Data Management, Exploratory Data Analysis, and Regression Analysis with 1969-2000 Major League Baseball Attendance

James J. Cochran
Louisiana Tech University

Journal of Statistics Education Volume 10, Number 2 (2002),
ww2.amstat.org/publications/jse/v10n2/datasets.cochran.html

Key Words: Classroom data; Exploratory data analysis; Regression analysis.

Abstract

The 1969-2000 Major League Baseball Attendance dataset contains Runs Scored, Runs Allowed, Wins, Losses, Number of Games Behind the Division Leader, and Home Game Attendance of each major league franchise for the 1969 through 2000 seasons. Also included for each franchise are its location, league affiliation (National or American), and division affiliation (East, Central, or West). These data have been used in a project-based modeling course to instruct students in basic data management, the use of exploratory data analysis to "clean" data, and construction of regression models. The dataset, which is both cross-sectional and time-series, is of a manageable size and easily understood. Furthermore, it provides a useful, interesting, and realistic classroom example for discussing many important statistical concepts.

1. Introduction

This paper deals with the use of the 1969-2000 Major League Baseball Attendance dataset in a capstone data analysis course designed for undergraduate and Master's level quantitative analysis students. Although most of the enrolled students were working toward business degrees, each had completed a full sequence of mathematical calculus (not business calculus) courses, a course in linear algebra, and two courses in computer programming using a high-level language. Additionally, the undergraduates had completed a three-course sequence in mathematical statistics. The graduate students had completed a two-course mathematical statistics sequence with similar coverage. Students at both levels had also completed a course in statistical methods.

To provide the students with a comprehensive data analysis experience, the course was designed around
an analytic project. At the onset of the course, the class was divided into groups of three or four students. Each student group was given a diskette containing a unique part of a dataset that the instructor had intentionally corrupted by introducing various anomalies including missing values, repeated observations, and misplaced decimals. The first eight weeks of the ten-week course were devoted to discussions on exploratory data analysis, data management, advanced topics in statistical modeling, and the role of a statistical consultant. During the last two weeks of the quarter, the instructor met individually with the student groups to help refine their analyses and polish their presentations. During their final-examination period, each group submitted a written report and presented their results to an audience composed of classmates, faculty, and Ph.D. students.

2. The Dataset

These data appear in most encyclopedic sources on baseball, including The Baseball Encyclopedia (1993) and Total Baseball (2001). The data include Runs Scored, Runs Allowed, Wins, Losses, Number of Games Behind the Division Leader, and Home Game Attendance for each major league franchise for each season from 1969 through 2000. Also included for each franchise are its location, league affiliation (National or American), and division affiliation (East, Central, or West). Many important events in major league baseball occurred over the period included in the data. A list of some of these events includes:

- Two franchise shifts:
  - After the 1969 season, the Seattle Pilots moved to Wisconsin and became the Milwaukee Brewers.
  - After the 1971 season, the Washington Senators moved to Arlington, Texas and became the Texas Rangers. When the Washington Senators moved to Texas, they shifted from the East Division to the West Division in the American League. The Milwaukee Brewers concurrently shifted from the West Division to the East Division of the American League.

- Four expansions:
  - In 1969 (the initial season included in the dataset), the National League added the Montreal Expos and the San Diego Padres, while the American League added the Seattle Pilots and the Kansas City Royals. Each league was divided into two divisions (East and West) and the League Championship Series (between the Division Champions for each league) was implemented.
  - In 1977, the American League added the Seattle Mariners (who replaced the departed Seattle Pilots) to its West Division and the Toronto Blue Jays to its East Division. This expansion also coincided with a fundamental change in the structure of the American League’s regular season schedule. At this point, the American League moved from what was referred to as an unbalanced schedule (in which teams play more games against their division rivals) to a balanced schedule. Thus, this change in the American league schedule is confounded with the league’s second expansion.
  - The National League followed suit in 1993 by adding the Colorado Rockies to their West Division, adding the Florida Marlins to their East Division, and adopting a balanced schedule.
In 1998, the American League added the Tampa Bay Devil Rays while the National League added the Arizona Diamondbacks. The Milwaukee Brewers were also transferred from the American League to the National League, increasing the National League to sixteen teams and maintaining the American League at fourteen teams. Each league was divided into three divisions (East, Central, and West), and the playoffs were expanded to include wild card teams (the division runner-up from each league with the best record).

- The American League adopted the "designated hitter" rule after the 1972 season. This rule - which allows a hitter who does not occupy another position in the lineup to bat for the (presumably weaker hitting) pitcher - was implemented to generate additional fan interest by creating more offense.

- The National League had used "turnstile counts" as its measure of attendance through the 1992 season, after which they began using tickets sold. The American League counted actual tickets sold throughout the period covered by the data. Thus, all American League attendance figures and National League attendance figures after 1992 include all tickets sold (including those not used).

- Seven work stoppages:
  - Thirteen days and eighty-six regular season games were lost when the players struck from April 1 through 13, 1972. These games were not rescheduled, and the remainder of the Major League Baseball season proceeded as scheduled.
  - Portions of spring training were lost to owner lockouts from February 8 through 25, 1973 and March 1 through 17, 1976.
  - Eight days were lost when the players struck from April 1 through 8, 1980. The cancelled games were rescheduled and played as part of the regular season.
  - Fifty days and seven hundred and twelve regular season games were lost when the players struck from June 12 through July 31, 1981, effectively eliminating approximately one-third of the 1981 Major League season. The team with the most pre-strike wins in each division was declared the “first-half” division winner, while the team with the most post-strike wins in each division was declared the “second-half” division winner. The “first-half” and “second-half” winners in each division then staged a three game “division playoff” to determine the ultimate division winners. The result of this convoluted plan was to award division championships to teams that did not win the most games within their division (in fact, the team with the most overall wins in the major leagues did not qualify for their division playoff).
  - Two days were lost when the players struck from August 6 through 7, 1985. The canceled games were rescheduled and played as part of the regular season.
  - A portion of spring training was again lost to an owner lockout from February 15 through March 18, 1990.
  - Finally, on August 12, 1994 the players initiated a strike that lasted two hundred and thirty-two days and resulted in the cancellation of nine hundred and twenty regular season games (as well as the entire 1994 post-season). Because this strike did not officially end until March 31, 1995, the 1995 spring training was drastically reduced and the 1995 regular
season schedule altered.

Each of these events may have had an impact on the total attendance of a franchise by reducing the number of games they played and antagonizing fans. These issues must be dealt with when modeling attendance. Additionally, several new stadia opened during this period:

- Riverfront Stadium (renamed Cynergy Field in 1998) in Cincinnati and Three Rivers Stadium in Pittsburgh during the 1970 season.
- Veteran’s Stadium in Philadelphia on the first day of the 1971 season.
- Royals Stadium (renamed Kauffman Stadium in 1993) in Kansas City on the first day of the 1973 season.
- Olympic Stadium in Montreal on the first day of the 1977 season.
- The Hubert H. Humphrey Metrodome in Minnesota on the first day of the 1982 season.
- The Skydome in Toronto during the 1989 season.
- Comiskey Park II in Chicago (American league) on the first day of the 1991 season.
- Oriole Park at Camden Yards in Baltimore on the first day of the 1992 season.
- The Ballpark at Arlington in Texas on the first day of the 1994 season.
- Jacobs Field in Cleveland on the first day of the 1994 season.
- Coors Field in Denver on the first day of the 1995 season.
- Turner Field in Atlanta on the first day of the 1997 season.
- Safeco Field in Seattle during the 1999 season.
- Pacific Bell Park In San Francisco, Enron Field in Houston, and Comerica Park in Detroit on the first day of the 2000 season.
- PNC Park in Pittsburgh and Miller Park in Milwaukee on the first day of the 2001 season.

Each of these new stadia could also have had an impact on both on-field performance measures and attendance. The above list does not include stadia whose opening or initial use for Major League baseball coincides with the inaugural season of a new or relocated franchise. Also note the use of artificial turf increased over the early years covered by these data and has more recently decreased.

3. The Course

During the first class meeting, each of the student groups was provided with a unique portion of the 1969-2000 Major League Baseball Attendance Dataset. The data were limited to the 1969 through 1992 seasons to somewhat simplify the modeling process (reducing the number of expansions, work
stoppages, and changes in scheduling format the students would have to consider). The subsets of the data that were distributed to the student groups did not overlap and did combine to form the complete dataset to be used for the course. None of these subsets contained data for all years to be utilized, nor did any subset contain data for every team.

Students were also given a description of the contents of the full dataset. The class was then instructed to use these data (and any other data that they cared to collect) to determine if baseball franchises in large markets had significantly higher attendance relative to their counterparts in relatively small markets. The definitions of large and small market were left to the students’ discretion.

As the class progressed through a discussion of exploratory data analysis, each group was expected to use the methods discussed in class to clean its unique portion of the data. The groups used various univariate measures (such as the maximum, minimum, range, and mean) to detect data that appeared erroneous. The students then consulted *The Baseball Encyclopedia* to ascertain the validity of these data and make changes as necessary.

As part of the ensuing coverage of data management issues, the groups exchanged their cleaned datasets and attempted to aggregate their accumulated data into a single, cohesive dataset. The student groups had to cooperatively interleave and concatenate their individual datasets to create a single comprehensive dataset.

Once the student groups successfully aggregated the data, the class began a discussion of statistical modeling. At this point, the course could move in whatever direction the instructor deemed appropriate. Depending on the class project chosen, the instructor could opt to cover any of a wide variety of topics in applied statistics. The course was designed to allow for reasonable coverage of topics such as regression analysis, design and analysis of experiments, time-series analysis, nonparametrics, or categorical data analysis from an applied perspective. Thus, the course could be tailored to address the specific interests of a particular class of seniors and graduate students. The first time the course was offered, the 1969-2000 Major League Baseball Attendance Dataset was utilized to facilitate coverage of statistical modeling with regression analysis. Topics covered in the course included model-building strategies, dummy variables, interactions, assumptions, and diagnostics. These topics were covered in much greater detail than had been provided in the students’ previous course work. Edward W. Frees’ *Data Analysis Using Regression Models (1996)* was used for the course text, and the students used SAS software to perform all data management tasks and analyses.

As each topic (including model building and selection, and inference and assumptions) was discussed, the student groups were encouraged to apply what they were learning to their model(s). In using the 1969-2000 Major League Baseball Attendance dataset, the students encountered minor problems with multicollinearity, moderate problems with heteroscedasticity, and substantial problems with autocorrelation (as would be expected with any large dataset that is both cross-sectional and time series in nature). Eventually each student group independently realized that the heteroscedasticity and autocorrelation were both easily corrected through inclusion of time as an independent variable. Student groups also had to develop their own definition of what was meant by small market (each group defined this characteristic somewhat differently). Finally, the student groups wrestled with the implications of franchise-related differences in the residuals (that is, models consistently underpredicted the Los Angeles Dodgers’ attendance but almost always overpredicted attendance for the Oakland A’s).

Eventually each student group developed a unique model with its own merits and shortcomings. While each model was unique, each group reached a similar conclusion regarding the large market/small market attendance issue - the models developed by the student groups consistently estimated that large market franchises outdrew small market franchises by a statistically significant 450,000 to 800,000 fans...
4. Conclusions

Having a single, well-defined course objective encouraged the student groups to consider practical uses for course material as the classroom discussion progressed. This resulted in a much greater level of classroom participation and student involvement. The students generally completed their reading assignments before class meetings and came to class with prepared questions. Furthermore, each student group collected additional demographic and socio-economic data on the various markets and used these data to classify markets as large or small. The student groups also used some of these data as independent variables in their final regression analyses.

The models produced by the student groups were quite diverse, but each was well conceived and thorough. Interestingly, each group reported a significant difference between large and small markets in attendance, but the magnitude of the estimated difference varied greatly. Their oral presentations were well received, and many of the faculty complimented the student groups on the quality of their analyses and presentations. The class has since been offered again (emphasizing generalized linear models) with similar results.

5. Getting The Data

The file MLBattend.dat.txt contains the uncorrupted data in the fixed column format provided in the Appendix. The file MLBattend.txt is a documentation file containing a brief description of the dataset.

Appendix - Key To Variables in MLBattend.dat.txt

| Columns | Variable                                      |
|---------|-----------------------------------------------|
| 1 - 4   | Major League Baseball franchise               |
| 9 - 10  | League affiliation (National or American)     |
| 16 - 19 | Division affiliation (East, Central, or West) |
| 25 - 26 | Season                                        |
| 32 - 38 | Home game attendance                         |
| 43 - 46 | Runs scored                                   |
| 51 - 54 | Runs allowed                                 |
| 59 - 61 | Wins                                          |
66 - 68    Losses
74 - 77    Number of games behind the division winner

Values are aligned and delimited by blanks. There are no missing values.

References
Munsey, P., and Suppes, C. (2001), *Ballparks* [Online]. (www.ballparks.com/baseball)
Frees, E. W. (1996), *Data Analysis Using Regression Models*, Englewood Cliffs: Prentice Hall.
*InfoPlease* (2001), [Online]. (www.infoplease.com/index.html)
Reichler, J. L. (ed.) (2001), *The Baseball Encyclopedia*, New York: MacMillan Publishing Company.
Thorn, J., and Palmer, P. (2001), *Total Baseball*, New York: HarperCollins Publishers.

James J. Cochran
Department of Economics and Finance
Louisiana Tech University
Ruston, LA 71272
USA
jcochran@cab.latech.edu