Residuals: Trash or Treasure?

Larry H. Ludlow, Ph.D.

Boston College

Introduction

One of the significant challenges to the successful teaching of statistics is that of giving the topic some relevant historical perspective. Such an effort is worthwhile because many students need a context within which they can understand why we are studying a given technique and how that technique came about. In fact, many of our traditional statistical procedures can be associated with specific interesting individuals or time periods. For example, the history of probability estimation (as a study of gambling behavior) and the development of correlation and regression (Galton and his study of individual differences) are familiar interesting examples. We can also paint a broader picture when we consider the development of schools of philosophy and their attendant methods of observation and analysis (a good overview is provided by Butterfield, 1957). The purpose of the present paper is to illustrate that progress in understanding the world around us can be grasped in terms of efforts to explain unexpected events under some existing theory and mathematical model — unexpected events that ultimately led to re-formulations of both theories and models. In particular, this paper will argue that current psychometric models of item response data not only have an interesting history in their own right but may also be considered the result of progressive efforts focused on explaining unexpected observations, i.e., item responses.

Unexpected observations are noticeable because they are discrepancies in behaviors expected under existing models. Technically, the difference in what is predicted by a model and what is actually observed is termed a residual. While some researchers seem to dismiss the residual as a nuisance and a distraction (i.e., garbage to be ignored), other researchers consider the residual a key to progress in the development of theories and models (i.e., a gold nugget). Evolution in scientific model building, from this perspective, may be characterized as an effort to reduce residual variation.

Science and Error

The Scientific Revolution, for example, brought about an appreciation that unaccounted phenomena could be important, occasionally more important than the original research. An unexpected result, experimental discrepancy, residue, or "residual phenomena" came to be defined as whatever remains outstanding and unaccounted for after "subducting the effect of all known causes, as well as the nature of the case permits, either by deductive reasoning or by appeal to experience" (Herschel, 1851, art. 158). Herschel underlines the role played by residuals: "Almost all the greatest discoveries in astronomy have resulted from the consideration of what we have elsewhere termed residual phenomena…" (Herschel, 1871, art. 856). In effect, the "expected unexpected" became sought after (Brannigan, 1981, p. 159).

A hallmark of the Scientific Revolution was the emphasis placed on testing one's theories experimentally. As the practice of experimentation became commonplace, the physical sciences began to make rapid progress in the formulation of laws of nature. Myths and Aristotelian logic

are examples of philosophical paradigms that were no longer taken as a substitute for experimentation. Early experiments, however, led to qualitative theories, e.g., the phlogiston theory of combustion (Roberts, 1989). Qualitative theories left qualitative residuals. Consequently, there was no way to assess quantitatively the surprise or importance associated with discrepancies from expected outcomes.

It was eventually recognized that qualitative statements allowed too much leeway in a stated relationship. As experimental techniques improved, qualitative laws were either transformed into quantitative laws (e.g., Boyle's Law replaced the "spring of the air") or discarded as wrong (phlogiston theory replaced by oxygen theory). The scientific attitude shifted: if the purpose of a law was to explain and predict observations, then the law had to be quantified. "Indeed, it is a character of all the higher laws of Nature to assume the form of precise quantitative statement" (Herschel, 1851, art. 116).

To arrive at quantified laws meeting Herschel's description requires precise and accurate observations. Kepler, for example, was familiar with, but could not use, the inaccurate data compiled by Ptolemy — his law of elliptic orbits required the accurate measures of Brahe. In fact, Ptolemy's data confounded the efforts of astronomers for generations because there were so many unusual values that increasingly complex equations had to be specified in order to account for them (a situation not unlike multi-parameter item response theory models). Francis Bacon was one of the early scientists to stress the importance of accurate measurement: "Truth emerges more readily from error than from confusion" (Bacon in Kuhn, 1970, p. 18). Lord Kelvin, too, attributed the major discoveries of science to "accurate measurement and long-continued labors in the minute sifting of numerical results" (Kelvin in Conant, 1954, p. 121).

As measuring became more precise, discrepancies between observed and expected phenomena attracted more attention. The inter-dependency of experimentation and quantification led to awareness that discrepancies from predictions, while expected to be of some magnitude, ought to be reasonably small and not systematic. A means of determining what might constitute "reasonable agreement" between data and theory became possible through replicated experimentation. Replications provide statistical estimates of standard errors and provide practical rules for defining "good fit," "outliers," and "random error." Replications establish precision that, if the precision is acceptable for the task, can expose inaccuracy because residuals can be used to expose lack of congruence between theory and data. The magnitude and direction of the residual then provides information useful for supporting, rejecting, or modifying the theory.

This quantitative test of fit of observations to an explanatory model is a primary distinction between myth and science. There are no residuals from a myth. The myth is simply revised to account for new phenomenon. Residuals appear *only* against the background of a scientific paradigm. "It is the presence of the research as a specifically motivated course of action that makes the event accidental or fortuitous in the first place" (Brannigan, 1981, p. 73). Paradigms change, but if a theory can lead to an assertion that a specific condition should produce an expected outcome, then residuals will occur.

The Problem of Residuals

Since residuals always occur in empirical research, the problem becomes "what to do with them?" Every researcher notices and then passes by residual variation. There is neither enough time nor effort available to analyze every residual. Even "large" residuals usually disappear under scrutiny, i.e., observational, instrumental, recording, and computational mistakes routinely

occur. When such mistakes are corrected, the residual usually becomes negligible. But significant residuals, relative to some measure of "reasonable agreement," do occur and, more important, recur. The cause of a residual may be trivial but it may also lead to an important discovery.

Unfortunately, it is seldom obvious whether a residual is the key to an old puzzle or the clue to a new direction of inquiry. But, as Pasteur said, "In the fields of observation, chance favors only the prepared mind" (Pasteur in Kuhn, 1961, p. 49). Recognition occurs only when the scientist "*knowing with precision* what he should expect, is able to recognize that something has gone wrong" (Kuhn, 1962, p. 65). The crucial step is taken when the observation, "Something has gone wrong…," is followed up by analysis.

Individual skill and insight and relevant instruments and concepts are the conditions necessary for recognizing consequential unexpected results. Examples abound in the history of science where one researcher has encountered a persistent residual phenomenon, but not pursued it, and another has come along and "discovered" it. Cavendish, for example, saw and measured a residual gas when nitrogen was removed from air. But it was Ramsay and Raleigh a century later who "discovered" argon.

When a significant residual persists it usually passes from the category of "novelty" to "anomaly." An anomaly is defined as "A recognition that nature has somehow violated the paradigm-induced expectations that govern normal science" (Kuhn, 1962, p. 52). A persistent peculiarity may be considered a novelty and ignored, an anomaly to be pursued and explained, or an anomaly to be noticed but left for future generations to explain. Many of the laws of science are the result of explanations of violations of expectation which could no longer be ignored (Ashall, 1994; Kantorovich, 1993).

Research may be undertaken to delimit the boundaries of a persistent residual anomaly. What are the circumstances under which the residual happened? Can it be reproduced? Will it always happen if the circumstances recur? Sometimes, as in chemistry, more refined experimental techniques eliminate the discrepancy. Some persistent residuals have been left as known anomalies. Newton's theoretical value for the speed of sound, for example, was a scientific anomaly for 50 years until refined measurements attained by Delaroche and Berard were used by Laplace to reduce the discrepancy from 20 per cent to 2.5 per cent (Kuhn, 1977, p. 196). Other residuals have been resolved by the discovery of a new phenomenon. The discovery of Neptune accounted for the inexplicable orbit of Uranus. The prediction of Neptune is a beautiful example of a new theory (Newtonian physics) that was more explanatory than it's predecessor (Kepler's laws).

Models and Expectations

Whatever the ultimate explanation, residuals can only be understood relative to the mathematical and theoretical model from which they result. Models provide the potential for expecting certain conditions to occur; they provide a background against which surprises stand out. But once a significant residual is identified and replicated — where does one search for the cause? Does one question the theory, data, instruments, or computations? It is not always obvious where adjustments should be made. The residuals from the Ptolemaic mathematical model led Tycho Brahe to obtain more accurate measurements of the universe but he still retained Ptolemy's mathematical model (based on circular motion) and geocentric model for the relation of the heavenly bodies to earth. The same residuals led Copernicus to reject the geocentric model in favor of

the heliocentric model but retain Ptolemy's mathematics. Kepler, in turn, retained the heliocentric model but rejected circular orbits in favor of elliptical ones.

In education the phenomena of interest are variables such as arithmetic ability, academic motivation, inductive reasoning skill, etc. A variable is not measured by a single question (item) on a test. A variable is measured by a set (sample) of questions written to cover (replicate) a single topic. Test items are selected to cover a range from a lesser to a greater degree of the intended variable, e.g., knowledge, ability, motivation. When a set of content-homogeneous items is used to define a variable operationally it is called a "scale." Administering the scale to students results in scale value estimates for students ("measures") and scale value estimates for items ("calibrations"). The comparison of person measures with item calibrations leads to estimates of performance expected when a person takes any item. The difference between the observed and expected performance is then a residual.

In educational measurement, replication is accomplished through the administration of a homogeneous set of items — a measurement instrument. This is not only because a set of homogeneous observations estimates more precisely the performance level of a student than does a single item, but also because the intended replications provide evidence for consistency when it is obtained and they expose inconsistency, the unexpected residual, when that occurs. A response pattern inconsistent with a modeled expectation of performance leads to doubts about the relevance of the measurement. But since observed response patterns are never perfectly consistent with their expectations, some subjectivity is unavoidable in the determination of how inconsistent a pattern must be before the measurement should be judged inaccurate.

Many questions arise as a test of fit is carried out. How well have the expectations of the model been met, i.e., were all statistical assumptions met? How accurate is the instrument, e.g., how many items were there and how were they scored? Does the instrument yield useful information about the people, e.g., was the instrument measuring a construct appropriate for the people? Did the instrument work for the task at hand, e.g., to what extent did it measure a wide range of variability in the construct? Have relevant extraneous variables been controlled, e.g., were testing conditions standardized and were there individuals or groups of people with extraordinary characteristics? Each question is a check that the observed relation between model and data is of the form where the simple difference between the observed response given by a person and the predicted response under the statistical model is a residual. If the data have been collected carefully and the model is useful for explaining them, then observed and modeled (or predicted) values should be similar and the residual should vary in magnitude and pattern like a random variable.

How Important is the Residual?

The problem is to determine whether a residual is negligible and occurred as expected or whether the residual suggests something more like:

Observed value - Modeled value = Random error + Significant outlier + Systematic relation.

In an educational application a significant outlier may result from a student with a relatively low estimate of ability who nevertheless succeeds surprisingly on one or more difficult items. Similarly, a systematic relation might be traced to a classroom of relatively capable students who surprisingly miss a set of linked items on a scale. In these examples the important concern is whether or not the measurement process has yielded quantitative estimates of student ability that

are useful. When residuals exhibit variation uncharacteristic of that expected for random error, the measures (if they can still be called "measures") are determined by influences not expected or accounted for in the model. And the question now becomes: "Do we maintain the measurement model as it is or do we alter the model to account for variation peculiar to the specific testing circumstances?"

My answer to this question relies on a distinction between measurement theory and statistical theory. Measurement theory, in my opinion, started with the early efforts of the German psychophysicists Weber and Fechner to formulate a quantitative relationship between mind and matter. Their efforts succeeded in establishing that physical stimuli and subsequent reactions can be formulated as quantitative laws. From Galton and Pearson we came to understand that statistical techniques can be applied to human behavior to reveal individual differences and that those individual differences lead to statistical discrepancies that can be modeled as random error. From Binet we began to quantitatively measure and statistically test for cognitive differences. From Thurstone we discovered we can quantitatively measure and statistically test for affective differences. From the deterministic models of Guttman we saw that variables may be constructed as linear hierarchies within which performance at one level of functioning presumes successful performance at lower levels of functioning. And, finally, through the probabilistic models of Rasch we understand that we can model the expected response a person gives to an item as the simple difference between the level of ability possessed by the person and the level of ability required by the item to be successfully responded to. A residual worth noting and thinking about, in the Rasch model, is either an unexpected success or unexpected failure on a given item.

Statistical theory, on the other hand, prospered through Galton and Pearson who established that human variability in one domain (variable) may be understood as a function of variability in one or more other variables. From this simple concept of co-relation came regression, or the opportunity to incorporate as many variables as one chose to use to explain behavior. At the root of early and present statistical theory, however, is the recognition that as one accounts for more and more variation in some outcome variable, one is also reducing the unaccounted-for-variation (error). Thus as each additional variable is added to the statistical model, residual variation is reduced. In general, this situation is desirable. The problem is that in any given analysis, particularly an exploratory one where the effort truly is to maximize the proportion of predictable variance (which is the same as minimizing error variation), the risk of "capitalizing on chance" greatly increases until one could, theoretically, reduce residual variation to nil but have a model with no likely generalizability. Consequently, statistical theory took a positive step forward with the introduction of structural equation models.

In these models the investigator specifies the variables to be employed based on theoretical reasoning, the relationships between the variables (and error components) are specified, and the data are tested to see the extent to which they fit the proposed structure. Residuals from this model are then useful for pointing out areas of the theory to be strengthened, modified, or discarded.

Trash or Treasure?

To a great extent, then, Rasch models are structural equation models to be confirmed through the data. If the data do not fit the model, then the data irregularities may be investigated through the residual patterns. Here, then, we have a measurement model (Rasch) where the meaning of a residual remains constant across all similar circumstances. Two and three parameter logistic (2/3

PL) item response theory models, in contrast, are exploratory models to be fit and modified as the local testing conditions suggest. Residuals, in those models, represent noise to be eliminated. That is, unlike the one parameter Rasch model, two and three parameter item response theory models estimate item parameters that subsume and mask variation that would otherwise be noticed and interpreted as problematic and undesirable for measurement.

The analysis of measurement-model residuals is important for many audiences. The classroom teacher, tester and evaluator need to know how useful their measurements are. The progress of a student or class and the effectiveness of a curriculum is typically measured by some form of performance level score. This score is usually assumed to measure the variable of interest and not something else irrelevant to the task. If the scale is not working as intended, then the measures may be meaningless. Fortunately, analyses of residuals can make the use of meaningless measures avoidable.

The following examples, from my own work, illustrate the types of problems that were revealed through the analysis of residual variation. I have found: (1) *start-up effects,* attributed to subjects being initially confused about the responses expected of them; (2) *speed effects,* where limited time led to identifiable guessing behavior; (3) *interviewer effects,* where interviewers had different impressions about how to score interviewee responses; (4) *instrument effects,* where scoring sheets were miskeyed, items had multiple correct answers, or items had no correct answers; (5) *dimensionality effects,* where items clearly addressed two or more constructs (dimensions); (6) *teacher effects,* where teachers used their own unique interpretation of how to administer a performance assessment to their students; (7) *classroom effects,* where students were not taught a specific curricular component taught on a statewide assessment; and (8) *special characteristic effects,* where students with unique cognitive (low verbal ability), affective (highly anxious), or physical characteristics (wheelchair user) did much worse or better than was expected of others with their same estimated ability level. In each case, the problem was identified and a change was made in the instrument or the testing conditions — the model itself remained the same.

In conclusion, then, I encourage everyone who engages in the development or application of measurement instruments to aggressively dig through the residual variation. What you find may be more important than what you set out to investigate in the first place.

References

Ashall, F. (1994). *Remarkable discoveries!* New York: Cambridge University Press.

Brannigan, A. (1981). *The social basis of scientific discoveries*. Cambridge: Cambridge University Press.

Butterfield, H. (1957). *The origins of modern science 1300-1800*. New York; The Free Press.

Conant, J. B. (1954). *Science and common sense.* New Haven: Yale University Press.

Herschel, Sir J. F. W., & Bart, K. H. (1851). *Preliminary discourse on the study of natural philosophy* (new ed.). London: Longman, Brown, Green & Longmans, Paternoster Row.

Herschel, Sir J. F. W., & Bart, K. H. (1871). *Outlines of astronomy* (11th ed.). London: Longmans, Green, and Co.

Kantorovich, A. (1993). *Scientific discovery: Logic and tinkering*. Albany, NY: State University of New York Press.

Kuhn, T. S. (1962). Historical structures of scientific discoveries. *Science, 136*, 760-764.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: The University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension*. Chicago: The University of Chicago Press.
Roberts, R. M. (1989). *Serendipity: Accidental discoveries in science*. New York: Wiley.

Notes

1. Key words: residual, scientific models, theory building, psychometrics.

2. Correspondence should be addressed to Larry H. Ludlow, Boston College, Lynch School of Education, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467.