to the gender of the student. For example, one study indicated that female students give lower ratings to female teachers than to male teachers (Tieman and Rankin-Ullock, 1985).

In these data, this interaction cannot be estimated because the gender of the teacher is unknown. Student gender is statistically significant ($\chi^2 = 34.08$, df=1, p-value < 0.0001). Figure 3.8 shows the partial logit effect of gender on teacher ratings is practically flat. This suggests that no real difference exists between the genders.

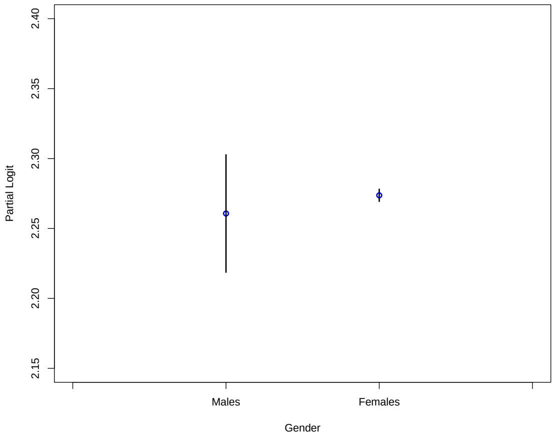


Figure 3.8: Partial logit for gender

3.3 Model Validation

Model validation is based on finding the misclassification rate for the overall data and also with cross-validation by using one year as a hold out. Two methods of prediction were considered. The first rule specifies that the predicted rating is determined by the rating with the largest predicted probability. If $\hat{\pi}_1 > max\{\hat{\pi}_2, \hat{\pi}_3\}$, the predicted rating is classified as "good." Similar rules were used to classify ratings as “average” and “poor.” The results for this are shown in Table 3.3.

Table 3.3: Classification for all data using majority rule

			Observed		
		Poor	Average	Good	Total
Predicted	Poor	0	0	0	0
	Average	11,544	26,634	23,703	61,881
	Good	58,638	251,323	541,506	851,467
	Total	70,182	277,957	565,209	913,348

Notice that this classification rule never gives a poor classification. The proportion of poor ratings in the data set is less than 8% and $\hat{\pi}_3$ is small. Another classification rule needed to be considered that would predict poor ratings. Consider a conditional rule. If $\hat{\pi}_3 > 0.15$, classify as “poor.” Otherwise classify as average if $\hat{\pi}_2 > 0.3$. Observations not classified as poor or average are classified as good. The results for this classification rule is shown in Table 3.6. The resulting misclassification rate is approximately 44%.

Cross-validation techniques were used to further validate the model. Each year was held out and then predictions were made to this year using the model. Classifications were made using the rules established above. The misclassification for each year was between 43 and 45%.

3.4 Prediction

Because of differences in college clusters, and course and instructor characteristics, benchmarks for student ratings could be computed from predictions from the model. For example, Figure 3.9 shows the predicted probabilities for each college cluster for male lowerclassmen earning a B in 100–200-level GE courses of size 25 taught in fall semester by a full-time faculty member. The green area represents the probability of receiving a good rating, yellow corresponds to average ratings, and red

Table 3.4: Classification for all data using the conditional rule

		Observed			
		Poor	Average	Good	Total
Predicted	Poor	9,720	21,624	18,717	50,061
	Average	31,424	106,728	153,755	291,907
	Good	29,038	149,605	392,737	571,380
	Total	70,182	277,957	565,209	913,348

Table 3.5: Observed marginal proportions for classification table across years

Year	Poor	Average	Good
1	2.22%	8.43%	88.11%
2	8.24 %	30.86%	60.91%
3	8.44%	31.51%	60.04%
4	6.36%	28.48%	65.17%
5	8.47 %	31.90%	59.63%

Table 3.6: Classification by year using the conditional rule

	Year 1	Observed					Year 2	Observed			
		Poor	Average	Good	Total			Poor	Average	Good	Total
Predicted	Poor	365	1,082	4,744	6,191	Predicted	Poor	1,917	3,739	3,365	9,021
	Average	764	2,766	22,375	25,905		Average	7,250	22,976	31,943	62,169
	Good	488	2,280	37,865	40,633		Good	7,381	35,279	87,052	129,712
	Total	1,617	6,128	64,984	72,729		Total	16,548	61,994	122,360	200,902
	Year 3	Observed					Year 4	Observed			
Predicted	Poor	2,184	4,648	3,677	10,509	Predicted	Poor	1,881	5,270	4,937	1,2088
	Average	8,965	27,914	37,632	74,511		Average	4,696	18,717	29,753	53,166
	Good	8,061	39,171	95,385	142,617		Good	4,260	24,545	76,371	105,176
	Total	19,210	71,733	136,694	227,673		Total	10,837	48,532	111,061	170,420
	Year 5	Observed									
Predicted	Poor	2,228	4,196	3,542	9,966						
	Average	9,156	28,224	37,503	74,883						
	Good	9,079	44,668	103,054	156,801						
	Total	20,463	77,088	144,099	241,650						

indicates a poor rating. The college clusters have clear differences. Faculty in Cluster A can expect a high rating, but faculty in Cluster D have half their student ratings in poor and average classifications.

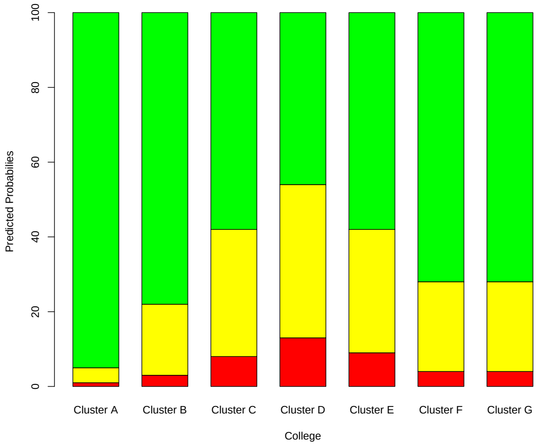


Figure 3.9: Predicted probabilities for college clusters for a male lowerclassman student earning a B in a 100–200-level non-GE course with enrollment of 25 taught by a full-time faculty member in fall semester

Figure 3.10 shows the predicted probabilities by grade for college Cluster F for male lowerclassmen in a 100–200-level GE course with enrollment of 25 taught by a full-time faculty member in fall semester. Graphs for all colleges are found in Appendix A. As grade increases, the probability of receiving a good rating increases. The probability of an instructor receiving a good rating from a failing student in cluster F is approximately 55%, and for an A student the probability is approximately 70%.

The predicted probabilities for each college cluster based on class level are shown in Figures 3.11 and 3.12. Cluster A stands out because of the high student ratings. Clusters C, E, and F show little difference in the predicted probabilities across class levels. The remaining clusters, B, D, and G, also show a similar trend in the predicted ratings for class levels. Classes numbered 300–400 have a slightly lower predicted

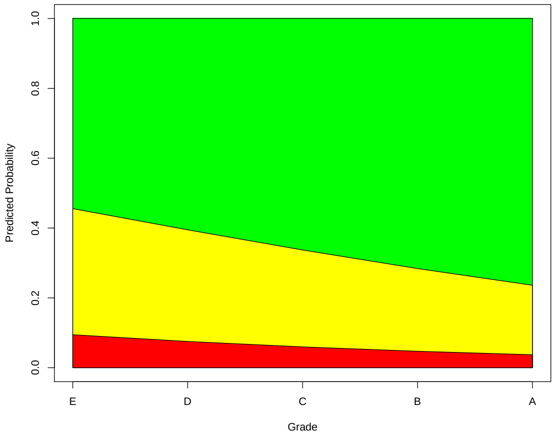
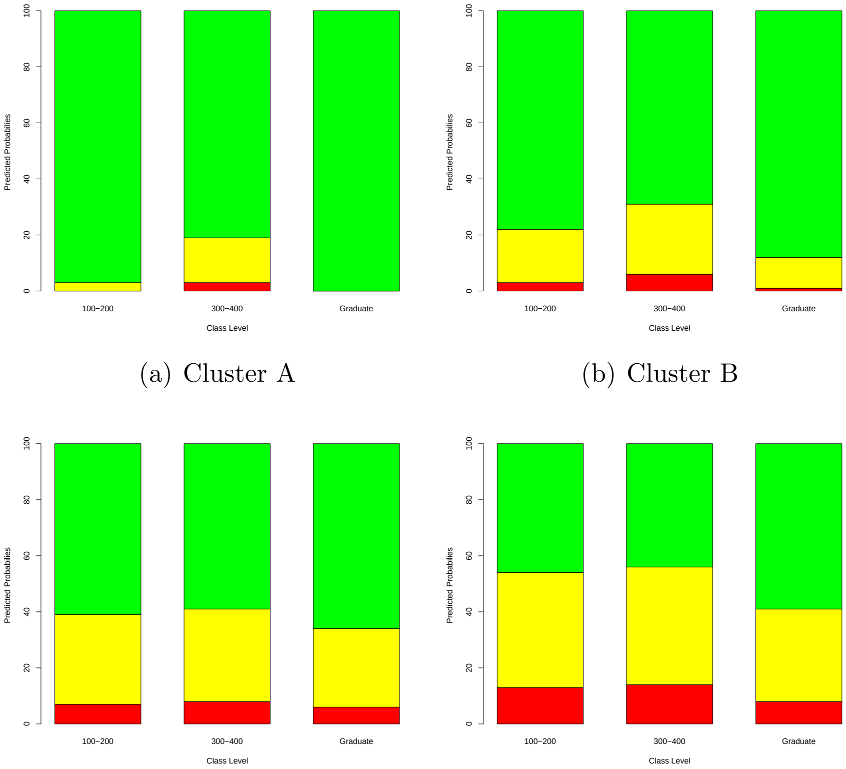


Figure 3.10: Predicted probabilities by student grade for a male lowerclassman student enrolled in a 100–200-level non-GE course with enrollment of 25 taught by a full-time faculty member in college Cluster F during fall semester

probability of a good rating when compared to both 100–200- and graduate-level classes. Also, 100–200-level courses show a slightly lower probability of a good rating than graduate-level classes. In Cluster D, this difference seems significant because teachers in these colleges have the lowest probability of receiving a good rating, and teaching graduate-level classes seems to make the probability increase to the overall average.



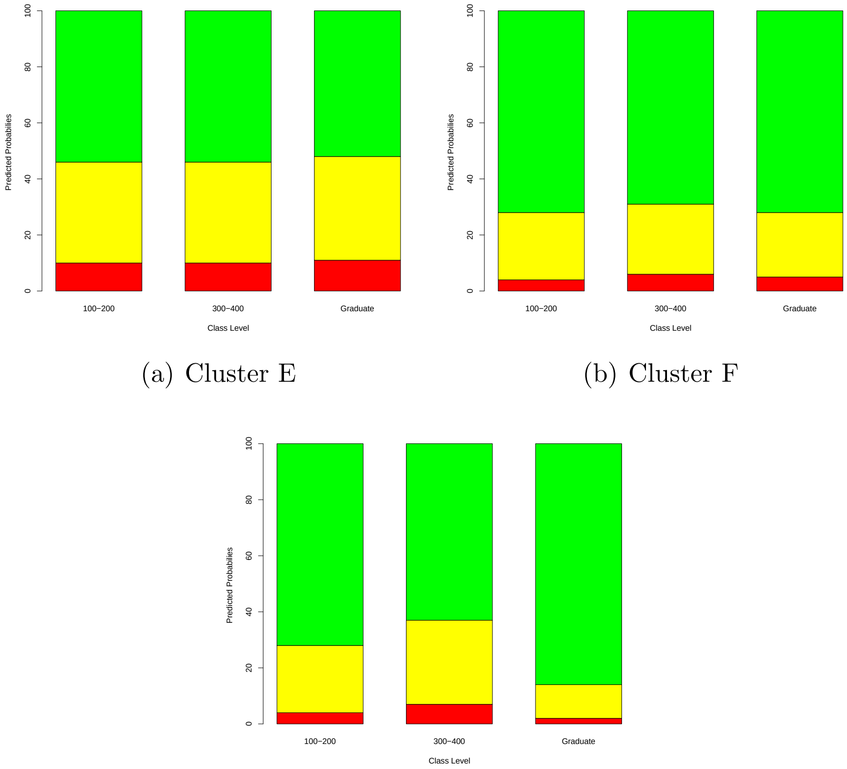
(a) Cluster A

(b) Cluster B

(c) Cluster C

(d) Cluster D

Figure 3.11: Predicted probabilities by class level for a male lowerclassman earning a B in a non-GE course with enrollment of 25 taught by a full-time faculty member in fall semester



(a) Cluster E

(b) Cluster F

(c) Cluster G

Figure 3.12: Predicted probabilities by class level for a male lowerclassman earning a B in a non-GE course with enrollment of 25 taught by a full-time faculty member in fall semester

4. CONCLUSION

This project demonstrates the use of ordinal regression to model student ratings data. Ordinal regression is a statistical tool used when the outcome is categorical with a natural ordering. Traditional tools used in linear regression do not provide the proper framework for this type of analysis because many of the assumptions made in linear regression do not apply. Ordinal regression allows for predicted probabilities of success to be calculated for each level of the response.

The data for this project includes over five years of student ratings data from Brigham Young University and contains multiple ratings by a given student over the course of these years. This data differs from other data collected in teacher evaluation analyses in several significant ways. First, the data contains all the evaluations a given student made over multiple courses and all the evaluations from a given instructor over multiple courses and students. In this analysis, the obvious correlation structure will be ignored because the purpose of the project is to investigate regression on ordinal data. Second, the data contains the actual grade the student received in the course. Most research uses the grade the student reports as their expected grade in the course. Overwhelmingly, the most important factors are student grade and college, where college includes the class level and total enrollment interactions with college.

The student grade factor indicates that as student grade increases, the probability that a teacher will receive a good rating also increases. However, the cause of this association cannot be explored. The effect of class level depends on the college cluster being considered. College 2 exhibits a very different pattern with regards to class level than the other colleges. Teachers of graduate classes will overwhelmingly receive a good rating regardless of other factors. Ratings in 100–200-level classes are slightly lower, and are still lower for 300–400 level-classes. Even considering the drop in ratings for 300–400-level classes, in College 2 the predicted probabilities of a good rating are consistently high. Most of the remaining colleges, like College 12, show a general U-shaped pattern in how class level affects rating. The ratings decrease from 100–200-level courses to 300–400-level courses and then increase again in graduate-level classes. This could be related in part to prior student interest inherent in taking upper-division classes.

The effect of total enrollment is influenced by not only college but GE class status. Because GE classes are typically larger than non-GE classes, the effect of total enrollment is confounded with GE class status. To account for this, a different spline was fit for each GE class status, with GE classes separated by honors and non-honors. Honors classes have a lower enrollment on average than other GE classes, which indicates that a separate spline must be fitted for these classes. In general, as class size increases for GE classes, the expected teacher rating decreases. One exception to this is College 12, which shows a dramatic increase in teacher ratings for larger GE classes.

Other variables such as gender, instructor status, term, and class standing show little practical difference although the groups are statistical different from each other.

BIBLIOGRAPHY

Cashin, W. (1990), “Students Do Rate Different Academic Fields Differently,” *New Directions for Teaching and Learning*, 43, 113–121.

Fernandez, J., Mateo, M. A., and Muniz, J. (1998), “Is there a Relationship between Class Size and Student Ratings of Teacher Quality,” *Educational and Psychological Measurement*, 58, 596–604.

Johnson, V. E. (2003), *Grade Inflation: A Crisis in College Education*, Springer.

McCullagh, P. (1980), “Regression Models for Ordinal Data,” *Journal of the Royal Statistical Society.Series B (Methodological)*, 42, 109–142.

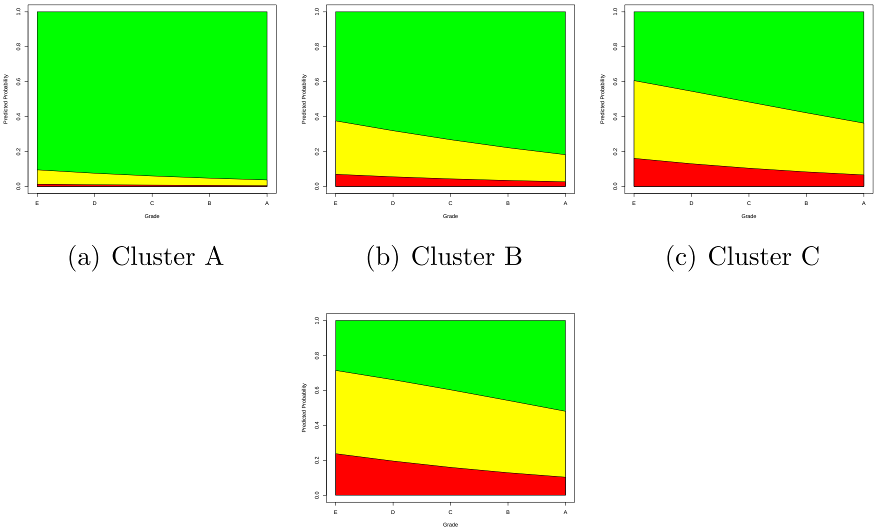
O’Connell, A. A. (2006), *Logistic Regression Models for Ordinal Response Variables*, Sage Publications Inc.

Seldin, P. (ed.) (1999), *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*, Anker Publishing Company, Inc.

— (2006), *Evaluating Faculty Performance: A Practical Guie to Assessing Teaching, Reasearch, and Service*, Anker Publishing Company, Inc.

Tieman, C. R. and Rankin-Ullock, B. (1985), “Student Evaluations of Teachers: An Examination of the Effect of Sex and Field of Study,” *Teaching Sociology*, 12, 177–191.

A. PREDICTED PROBABILITIES BY STUDENT GRADE FOR EACH COLLEGE CLUSTER



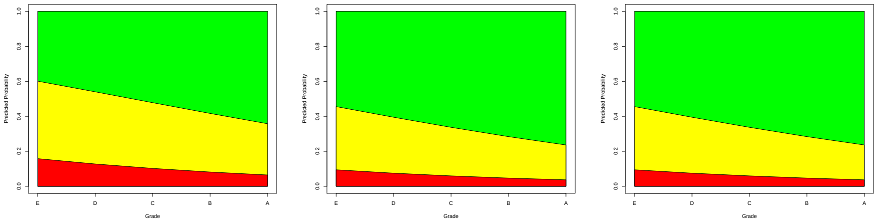
(a) Cluster A

(b) Cluster B

(c) Cluster C

(d) Cluster D

Figure A.1: Predicted probabilities by student grade for Clusters A through D



(a) Cluster E (b) Cluster F (c) Cluster G

Figure A.2: Predicted probabilities by student grade for Clusters E through G