CHAPTER 3

EXPLORATORY DATA ANALYSIS

Paul F. Velleman and David C. Hoaglin

Exploratory data analysis (EDA), pioneered by John W. Tukey (1915–2000), introduces a variety of innovative techniques and combines them with five important principles of data analysis: display, re-expression, residuals, resistance, and iteration. Many of the techniques that Tukey pioneered have become familiar: stem-and-leaf display, five-number summary, boxplot, and a rule for flagging potential outliers in batches of data. Computing methods have extended EDA to larger data sets and higher dimensions, and diagnostic statistics have extended the EDA approach to include more traditional statistical methods.

Although its innovative methods have received much attention, the principal contribution of EDA is philosophical. EDA advocates exploring data for patterns and relationships without requiring prior hypotheses. The principle of *resistance* calls for identifying extraordinary cases and then setting them aside or downweighting them. *Re-expression* uses mathematical transformations to simplify patterns in data. EDA suggests that analyses are more scientifically useful and productive when data have been transformed to agree better with basic assumptions. *Residuals* come from summarizing the patterns found so far and subtracting that summary from the data, to reveal departures and additional patterns. EDA often works with residuals to refine or extend models fitted to data. Frequent use of graphical *displays* maintains contact with data, residuals, and summaries, and it often reveals unexpected behavior. EDA approaches do not terminate with a hypothesis test. Effective data analysis is *iterative*, finding and summarizing patterns and then probing more deeply.

These approaches stand in contrast to the formalistic scientific method paradigm of first stating a hypothesis based on prior theory, then collecting data, and finally applying a statistical test of the hypothesis. Proponents of the EDA philosophy maintain that the EDA approach is more likely to discover new and interesting patterns and relationships, in much the same way that science has traditionally made progress. Exploratory analyses can incorporate methods of statistical inference, but use them more as indicators of the strength of a relationship or the fit of a model than as confirmation of a hypothesis.

In this chapter, we elucidate the EDA approach, illustrating it with examples. We hope to convince the reader that this approach should be a standard part of anyone's analysis of data. For many experienced data analysts, an EDA approach forms the main ingredient of their analyses, with only the occasional "seasoning" of formal hypothesis testing.

WOES OF TRADITIONAL STATISTICS

The discipline of statistics offers a wide variety of ways to formulate and test hypotheses. But all rely on assumptions about the pattern of behavior in the data and about the distribution of variations around that pattern. For example, fitting a simple regression line is appropriate when the relation of y to x

The authors are grateful to Gloria Gogola for permission to use the data on functional dexterity trials and to Katherine Freier for preparing Figure 3.17.

DOI: 10.1037/13621-003
APA Handbook of Research Methods in Psychology: Vol. 3. Data Analysis and Research Publication, H. Cooper (Editor-in-Chief)

resembles a straight line and the fluctuations around a line all have the same variance. Methods that compare groups may require that the groups share the same variance. The basic model in a two-factor analysis of variance (ANOVA) expresses the response as an additive combination of the contributions of the two factors and assumes that the error variance is the same for all treatment levels. In logistic regression, the individual outcomes are usually assumed to follow binomial distributions, whose success probabilities (in the logit scale) follow the pattern specified in the linear predictor. Virtually every standard method makes the tacit assumption that the underlying data are homogeneous—that is, that they are all consistent measurements of the same things about the same kinds of individuals for whom the same model, analysis, or comparison is appropriate.

These assumptions are frequently violated by otherwise perfectly ordinary data. Often, to check, we need only make an appropriate display. As the famous philosopher (and baseball Hall of Famer) Yogi Berra said, "You can learn a lot by looking."

Ironically, although statistics software makes displaying data simple, it also abets the tendency to rush to a hypothesis test without pausing for displays. The traditional approach to statistics has an even more fundamental weakness. By focusing on testing hypotheses, we fail to ask the far more fundamental and important question of our data: "Is anything going on that I didn't expect?" Isaac Asimov is commonly credited[1] with saying,

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but "That's funny . . ."

EDA increases our chance of a "That's funny . . ." insight about our data—and those are the events that lead to new theories and breakthroughs. For example, Leonard Mlodinow, in his 2008 book *The Drunkard's Walk,* tells the story of the "That's funny . . ." moment that led Daniel Kahneman to begin his revolutionary research with Amos Tversky into the psychology of how humans misperceive random events—work that led to his 2002 Nobel Prize in Economics. As a junior professor of psychology at Hebrew University, Kahneman was lecturing to a group of flight trainers. The trainers insisted that when they chastised a flyer for a poor maneuver, the flyer improved, but when they complimented a good one, the flyer usually did worse the next time. The trainers had concluded that negative reinforcement worked and positive reinforcement did not. Kahneman knew that psychology had shown otherwise. Rather than ignoring this observation as just a strange aberration, he searched for the explanation. His insight that the often-counterintuitive result known as *regression to the mean* was responsible for the trainers' misperception started him on his research path.

Exploratory data analyses incorporate techniques, such as boxplots and stem-and-leaf displays, and also commonly use such traditional methods as least squares regression, ANOVA, analysis of covariance (ANOCOV), and logistic regression. So EDA should be viewed primarily as an enhancement of, rather than a replacement for, traditional methods. By using both wisely, you will learn more from your data.

EXPLORING

The tradition of the scientific method requires researchers to follow a straight and narrow path: theory first, then hypothesis, then data collection, and finally statistical tests of the stated hypothesis (and no others).

Exploring often requires that we leave the beaten path. But it does not require us to travel without a compass—or a GPS. The challenge is less to know where we are than to see where we are going. In this chapter, we present guidelines and best practices to help you make progress without wandering aimlessly. Our guiding principles are display, re-expression, resistance, residuals, and iteration.

Lacking a map, we need to take frequent sightings. We make many displays of the data, and we continue to do so all along the way. We are not trying—or even hoping—to confirm that the required assumptions for a hypothesis test have been met. We are looking for new territory to explore. Like all explorers, we are seeking—and expecting—the unexpected. Fortunately, displays

[1]This is commonly, and plausibly, credited to Asimov, but we have not found a reference.

are particularly good tools for this task. Appropriate displays readily show outliers, unexpected clumps and clusters, and relationship patterns that we might not have anticipated.

Nor are we obliged to follow steep or thorny paths when gentler, clearer ones are available. We re-express data to make it easier to summarize and compare them and to fit models that describe relationships. It is easier and more scientifically useful to fit a simple linear model to the logarithm of a variable than to fit an exponential model to the original data. Speed (reciprocal time) is often a more useful variable than duration.

Resistant methods protect against undue influence by outlying cases or even small clusters of cases. Some statistics, such as the median, are inherently resistant. Others, such as least squares regression, can be made resistant by diagnostics that identify potentially influential cases so they can be dealt with separately.

Whenever we fit a model to data, it is wise to examine the residuals—the differences between the model and the data. EDA suggests displays such as quantile–quantile plots, partial regression plots, and partial boxplots that are particularly well suited to understanding what residuals tell us about our data and models. In the names of these displays, *partial* refers to the removal (*partialing out*) of the contributions of the other predictors or factors.

EDA anticipates that what we learn from examining the residuals will lead us to improve our understanding and our models. We may decide to re-express variables to mitigate violations of assumptions revealed only in the residuals, set aside outliers identified in the residuals, or introduce additional variables suggested by the residuals.

DISPLAYS AND SOME OF WHAT THEY REVEAL

Every data analysis should begin with graphs. Of course, this idea is not unique to EDA. Every introductory statistics text starts with data displays (see Chapter 4 of this volume). But often, displays of the data appear only in the introductory chapters, and the figures in the chapters on statistical inference show only normal or *t*-distribution curves. EDA teaches that we should make graphs early and often.

Some data displays are taught in every statistics class. We look at histograms to get a picture of how the data are distributed and to check for multiple modes and outliers. We look at scatterplots to see how one variable changes in relation to another, focusing on the direction, form, and strength of that relationship.

To illustrate, consider a standard test of manual dexterity, which records the time it takes subjects to invert 16 cylinders with one hand (Aaron & Jansen, 2003). This test can form the basis for studying cross-over training: whether training one limb can benefit the other and what underlying mental processes might support the benefit. Gogola, Lacy, Morse, Aaron, and Velleman (2010) studied the performance of 175 subjects ranging in age from 4 to 16. A histogram of their times (see Figure 3.1) is skewed to the right and suggests that two subjects had unusually long times.

Modes

Histograms such as the one in Figure 3.1 may be the most common display for examining the distribution of a variable. Exploratory analyses often start with histograms. But EDA teaches that histograms can be a call to action. Histograms offer the opportunity to detect *inhomogeneity* in the data. All statistical methods make the tacit assumption that the data are homogeneous—that is, that we are dealing with measurements of the same thing made on members of a coherent population. Without homogeneity, it is difficult to imagine what—or whom—a summary of the data would be about.

One clue that the data may not be homogeneous is the presence of two or more modes in a histogram.

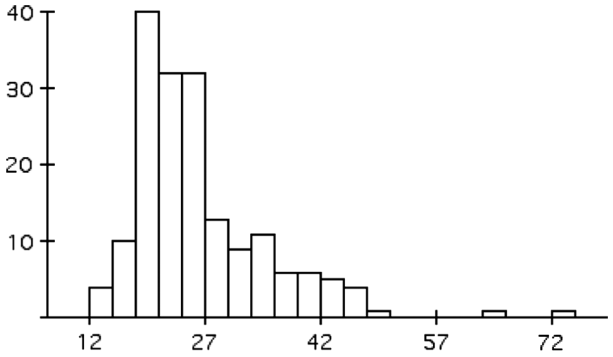

FIGURE 3.1. Histogram of the times of 175 subjects on a test of manual dexterity.

For example, the same researchers also measured the dexterity of patients who had had surgery, using a variety of measures. One of these measures, the Jebsen Large Heavy Object Test (Jebsen, Taylor, Trieschmann, Trotter, & Howard, 1969), records the time (in seconds) to lift and move a weighted can. In the histogram of the times for 34 subjects, shown in Figure 3.2, the mode around 9 s consists of patients who had had surgery on both limbs, who may differ from other patients in important ways.

Outliers

A related issue is the identification and treatment of outliers. Values that stray far from the rest of the data, or that stand apart from important patterns in the data, demand our attention. They may be particularly informative by clarifying the limits of our data or pointing out special cases, or they may be errors in need of correction or removal. Regardless of the reason, they should not be allowed to distort subsequent analyses of the data. EDA teaches that if we cannot correct an outlier, we should set it aside or use methods that are immune to its effects.

The argument that outliers should be prevented from subverting a data analysis reflects the philosophical foundations of EDA. Some analysts are reluctant to set aside any legitimately recorded values, fearing that doing so could bias subsequent analyses. But standard statistical methods are notoriously sensitive to outlying values and are likely to be invalid when applied to data that include outliers. Data containing outliers are, almost by definition, not homogeneous. So statistical models intended for homogeneous data are likely to strike a clumsy compromise between the outliers and the rest of the data rather than describe patterns and relationships in the bulk of the data.

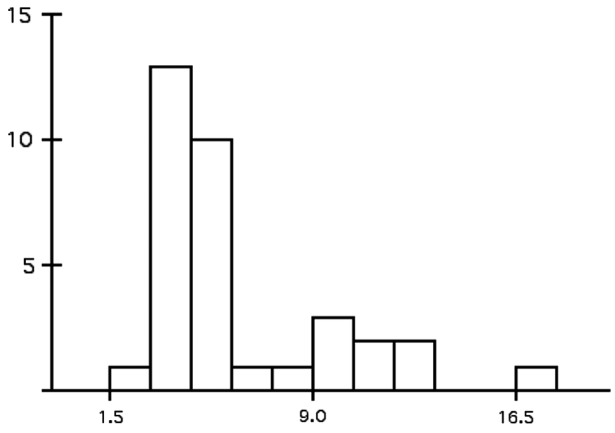

FIGURE 3.2. Times (in seconds) on the Jebsen Large Heavy Object Test for 34 subjects who had had surgery.

Boxplots

One tool that can be helpful in nominating extreme values for consideration as outliers is the boxplot, introduced by John Tukey in his pathbreaking 1977 book *Exploratory Data Analysis*. The standard boxplot uses a rule to identify possible outliers (as *outside* or, if extreme enough, *far outside*) and plot them individually. Hoaglin, Iglewicz, and Tukey (1986) studied the rule's performance, and Hoaglin and Iglewicz (1987) refined it.

A boxplot of the task times for the 175 normal subjects, shown in Figure 3.3, identifies five task times as outside and two as far out. The records for these subjects should be examined for possible explanations of their particularly slow performance.

The boxplot's outlier nomination rules should not be taken as a *definition* of outliers. The decision to treat a case as an outlier is a judgment call

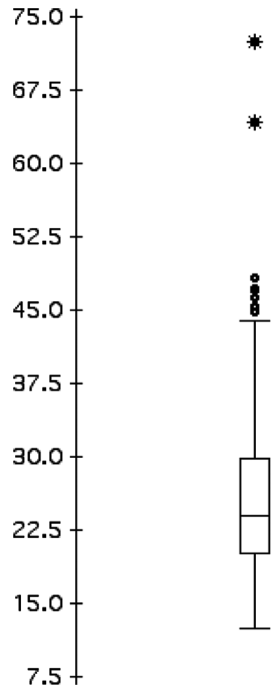

FIGURE 3.3. Boxplot of the times (in seconds) of the 175 normal subjects on the test of manual dexterity.

that the data analyst must make. But boxplots can help by directing attention to cases that deserve consideration as outliers. The standard boxplot calls attention to outside observations fairly often. Hoaglin et al. (1986) found that in well-behaved (i.e., Gaussian) data the percentage of samples that contain one or more outside observations varies between 33% and 14% for $5 \le n \le 20$. The corresponding percentage for far outside lies between 1% and 5%.

Boxplots are also helpful for comparing groups. Because they show the median and quartiles of each group, they make it easy to compare centers and spreads (as interquartile ranges) among groups. Because they suppress details of the distributions, they minimize distractions that can make it difficult to compare several histograms.

Figure 3.4 shows the results of repeated dexterity trials by the same 175 subjects. Except for a few high values, times decreased from Trial 1 to Trial 3 as subjects practiced the task, but times stabilized after Trial 3. Because boxplots isolate outliers, we can more easily ignore their influence when judging the performance pattern of most of the subjects.

Stem-and-Leaf Displays

Another display introduced by John Tukey (in the late 1960s) is the stem-and-leaf display, which offers a histogram's view of distribution shape while preserving the individual data values. Each digit to the right of the vertical line represents a data value (e.g., 12 s and 14 s on the first line). A stem-and-leaf display of some of the dexterity times (Figure 3.5) shows individual values as well as the overall distribution shape. Stem-and-leaf displays are particularly useful as pencil-and-paper tools for a quick look at modest collections of data values.

Two-Variable Relationships

The EDA approach guides the consideration of relationships between pairs of variables. Of course, we start with a display. Now we look at the overall *direction*, *shape*, and *strength* of the relationship. The scatterplot of dexterity task time versus age in Figure 3.6 shows a negative direction with older subjects taking less time, a curved shape, and a reasonably consistent pattern.

Because exploratory analyses rely on displays, they often push common methods a bit farther. For example, points in scatterplots can be assigned

```
1|24
1|67888999999
2|0000111123344
2|555555666677889
3|0023333
3|5567889
4|0012
4|57
5|
5|
6|4
6|
7|2
```

FIGURE 3.5. Stem-and-leaf display of the task times (in seconds) for 63 subjects.

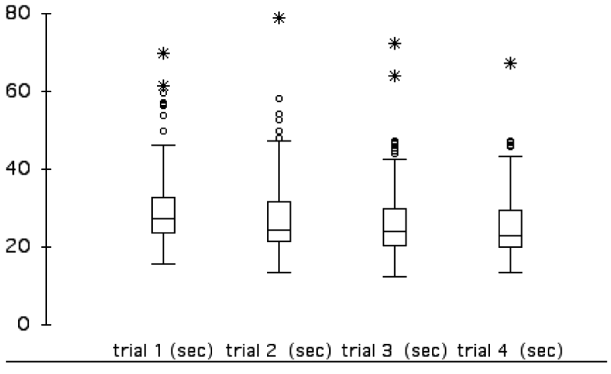

FIGURE 3.4. Boxplots of the times (in seconds) of the 175 normal subjects on four consecutive dexterity tests.

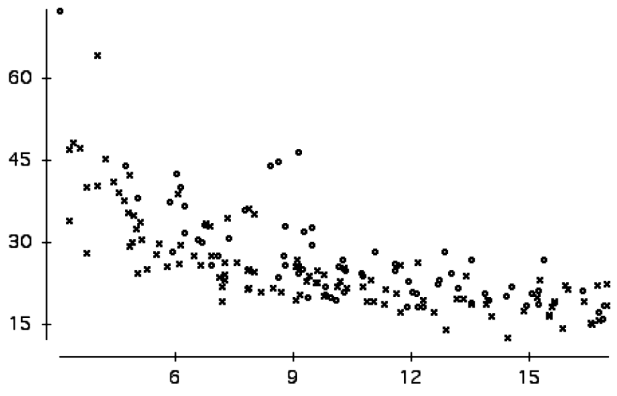

FIGURE 3.6. Scatterplot of task time (in seconds) versus age (in years). The plotting symbol distinguishes the dominant hand (×) from the nondominant hand (o).

colors or symbols according to values of a third categorical variable. The symbols in Figure 3.6 record whether the tested hand was the subject's dominant (×) or nondominant (o) hand.

If the form of the relationship were straight, then lines fitted separately to each group could be compared. That is not feasible with a curved plot such as this one. Modern statistics software often supports the ability to "touch" a point in a plot to ask for its identity—a valuable tool for identifying and understanding outliers and subgroups in the data.

Other Displays

Other displays and display methods are less common or depend on computers. Normal probability plots provide a better way to compare a variable's distribution with the normal than does a histogram with a normal curve overlaid. One drawback of such histograms is related to Winsor's Principle: All distributions are normal in the middle (Tukey, 1960). Although the principle is clearly not universally true, it does correctly—and memorably—advise us to focus attention on the tails of a distribution.

A scatterplot matrix (called a SPLOM by some programs) is an array of scatterplots laid out in the same pattern as correlations in a correlation table. *Plot brushing* highlights the same cases in each plot simultaneously as the viewer passes a rectangular "brush" over any one of them, so that relationships among several variables can been seen. An alternative approach to displaying three variables together is an animated *rotating plot*, offered in several statistics programs. Some programs, such as Data Desk (Velleman, 2004), can display data that have up to nine dimensions.

Another approach to displaying high-dimensional data is a parallel-coordinate plot (Inselberg, 2009). Figure 3.7 shows the four trials of the dexterity experiment seen in the boxplots of Figure 3.4. In a parallel-coordinate plot, lines connect the times for each subject. Here we can trace the performance of individual subjects—for example, to see that the outlying slow performances were by the same subjects.

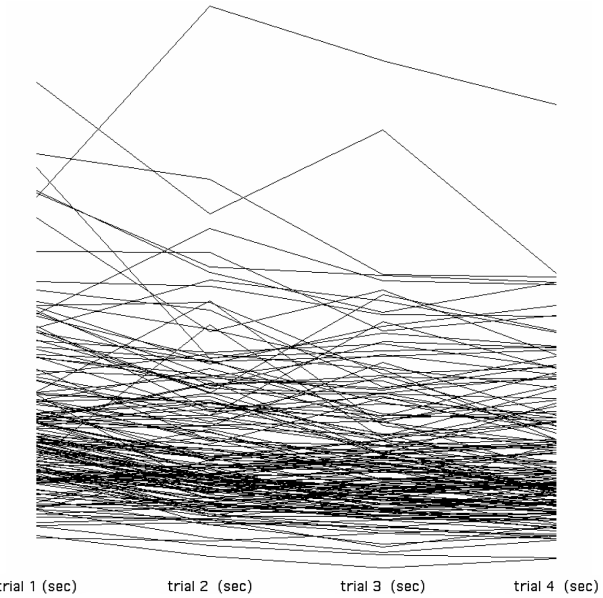

FIGURE 3.7. Parallel-coordinate plot of the times (in seconds) of the 175 normal subjects on four consecutive dexterity trials.

RE-EXPRESSION

One of the most versatile techniques in the data analyst's toolkit, re-expression applies the same mathematical function (such as the logarithm or the square root) to each data value of a variable. This transformation smoothly changes the relative positions of the data, with the aim of simplifying the analysis.

Some researchers mistakenly believe that transforming the data is wrong, that the analysis must work with the data in their original scale. Those who hold this view often fail to find important features of their data. In fact, many measures in science and social science involve a transformation. Some of these are familiar in everyday experience (Hoaglin, 1988).

Reports of earthquakes usually give the magnitude on the Richter scale, which expresses the strength of the earthquake's motion in logarithmic units (base 10).

Measurements of the intensity of sounds customarily produce results in decibels. The fundamental quantity actually measured, however, is sound pressure (in dynes per square centimeter, e.g.), and the "sound pressure level" (in decibels) is related logarithmically to sound pressure.

Although ratings of automobiles' fuel economy are usually given in miles per gallon in the United States, these often come from measurements that

determine the number of gallons used on a standard test course. Thus, the familiar miles-per-gallon figures result from a reciprocal transformation. In fact, throughout the rest of the world, fuel efficiency is reported in units of liters per 100 km—the reciprocal (gasoline volume per distance) of miles per gallon. (The constant multiple needed to convert from metric to old British units does not affect the distribution.)

EDA teaches that we should always consider whether an alternative form of a variable might allow a simpler model or description. This aspect of EDA draws on more than 70 years of work in statistics (Bartlett, 1947; Tukey, 1957; Box & Cox, 1964; Kruskal, 1968; Emerson, 1983; Emerson & Stoto, 1983). This work has identified a variety of benefits of a wisely chosen re-expression:

1. The distribution of data and residuals can be made more symmetric and more nearly normal.
2. The variances of several groups to be compared can be made more nearly equal.
3. The relationship between two variables can be made more nearly linear.
4. The variation of points around a regression line can be made more nearly equal across the span of the data.
5. A linear relationship between two variables can be made more nearly parallel for groups of values in the data.
6. The appropriateness of an additive model for two or more factors can be improved (and the need for interaction terms reduced or removed).

In traditional terms, the symmetry achieved by Benefit 1 is necessary for the mean to summarize the center of the distribution. Normality is expected for *t* tests, and normal fluctuations are assumed for linear models such as regression and ANOVA. The equality of variance among groups (Benefit 2) is assumed by ANOVA and ANOCOV models as well as for a pooled *t* test. Linearity (Benefit 3) is fundamental to regression methods, which also require the residuals to have constant variance everywhere (Benefit 4) for common inference methods. Both the use of dummy variables in regression and the generalization of that idea to ANOCOVs require that linear models for subgroups be parallel (Benefit 5). And the ANOVA model calls for additivity (Benefit 6).

One remarkable insight noted by the authors is that re-expressions that improve one aspect of the data often improve several—or even all—of the others.

A Re-Expression Example

The dexterity experiment offers a good example. Many task-based measures of performance record the time for a subject to complete a task. These include the classic mouse-running maze tasks as well as cognitive function tests, such as the Trail Making Test (Reitan, 1958) and the Stroop Test (Stroop, 1935).

Thinking about re-expression encourages us to consider constraints on the distribution of the data values. For *time per task*, it is not possible to get a value below some lower limit—certainly not less than zero. But there is no upper limit. Indeed, some subjects may not be able to complete some tasks at all. (The dexterity measure is used in cognitive-based training of postsurgery patients who have suffered severe hand or arm injuries, some of whom cannot complete the task in any reasonable amount of time. In addition, mice under sufficiently stressful conditions may give up on solving a maze.) These constraints introduce two problems for data analysis. First, the distribution of values is almost certain to be skewed to the high end—and the stem-and-leaf, histogram, and boxplot in Figures 3.5, 3.1, 3.3, and 3.4 show the expected skewness. And, second, we must cope with *infinities* for subjects who do not complete the task at all.

Researchers have often ignored skewness, hoping it would not be severe enough to invalidate statistics applied to the data. A variety of ad hoc methods have been suggested for the infinities, including just substituting "some large value" and omitting them from the data. Neither ignoring the data nor assigning some large value is suitable, because either approach can severely distort statistical analyses. Some authors have suggested that nonparametric methods would be more appropriate because they are insensitive to both skewness and outlier problems. But these methods often restrict what we can see in the data.

Exploratory data analysis suggests that, by reconsidering the way the data are recorded, we may gain additional insights and understanding. And, along the way, we can obtain simpler models for how the data behave. For the dexterity results, EDA-trained analysts would immediately think of looking at the *reciprocal* of the recorded times. These values have units of *tasks per second*—a measure of speed rather than duration. To make the units easier to deal with, we might multiply by 60 to obtain *tasks per minute*. A scale change or a shift in values from adding or subtracting a constant has no effect on the pattern of the values.

The reciprocal addresses both of the problems we noted. Infinities are now dealt with rationally as a speed of zero tasks per second. And re-expressing data as $1/x$ has the effect of pulling in the high tail relative to the low one. For example, Figure 3.8 shows the histogram and boxplot of the speeds that correspond to the durations in Figures 3.1 and 3.3. The boxplot looks much more symmetric, and one data value is outside at the upper end—a 14-year-old whose speed was 5 tasks per min.

As most discussions of re-expression have noted, appropriate re-expression can improve the relationships among variables as well as their individual distributions. Figure 3.9 shows the scatterplot of speed versus age, with symbols assigned according to hand as in Figure 3.6.

Now the relationship is nicely linear—suitable for a regression model. The slope has changed from

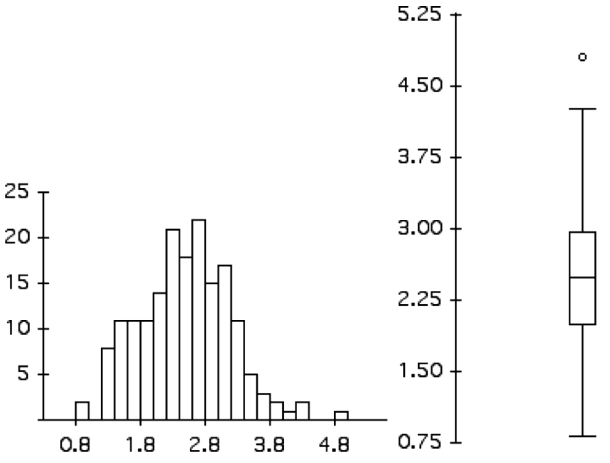

FIGURE 3.8. Histogram and boxplot of the speeds (in reciprocal seconds) of the 175 normal subjects on the test of manual dexterity.

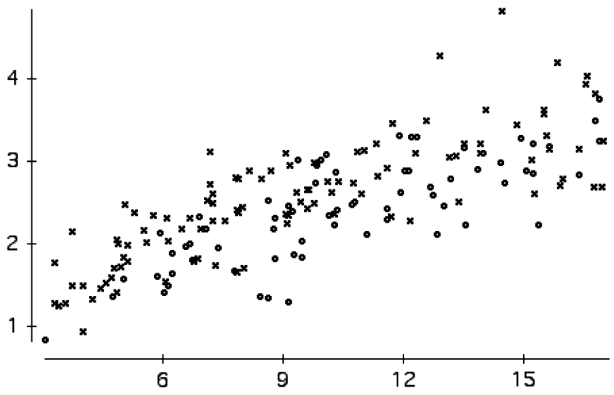

FIGURE 3.9. Scatterplot of speed (in reciprocal seconds) versus age (in years). The plotting symbol distinguishes the dominant hand (×) from the nondominant hand (o).

negative to positive, corresponding to the change in the meaning of the response variable. Durations are shorter for older subjects, so speed improves with age. The variation in speed is roughly constant across the age range.

Occam

William of Occam (c. 1288–c. 1348) is known for asserting that the simpler explanation that accounts for the facts is generally better. Years of experience in consulting and data analysis, along with the advice of prominent statisticians who have amassed far more experience, have convinced us that this is excellent guidance for analyzing data. Simpler models are not only easier to understand and explain; they are more likely to lead to future advances.

We often choose a re-expression such as the logarithm or square root because it works to simplify the analysis. Appropriate re-expression has merit for this reason alone. Models built to fit appropriately re-expressed data are both easier to understand and more likely to lead to further advances.

The cynical (but quite correct) view is that proper re-expression and data exploration are more likely to answer the most important question a researcher can ask: "What should my next grant proposal be about?" For the clinician, an analysis of appropriately re-expressed data is likely to give simpler diagnostic rules and guidelines.

For example, in Figure 3.10 the original scale (time) does not make it easy to summarize the effect of dominant versus nondominant hand. In the

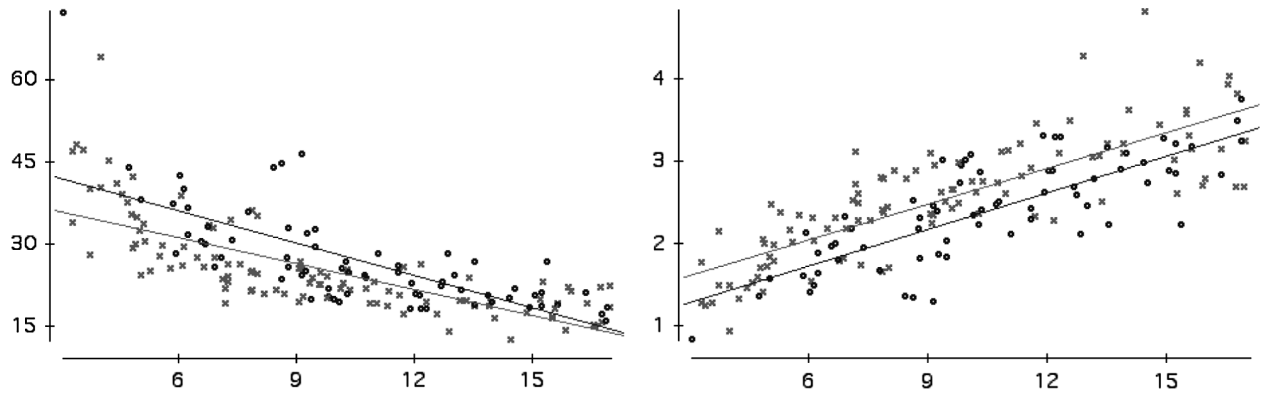

FIGURE 3.10. Scatterplots with separate regression lines for dominant (×) and nondominant (o) hands. (a) Time (in seconds) versus age (in years). (b) Speed (in reciprocal seconds) versus age (in years). The dominant hand takes less time (lower line in the left plot) and is, correspondingly, faster (upper line in the right plot).

reciprocal scale (speed), the patterns for the two hands are linear and parallel, offering the summary that the dominant hand is about 0.3 tasks per minute faster at any of the ages studied.

EXTENDING THE ANALYSIS

One purpose of the dexterity trials was to investigate the phenomenon of crossover training. As we have seen, subjects improve with training up to about the third session. Researchers wanted to investigate whether performance in one hand is improved by training the *other* hand. The question is of importance to cognitive researchers because it addresses issues of how learning takes place. It is of clinical importance to therapists, who do not want to tire out an injured limb by training and wonder whether training the other, uninjured limb would be an appropriate protocol.

In Figure 3.11 the boxplots show that, overall, training the opposite hand does lead to improvement. A parallel-coordinate plot of the speeds in the tested hand before and after training the other hand shows the individual improvements and would support investigating, for example, whether handedness matters.

DATA ANALYSIS ETHICS

We set outliers aside for special consideration and re-express to improve the ability of models to fit the data. Both of these practices can improve the *p*-values of hypothesis tests. Are we cheating?

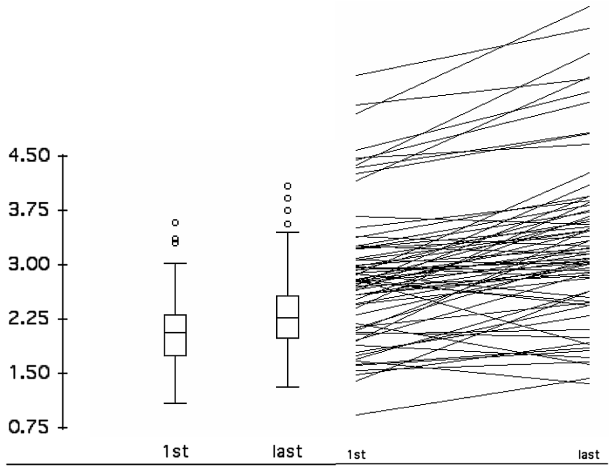

FIGURE 3.11. Boxplots and parallel-coordinate plot of speeds (in reciprocal seconds) before and after training the opposite hand.

This question cuts to the core of the difference in philosophy between EDA and the way traditional statistical inference is often used in psychology. Those who look to statistical methods to guard against the misuse of data may be scandalized by our proposal that a scientist should be free to seek the most appropriate re-expression and should focus on modeling the main body of the data and not the occasional extraordinary case. We reject the idea that statistical methods can or should serve as enforcers of honest data analysis. The goal should be truth about the world. Truth is, of course, a challenging goal, and one that we may rarely, if ever, attain. Nevertheless, it should be our guiding principle. (Philosophers call such a goal a *regulative ideal*.)

Methods alone can never approach this ideal. It requires the honest efforts of the scientists and social scientists who do the research and analysis. In short, if a researcher wants to cheat, no amount of standardized statistical practice can stop it. Indeed, it may be easier to hide misconduct by following so-called standard practice and reporting only the usual summary statistics. We should not constrain honest researchers in an attempt to restrain less honest ones. Moreover, we should teach the ethics of honest research as a fundamental part of the education of researchers in all fields (Velleman, 2008).

Even honest researchers share the well-known human tendency to imagine patterns in data. In a provocative essay, Diaconis (1985) examined this aspect of EDA and discussed a variety of remedies:

- Publish without *p*-values
- Try to quantify and correct for the distortion
- Try it out on fresh data
- Borrow strength[2]
- Cross-validate
- Bootstrap the exploration.

He placed EDA in the context of theories of data analysis and concluded that "the new exploratory techniques seem to be a mandatory supplement to more classical statistical procedures" (p. 32).

MEASUREMENT SCALES

One tradition in psychology categorizes measurement scales as nominal, ordinal, interval, or ratio. The work in measurement arose from a debate about whether psychology could be a true science when many measurements made by psychologists were not comparable to those made by physicists. Although the argument for psychology as a science has largely been won, it has led some to suggest that the measurement scale of a variable should constrain the appropriate analysis methods. Often, for example, this line of reasoning has been used to argue in favor of nonparametric methods.

But this approach misunderstands measurement scales in many situations. Velleman and Wilkinson (1993), for example, noted that the measurement scale of a variable is not a property of the variable itself, but rather a property of how it is used. They offer a number of examples of common variables that may be viewed as having one measurement scale in one context and a different scale in another.

A little thought reveals many common examples. Playing card suits appear to be nominal, but are ordered in bridge. Playing card values can be ratio scaled in some games (casino), ordinal in others (poker), and nominal in still others (Go Fish). Velleman and Wilkinson (1993) offered the example of the consecutive numbers on tickets for a door prize handed out as attendees arrive at a meeting. These are nominal when selecting the winner but could be used (in interval scale) to count the number of attendees or (in ordinal scale) to record the order of their arrival.

We approach measurement scales by exploring data without any assumptions but then asking whether the best models and descriptions found for the data can be supported by the ways in which the data were measured. Surprisingly often, we have discovered a richness in data that was not evident at first. In short, exploratory data analyses deemphasize measurement scales—at least, until the final summary of the analysis.

Kinds of Data

Although it rejects measurement scale as a constraint, EDA does categorize data by types to offer guidance for re-expression. Mosteller and Tukey (1977, Chapter 5) suggested the categories and recommended re-expressions in Table 3.1.

REGRESSION

The EDA approach extends beyond elementary data summaries. Analyses that use multiple regression or ANOVA should consider the benefits of re-expression and of removing outliers.

Most statistics programs compute diagnostic statistics that can help identify both influential cases

[2]Borrowing strength is a general term that refers to seeking support for estimating values and, especially, variances from other parts of the data or other sources. When we fit a regression, we borrow strength from all the data (and the assumption that a linear model is appropriate) to enable more precise estimates and predictions for individual values.

TABLE 3.1

Summary of Re-Expressions Suggested for Various Types of Data

Type of data	Suggested re-expressions
Amounts	Nonnegative real values. Logarithms are a good first guess. Natural and base-10 logs differ only by a constant multiplier, but base-10 logs are usually easier to interpret. (Times and rates are examples of amounts although, as we have noted, rates—because they are ratios—often benefit from a reciprocal transformation.)
Counts	Nonnegative integer values whose units are "number of. . . ." Square roots and logs are a good place to start.
Balances	Real values that may be positive or negative. Often these arise as a difference between two amounts (and then re-expressing those amounts may be helpful). Balances often need no re-expression.
Counted fractions	Fractions of a whole such as percentages (100 × number counted in a group/total group size). Special re-expressions that acknowledge the boundaries of these data at both ends, such as the logit, may be helpful.
Ranks	Integer values recording order. Treat like a counted fraction.
Grades	Ordered groups such as Freshman/First-Year, Sophomore, Junior, Senior. Little need to re-express.
Names	Nominal data. If the data simply name individuals, re-expression offers no advantage.

and unexpected patterns. We recommend examining the following:

- Leverage. The leverage of any case in a simple or multiple regression is the amount by which that case's predicted value would change if the dependent value of the case were changed by one unit and the regression recomputed. Cases with particularly large leverage can dominate the fitting of a regression model.
- Studentized Residuals. The standard deviation of the sampling distribution of a regression residual depends on the case's leverage. Cases near the multivariate mean of the *x*'s have smaller variance than those far from the center of the data. Studentized residuals adjust for this. A scatterplot of the studentized residuals versus the predicted values has had the linear effects of the predictors removed and is adjusted for differences in the underlying variation of the prediction errors. Consequently, it displays any underlying patterns more clearly and vividly. This makes it an effective tool for assessing whether the relationship of the response variable to the predictors is linear (and, thus, whether the regression model is appropriate).
- DFFITS. The DFFITS statistic for each case shows the impact of omitting that case on the corresponding predicted value. This leave-out-one diagnostic combines leverage and studentized residual in a single measure.
- DFBETAS. When the *coefficients* of predictor variables are of interest, the DFBETAS statistics provide a suitably scaled measure of the change in each coefficient associated with omitting each case.

Most statistics packages have options to report all four of these statistics for a regression analysis. Each can be used to identify cases that may deserve special attention because of their undue influence on the regression model or because they deviate sharply from the pattern fitted by the regression model.

Diagnostic plots can be especially helpful for exploring multiple-regression models. Displays of residuals are commonly offered by all statistics software and should be examined for the same features that one would look for in a single variable.

In a multiple regression, it is common to make a scatterplot of the residuals against the predicted values. Displays of studentized residuals are more useful because of their stabilized standard deviations.

When a coefficient is of interest, it is important to interpret it correctly: The coefficient for a particular predictor indicates how the dependent variable changes in response to change in that predictor after adjusting for the linear effects of the other predictor variables (in the data at hand). In general, no simpler language is satisfactory. We often read (even in textbooks) that a regression coefficient estimates the amount of change in the dependent variable when its predictor changes by one unit and all other

predictors are held fixed. It is straightforward to show mathematically that such an interpretation is incorrect—unless the data have been collected in a way that explicitly holds the other predictors fixed (e.g., in a well-controlled experiment, to which we turn next). More important, EDA offers displays that help to interpret regression coefficients correctly.

One particularly useful exploratory display, offered by most modern statistics programs, is the *partial regression plot*. This scatterplot displays the relationship between the response variable and any selected predictor variable after adjusting for simultaneous linear change in the other predictor variables. The plot thus displays exactly what the coefficient of the selected predictor means. Specifically, the plot has a least squares slope equal to the coefficient of the selected predictor in the multiple regression model, and it has the same least squares residuals as the full multiple regression. It is an excellent tool for understanding how consistently the selected predictor fits the response (by judging the variation in the residuals) and for diagnosing unusual behavior in the data that may affect that particular coefficient (Cook & Weisberg, 1982; Velleman & Welsch, 1981).

ANALYSIS OF VARIANCE

The term *analysis of variance* customarily describes analyses of data that involve two or more factors and have one or more observations for each possible combination of the levels (or versions) of all the factors. Beyond such *factorial designs*, a wide variety of designs use balanced subsets of the possible combinations.

EDA emphasizes analyzing the data first and only later summarizing the contributions of various sources of variability (Hoaglin, Mosteller, & Tukey, 1991). Natural initial steps include looking at the data and considering the possibility of re-expression. We illustrate these ideas in the context of a classical set of data from a difference limen experiment for which a variety of illustrative analyses have been published on several occasions (Green & Tukey, 1960; E. G. Johnson & Tukey, 1987; P. O. Johnson, 1946; P. O. Johnson & Tsao, 1944).

The experiment involved eight subjects, "two persons in each cell of a 2 × 2 design for male versus female and sighted versus [congenitally] blind." Subjects were asked to detect a change in the weight pulling on a ring. The controlled treatments consisted of "two *Dates* (1, 2), four *Rates* (50, 100, 150, and 200 grams per 30 seconds), and seven (initial) *Weights* (100, 150, 200, 250, 300, 350, and 400 grams). The experimental procedure involved attaching a pail by a lever system to a ring on the subject's finger. One of the seven initial *Weights* was placed into the pail, and then water was allowed to flow into the pail at one of four constant *Rates* until the subject reported a change in pull on the finger. The intended response, the difference limen (*DL*), was measured by the amount of water added [up to the] time of report. Five determinations were made for each of the 28 *Rate* × *Weight* combinations, and the average of these values was used as the response. The entire experiment was conducted, for each person, at each of two *Dates*, one week apart" (E. G. Johnson & Tukey, 1987, p. 174).[3] The design calls for an analysis predicting *DL* from *Date, Rate, Weight, Sight, Sex,* and *Subject* (nested within the combination of *Sight* and *Sex*). That ANOVA is shown in Table 3.2.[4]

EDA teaches that we should *always* examine the residuals in an analysis because subtracting a summary of the patterns found so far is likely to reveal additional patterns or further evidence about the current summary. Plotting the residuals from the ANOVA of Table 3.2 against the predicted values (Figure 3.12) reveals that *DL* as originally measured does not satisfy the assumptions of ANOVA. In particular, it is clear that the variance of the response is not constant across the values of the factors studied. The pattern resembles a

[3]In this quotation, we have added italics and initial capitals to names of factors for consistency with usage in this chapter.

[4]Our results do not exactly match those of P. O. Johnson and Tsao (1944) or those of Green and Tukey (1960). P. O. Johnson (1946) used these data to illustrate complex ANOVA calculations and found a model with all interaction terms but without a *Subject* term. Green and Tukey followed a path closer to the one we are about to discuss. All of the previous authors had to perform these rather daunting calculations by hand, so errors may have crept in.

TABLE 3.2

Analysis of Variance Table for Difference Limen (DL)

Source	*DF*	Sum of squares	Mean square	*F*-ratio	*p*-value
Const	1	276799	276799	46.296	0.0024
Date	1	115.629	115.629	1.6582	0.1985
Rate	3	51077.6	17025.9	244.16	≤ 0.0001
Weight	6	1913.63	318.939	4.5737	0.0002
Sight	1	11816.3	11816.3	1.9763	0.2325
Sex	1	31543.7	31543.7	5.2758	0.0832
Subject (Sight × Sex)	4	23915.6	5978.89	85.740	≤ 0.0001
Sight × Sex	1	14136.8	14136.8	2.3644	0.1989
Error	430	29985.3	69.7332		
Total	447	164504			

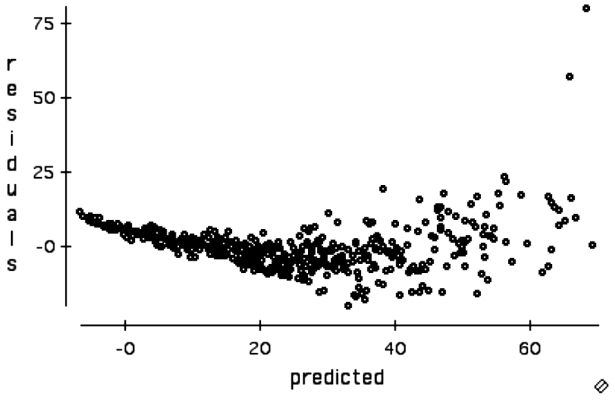

FIGURE 3.12. Residuals from the analysis of variance for *DL* on *Date*, *Weight*, *Rate*, *Sight*, *Sex*, and *Subject* (*Sight* × *Sex*) show nonlinearity and a fan shape that calls for re-expression or reformulation.

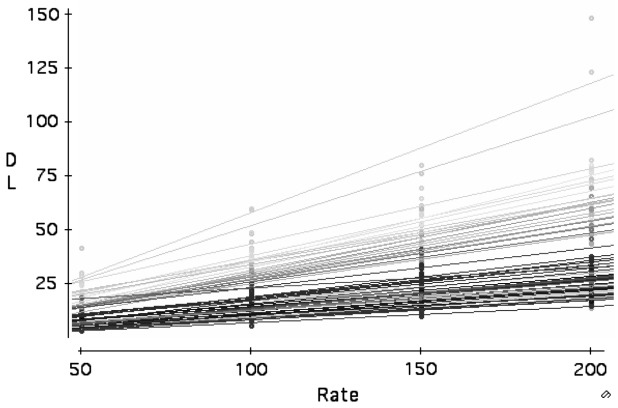

FIGURE 3.13. *Difference Limen* plotted against *Rate*. Lines are least squares fits for each *Subject* × *Weight* × *Date* combination.

fan, opening to the right (larger residuals tend to belong to larger predicted values), and also shows some curvature.

Green and Tukey (1960) discussed the importance of the scale used for the response variable:

> We want to choose a scale that will yield the simplest relations with the independent variables. By simplest relations we mean, for example, fewer important interactions, and larger main effects relative to the error variance. A change of variable that nearly removes a particular main effect also usually leads to very revealing results. Secondarily, we would like the dependent variable to have approximately homogeneous variance within cells of the design. (p. 128)

When variance increases with predicted value (as in Figure 3.12), functions such as the square root, logarithm, and reciprocal are likely to help. For these data, the units of the variables suggest an alternative way to address the first goal mentioned by Green and Tukey. The *Difference Limen* is measured in grams. The *Rate* factor is measured in grams per 30 sec. If we plot the *DL* against *Rate* and add lines to the plot for each *Subject* × *Weight* × *Date* combination, we obtain Figure 3.13.

But if we reformulate *Rate* as *Time* (in seconds) until the subject declares a felt change, we obtain Figure 3.14. (The calculation is *Seconds* =

(*DL/Rate*) × 30.) The fact that response *Time* is virtually constant for the four rates of increase (for each combination of *Subject*, *Weight*, and *Date*) suggests that one can get whatever difference limen one wants by choosing the *Rate* appropriately. That observation argues strongly that converting the average *DL* to an average *Time* (in seconds) will produce a simpler analysis. P. O. Johnson and Tsao (1944) found this relation (as an interaction in their analysis of average *DL*). Green and Tukey found it and made the argument for using *Time* as the response that we recount here. Technically, this type of change is sometimes called *reformulation* rather than re-expression because it does not involve applying a simple mathematical function to each data value. When available, it is another useful technique in the analyst's toolkit.

Although the summary is simpler for the data in seconds, the residual plot in Figure 3.15 still shows a bend, increasing spread from left to right, and possible subgroups. An analysis of Log_{10} *Seconds* yields the residual plot in Figure 3.16. This pattern is more like what we hope for: roughly flat, with similar variation in the residuals across the range of predicted values and no bends. The overall pattern in Figure 3.16 is horizontal and homoskedastic. The residuals form clusters. Further checking shows that each subject is a cluster, although some subjects' clusters overlap. The ANOVA table presented in Table 3.3 shows significant effects for *Date* and *Weight*. Because this ANOVA treats *Subject* as a random factor (and the other factors as fixed), we do not focus on the significance of the effects for the individual subjects; the *p* value for *Subject* arises from comparing the *Subject* mean square against the residual mean square. Although they have much larger mean squares than *Date* and *Weight* (and *Rate*), *Sight*, *Sex*, and *Sight* × *Sex* are far from significant because the denominator for their *F* ratios is the *Subject* (*Sight* × *Sex*) mean square. Green and Tukey (1960) discussed choices between fixed and random for the various factors and the reasons for them.

From an EDA perspective, the analysis does not end with an ANOVA table. The customary next step examines the individual values for each line in the table: the main effects for *Date*, *Rate*, *Weight*, *Sight*, and *Sex*; the interaction effects for *Sight* × *Sex*; the

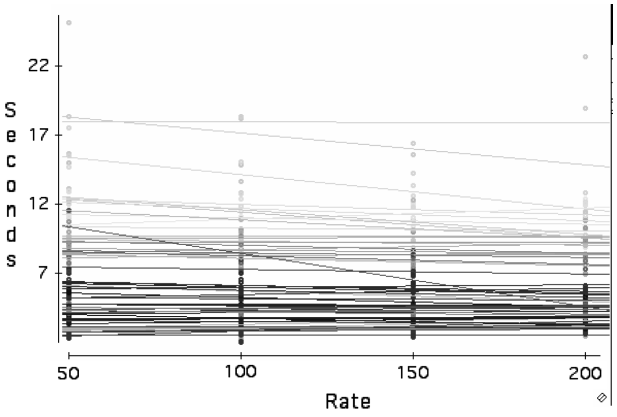

FIGURE 3.14. Response in seconds is almost constant versus *Rate* for each *Subject* × *Weight* × *Date* combination.

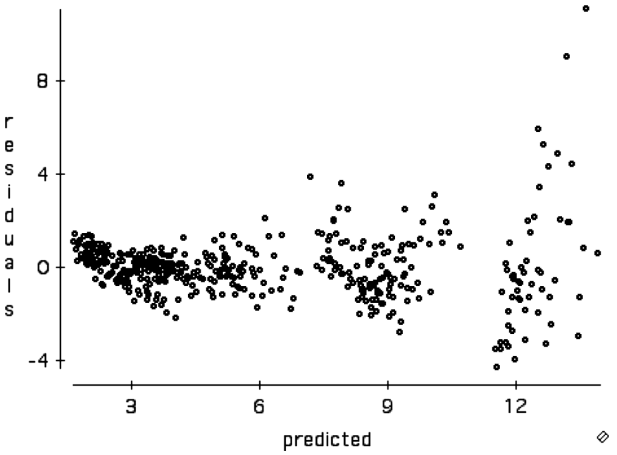

FIGURE 3.15. Plot of residuals versus predicted values for the ANOVA of *Seconds* on *Date*, *Weight*, *Rate*, *Sight*, *Sex*, and *Subject* (within *Sight* × *Sex*).

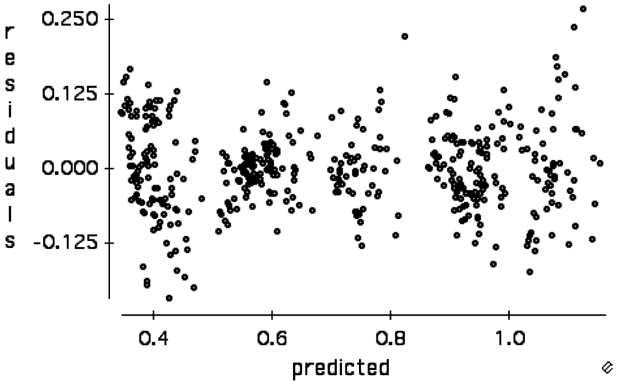

FIGURE 3.16. Plot of residuals versus predicted values for the ANOVA of Log_{10}*Seconds* on *Date*, *Weight*, *Rate*, *Sight*, *Sex*, and *Subject* (within *Sight* × *Sex*).

TABLE 3.3

Analysis of Variance Table for Log_{10} *Seconds*

Source	*DF*	Sum of squares	Mean square	*F*-ratio	*p*-value
Const	1	223.836	223.836	80.238	0.0009
Date	1	0.093378	0.093378	17.592	≤ 0.0001
Rate	3	0.012418	0.004139	0.77983	0.5057
Weight	6	0.332250	0.055375	10.433	≤ 0.0001
Sight	1	3.79149	3.79149	1.3591	0.3085
Sex	1	7.70730	7.70730	2.7628	0.1718
Subject(Sight × Sex)	4	11.1586	2.78964	525.56	≤ 0.0001
Sight × Sex	1	3.40870	3.40870	1.2219	0.3310
Error	430	2.28240	0.005308		
Total	447	28.7865			

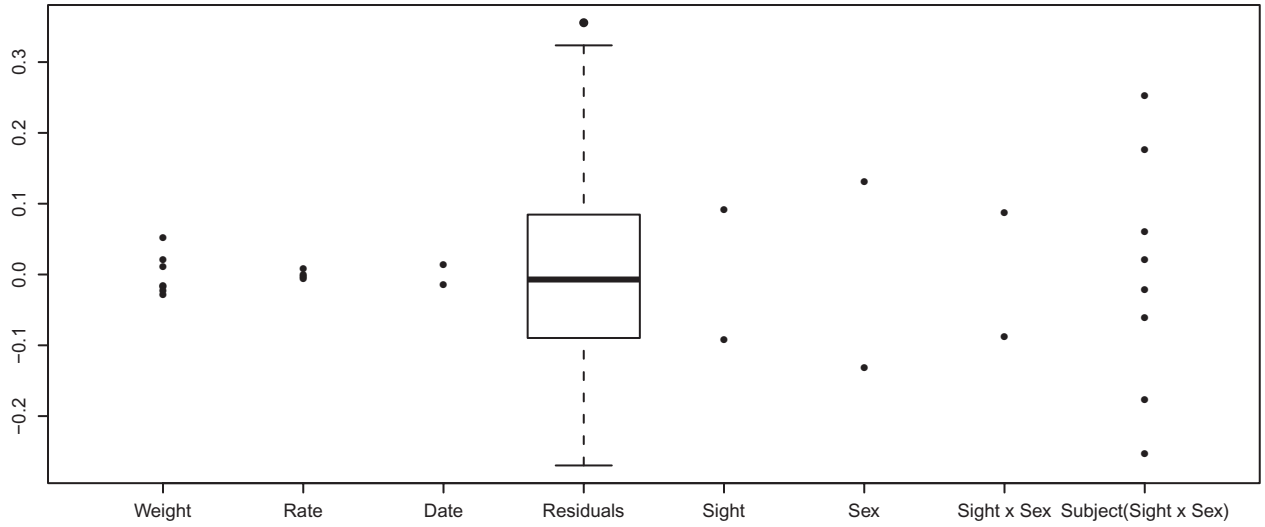

FIGURE 3.17. Effects and residuals for the ANOVA of Log_{10} *Seconds* on *Date*, *Weight*, *Rate*, *Sight*, *Sex*, and *Subject* (within *Sight* × *Sex*).

effects for *Subject* (nested within *Sight* × *Sex*); and the residuals. Figure 3.17 shows dot plots of the various effects and a boxplot of the residuals; we chose the order because the residuals provide the denominator for *Weight*, *Rate*, and *Date* and the *Subject* effects provide the denominator for *Sight*, *Sex*, and *Sight* × *Sex*. Because each set of effects sums to zero (in two ways for *Sight* × *Sex*), the effects for the lines other than *Weight* and *Rate* consist of positive and negative values with the same magnitude. Such a display makes it easy to compare the sizes of the effects for the various parts of the model.

The variation among the residuals and among the *Subject* effects dominates the display. The effects for *Date* appear small alongside the residuals, but each of those effects represents a mean of 224 observations. On the other hand, the effects for *Sight*, *Sex*, and *Sight* × *Sex* do not seem much smaller than the *Subject* effects, but each *Subject* effect represents a mean of 56 observations. Relative to the variation in the *Subject* effects, the effects for *Sight*, *Sex*, and *Sight* × *Sex* are not large enough to reach significance.

Of the two experimental factors, different *Weights* have noticeable effects, although they are not very large compared with the variation in the residuals. By contrast, differing *Rates* have little effect at all.

As Figure 3.18 shows, the effects for *Weight* are systematically related to the level of weight. Subjects were able to more quickly discern a change when

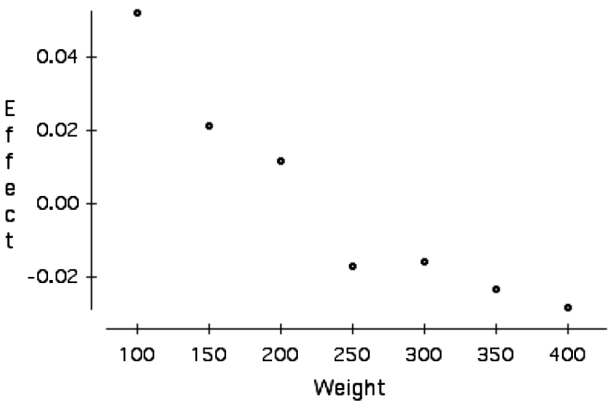

FIGURE 3.18. The effects of *Weight* generally fall consistently with increasing weight.

the initial weight was small than when it was larger—a result consistent with Weber's law but not a particularly exciting outcome for so complex an experiment.

Exploratory ANOVA can go further (Hoaglin et al., 1991), but we stop here for this chapter. The choice to reformulate the response as seconds and then re-express it as log-seconds may seem specific to this analysis and difficult to generalize, but other analyses present similar opportunities. And for re-expression, EDA has specific techniques for letting the data guide the choice (Emerson, 1983; Emerson & Stoto, 1983).

THE EDA PROCESS

Data exploration is a diagnostic procedure. The data analyst attempts to detect and characterize any aspect of the data that may help in understanding the underlying pathologies. Diagnosis requires an open mind and a willingness to expect the unexpected. Checklists have proved helpful in clinical medicine (e.g., Gawande, 2010). In that spirit, we offer a data exploration and diagnosis checklist. As with many checklists for complex processes, this one is divided into shorter, focused checklists for each of the large steps of analysis.

Checklist I: Display the Data

A. Look at the distributions of the variables with a stem-and-leaf display, histogram, or dot plot.
1. Be alert for multiple modes
   a) Consider splitting data into subgroups
2. Be alert for outliers
   a) Identify, understand, correct (if possible), or set aside
3. Check for skewness
   a) Consider re-expression to improve symmetry

B. Compare groups with boxplots.
1. Check for skewness within groups
   a) Consider re-expression to improve symmetry
2. Check that groups have similar spreads
   a) Consider re-expression to promote constant spread (spread-vs.-level plot)
3. Check for outliers
   a) Identify, understand, correct (if possible), or set aside

C. Consider bivariate relationships with scatterplots.
1. Check for approximate straightness
   a) Consider re-expression to improve straightness
   b) Usually re-expressing *y* is preferred, as it is likely also to make the variability of *y* more nearly constant across the range of *x*
2. Check for constant variance of *y* for all *x* values (homoskedasticity)
   a) Consider re-expressing *y*, especially if larger values of *y* are more variable
3. Check for local clusters, parallel trends, or other evidence of subgroups
   a) Consider splitting the data into subgroups
4. Check for outliers
   a) Consider both *y*-direction outliers and *x*-direction high-leverage points
      i) The influence of a case depends on both
      ii) Identify, understand, correct (if appropriate), or consider setting aside or treating specially

D. Consider multivariate relationships with scatterplot matrices and rotating plots.
1. Check for approximate straightness
   a) Consider re-expression to improve linearity
   b) Scatterplot matrices make it easy to notice variables that would benefit from

re-expression in several bivariate relationships; re-express those variables first
2. Check for local clusters, parallel trends, or other evidence of subgroups
   a) In rotating plots, rotate so that the main trend is perpendicular to the screen to look for subgroups
   b) In scatterplot matrices, use plot brushing to identify clusters in one scatterplot and check whether they stand out in other scatterplots
   c) Consider splitting the data into subgroups
3. Check for outliers
   a) In scatterplot matrices, select points that stand away from the other data to see whether they may be unusual in other scatterplots; they may be outliers in a multivariate sense
   b) Be alert for points that may be multivariate outliers but are not unusual on any univariate or bivariate display
      i) Diagnostic statistics are available to help with this (see discussion of regression)

If the exploratory displays in the first checklist suggest that it is appropriate to summarize variables or model their behavior, then use Checklists II through V, depending on the number and structure of the variables.

Checklist II: Summarize and Describe Individual Variables

A. Summarize individual variables with resistant measures.
1. Median
2. Quartiles
   a) Several definitions of quartiles are in use, but it matters little which you use
3. Q-spread (interquartile range [IQR])
4. Extremes (minimum and maximum)

B. Summarize individual variables with traditional maximum-likelihood methods.
1. Mean
2. Standard deviation
3. Confidence interval
   a) But check displays to be sure assumptions are plausible

Checklist III: Compare Multiple Groups

A. Use ANOVA to compare group means.
1. Consider the form of the response variable
   a) Use spread-vs.-level plot to check for need to re-express
   b) Plot residuals against predicted values and look for violations of homoskedasticity and for curvature
2. Check residuals for outliers

Checklist IV: Summarize and Describe Relationships Between Pairs of Quantitative Variables

A. Use resistant smoothing to reveal general trends.

B. Fit linear models with least squares regression and check diagnostic displays and statistics.
1. Plot studentized residuals vs. predicted values; check for the following:
   a) Curvature
      i) If so, go back to Checklist I.D.1, consider re-expression, and fit again
   b) Subgroups, especially parallel patterns
      i) Parallel patterns suggest an analysis of covariance approach or the introduction of indicator variables and a multiple regression
   c) Outliers and high-leverage points
      i) Consider correcting, omitting, or treating specially; one possibility is introducing an indicator variable for each errant case

Checklist V: Summarize Multivariate Relationships Involving a Single Quantitative Response Variable and Multiple Potential Quantitative Predictors

A. Fit a least squares multiple regression and compute and examine the residuals.
1. Plot studentized residuals against predicted values
   a) Look for bends—consider re-expressing *y*
   b) Look for heteroskedasticity—consider re-expressing *y*
   c) Look for parallel patterns—consider indicator variables for groups

   d) Look for outliers—consider correcting, setting aside, or treating specially (e.g., with an indicator variable for each)
2. Examine leverage and influence measures; identify influential cases
   a) Histograms and stem-and-leaf displays help to identify extreme values; boxplots can do that conveniently for large numbers of values; leverages are not likely to be normally distributed because they are bounded by zero and one
   b) Set aside influential cases and repeat the analysis to assess their true influence on your conclusions
3. Examine partial regression plots—especially for coefficients that are of particular interest
   a) Check as for simple regressions in Checklist IV

Regression model building is an exploration of several aspects at once. We simultaneously seek effective functional forms to model the data and appropriate variables with which to build our models. Along the way, we identify extraordinary cases and prevent them from dominating our decisions. There is no fixed path. You may choose to examine a response variable plotted against each factor or predictor before moving on to models with multiple factors and predictors. The most important aspect of the EDA approach is that the human analyst is intimately involved at each step, using discipline-based knowledge to guide decisions. For example, one path for building a multiple regression model might go according to Checklist VI.

Checklist VI: Build a Multiple Regression Model With Interactive Steps

A. Choose a promising predictor variable and fit a simple (*y* vs. *x*) regression. Look at plots and diagnostics as in Checklist V. Use the information from displays of this relationship to reconsider the form of the model, possible re-expressions of the variables, isolation of possible outliers, and whether the selected predictor should be replaced with an alternative predictor.

B. Consider the relationships between the residuals and remaining available predictors. Be alert to opportunities for re-expression and the possibility that other cases should be isolated as outliers. Select one or more of the available predictors (after possible re-expression and outlier deletion) to add to the model. Diagnose as in Checklist V.

1. Isolate extraordinary cases by assigning individual indicator variables to them and including those in the model

C. Iterate. At each iteration it may be appropriate to proceed stepwise—that is, introduce one predictor at a time or introduce collections of conceptually related predictors together. The most important idea is to continually monitor and diagnose the developing model, exploring for unanticipated relationships, clusters, outliers, and violations of assumptions.

Designed experiments are likely sources of data with a quantitative response and two categorical factors. In such cases the factors are defined and controlled. However, this form of data can also arise from observational studies, in which case the factors may be observed and not under control.

Checklist VII: Summarize Multivariate Relationships of a Single Quantitative Response Variable and Two Categorical Factors

A. Median polish.

B. Look for re-expressions to improve additivity.

1. Use the diagnostic plot for a two-way table to suggest a function for re-expression[5]

C. Plot residuals. If they are reasonably symmetric and not heavy-tailed, consider fitting the ANOVA model (i.e., summarize by means).

1. Diagnose ANOVA residuals as in Checklist VI; check for heteroskedasticity and curvature
2. Diagnose possible outliers

[5]Emerson and Hoaglin (1983) and Emerson (1983) discussed the diagnostic plot for a two-way table and its background. The plot is a graphical descendant of Tukey's one degree of freedom for nonadditivity (ODOFFNA; Tukey, 1949).

Checklist VIII: Summarize Multivariate Relationships of a Single Quantitative Response Variable and Multiple Categorical Factors

A. If data come from a designed experiment, explore the ANOVA corresponding to the design.

1. Examine boxplots of the residuals at each level of each factor and possibly for combinations of factors
   a) Look for outliers—treat them as before
   b) Look for skewness—consider re-expressing *y*
2. Explore for evidence of nonadditivity

B. Data from designed experiments can be explored in much the same way as for multiple regression. EDA supports the idea that factors and interactions can be included or removed from a developing model. At each step, the residuals should be checked, and unanticipated clusters, outliers, nonlinearity, or (for analyses of covariances) lack of parallelism should be addressed.

These EDA principles and approaches extend naturally to more complex situations. The general rules are to make displays of individual variables first and deal with any special issues. Then check any models for the data by similarly examining residuals to see whether they reveal anything worthy of special attention. Always be alert to the opportunity to re-express variables to simplify the models or to better satisfy model assumptions. Be willing to isolate outliers and influential cases by setting them aside or by fitting special terms just for them.

COMPETING MODELS

When we explore data with many variables, it is often productive to entertain multiple, competing models, developing each, comparing results, and learning from one to inform the others. This strategy might lead simply to offering multiple alternative models for the data. Or it might develop into a Darwinian competition among the models, with those that develop into less useful or successful forms being abandoned in favor of the more successful models—a "Survival of the Best Fit."

When researchers apply statistics primarily to test hypotheses, they often have a sense of completion. We have stated and tested the hypotheses, reached conclusions about them, and can move on to other topics. EDA does not adopt that attitude. There is never a natural stopping place. As with the larger corpus of science, an exploratory data analysis is never done. There is always the possibility that new data or new understanding can lead us to modify or develop an analysis further. Of course, you can reach a point at which everyone involved agrees that there is little gain from doing more with the data you have. But experience has shown that, even then, a new idea or suggestion can reopen the question.

References

Aaron, D. H., & Jansen, C. W. S. (2003). Development of the Functional Dexterity Test (FDT). *Journal of Hand Therapy, 16*, 12–21. doi:10.1016/S0894-1130(03)80019-4

Bartlett, M. S. (1947). The use of transformations. *Biometrics, 3*, 39–52. doi:10.2307/3001536

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B, 26*, 211–252.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York, NY: Chapman & Hall.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 1–36). New York, NY: Wiley.

Emerson, J. D. (1983). Mathematical aspects of transformation. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 247–282). New York, NY: Wiley.

Emerson, J. D., & Hoaglin, D. C. (1983). Analysis of two-way tables by medians. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 166–210). New York, NY: Wiley.

Emerson, J. D., & Stoto, M. A. (1983). Transforming data. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 97–128). New York, NY: Wiley.

Gawande, A. (2010). *The checklist manifesto.* New York, NY: Metropolitan Books.

Gogola, G. R., Lacy, B., Morse, A., Aaron, D., & Velleman, P. F. (2010, June). *Hand dexterity values for 3 to 17 year-old typically developing children*. Paper presented at the Eighth Triennial Congress of the International Federation of Societies for Hand Therapy, Orlando, FL.

Green, B. F., & Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika, 25*, 127–152. doi:10.1007/BF02288577

Hoaglin, D. C. (1988). Transformations in everyday experience. *CHANCE*, *1*(4), 40–45.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82*, 1147–1149. doi:10.2307/2289392

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81*, 991–999. doi:10.2307/2289073

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). *Fundamentals of exploratory analysis of variance*. New York, NY: Wiley.

Inselberg, A. (2009). *Parallel coordinates: Visual multidimensional geometry and its applications*. New York, NY: Springer.

Jebsen, R. H., Taylor, N., Trieschmann, R. B., Trotter, M. H., & Howard, L. A. (1969). An objective and standardized test of hand function. *Archives of Physical Medicine and Rehabilitation, 50*, 311–319.

Johnson, E. G., & Tukey, J. W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In C. L. Mallows (Ed.), *Design, data, and analysis by some friends of Cuthbert Daniel* (pp. 171–244). New York, NY: Wiley.

Johnson, P. O. (1946). *Statistical methods in research*. New York, NY: Prentice Hall.

Johnson, P. O., & Tsao, F. (1944). Factorial design in the determination of differential limen values. *Psychometrika, 9*, 107–144. doi:10.1007/BF02288717

Kruskal, J. B. (1968). Statistical analysis: Transformations of data. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 15, pp. 182–193). Chicago, IL: Macmillan & The Free Press.

Mlodinow, L. (2008). *The drunkard's walk*. New York, NY: Pantheon Books.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*, 271–276. doi:10.2466/PMS.8.7.271-276

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662. doi:10.1037/h0054651

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics, 5*, 232–242. doi:10.2307/3001938

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics, 28*, 602–632. doi:10.1214/aoms/1177706875

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 448–485). Stanford, CA: Stanford University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Velleman, P. F. (2004). *Data Desk*. Ithaca, NY: Data Description.

Velleman, P. F. (2008). Truth, damn truth, and statistics. *Journal of Statistics Education*, *16*(2). Retrieved from http://www.amstat.org/publications/jse/v16n2/velleman.html

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35*, 234–242. doi:10.2307/2683296

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*, 65–72. doi:10.2307/2684788