Some Favourite Data Sets : Using the Computer on Real Data in Class

Robin H Lock - Canton, New York, USA

1. Introduction

Increased availability of computers and easy-to-use statistical software has greatly enhanced the ability to efficiently use large sets of real data for motivation and illustration of statistical concepts in applied courses. Many statistics educators, for example, Moore and Roberts (1989), have asserted the merits of teaching "data-driven" courses where the questions which naturally arise from an interesting set of data stimulate the discussion and development of statistical techniques. Several examples of such data and their use in a variety of courses are given below.

2. The first day survey

Following an idea of Loyler (1987), students in an introductory applied statistics course are given a survey on the first day of class. Items include GPA, SAT scores, height, weight, number of siblings, pulse rate, smoking habits, fraternity or sorority membership, exercise, and TV watching habits. Data are pooled from several sections of the course into a common file to give responses from about a hundred students. The data are revisited throughout the semester as needed to provide illustrative examples.

One of the first uses of these data is as part of an introductory demonstration of MINITAB. Students have already seen some descriptive statistics and graphical methods worked out by hand on smaller data sets. To begin to appreciate the versatility and speed of doing statistics on a computer, we spend a class period learning MINITAB basics and exploring facets of the class survey data. A projection system is used for display from a personal computer which is connected to our campus network giving access to the software and data.

Although the instructor is "driving" the computer, the directions of inquiry are frequently guided by student suggestions and requests. Graphical comparisons, such as adjacent boxplots or dotplots, are particularly popular to compare items such as grade point average (GPA) between males and females, different class years, fraternities and sororities, or smokers and non-smokers. Another favourite is a scatterplot of GPA versus scores on the mathematics portion of the Scholastic Aptitude Test (SAT) which generally shows almost no relation between the two variables (see Figure 1). The session invariably ends with many "what about ...?" questions still pending and a very high interest level in the data.

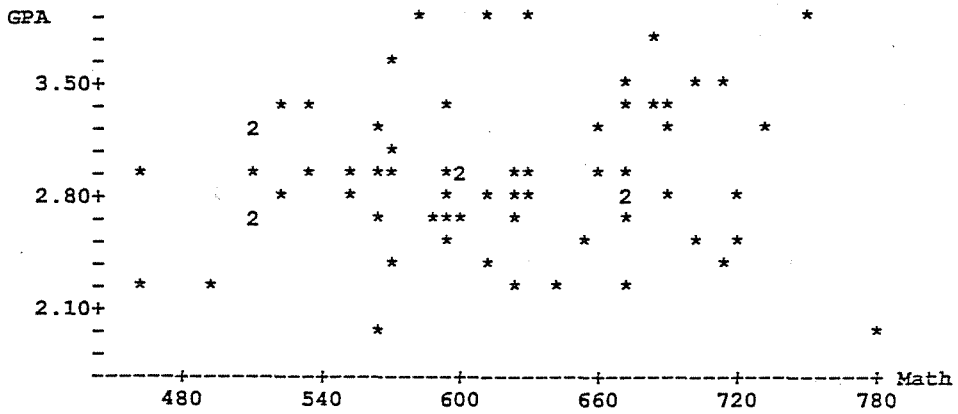

FIGURE 1
Grade point average vs Math SAT scores

Fortunately, we come back to the data many times during the semester when questions can be posed more precisely and more formal techniques are available for addressing them. For example, the survey asks for students to generate random three digit numbers. A chi-square test usually shows their numbers are not so random, although the last four digits in their social security numbers generally are. Pulse rates often show interesting relationships with gender, height, weight, or smoking and we get a nice paired data example by collecting pulse rates a second time during a quiz or examination.

Student interest in such examples is naturally heightened when they are the source for the data and the questions relate directly to themselves. Part of the appeal is also the element of unpredictability - particularly during a "live" demonstration when they are aware that the instructor isn't sure what's coming next.

3. Baseball salaries

These data were used for the Statistical Graphics Exposition at the American Statistical Association's summer meetings in 1988. They consist of various statistics for most major league baseball players in the year 1986. Career information and team data are also available. One other key piece of information is each player's salary at the

start of the 1987 season. The question posed at the ASA session was "Are baseball players paid what they are worth?". I have used these data on two occasions as a take-home final examination for a second level course in applied statistics. The basic question for the examination is the same as that given to ASA participants, along with instructions that students are encouraged to display the full range of what they have learned during the course as part of their analyses.

Students are initially uncomfortable with this sort of open-ended assignment, particularly when the data set contains over 500 individuals (322 hitters and 208 pitchers) and more than 20 variables for each. As they begin to explore the data they are generally able to focus on a set of questions which are relevant to their interests - often creating new variables (such as dividing career statistics by number of years played or working with log(salary)). Some concentrate on hitters or pitchers alone, others look for differences between the two leagues or among positions. Most use regression and ANOVA techniques which reflect the course syllabus.

The size and variety of these data play an important role in enabling students to pursue diverse methods for attacking the problem - even if they aren't baseball fans. They get a feel for how a simply stated question combined with a rich data set can lead to a labyrinth of new questions and areas to explore. Fostering this ability to develop and test their own conjectures is a significant, and often neglected, part of learning to do applied statistics.

4. Wisconsin restaurants

These data are provided as part of the MINITAB Handbook (1985). They are based on a survey of restaurants undertaken by the University of Wisconsin's Small Business Development Center. 279 eating establishments are included, with data such as the number of employees, yearly sales, type of restaurant, number of seats and expenditures for equipment, food, wages, and advertising (13 variables in all). They were used in a second level course in applied statistics which focused mainly on applied regression. We chose to consider models for predicting the sales of a restaurant based on the other variables. Each student was given a randomly chosen subset of 125 restaurants from the entire data file. The main intent was to compare models from the different groups to see how estimates of model parameters and relative importance of the predictors might change from sample to sample.

Students were originally very puzzled at the results. The best models for some students might have an R^2 value around 40%, while other students could "explain" 80% of the variability in sales for their restaurants with almost any model they tried! Also, the variables which were "important" for predicting one set of restaurants were often completely different from the best predictors for another set, even though there was considerable overlap between the two sets of restaurants.

Based on their past experience in the course, the students suspected two reasons for these phenomena. The data contains a fairly large number of missing values for certain variables. Thus the number of restaurants which were actually used in any regression calculations might vary considerably depending on which variables were included. Also, as one would expect, there was a high degree of collinearity among many of the potential predictors which often depended mostly on the size of the

restaurant. However, these factors alone did not sufficiently explain the huge fluctuations in models between different subsets of restaurants.

After digging around in the data, the class eventually found the key to the mystery. It was restaurant No 219. If No 219 was in your data, you were blessed with high R^2's for almost any model. If not, the results were invariably less spectacular. What's so special about No 219? It had over $8 million in sales, more than double any other restaurant and well above the median of $0.2 million. With 250 full-time employees, 550 seats, and an appraised value of $12 million it is, by almost any comparison with the other Wisconsin restaurants, a clear outlier.

Restaurants #100-#224
107 cases used, 18 cases contain missing values

Predictor	Coef	Stdev	t-ratio	P
Constant	12.48	59.49	0.21	0.834
SEATS	1.8415	0.9044	2.04	0.044
FT.EMPL	26.410	2.372	11.13	0.000

s = 383.4 R-sq = 80.3% R-sq(adj) = 79.9%

Restaurants #100-#224 (except #219)
106 cases used, 18 cases contain missing values

Predictor	Coef	Stdev	t-ratio	P
Constant	73.12	59.54	1.23	0.222
SEATS	2.0912	0.8658	2.42	0.017
FT.EMPL	13.484	4.459	3.02	0.003

s = 365.7 R-sq = 21.4% R-sq(adj) = 19.9%

FIGURE 2
Typical effect of restaurant No 219

This led very naturally into a discussion of influence and leverage points. Having already experienced the effects of an extreme value, the class was eager to discover more formal techniques for quantifying the impact of individual data values on a regression fit. They were then able to identify several other potential outliers among the Wisconsin restaurants and explore methods for handling them in subsequent regression models.

5. Northup challenge

A text by H Roberts (1988) contains quarterly data over a 10 year period from a capital equipment manufacturer (identified as "Northup" in the filename). Variables include quarterly sales, inventory, total industry sales, plant expenditures, and national GNP. These data were used in a course on Time Series Analysis as a bonus assignment

after covering topics in multivariate regression and Box-Jenkins methods. The data were made available on a computer network and students, working individually or in pairs, were asked to develop a model and forecast Northup sales for the next two quarters.

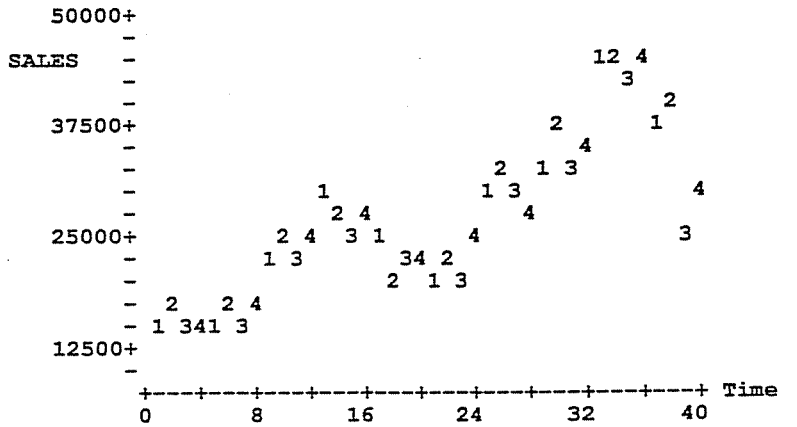

FIGURE 3
Quarterly Northup sales (10 years)

After several days, each "team" presented their model to the rest of the class with the students serving as judges for the competition. In addition, the instructor presented a "middle-of-the-road" base model to which each of the student models could be compared. Students received points for presenting a model and additional bonus points for "beating" the base model (which all participants did). Finally, the class voted on the "best" model from among those submitted for further extra credit.

One of the surprising benefits of this exercise was the lively discussion which took place over the relative merits of the competing models. For example, the conflict between obtaining a high R^2 and the principle of parsimony was hotly debated. Students had to develop their own means for comparing models such as a lagged regression on related series with a univariate ARIMA model. They got some sense of the variety of models which might be appropriate for the same data and that there is not always a "correct" answer to an applied statistics problem which can conveniently be found in the back of the text.

There's nothing unique about this particular company or data set - similar data from a local company with which the students were familiar would probably be even more effective. Also, in retrospect, it would have been interesting to have available the true "future" values so we could see how well the forecasts actually worked. However, the element of competition, the experience with presenting statistical results to the rest of the class and seeing the variety of methods which could be applied to the same data made this a distinctive and memorable activity for most of the students.

6. Conclusion

While real data sets have always been an important part of courses in applied

statistics, the computer has substantially enriched our ability to use data in class. A live demonstration that can react to student input is likely to be much more effective than prepared handouts or textbook examples. The computer allows students to experiment, try lots of models, and work with many variables with comparatively little computational effort. Exploring a substantial set of data via computer can provide students with important cues for organising statistical concepts via associations with interesting features of the data. A student who has struggled with the Wisconsin restaurant data will very likely remember restaurant No 219 when he or she again encounters the concept of leverage.

Electronic media also facilitate the distribution of data. Textbooks are now accompanied by data diskettes, researchers swap data instantly via electronic mail, and many government generated databases are readily available on a floppy disk. Campus computer networks allow students easy access to common data sets and software. Electronic "classrooms" can be equipped with projection facilities and desktop work-stations.

The purpose of this paper has been to relate some examples of real data which have seemed to work well in class. Certainly, many ICOTS participants have similar experiences and favorite data sets. I would like to encourage us to take advantage of computer technology, share interesting data with each other, and help pass along to our students the enthusiasm for messing around with data.

References

Loyler, M W (1987) Using classroom data to illustrate statistical concepts. *Proceedings of the Second (Oneonta) Conference on Teaching Statistics*, State University of New York at Oneonta, 43-73.

Moore, T L and Roberts, R A (1989) Statistics at liberal arts colleges. *The American Statistician* **40**, 80-85.

Roberts H (1988) *Data Analysis for Managers (with MINITAB).* The Scientific Press, Redwood City, California, p320.

Ryan, B F, Joiner, B L and Ryan, T A (1985) *MINITAB Handbook* (2nd ed). Duxbury Press, Boston, 321-328.

1988 Proceedings of the Section on Statistical Graphics, American Statistical Association, contains several papers analysing the baseball data, 76-137.