Rethinking Practice: Using Faculty Evaluations to Teach Statistics

Larry H. Ludlow
Boston College

Journal of Statistics Education Volume 10, Number 3 (2002), ww2.amstat.org/publications/jse/v10n3/ludlow.html

Copyright © 2002 by Larry H. Ludlow, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: General linear models; Longitudinal data; Teaching effectiveness.

Abstract

This article explains why and how a course in general linear models was restructured. This restructuring resulted from a need to more fully understand traditional teaching evaluations, coupled with a desire to introduce more meaningful data into the course. This led to the incorporation of a longitudinal dataset of teaching evaluations into the lecture material and assignments. The result was a deeper appreciation of how students perceive my teaching, specifically, and a greater understanding of how statistics courses, in general, can be taught more effectively.

1. Introduction

The Lynch School of Education at Boston College has recently undertaken a study of faculty teaching effectiveness. The study aims to understand better, among other things, what the faculty mean by “effective teaching” and how they try to provide it. One very simple, yet widely used measure by which faculty teaching is judged effective is the course evaluations that are gathered from students at the end of each semester. One reason these evaluation ratings are taken seriously by many of our faculty is the statistical work conducted by graduate students in the Lynch School’s General Linear Models (GLM) course. The purpose of this article is to explain (a) the purpose and structure of the course, (b) the faculty evaluation dataset used for course examples and assignments, and (c) how the students, faculty, and administration benefit from the work produced by the students in the GLM course. The specific results of the student’s analyses are not, however, generalized to other faculty or institutions.

2. Course Background

The basic purpose of our GLM course is to lay a firm foundation for the independent study of multivariate data. This means that considerable time is spent on testing the assumptions of models, analyzing residuals, identifying multicollinearity problems, and representing the multivariate data geometrically through matrix operations. In addition to this general preparation, diverse topics are woven into what is essentially an ordinary least squares regression course. The lectures emphasize a deliberate approach to hierarchical model building—start slowly and carefully, one variable at a time; understand simple relations and models before building more complex ones; understand why results change as models become more complex; and justify each step along the way.

A second purpose of the course is to prepare some of the students to be future teachers of statistics. To that end, the GLM course assignments require quite detailed work—work that cannot be assessed simply through traditional examinations. Similar in principle to the TATTO program at Emory University (Hertzberg, Clark, and Brogan 2000), the GLM course aims to illustrate to the students that teaching and research are complementary intellectual activities and that, regardless of their specific career goals, they will acquire technical communication skills that will carry over to future professional writing endeavors.
As a teacher, I have always found the course an interesting challenge. About four years ago, however, I became dissatisfied with the data used in class partly because the “canned” datasets from textbook publishers were boring, both for me and for students. I wanted data in which I had some professional and personal interest. I also wanted students to work with data that would interest and motivate them enough to perform not only to the best of their ability on each assignment, but also to take an assignment further than required because of their own desire to learn more than what was presented in class. The “canned” data were simply too clean to accomplish either of these objectives.

My reasons for wanting to change the course structure and data sets were consistent with the advice and suggestions of many writers on statistics education. For example, Singer and Willett (1990), Wilson (1992), Sowey (1995), and Smith (1998) suggest that interesting data, challenging intellectual assignments, extensive communication exercises (whether written or oral), and enthusiasm on the part of the instructor are necessary ingredients in creating a memorable statistics experience for students. Furthermore, Snee (1993) and Gardner and Hudson (1999) argue that effective teaching and learning of statistics are fostered when the experience is relevant and meaningful to both the student and the instructor.

A possible solution to what I thought the students needed was a dataset that I had been building for and about myself -- specifically, my effectiveness over time as a university teacher. My motivation for creating this individualized teaching evaluation dataset had come initially from my dissatisfaction with the way course evaluations traditionally are employed at my institution. Specifically, an instructor’s combined “excellent + very good” overall course ratings are compared to ratings summarized across the school as a unit or the entire aggregated University. This comparison of overall ratings tells nothing, however, about what might have led to a given individual’s ratings.

Coupling my desire to revamp the GLM course with my dissatisfaction with the traditional use of course evaluations, I began to use my course evaluations as the dataset for the GLM class. Using these data became, in a very real sense, an experiment in “classroom research” (delMas, Garfield, and Chance 1999; Hollins 1999) or, a piece of “teacher research” wherein the teacher poses and engages in systematic intentional inquiry about his or her own classroom and then opens interpretations and analyses to public scrutiny by colleagues (Cochran-Smith and Lytle 1993). Basically, I wanted to see what would happen when I introduced these data into the course. Would the use of the data influence what and how I taught? What would be the effects on the students?

### 3. Instructional Activities

#### 3.1 Students

Typically the GLM course is taken by doctoral students in the Lynch School of Education as their third or fourth semester of applied statistics. Students in the Educational Research, Measurement, and Evaluation (ERME) program and the Counseling and Developmental Psychology program are required to take the course. More recently the course also has had an influx of students from the schools of nursing, psychology, and social work who take the course as an elective. Generally, motivation to learn and anxiety over statistics are not issues for these students; they want to be there. Although the characteristics of these students might seem to limit the generalizability of the effectiveness of the teaching approach described herein, even they still need meaningful and interesting examples, and creative and challenging assignments to push them to their fullest potential.

#### 3.2 Data

The data set presently consists of 89 separate records of course evaluations for classes taught from the fall semester, 1984, through the fall semester, 2001. These courses range from entry-level “Child Development” to a capstone doctoral “Seminar in Statistical Methods.” Most of the evaluations are for graduate courses in “Interpreting and Evaluating Research,” “Introductory Statistics,” “Intermediate Statistics,” “General Linear Models,” “Multivariate Statistics,” “Design of Experiments,” and “Psychometrics.” In all, the data set represents evaluations by some 2100 students.

The variables included in the dataset are extracted from the end-of-semester evaluation summaries distributed each semester to faculty by the University. There are a variety of different evaluation forms faculty may choose depending on the nature of their course (seminar, laboratory, or general lecture). However, all have a common core of questions which which are Likert scored from “strongly agree” to “strongly disagree.” Instructors receive a report of the percent of students in each class who marked the various options.

For the GLM students, the fundamental analytic question is, “What variables can be used to construct models that will explain the evaluation ratings?” Furthermore, “Are any of those variables under the control of the instructor?” For example, class size may be out of my control but if availability outside of the classroom is a significant predictor of ratings, then I can...
do something about that variable. Thus, the data include variables that address important events in my personal and professional life, and characteristics about my classes that are unique to me. Their importance to the students is that these personal variables offer a rare opportunity for students to gain statistical expertise using data they are familiar with professionally, and have experienced personally to some extent. Thus, they can propose and test models that are justifiably confirmatory, rather than exploratory, in nature. Again, it is important to recognize that the generalizability component of this approach is through the utility of such a data set to facilitate instruction and learning, not the replication of variables unique to me nor the specific statistical results.

The data exist as a simple SPSS course-by-variable flat file (see Appendix 1). The dataset is updated each fall with the evaluation ratings from the previous fall, spring, and summer courses. When an entirely new variable is created I work back through the dataset to add the appropriate observation for each course. The dataset presently consists of the following four relatively distinct categories of variables: (a) administrative/course characteristics (such as year taught and class size), (b) student perceptions (such as percent of time spent on the course and extent to which they acquire factual information), (c) personal variables (such as tenure and marital status at the time the class was taught), and (d) overall evaluation ratings (such as the percent who marked excellent, very good, good, acceptable, or poor). Appendix 2 contains a detailed description of the variables and their scoring options.

Some data are missing for five classes that were taught and repeated in the summers of 1996 and 1997 (one course was taught twice, the other three times). This was because the University was not using the regular evaluation forms at that time. Those missing values serve as an instructional opportunity to discuss, run, and compare a variety of missing data options, that is, what the different options do to replace the missing values and how they affect the estimates. They are left as blanks in the attached data file.

3.3 Procedure

The first time (1998) I used this dataset in my GLM course, I really did not know how well the new data and assignments would work out. I wasn’t sure what analyses would be more or less appropriate for the data. I wasn’t sure which, if any, of the course topics I felt were particularly salient would be well-demonstrated by the data. I wasn’t sure whether the students would find the data and assignments useful or even interesting.

In addition, I wasn’t sure how I would feel about students analyzing my “confidential” teaching record (faculty evaluations are not made public at my university). Specifically, would I feel threatened by their analyses and interpretations? What would I learn about my teaching (and myself) from their analyses? Would I feel tempted to “lose” or “correct” some data? Would I believe the results strongly enough to modify anything in my teaching?

Finally, would students submit self-serving analyses to flatter me for the purpose of seeking a good grade? This important question about potential bias (in their writing and my being affected by it) strikes at the validity of their assignments. Specifically, could the assignments be constructed in such a way to prevent deliberate and overt flattery? This potential problem was openly discussed at the start of each new offering of the course in an attempt to minimize its occurrence. My impression is that this has not been a problem with the work they submitted (either in their attempting flattery or my being swayed by any such thing).

During the first year, the assignments were relatively simple, restricted in scope, and written in a traditional style that led to a single “correct” answer to each question. For example, “regress excellent ratings on class size and interpret the result.” Basically, there was not much that students were asked to do with the data and there was not much that provided an opportunity for individual initiative. Although the exercises were relatively simple, there still were some interesting results from that first year. Students discovered that ratings tended to: a) rise over time, b) decrease as class size increased, and c) not be affected by the amount of time students spent on the course. In addition, there were some multicollinearity problems that provided useful teaching examples, and there were outlier ratings that illustrated the value of residual diagnostic tests (see Appendix 3).

For the second year, I added student-level variables from the evaluations (for example, “extent to which students acquired principles and concepts,” whether the course was for undergraduate or graduate students, and “extent to which class attendance was necessary”). At that point the assignments expanded because more complex models could be tested and, more importantly, more opportunity was provided for students to build, test, and justify their own independent models.

The final assignments from this second year were so creative and interesting, and of potential use to the university, that we gathered them into an edited monograph (Ludlow 2000). The results reported in this monograph include the finding that the strongest predictors of excellent ratings were the extent to which students understood principles and concepts, acquired factual information, and felt I was available outside of class. Each student in the class received a copy of the monograph.
Perhaps the most important thing that I discovered from this final exercise was how seriously the students took an open-ended assignment when the work product was valued both by themselves and the instructor.

The third year continued the evolution of the dataset in a particularly dramatic fashion: personal data were added to the file, including changes in my marital status, whether I was tenured or not at the time the course was taught, and whether the course was taught before or after I took a medical leave. The opportunity for meaningful model building was limited from that point on only by the students’ interest and initiative. The assignments still required a common response to specific technical questions, but the variables used to answer the general research question (“what would ‘best’ explain the ratings?”) were left to the choice of the students. This meant that they not only had to perform and write up the analyses correctly, but they also had to explain and justify the models they built for each assignment, rather than just the final cumulative assignment.

It is important to recognize that the students were not submitting neutral or passive interpretations of the results. Their comments about what they thought about my teaching were based on their years of exposure to university-level teaching in general, and statistics in particular. Their final analyses found that the ratings were, among other things, a function of my marital status, tenure status, and medical leave. Students’ interpretations of why these variables had statistically significant effects upon the ratings were, to say the least, insightful, provocative, and sometimes uncomfortable to read. One of the interesting challenges of the course centers around my internal struggle with how much personal information should be shared with them. This internal struggle, too, is shared with students in open discussions about the analyses and interpretations of their models.

Papers from this course were also collected into an edited monograph (Ludlow 2001). This monograph, too, was distributed to the students and various administrative officers in the university (as examples of how teaching evaluations can be analyzed).

### 3.4 Assessment

The teaching examples and student assignments follow what may be called an “active case study” (Weinberg and Abramowitz 2000). That is, each incoming class of students faces a real, professionally meaningful dataset that continually expands with 5 to 7 additional course evaluations, depending on how many courses I teach from that current semester up to when the course is next taught.

No exams are administered. It is my philosophy that since students are being trained as consultants, future academics, and professional researchers, they must be able to conduct comprehensive, independent, creative, and resourceful analyses. To this end they are not just expected to run some statistical software and then submit a reasonable interpretation of the results. Rather, the write-ups must be comprehensive (including problem statements, hypotheses, tests of assumptions, and explanations of the statistical procedures), equation editors must be used, and graphs and tables must be imported from the statistical software. They are encouraged to submit each assignment in sufficient detail and breadth that they could teach the specific topic themselves at some later date in their careers, drawing on the information included in their assignments.

The sequence of required analyses proceeds in a linear, hierarchical fashion in that the assignments require students to draw upon work performed previously. The first assignment has the modest objective of ensuring that the basic principles of simple ordinary least squares regression are well established. The outcome variable is the “percent excellent ratings” and the predictor is the “course code number.” The second assignment keeps the model simple but changes the predictor and focuses on the detail of residual analysis. The third assignment requires a multiple regression model using variables of their own choosing and focuses on the order of entry of predictors and recognizing collinearity effects.

The fourth assignment consists of matrix algebra exercises that demonstrate understanding of matrix operations, the relationship between a determinant and the inverse of a matrix, the geometric representation of a determinant as a generalized variance, the role that first and second derivatives play in the principle of least squares and maximum likelihood estimation, and the geometric relationships among angles, correlations, and distances between vectors.

The final assignment gives students the opportunity to put it all together. The full assignment is attached (see Appendix 4) since its success as a teaching tool led to continuation of the course and the development of the two monographs. In brief, its purpose is to provide them with the opportunity to build their “best” model however they define “best.”

The first four assignments are graded on a 100 point system (the final one is worth 150 points). In this system everyone starts with the maximum and points are then deducted for equation mistakes, confusing explanations, errors in interpretation or execution of the analysis, and for parts of the assignment that are missing. Such problems are pointed out and
clarifications are provided. In addition, extensive positive feedback is provided for particularly creative analyses, insightful interpretations, and well-stated explanations of technical detail.

4. Outcomes and Interpretations

4.1 Student benefits

The change to a new and continuously evolving dataset and sequence of assignments has led to a remarkable level of interaction with the students, both in and outside of class. The students enjoy the “live” aspect of the data, the honesty and trust displayed by sharing and discussing these data, and the product they walk away with at the end of the semester. This latter point is evident in their pride of authorship in the two monograph volumes and the positive feeling of self-accomplishment in tackling and meeting a difficult and challenging course. Similar to results reported by Gourgey (2000) in her use of “real-life examples,” the GLM students have demonstrated gains in their understanding and application of statistical procedures doing hands-on analysis of meaningful data (as examples, recent dissertations have followed the structure of the final project, and conference papers have been presented based on principles acquired as a direct result of this course).

Anecdotal evidence of teaching effectiveness and educational accomplishment is always somewhat problematic (who really shares the worst comments that students have about you?), but I think their course evaluation statements have been quite supportive of the dataset and their sense of accomplishment. A selection of comments from students over the past three years is attached (see Appendix 5).

4.2 Faculty benefits

The enthusiasm and persistence with which GLM students have tackled the assignments and the skills they have acquired have been enormously satisfying insights for me. I now understand, for example, that students will tackle and succeed on challenging tasks not because that is what is expected of them, but because the task is meaningful and the work product is valued and appreciated. The challenge I now face is how to take this continually evolving case study approach combined with an integrated authentic assessment system and apply it in my lower-level introductory and intermediate statistics courses. Such a challenge has been a prominent, recurring feature of the JSE and always will be a challenge when we are faced with reluctant students who bring into our classes the baggage of mathematics, statistics, and computer anxiety.

The faculty evaluation analyses conducted by the students in the GLM course have been presented to the Lynch School of Education faculty on two separate formal occasions. The response from faculty has been pleasantly surprising. Not only have faculty been pleased that sophisticated analyses are being conducted on an important issue in which they all have a vested interest, but they also are pleased their students are acquiring realistic experiences using real data. Furthermore, a number of faculty have asked for assistance in the creation and analysis of their own evaluation data sets.

4.3 Administrative benefits

The Lynch School of Education recently formed a committee to look at how teaching effectiveness is understood by the faculty, how it is assessed, and how it may be enhanced. The University, too, recently formed a committee to look at how faculty evaluation results currently are used and how the evaluations and the results extracted from them might be improved. Both committees have copies of the student analyses contained in the monographs described in this article. Any educational program that practices self-reflection in teaching practice could undertake similar statistical analyses of teaching evaluations.

5. Conclusions

One of the useful aspects of this dataset is that it changes each year. For one thing, this means there is no opportunity for a student simply to use results from a friend in a previous class. More importantly, I look forward to each new round of analyses because I do not know what the new results will look like. For example, one particular zero-order correlation was not statistically significant in 1998 yet became statistically significant in 2000, even though it retained essentially the same magnitude. This finding led to an excellent discussion around power, sample size, and statistical significance.

The use of my own evaluation ratings has not been without some risk -- not all my evaluations have been stellar, some have been pitiful -- but all of them are included and are all are fair game for analysis and interpretation. I do know, however, that such an open analysis of these data has been beneficial to both the students and me. I have been re-energized as a teacher and the students have developed excellent statistical and communication skills that will serve them well in their future as educators and researchers.
Acknowledgements

I would like to thank Marilyn Cochran-Smith, Camelia Rosca, and Ann Kennedy for their critical reviews of this paper.

Appendix 1

Link to an SPSS version of the dataset: ludlow_jse_data.sav

Link to an Excel version of the dataset: ludlow_jse_data.xls

Appendix 2

The dataset presently consists of the following variables broken into four relatively distinct categories.

Administrative/Course predictor variables:

- sequence order of the class (1 to 89);
- year the course was taught (1984 to 2001);
- semester the course was taught (spring, summer, fall);
- university course code (030 to 960);
- number of times the course was taught (1 to 18);
- class size (3 to 52);
- category of class (e.g., research methods, introductory statistics, psychometrics); and
- level of the class (undergraduate, primarily masters, primarily doctoral).

Student predictor variables:

- extent to which students thought that regular class attendance was necessary for learning the required content (percent of those who marked SA, A, D, SD);
- extent to which students thought they acquired factual information (SA, A, D, SD);
- extent to which students thought they understood principles and concepts (SA, A, D, SD);
- extent to which students thought they acquired academic skills (SA, A, D, SD); and
- percent of time that students spent on my course compared to other courses that semester (0% to 100%).

Instructor predictor variables:

- extent to which students thought I was available for help outside of class (SA, A, D, SD);
- tenure status when I taught the course (pre-post tenure);
- medical leave status when I taught the course (pre-post leave);
- marital status when I taught the course (married, separated/divorced, remarried);
- number of publications I had at the time the course was taught (from 8 to 59); and
- whether or not small group interactions were required as part of the class period.

The Outcome variables:

- my overall evaluation ratings (percent of those who marked excellent, very good, good, acceptable, or poor).

Appendix 3
The ratings follow a general upward trend over time across all classes. Note that at the time that 668, 669, and 960 were taught for the first time (.01), their ratings were much higher than expected based on other classes up to that point. The lowest rated courses in this graph, and those that follow, were for freshman Child Development classes taught the first four years of my career. One of the analysis options students have is to delete those specific courses and run their models just on the methods and statistics courses that I continue to teach. Descriptions of the specific courses are provided in the attached dataset.
There is a clear, unmistakable relationship between the excellent ratings and class size—as the enrollment increases, the excellent ratings tend to drop. The drop is actually at the rate of a decrease of 1% in excellence ratings for each additional student added to the class. Introductory Statistics 468.01 is particularly interesting—it was the first time it was taught by me, it attracted a crowd, and it differed substantially from the way it had been previously taught. A quadratic fit is actually statistically significant here and it makes sense because my classes have increased in size over the past 4-5 years.
The term “spillover-effect” is usually applied to the spillover of pressures from work-to-home. Here, it is used to refer to spillover from home-to-work. Specifically, during the early phase of marriage (M) and work at Boston College the ratings show an upward trend. The ratings, however, start to fall off prior to and continuing into the period of separation and divorce (S/D). During this period they again change direction and begin to recover prior to and continuing into the current remarried stage (RM). This is a nice example of a statistically significant cubic relationship. I also call this the “wife effect” in honor of my wife.
There is an unmistakable positive relationship between the extent to which students strongly agreed (SA) they were taught principles and concepts (PRIN+CONCEPTS) and the percent of excellent ratings they gave. The 668 course is Multivariate Statistics, the 669 course is Psychometrics, the 216 course is undergraduate Research Methods. This finding is the strongest of any of the predictors in the dataset.

Figure 5. An example of variable entry order and collinearity effects.

Step A: the “percent excellent ratings” are regressed on the percent who strongly agreed that principles and concepts were taught.

| Model | R    | R Square | Adjusted R Square | Std. Error of Estimate |
|-------|------|----------|-------------------|------------------------|
| 1     | .804 | .646     | .641              | 14.0420                |

Note the large adjusted R2 with only one predictor.
Now we observe that there is a nearly one-to-one relationship between the percent who strongly agreed that principles and concepts were taught and the percent who gave excellent ratings.

Step B: Now the excellent ratings are regressed on the percent who strongly agreed that factual information was taught.

We again observe a strong, statistically significant relationship that is very similar to the previous one.

Step C: Now the excellent ratings are regressed on both of the predictors and the point is to see the effect on the overall solution and the individual level statistics.

Somewhat surprisingly the overall solution is not much better than the one using just principles and concepts.
Now the real point of the exercise is apparent. The regression estimates are greatly affected by the covariation of the two predictors, that covariation inflated the standard errors, those standard errors then diminished the magnitude of the t-statistics, and factual information as a predictor adds nothing to the model. Note also how the partial and zero-order correlation for the factual information variable are extremely different. This is a simple, easy to understand example of why we don’t want highly correlated predictors.

Figure 6: What is the best model for explaining and predicting the variation in the excellence ratings?

The following discussion serves as one of many possible “final models” available to address this question. Here a multiple regression was run with variables entered that I thought made theoretical sense. The outcome variable was the percent excellent rating. The predictors were the Boston College course code as a proxy for level of difficulty, class enrollment, class enrollment squared (SIZE2), percent who strongly agreed that principles and concepts were taught, and the amount of time students spent relative to other classes.

This table tells us that an extraordinary percent of variation in the excellent ratings can be accounted for by the five variables: 67.8%.

| Model | Unstandardized Coefficients | Standardized Coefficients | Correlations |
|-------|-----------------------------|---------------------------|--------------|
| 1     | (Constant)                  | -18.557                   | -3.556       |
|       | Understood principles       | .884                      | .519         |
|       | Acquired factual information| 6.596E-02                 | .349         |

a. Dependent Variable: PERCENT EXCELLENT RATING
This table tells us that the variables were entered into the model as blocks representing separate aspects of the classroom environment, each block accounted for statistically significant variation in the ratings, and each variable was statistically significant. Extensive time would now be spent on interpreting the coefficients.

The equation to predict excellent ratings for any future class is:

\[
\text{Predicted excellent rating} = 0.024 \times \text{Course Code} - 1.37 \times \text{class size} + 0.028 \times \text{class size}^2 + 0.83 \times \text{princ+concepts} - 0.11 \times \text{time}.
\]

This means

- for a 100 unit change in course code there is an expected change of +2 excellence points,
- for each additional student there is an expected change of –1.4 excellent rating points,
- when class size reaches a certain upper limit there is, however, an expected positive change in excellence points,
- for each additional increase of 1% in the percent of students who SA that they received principles+concepts there is nearly a 1 unit increase in the percent excellence points, and
- for each additional 11% more time spent on the course relative to others there is a drop of 1% in the excellent ratings.

To see how this works, consider the following situation:

if the next offering of 669, had 10 students, and they had an 80% Strongly Agree response that they received principles and concepts, and they felt the time they spent relative to others was 90% more on 669, then the predicted excellence rating would be

\[
0.024 \times 669 - 1.37(10) + 0.028 \times 100 + 0.83 \times 80 - 0.11 \times 90 = 61.7% 
\]

which compares quite favorably to the past mean of all 669 classes (60.67%).

### Appendix 4

The final assignment gives them the opportunity to put it all together. The full assignment is presented since its success as a teaching tool led to the continuation of the course and the development of two monographs on the longitudinal analysis of
faculty evaluation data.

Given the course evaluation data you have been analyzing this semester, your problem is to formulate and test the model that is your choice as the “best” for understanding the evaluation ratings. “Best” is defined by the purpose you choose for building the model. This exercise is primarily concerned with your thinking about a model, explaining the model, and justifying the steps you took to arrive at a final solution. Everyone will likely have a different final model. I particularly want to know how you arrived at your final choice.

I would like you to think of this final as a cumulative project. This means you may cut-and-paste from your previous assignments. For example, the description of the data set, the OLS assumptions, and general diagnostic analyses should be included. But, I’m not interested in your repeating and including everything you have already done.

In general, you may take either a predictive/exploratory approach or a theoretical/confirmatory approach. Furthermore, you may want to think of the data set as consisting of three general sets of variables: instructor variables, student variables, and institutional variables. Whichever way you choose and however you think of the data set, explain the rationale for the model you propose. This explanation should include

- the variables you selected (why did they interest you),
- their order of entry (if it is a confirmatory model, then why did you put them in a specific order),
- the statistical variable selection procedure you used (if it is a predictive model, then why did you chose either forward, backward, or stepwise),
- you may enter variables as blocks and test their block effects, not necessarily their individual effects (this would hold for both predictive and confirmatory models),
- your reasons for using and ultimately discarding different variables and approaches, and
- your final conclusion about these data That is, for this set of data what variables seem to have influenced the student’s ratings?

You must include the following:

Create a three-level indicator variable (e.g. marital status, level of degree, type of course, semester taught, day of the week taught, etc).

- This variable may be dummy, effect, or orthogonal coded--the choice is up to you, as is the way the codes are assigned. You will have to explain why you chose the approach you did.
- Explain the results and the coefficients:
  - What group comparisons were formed?
  - What is the overall effect?
  - What do the a and b’s mean?

You have the following options available to you as you think about your model:

- you may use the “excell” ratings or any other combination of other variables to arrive at a “rating” variable as your outcome (e.g. “excell” + “vgood”, or you may look at “poor”),
- you may use any of the predictors,
- you may use any interactions between predictors (e.g. TIME * SIZE),
- you may use any transformations (e.g. log, square, square root), and
- you may use any regression diagnostics and plots to aid your choice for your “best” model.

You do want to address OLS residual assumptions and comment on the presence of influential points. Whatever choices you make you do need to explain them to me. When you re-read your paper, ask yourself “Is he going to ask why I did this?” If the answer is “yes”, then make sure you have an answer.

I think it would be useful if you write up your project following this article style: purpose/introduction, method (sample, instrument, procedure), results (assumptions tested, analytic procedures and their rationale, statistical results), discussion (interpretation of the statistical results, practical benefit of the results--”so what and who cares”). In terms of format, you might think of including a table of contents containing the various sections addressed by your paper.

Finally, your class presentation the last night of class should be planned for no more than 10 minutes a person--practice your timing. The presentations should be short and they should emphasize the relevant point(s) you most want to make. Think about what you would like the listener to walk away thinking about. Think about a relevant handout and overhead to
emphasize your main points. Think of this as a conference presentation.“

Appendix 5

I think the following course evaluation statements are noteworthy and representative of students from the past three years. They fall into two categories.

Teaching Effectiveness:

- “Use of the data was great. It was original, unique and helped me conceptualize the material.”
- “SPSS output and actual examples were very helpful. Using a “real” data set reinforced the practical aspects of the course content”.
- “The use of the same data set--a data set students have some familiarity with was helpful.”
- “Using a ‘real’ data set for in-class examples and having relevant assignments demonstrated concepts/problems in the data.”

Professional Growth:

- “After some initial fear and dread I really enjoyed this class. I feel I now have the basic skills and understanding with which to progress.”
- “Challenging, made me want to do my best. I am not the same person who entered this class in September.”
- “The output examples were invaluable in sealing my understanding of the analyses. I was inspired to be a better student.”

References

Cochran-Smith, M., and Lytle, S. L. (1993) Inside/Outside: Teacher Research and Knowledge, New York: Teachers College Press.

delMas, R. C., Garfield, J., and Chance, B. L. (1999), “A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students’ Statistical Reasoning,” Journal of Statistics Education [Online], 7(3). (ww2.amstat.org/publications/jse/secure/v7n3/delmas.cfm)

Gardner, P. L., and Hudson, I. (1999), “University Students’ Ability to Apply Statistical Procedures,” Journal of Statistics Education [Online], 7(1). (ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm)

Gourgey, A. F. (2000), “A Classroom Simulation Based on Political Polling to Help Students Understand Sampling Distributions,” Journal of Statistics Education [Online], 8(3). (ww2.amstat.org/publications/jse/secure/v8n3/gourgey.cfm)

Hertzberg, V. S., Clark, W. S., and Brogan, D. J. (2000), “Developing Pedagogical and Communications Skills in Graduate Students: The Emory University Biostatistics TATTO Program,” Journal of Statistics Education [Online], 8(3). (ww2.amstat.org/publications/jse/secure/v8n3/hertzberg.cfm)

Hollins, E. R. (1999), “Becoming a Reflective Practitioner,” in Pathways to Success in School: Culturally Responsive Teaching, eds. E.R. Hollins and E.I. Oliver, Mahwah, NJ: Lawrence Erlbaum Associates.

Ludlow, L. H., Alvarez-Salvet, R. and Rosca, C. (2000), A Longitudinal Analysis of One Professor’s Course Evaluations (Vol. I), Chestnut Hill, MA: Boston College Press.

Ludlow, L. H. and Rosca, C. (2001), A Longitudinal Analysis of One Professor’s Course Evaluations (Vol. II), Chestnut Hill, MA: Boston College Press.

Singer, J. D., and Willett, J. B. (1990), “Improving the Teaching of Applied Statistics: Putting the Data Back Into Data Analysis,” The American Statistician, 44, 223-230.
Smith, G. (1998), “Learning Statistics By Doing Statistics,” *Journal of Statistics Education* [Online], 6(3). (ww2.amstat.org/publications/jse/v6n3/smith.html)

Snee, R. D. (1993), “What’s Missing in Statistical Education?,” *The American Statistician*, 47(2), 149-154.

Sowey, E. R. (1995), “Teaching Statistics: Making It Memorable,” *Journal of Statistics Education* [Online], 3(2). (ww2.amstat.org/publications/jse/v3n2/sowey.html)

Weinberg, S. L., and Abramowitz, S. K. (2000), “Making General Principles Come Alive in the Classroom Using an Active Case Studies Approach,” *Journal of Statistics Education* [Online], 8(2). (ww2.amstat.org/publications/jse/secure/v8n2/weinberg.cfm)

Wilson, W. J. (1992), “Statistical Consulting Is Scholarship,” *The American Statistician*, 46, 295-298.

---

Larry H. Ludlow  
Lynch School of Education  
Boston College  
Chestnut Hill, MA 02467-3813  
USA  
ludlow@bc.edu