This department publishes articles of interest to statistical practitioners. Innovative applications of known methodology may be suitable, but sizable case studies should be submitted to other journals. Brief descriptions and illustrations of new developments that are potentially useful in statistical practice are appropriate. Acceptable articles should appeal to a substantial number of practitioners.

A Critical Look at Some Analyses of Major League Baseball Salaries

David C. HOAGLIN and Paul F. VELLEMAN

At a data analysis exposition sponsored by the Section on Statistical Graphics of the ASA in 1988, 15 groups of statisticians analyzed the same data about salaries of major league baseball players. By examining what they did, what worked, and what failed, we can begin to learn about the relative strengths and weaknesses of different approaches to analyzing data. The data are rich in difficulties. They require reexpression, contain errors and outliers, and exhibit nonlinear relationships. They thus pose a realistic challenge to the variety of data analysis techniques used. The analysis groups chose a wide range of model-fitting methods, including regression, principal components, factor analysis, time series, and CART. We thus have an effective framework for comparing these approaches so that we can learn more about them. Our examination shows that approaches commonly identified with Exploratory Data Analysis are substantially more effective at revealing the underlying patterns in the data and at building parsimonious, understandable models that fit the data well. We also find that common data displays, when applied carefully, are often sufficient for even complex analyses such as this.

KEY WORDS: Data analysis; Outliers; Regression; Transformation; Variable selection.

1. INTRODUCTION

Before the 1988 Annual Statistical Meetings, the Statistical Graphics Section of the American Statistical Association made available salary data for 439 major league baseball players, along with various career and 1986 performance statistics and team attendance figures, challenging members of the ASA to analyze the data and present their analyses at a poster session. This exposition was announced at the 1987 Annual Meeting and in the September–October 1987 issue of *Amstat News*. No detailed instructions were distributed, other than a challenge to answer to question: "Are players paid according to their performance?"

One hundred twenty-seven groups asked for the data, and 15 of these presented analyses. Subsequently, Colin Mallows suggested to the authors (who had not participated in the exposition) that we synthesize lessons from the experience. The 15 presenters were contacted and asked to complete a summary questionnaire and to provide their papers (as prepared for the *1988 Proceedings of the Section on Statistical Graphics*).

In this article we review the methods used in those 15 analyses to find a statistical model that responds to the question of whether players are paid for performance. We aim to learn which methods and approaches seem most successful in revealing the structure of these data. Those parallel analyses offer a unique opportunity to compare and constrast a variety of approaches to data analysis. We hope this comparison can provide guidance to others who have data to analyze.

The exposition was not a competition. The groups took many different approaches, and some of these were experimental, reflecting the goal of the exposition to encourage participants to try a range of new methods. Indeed, some of the least "successful" analyses have taught us the most about choosing methods of analysis.

We stress that the use of baseball as a source of data is a choice of convenience. Others have noted that professional baseball offers an unusually rich source of data that are quite complete over a long time period. The ASA challenge took advantage of this wealth of data to present one set of data within a larger framework to many teams of data analysts. The present review examines the methods used by the statisticians; our goal is not to reach a deeper understanding of baseball salaries. We do not propose that any of these analyses would be particularly appropriate for understanding or arbitrating baseball salaries. Indeed, none of the participants professed any sophisticated knowledge of baseball or of previously published analyses of baseball players' salaries. In this, the participants more closely resembled statistical consultants, who often are

David C. Hoaglin is Senior Scientist, Abt Associates Inc., 55 Wheeler St., Cambridge, MA 02138; e-mail: dave_hoaglin@abtassoc.com. Paul F. Velleman is Associate Professor of Economic and Social Statistics, 358 Ives Hall, Cornell University, Ithaca, NY 14853 and President of Data Description, Inc.; e-mail: pfv2@cornell.edu. The authors are grateful to the participants in the data analysis exposition for their substantial efforts in analyzing these data, without which this overview would not have been possible, and for their comments on earlier versions. The authors also thank James M. Landwehr, Colin L. Mallows, Carl N. Morris, and Frederick Mosteller for helpful comments and suggestions. An earlier version of this paper appeared in the *1991 Proceedings of the Section on Statistical Graphics*. Preparation of this article was facilitated by NSF Grant SES 8908641 to Harvard University.

not expert in the discipline from which the data arise, but who nevertheless are called on to advise and assist in an analysis.

The question of whether salary reflects performance provides a focus for our review. We prefer models that account well for the relationship between performance and salary, and we seek models that are both parsimonious and interpretable. In this discussion we consider the models that were most parsimonious, most interpretable, and best fitting to be the most successful, because these are often good criteria for statistical analyses. Other analyses of baseball players' salaries may have other goals and might therefore lead to other models. We seek to identify the methods that led to the most parsimonious and best fitting models and to understand why these methods worked better than others.

All the analyses were performed with commercially available software. This could not have happened even five years before, and it highlights an important aspect of this review: all the methods used in these analyses are readily available.

2. THE DATA

Salary data for 439 major league players (263 hitters and 176 pitchers) came from the April 20, 1987 issue of *Sports Illustrated*; various career and 1986 performance statistics came from the *1987 Baseball Encyclopedia Update*; 1986 team attendance figures were obtained from the Elias Sports Bureau. The full data set included data on pitchers separately, but most respondents dealt only with the hitters; we likewise restrict our attention to the hitters here. Table 1 lists the specific data items. Lorraine Denby, 1988 Program Chair for the Graphics Section, made the data available by electronic mail and on floppy disk. They are currently available by electronic mail from the StatLib retrieval system (Meyer 1991, 1993) at statlib@lib.stat.cmu.edu.

Many of the analyses found outliers, and several identified errors in the data. We present some of these findings as we go along. The Appendix lists all the corrections known to us. The corrections are also available from StatLib.

Table 1. Data Items in the Baseball Data Set (as Distributed) for Hitters and Teams

1986	*Career*	*Team Data*
at bats	at bats	team name
hits	hits	league
home runs	home runs	division in 1986
runs scored	runs scored	final standing
runs batted in	runs batted in	wins
walks	walks	losses
league	years in major leagues	home attendance
division	player's name	away attendance
team	1987 annual salary	1987 average salary
position(s)	league at start of 1987	
put outs	team at start of 1987	
assists		
errors		

NOTE: The data set also contained data on pitchers, which are not analyzed in this article.

3. INITIAL DISPLAYS AND REEXPRESSIONS

Almost everyone began by displaying the data. (After all, the session was sponsored by the Statistical Graphics Section.) Most groups noted that the dependent variable, *salary*, was skewed (see Fig. 1) and that plots of *salary* against the most likely predictors were not linear. Almost all chose to work with log(*salary*); analyses of the raw data were less successful. Reasons for reexpressing to logs (explained in response to the questionnaire) included making the distribution more nearly symmetric, stabilizing the variance, obtaining a better fit, taking into account the tendency for larger raises to go with larger salaries (hence a multiplicative model), and simply accepting it as the common thing to do with salary data.

Some groups tried ranks or classes, either instead of logs or in addition to them.

Several of the groups transformed the explanatory variables, but they usually discussed these transformations differently from the choice of reexpression for *salary*. Indeed, most said little about the distribution shape of the predictors. All the variables relating to hitters are strongly skewed. A square root reexpression makes most of them more nearly symmetric and more nearly linear with log(*salary*). This result is not surprising when we recognize that almost all of these variables are counts; counted data often benefit from a square root reexpression. (See, for example, Mosteller and Tukey 1977.) Still, none of the groups chose this approach.

Instead, several of the groups constructed new predictors from the career totals by dividing totals by *years in the major leagues* or, in one instance, by *career at-bats*. One

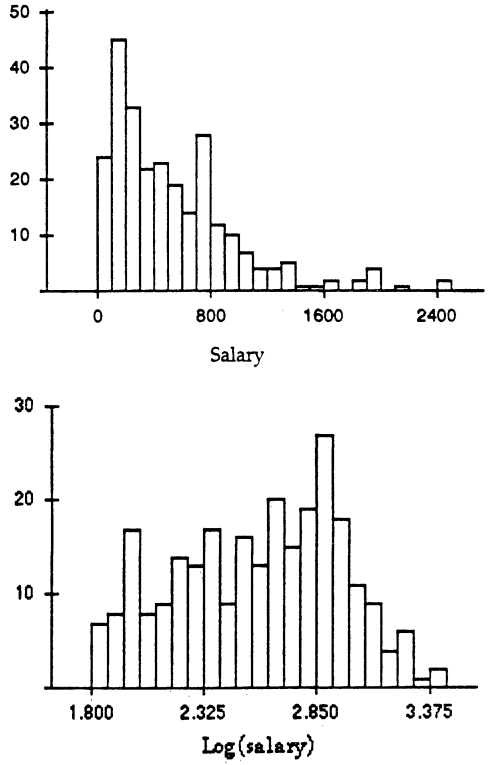

Figure 1. Histograms of Salary and log(salary), Indicating That a Logarithmic Reexpression Improves the Symmetry of the Salary Variable.

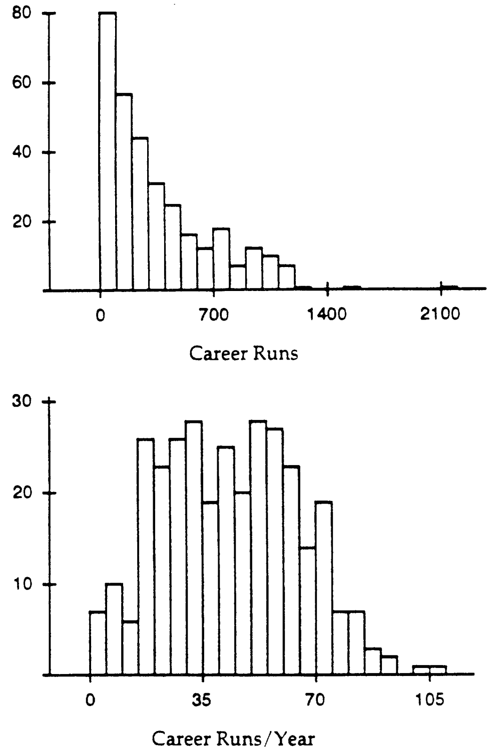

Figure 2. Expressing Career Runs as an Annual Rate, Career Runs/Year, Makes the Distribution of This Predictor More Nearly Symmetric.

group noted that the 1986 figures were included in the career totals, and chose to subtract them (leaving 0 career totals for rookies). Another group (Henry, Bauer, Johnson, and Noble 1989) cited an established expert, Earnshaw Cook (1966), and constructed a measure of Total Runs Produced as

$$\text{TRP} = (\text{runs scored} + \text{runs batted in} - \text{home runs})/\text{years}.$$

Any of these transformations creates variables that have reasonably symmetric distributions and are relatively linear with log(*salary*). In subsequent private communications some experts in baseball data have suggested that more appropriate measures of a player's contribution at the plate are now known. For the purposes of our review these would make little difference, although we would be interested to know whether they lead to symmetrically distributed values that are linear with log(*salary*). We are chiefly concerned with statistical practice and, in particular at this stage of the analysis, with the value of transformations and constructions that improve symmetry and linearity.

Transformation of career totals to annual rates (as in Fig. 2) seems more interpretable and more natural than taking square roots.

None of the groups transformed the 1986 performance figures, although these, too, show skewed distributions.

An overview of the analyses indicates that those working with log(*salary*) and with annual rate predictors fared better than those who worked with the raw forms of these variables. The models built were more successful at prediction, at identifying errors in the data, and at providing interpretable choices of predictors.

4. OUTLIERS AND ERRORS

As often happens, these data contained both errors and points that, although correct, were outliers for some models. Several groups found errors in the data and corrected them. Some groups omitted suspicious points, but did not identify them as erroneous. Other groups noted the anomalies but elected to continue with the data as given. Some groups did not report seeking or finding outliers.

An overview of many analyses provides an unusual opportunity to evaluate alternative approaches to outliers. Some statisticians have argued that one should not omit outlying observations without accounting for why they are extraordinary (and the questionnaire responses of some of the groups suggest this reason for not omitting outliers). Other statisticians advocate omitting outliers even if their distinctive behavior has no identifiable cause.

It is clear from the analyses surveyed here that we learn more about these data by omitting or correcting the outliers. The resulting models are more parsimonious, the coefficients are more readily interpretable, and the R^2 values are substantially higher. Because this moderately large data set contained relatively few errors, models fitted after omitting erroneous values were not very different from those fitted after correcting them.

The original sources for these data are publicly available, so it is relatively easy to determine which outliers are erroneous and which are truly extraordinary. Other data may pose greater difficulties; after eliminating clerical errors, it may not be possible to redo an experiment or resurvey a respondent. The analyses considered here suggest that it is generally worthwhile to look for outliers and that, if any are found, it is wise at least to reanalyze the remainder of the data—possibly as an alternative to the analysis of the entire data set.

If they were left unaltered in the data, the outliers were particularly damaging to automated attempts to build a model. Variable-selection methods such as stepwise regression select an entirely different set of predictors when even a few outliers are present than they would if the outliers were omitted. As we shall see, some of the outliers were discovered most easily in the process of building regression models, so the groups using stepwise regression faced a Catch-22; the method could not find a good regression model in the presence of outliers, but the outliers could not be found without a good model.

Similarly, multivariate methods such as principal components and factor analysis are well known to be sensitive to outliers. (See, for example, Gnanadesikan and Kettenring 1972; Devlin, Gnanadesikan, and Kettenring 1981; Seber 1984, pp. 171, 187.) The estimated covariances, which are central to these calculations, are easily perturbed by even one or two extraordinary values.

One indication of how hard it can be to identify errors is that the groups that reported specific errors found different ones. We present here some methods and displays that successfully reveal outliers in these data, with the caveat that we had already learned of these errors from the original 15 analyses. Other methods may work better with other data, but the lesson should not be lost in the methods; Looking for outliers, isolating them, and either correcting them or

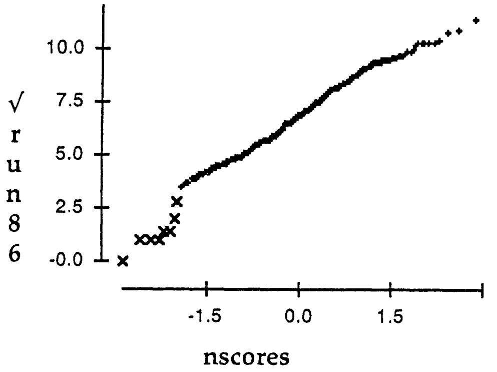

Figure 3. A Normal Probability Plot of $\sqrt{(runs\ scored\ in\ 1986)}$ Suggests a Normal Distribution With Some Extraordinary Points (Plotted as "x's"). All of these turn out to be errors.

omitting them (at least in an alternative analysis) form an essential step of successful data analysis.

Errors in the predictors can be especially difficult to identify. Histograms, stem-and-leaf displays, and normal probability plots of these variables all show some extraordinary points that warrant attention. If the data are transformed to improve symmetry, most of the errors become clearer. This outcome illustrates an important distinction between outliers and the values that happen to lie in the longer tail of a skewed distribution. Reexpressions that make a skewed distribution more nearly symmetric generally pull in the values in the longer tail so that they fit with the overall distribution. Outliers, however, tend to separate from the distribution when the main body of points is reexpressed for symmetry. For example, Figure 3 shows a normal probability plot of $\sqrt{(runs\ scored\ in\ 1986)}$. The shape seems close to a straight line indicating normality, except for a tail of suspicious points at the left. All of these points turn out to be errors in the data. Many led us to entire rows of data that had been entered wrong (most often by transcribing an adjacent row from one of the data sources) and thus identified several errors at once.

Some of the errors are in *number of years in the major leagues*, and these contaminate any annual rate variable, constructed as performance/year. This contamination affected several of the multivariate methods.

A normal probability plot of log(*salary*) finds no outliers among the salary values. As we shall see, the salary figures have errors as well, but they are revealed only in the course of constructing and examining models for the relationship of log(*salary*) to various predictors.

5. MODEL-SELECTION STRATEGIES

With 15 groups of data analysts at work it is no surprise that they used a variety of methods to find a suitable model. Most tried stepwise regression. None of the stepwise regressions worked, in part because of the influence of outliers noted earlier, and in part because of a failure of the model assumptions (discussed in Sec. 8). One group noted on the questionnaire that, although they had not used stepwise regression, they would not admit it even if they had. We agree that stepwise methods are unlikely to work well in most real-data situations and should be used only with great caution. Others have raised objections to the use of stepwise regression for building models. See, for example, Henderson and Velleman (1981).

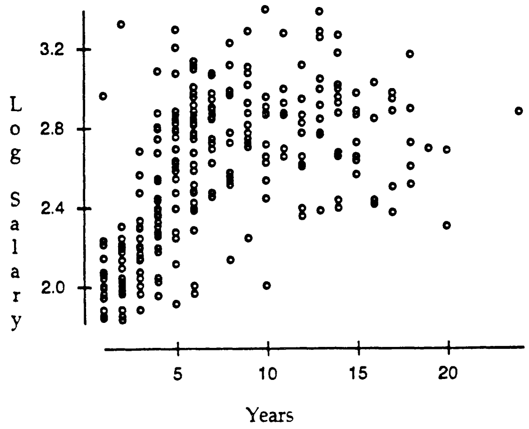

Figure 4. Log(salary) is Not Linear With Years.

Several groups used principal components and factor analysis, sometimes as a way to construct new predictors. These analyses were not successful, partly because the linear combinations of predictors found by the principal-components analyses were neither symmetrically distributed nor linearly related to the dependent variable, and partly because of the effects of outliers in the data. We discuss some of these effects in Section 6.

One group used recursive partitioning (CART) (Breiman, Friedman, Olshen, and Stone 1984), obtaining a model with a very different structure, but little intuitive appeal and low predictive ability. Another group introduced much additional data from earlier years and used time-series methods.

The method used by most of the groups (probably motivated by the initial question about whether salaries reflect performance) was regression. Some groups worked with the original data, others transformed the data, and still others constructed predictor variables via principal components or other multivariate methods.

The form of the relationship between log(*salary*) and *years in the major leagues* attracted considerable attention (see Fig. 4). Several groups noted that the relationship was not linear, and they offered two different opinions about its nature. Some groups chose to include $(years)^2$ in their regression models. One group saw the relationship of log(*salary*) to *years* as piecewise linear, growing linearly up to seven years and leveling off after that. It is plausible that salaries rise with increasing experience but then level off or decline as experienced players are signed to one-year free-agent contracts. Of the groups that considered the form of the relationship between log(*salary*) and *years*, most chose to introduce a quadratic term in *years*. On substantive grounds, however, the piecewise linear pattern "$years \leq 7$" may be more appropriate.

Figure 4 also reveals two additional errors in the data. Two salary values for players in their first and second years in the majors, plotted in the upper left corner of the display, seem extraordinarily generous for rookies. In fact, these points reflect errors in *years* for these two players, Terry Kennedy and Mike Schmidt. This finding is methodologically interesting because we can recognize these outliers only by displaying the two variables together; the data for these players are extraordinary in neither *years* nor log(*salary*), but they stand out clearly when the variables are combined. It is also interesting that we need not select a particular functional model relating log(*salary*) to *years* to see that these two points deserve a second look. Several groups identified these two points. Henry et al. (1989) discovered them in boxplots of log(*salary*) for each year, where they are even more visible than in the scatterplot.

Our baseball experts expressed some dismay that variables were not selected on the base of intimate knowledge of the game. Although this certainly could have been done, our concern here is for the relative success of the statistical methods that were used. We hope to learn about the effectiveness and pitfalls of these methods, so that we can use them more effectively on other data. In comparison, we are less concerned with building the most intuitive model of baseball salaries. Nevertheless, the best models found by these methods (i.e., those with the highest predictive ability and fewest predictors) cannot be improved by adding any of the variables available in the data.

6. OUTLIERS AND ERRORS IN THE DEPENDENT VARIABLE

Each of the salary figures is plausible, and (as we saw in Fig. 1) the distribution of log(*salary*) is reasonable. However, once we start to build a model to describe the relationship between log(*salary*) and the available predictors, some of the salary values stand out.

For this discussion of errors, we have chosen one of the simpler and more successful regression models, but the residuals from many of the proposed models (provided log(*salary*) is the dependent variable) show essentially the same extreme points. It is easiest to locate errors in salary when the errors in the explanatory variables are corrected or omitted first. We have omitted the points identified as outliers in Figures 3 and 4. The regression model (Table 2) predicts log(*salary*) from *years*, $(years)^2$, and *career runs/year*.

A plot of residuals versus predicted values shows three extreme points, labeled in Figure 5. The salary values for Jeffrey Leonard and Steve Sax are grossly in error. Pete Rose's salary is correct.

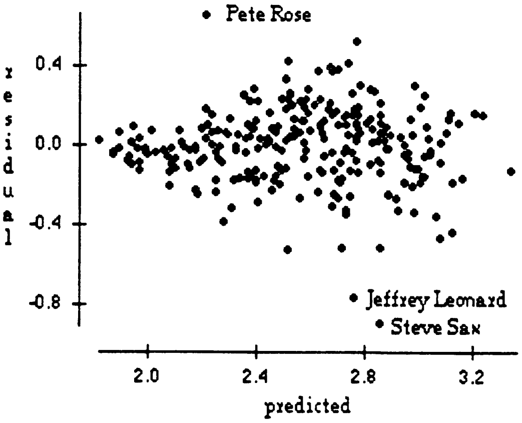

Figure 5. Residuals Versus Predicted Values for a Regression in Which Errors in Predictor Variables Have Been Corrected.

Several of the groups identified these points. Some corrected the salaries for Leonard and Sax, others omitted both players, and a few noted the errors but left them in the data.

Pete Rose is a special case. It can be argued that, as a playing manager, his salary was not based primarily on his performance on the field. A good argument can be made for treating him specially (perhaps with an indicator variable) in this data set.

7. SUCCESSES

The most successful models were regression-based models predicting log(*salary*). The most successful predictors included *years*, either $(years)^2$ or $(years \leq 7)$, *career runs/year*, and a performance measure from 1986 (*runs* seems to work well, but so do several others, including *rbis*). Some groups constructed new variables from these or from other predictors related to them. In general, models of this form with the errors corrected or omitted have R^2 just over 80% and *t* statistics that indicate a significant contribution from each of the predictors—evidence that hitters are indeed paid according to their performance (in addition to receiving initial salary increases for experience).

Table 2. One Illustrative Regression for Hitters, Predicting Log Salary from Three Predictors

Dependent variable is: Log Salary
Includes only players not extraordinary in probability plot
322 total cases, of which 62 are missing

R-squared = 74.5% *R*-squared (adjusted) = 74.2%
s = 0.1954 with 260 − 4 = 256 degrees of freedom

Source	*Sum of Squares*	*df*	*Mean Square*	*F Ratio*
Regression	28.4895	3	9.49650	249
Residual	9.77225	256	.038173	

Variable	*Coefficient*	*s.e. of Coeff.*	*t Ratio*	*Prob.*
Constant	1.48267	.0429	34.6	≤.0001
Years	.162324	.0091	17.9	≤.0001
Years^2	−.006955	.0005	−14.8	≤.0001
Career Runs/Year	.009364	.0006	15.3	≤.0001

NOTE: The hitters who were extraordinary in the normal probability plot of Figure 3 are omitted from this regression (as are players missing data on salary, years, or career runs).

Models that used principal components or factor analysis were generally less successful than regression-based analyses for two main reasons. First, the linear combinations of predictors found by these methods were not linearly related to the dependent variable—often in part because they included *years* linearly. In some instances this problem was exacerbated by analyzing *salary* rather than log(*salary*). Principal components and factor analysis do not help the data analyst to consider the possibility of nonlinear relationships among the variables. Second, these analyses were adversely affected by the outliers and errors in the data. Outliers altered the loadings substantially and, in the process, disguised their presence.

In general, the lack of effective multivariate display methods may have discouraged those groups who chose multivariate methods from drawing the simple displays that revealed these problems. There is a lesson here for all who use multivariate methods: Data display is essential. Even simple bivariate displays can often reveal key violations of model assumptions.

The description produced by recursive partitioning was perhaps the most strikingly different from the regression models found by most groups. Although it showed resistance to the effects of the errors and outliers in the data, it still had limited predictive ability. Most of the other models had R^2 values between 40% and 60%.

After performing transformations and correcting or omitting outliers, the most successful models did not require many predictors or a complex form. When the data for Sax and Leonard are corrected and Pete Rose is either omitted or treated specially, the regression model of Table 2 is quite successful.

Our expert discussants expressed some surprise that other variables (such as fielding position) played no role in these models. We note that the most successful of these models is improved only slightly by adding any of the other predictor variables available. Future work comparing these models to the best available from other sources may yield further insights.

8. WHAT WORKED

The experience of analyzing so challenging a data set and the opportunity to learn from the collection of analyses have been enlightening and productive. We hope that the Section on Statistical Graphics and other sections of the ASA will continue to offer similar challenges. Our discussion here, as previously, concentrates on what worked *statistically*, rather than on how baseball salaries are arrived at. We aim to identify successful data analysis strategies in the hope that they will apply broadly to other data sets.

We were struck by the degree to which the analyses integrated statistical graphics with sophisticated methods of analysis. The time is past when statistical graphics can consist only of a few static histograms or scatterplots as a preliminary to the "real" data analysis. Modern statistics programs make it relatively easy to move from graphics to analyses and back again as needed.

We were also impressed by the power now available on the data analyst's desktop. Many of the analyses were performed on desktop machines, and the groups that used mainframes generally could have accomplished the same tasks with personal workstations. (The analyses and displays in this paper were produced by Data Desk 4.2 [Velleman 1993].)

For the data on hitters' salaries we are able to describe what worked:

First, make simple univariate displays (histogram, stem-and-leaf display, probability plot, etc.). Look for symmetry (and a need to reexpress), outliers, and multiple modes.

Reexpress *salary* to log(*salary*), or there is little hope of further success.

Keep the guiding question in mind: *Are players paid according to performance*? Analyses that forged ahead blindly tended to lose sight of the initial focus for the analysis. The most impressive and powerful tools do little good if they do not address the questions at issue.

Make plots based on these goals. Simpler displays such as scatterplots and plots of residuals were usually more to the point than fancier displays.

Actively seek outliers, and either omit them from the data or correct them. Outliers severely affected almost every method of fitting models to these data. Correcting or omitting outliers often made it possible to discover still further problems in the data and deal with them.

Do some modeling or smoothing to study the relationship of log(*salary*) and *years*. Regression analysis assumes a linear relationship between the predictors and the response variable. Graphics are an excellent way to check this assumption, and in these data we find an important failure of the assumption. Once a group missed the nonlinearity of this relationship, the rest of its model selection could not recover. Stepwise methods stumbled on this problem as well.

Diagnose possible models. Each potential model deserves plots of residuals, diagnostic statistics, and (for regression models) partial regression plots.

Work with the understanding that no simple model may be "best" for the data. Several groups found excellent fits to the data. All of the best fits corrected or omitted errors, predicted log(*salary*), transformed career performance statistics, and allowed for the nonlinearity of log(*salary*) with *years*. All of these models found that run production was a key predictor. But no two groups found exactly the same model.

9. BEYOND THE ORIGINAL ANALYSIS

As we have worked with these data and talked with others about them, we have learned some things that none of the original analyses mentioned. We discuss some of them here for completeness.

9.1 Form of the Model

We noted earlier that one of the groups fitted the nonlinear relationship between log(*salary*) and *years* with two straight lines, and others chose a quadratic. A careful examination of Figure 4 shows a short "tail" in the first two or three years. Part of this flattening, and other aspects of the pattern, may reflect the major league contracts in use at the time. Among other features (Mann 1989), those contracts allowed the team to set a player's salary during the first two years, subject to a guaranteed minimum.

Table 3. A Regression Model Omitting the Erroneous Points Identified in the Appendix, and All Players With Missing Data on These Variables

Dependent variable is: Log Salary
Omits cases with known errors or missing values
322 total cases, of which 64 are missing
R-squared = 81.7% R-squared (adjusted) = 81.4%
s = 0.1648 with 258 − 5 = 253 degrees of freedom

Source	*Sum of Squares*	*df*	*Mean Square*	*F Ratio*
Regression	30.6773	4	7.66931	282
Residual	6.87284	253	.027165	

Variable	*Coefficient*	*s.e. of Coeff.*	*t Ratio*	*Prob.*
Constant	1.53299	.0494	31.0	$\leq$.0001
Career Runs/Year	.007034	.0008	9.35	$\leq$.0001
$\sqrt{}$run86	.036205	.0088	4.10	$\leq$.0001
yrs 3 to 7	.149806	.0066	22.7	$\leq$.0001
years $>$ 7	−.017340	.0040	−4.31	$\leq$.0001

NOTE: The variables *yrs 3 to 7* and *years* $>$ *7* are designed to describe the three-line pattern of salaries versus years described in the text.

Players with at least two years but less than six years in the major leagues had a right to have their salaries determined through arbitration, and a player with at least six years in the major leagues could declare himself a free agent and sell his services to any club. In practice, the teams signed some players (after their first two years) to multiyear contracts, at salaries ranging up to the level of free-agent salaries.

Thus, we might prefer to summarize the trend in Figure 4 with three line segments, one for the first two years, another for the succeeding five years, and a third for the remaining years. We could introduce two new variables:

yrs 3 to 7 defined by the expression:
if years $\leq$ 2 *then 0 else*
if years $\leq$ 7 *then (years* − *2)*
else 5

and *years* $>$ 7 defined as
if years $\leq$ 7 *then 0 else (years* − *7).*

The regression of log (*salary*) on these two variables along with *runs86* and *career runs/year* fits slightly better than the regression with the two-line or quadratic alternatives for *years*, and makes substantive sense. Table 3 shows the resulting regression model after omitting the outliers identified thus far except for Pete Rose.

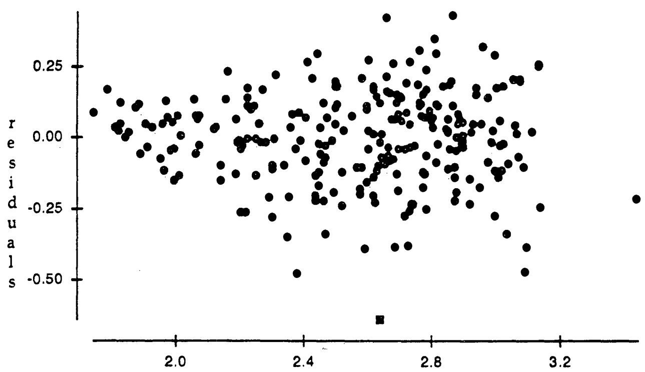

Figure 6. A Plot of Residuals From the Regression Model in Table 3 Suggests That Steve Balboni (Plotted With an "x") May Be an Outlier, But Checks on the Data Set Show no Error.

9.2 Diagnostics

A plot of residuals for the model of Table 3 (Fig. 6) reveals yet another outlier. The point (shown in the plot with an x) is Steve Balboni. A check of the original sources shows no error in Balboni's data. An outlier such as this can raise new questions. If, for example, Balboni's salary was part of a multiyear contract, we could modify Figure 6 by highlighting all players in the same year of multiyear contracts, but the current data do not provide this information.

10. CONCLUSIONS

This study considers analyses in an unusual setting. Although participants worked much like statistical consultants, they had to work without the consultant's usual access to a subject-matter expert. Nor can we consider voluntary participants representative of the population of practicing statisticians.

Our focus, however, is on the methods rather than on the analysts. The participants applied a broad variety of methods to a single data set and question. They generally applied these methods skillfully and according to common practice. Where a method failed to reveal the underlying structure of the data, that failure can usually be attributed to the method's inability to deal with the data rather than to misuse of the method by the analysts.

Thus, despite its special circumstances, this study can help us to understand the relative effectiveness of commonly used methods and their potential to lead data analysts astray when their assumptions are not met. Some of the methods that worked best on these data are not commonly taught in introductory statistics courses and are not widely used. Their success on these data recommends them for increased attention.

To summarize our conclusions about the methods:

(1) Graphics are essential to good data analysis and must be integrated with the entire process of data analysis. Many authors have emphasized the importance of good data graphics (Chambers, Cleveland, Kleiner, and Tukey 1983; Tufte 1983). We extend that recommendation to suggest that graphics be an integral part of *every*

stage of the analysis. Thus graphics are essential in assessing the need to reexpress variables and in the hunt for outliers and errors. Graphics are needed for evaluating model assumptions such as linearity and homoscedasticity, and in evaluating steps taken to deal with violations when they are found. Finally, graphics are the heart of good diagnosis and post-analysis searches for additional patterns.

(2) Data reexpression is essential. It may be that the data can be analyzed in their original form, but in our experience this is relatively rare. Mosteller and Tukey (1977, sec. 5H) advocate reexpressing data even before examining them. A slightly less radical approach calls for initial displays of every variable to assess symmetry, homoscedasticity, and linearity with the dependent variable, and encourages reexpressions that improve these aspects of the data.

(3) Outliers and errors can do great harm to most standard analyses. A systematic search for outliers should be routine at every stage of the analysis. Although some outliers and errors may be obvious from initial displays, others will be evident only after reexpressing the data or after fitting an initial model to the data. Even the "final" residuals may reveal unusual data points.

(4) Simple graphics are often more successful than sophisticated methods. The best analyses of the hitters' salaries needed only histograms (or stem-and-leaf displays), probability plots, and scatterplots (including residual plots and partial regression plots).

(5) Attempts at automated analysis are prone to substantial failures. Most such attempts for these data generated complex models that nonetheless fit poorly and offered no insight. Such methods included stepwise regression and CART.

(6) Blind application of multivariate methods is similarly dangerous. For these data such analyses produced complex models that fit poorly and offered no insight. Such methods include principal components and factor analysis, and combinations of these with regression.

(7) Poorly chosen graphics obscure rather than clarify. Velleman and Hoaglin (1992) discuss related issues, with an emphasis on principles and philosophy of data analysis.

We hope that practicing statisticians find these results helpful. By applying diverse methods to the same data, the groups of participants in the exposition have provided a valuable opportunity to advance our comparative understanding of the methods. We encourage the Statistical Graphics Section and the Section on Statistical Computing to continue to sponsor similar events.

APPENDIX: CORRECTIONS TO THE HITTERS' DATA

Original data are on the first line; corrected data are on the second line.

Name	salary	bat86	hit86	hr86	run86	rb86	wlk86	yrs	batcr	hitcr	hrcr	runcr	rbcr	wlkcr	po86
Tony Armas	NA	16	2	0	1	0	0	2	20	4	0	1	0	0	CF
Tony Armas	NA	425	112	141	40	58	24	11	4513	1134	224	542	727	230	CF
D. Baker	NA	24	3	0	1	0	2	3	159	28	0	20	12	9	OF/1B
D. Baker	NA	242	58	4	25	19	27	19	7117	1981	242	964	1013	762	OF/1B
Bob Boone	NA	22	10	1	4	2	1	6	84	26	2	9	9	3	C
Bob Boone	NA	442	98	7	48	49	43	15	5982	1501	96	555	702	533	C
Dave Henderson	325	388	103	15	59	47	39	6	2174	555	80	285	274	186	OF
Dave Henderson	525	388	103	15	59	47	39	6	2174	555	80	285	274	186	CF
Cliff Johnson	NA	19	7	0	1	2	1	4	41	13	1	3	4	4	OF
Cliff Johnson	NA	336	84	15	48	55	52	15	3945	1016	196	539	699	568	OF
Ricky Jones	NA	33	6	0	2	4	7	1	33	6	0	2	4	7	OF
Ruppert Jones	NA	393	90	17	73	49	64	11	4223	1056	139	618	551	514	OF
Jeffrey Leonard	100	341	95	6	48	42	20	10	2964	808	81	379	428	221	LF
Jeffrey Leonard	900	341	95	6	48	42	20	10	2964	808	81	379	428	221	LF
Terry Kennedy	920	19	4	1	2	3	1	1	19	4	1	2	3	1	C
Terry Kennedy	920	432	114	12	46	57	37	6	3374	915	83	345	475	238	C
Mike Schmidt	2127.3	20	1	0	0	0	0	2	41	9	2	6	7	4	3B
Mike Schmidt	2127.3	552	160	37	97	119	89	15	7292	1954	495	1347	1392	1354	3B
Ronn Reynolds	190	126	27	3	8	10	5	4	239	49	3	16	13	14	LF
R.J. Reynolds	190	402	108	9	63	48	40	4	1034	278	16	135	125	79	LF
Steve Sax	90	633	210	6	91	56	59	6	3070	872	19	420	230	274	2B
Steve Sax	740	633	210	6	91	56	59	6	3070	872	19	420	230	274	2B

[*Received February 1994. Revised February 1995.*]

REFERENCES

Armstrong, M. A., Tekawa, I. S., and Johnson, R. A. (1989), "Analysis of Major League Baseball Salaries," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 76–80.

Becker, R. A., Clark, L. A., Denby, L., Landwehr, J. M., Mallows, C. L., Nair, V. N., and Wilks, A. R. (1989), "Analysis of Baseball Salary Data," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 81–86.

Boling, J. C., Brocklebank, J. C., and Powers, W. (1989) "Analysis of Major League Baseball Players Salaries," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 87–92.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth International Group.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Chatterjee, S., and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, New York: John Wiley.

Cleary, R. J., and Lock, R. H. (1989), "Why Are They Missing?," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 93–97.

Conlon, M., and Meyer, J. (1989), "Baseball Performance and Salaries," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 98–103.

Cook, E. (1966), *Percentage Baseball* (2nd ed.), Cambridge, MA: MIT Press.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association*, 76, 354–362.

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics*, 28, 81–124.

Henderson, H. V., and Velleman, P. F. (1981), "Building Multiple Regression Models Interactively," *Biometrics*, 37, 391–411.

Henry, N. W., Bauer, D. F., Johnson, R. E., and Noble, J. H. (1989), "Hits, Runs and Dollars," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 104–109.

Kern, D. M., and Reeves, J. H. (1989), "Graphical Analysis of Baseball Salary Data," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 110–114.

Kim, Y.-K., and Kim, C. (1989), "Analysis of Major League Baseball Salary Data," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 115–119.

Lomax, R. G., and Chou, T. (1989), "The Relationship Between Performance and Salary of Baseball Players," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 120–125.

Lombardi, D. A. (1988), "Why They Make What They Make—An Analysis of Major League Baseball Salaries," presented at the Joint Statistical Meetings, New Orleans, LA.

Mann, S. (1989), "The Business of Baseball," in *Total Baseball*, eds. J. Thorn and P. Palmer, New York: Professional Ink, Inc., pp. 628–641.

Meyer, M. M. (1991), "StatLib Offers System for Distributing Statistical Software and Data," *Amstat News*, July, p. 16.

——— (1993), "StatLib@lib.stat.cmu.edu:—A System for Distributing Statistical Software and Data," *The IMS Bulletin*, 22, no. 1, 6–7.

Mosteller F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Pruent, D. A., and Truss, L. T. (1989), "Why They Make What They Make—An Analysis of Major League Baseball Salaries," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 126–131.

Ronser, B., and Woods, C. (1989), "Autoregressive Modeling of Baseball Performance and Salary Data," in *ASA 1988 Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 132–137.

Seber, G. A. F. (1984), *Multivariate Observations*, New York: John Wiley.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Tukey, J. W. (1972), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft, Ames, IA: Iowa State University Press, pp. 293–316.

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Velleman, P. F. (1993), *Data Desk 4.2*, Ithaca, NY: Data Description.

Velleman, P. F., and Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Boston, MA: Duxbury Press.

——— (1992), "Data Analysis," in *Perspectives on Contemporary Statistics*, eds. D. C. Hoaglin and D. S. Moore, Washington, DC: Mathematical Association of America, pp. 19–39.

Velleman, P. F., and Welsch, R. E. (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician*, 35, 234–242.

Statistical Artifacts in the Ratio of Discrete Quantities

Roger G. Johnston, Shayla D. Schroder, and A. Rajika Mallawaaratchy

The ratio is a familiar statistic, but it is often misused. One frequently overlooked problem occurs when ratioing two discrete (digitized) variables. Fine structure appears in the histogram of the ratio that can be very subtle, or can sometimes even dominate the histogram. It disappears when the numerator and/or denominator become continuous. This statistical artifact is not a binning error, nor is it removed by taking more data. It is important to be aware of the artifact in order to avoid misinterpretation of ratio data. We provide examples of the statistical artifact (including one from baseball) and discuss ways to avoid or minimize the problems it can cause.

KEY WORDS: Analog-to-digital conversion; Binning errors; Digitization; Histogram, Ratio; Statistical artifacts.

1. INTRODUCTION

The ratio is one of the most frequently used statistics. It is also one of the most frequently misused statistics (Schor

Roger G. Johnston is Technical Staff Member, Los Alamos National Laboratory, Los Alamos, NM 87545. Shayla D. Schroder is a student, Computer Sciences Department, Eastern New Mexico University, Portales, NM 88130. A. Rajika Mallawaaratchy is a student, Electrical Engineering Department, Iowa State University, Ames, IA 50012. This work was performed under the auspices of the U.S. Department of Energy. The authors thank John Martin, Jim Jett, and Bill Baird for their discussions.