Research Report

Orthogonal Regression, the Cleary Criterion, and Lord's Paradox: Asking the Right Questions

ETS RR–20-14

Michael T. Kane
Andrew A. Mroch

December 2020

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Orthogonal Regression, the Cleary Criterion, and Lord's Paradox: Asking the Right Questions

Michael T. Kane[1] & Andrew A. Mroch[2]

1 Educational Testing Service, Princeton, NJ
2 National Conference of Bar Examiners, Madison, WI

Ordinary least squares (OLS) regression and orthogonal regression (OR) address different questions and make different assumptions about errors. The OLS regression of Y on X yields predictions of a dependent variable (Y) contingent on an independent variable (X) and minimizes the sum of squared errors of prediction. It assumes that the independent variable (X) is an observed score that is known without error, and all of the error is assigned to the dependent variable (Y). OLS is not designed to estimate underlying functional relationships, and if both variables contain error, OLS regression tends to yield biased estimates of such functional (or true-score) relationships. OR models, including the errors-in-variables (EIV) and geometric-mean (GM) models, assume that both variables contain error and seek to identify the line that minimizes squared deviations of the data points from the line in both the X and Y directions. OLS and OR address different questions and serve different purposes. If one wants to predict one variable from another variable, OLS regression is an optimal approach and OR is less efficient; in examining the functional relationship between two variables, OR provides a more plausible model. The OR models are hard to apply in many contexts because they depend on strong assumptions about sources of error. As examples of cases where OR can shed light, we examine its use in analyzing test bias as distinct from predictive bias and in making sense of Lord's paradox.

Keywords Orthogonal regression; Cleary criterion; Lord's paradox; equation errors; test bias; predictive bias

doi:10.1002/ets2.12298

This paper provides an introduction to orthogonal regression (OR) models, specifically the general errors-in-variables (EIV) model and the geometric-mean (GM) model. It describes these models and the commonly used ordinary least squares (OLS) regression models within a least-squares framework using the same notation for all models in order to be clear about the assumptions made by the different models and the uses best served by the different models. These comparisons suggest that OLS regressions are optimal if the goal is to predict one variable from the other, and an OR can be appropriate when the goal is to estimate a functional, or true-score, relationship between the two variables. All of the models discussed in this report are well known; this paper focuses on the differences in the assumptions made about errors in the different models and on the resulting differences in the interpretations of the lines yielded by the different models. We seek to identify how the different models can most usefully be applied.

Any more-or-less linear relationship between two variables can be represented by a straight line, and least-squares analyses provide powerful tools for identifying the line that best represents the relationship. Although researchers typically do not know that the relationship is linear and may even have reasons to expect some departures from linearity, a linear relationship is often "considered a reasonable approximation" (Fuller, 1987, p. 35). That the relationship is at least approximately linear over the range of Y and X values that are of interest is a basic assumption in all the models discussed in this paper.

The most commonly used least-squares model, the OLS regression model, determines the line that minimizes the sum of the squared differences between the observed and estimated values of a dependent variable, contingent on the values of a fixed independent variable. The independent variable is a known observed score, and all the variability around the line is attributed to random errors in the dependent variable. OLS regression assigns dramatically different roles to the two variables; it is not symmetric, and for OLS regression, the Y on X line can be quite different from the X on Y line. OLS regression is particularly appropriate if the goal is to predict an unknown dependent variable from a known independent variable. Typically, a data set with both Y and X is available, and OLS is used to estimate a prediction equation, which is then used to predict Y from X in subsequent data sets where only X is available.

Corresponding author: Michael T. Kane, E-mail: mkane@ets.org

This asymmetric approach makes sense if one of the variables (taken to be the dependent variable) is subject to substantially more error than the other (taken to be the independent variable). It also makes sense if the independent variable (e.g., a test score) has already been determined (i.e., has a fixed, observed value) and the goal is to predict an unknown dependent variable (e.g., some future performance criterion) from the known independent variable. That is, the analysis is contingent on observed values of the independent variable. However, if both variables contain error and the goal is to estimate a functional, true-score relationship between the two variables, OLS regression yields a biased estimate of the true-score relationship even if the true-score relationship is, in fact, perfectly linear and the errors are purely random (Berry, 1993; Berry & Feldman, 1985; Draper & Smith, 1998; Lewis-Beck, 1980; Pedhazur & Schmelkin, 1991).

A number of authors (Adcock, 1877; Bartlett, 1949; Deming, 1943; Draper & Smith, 1998; Fuller, 1987; Madansky, 1959; Mandel, 1964; Nievergelt, 1994; Pearson, 1901; Riggs et al., 1978; Wald, 1940) have suggested methods for estimating what Sprent (1969) calls a "functional" or "law-like" (p. 31) relationship (or in psychometric terms, a *true-score relationship*) between two variables, when both variables are subject to error. These models are generally referred to as OR models, because the deviations being minimized include both the X and Y deviations from the line (with suitable weighting of the two components). For example, a maximum-likelihood estimate of the underlying error-free relationship between two variables minimizes a "weighted orthogonal distance" (Carroll & Ruppert, 1996, p. 2) with the weights specified by the ratio of the error variances for the two variables. This model is referred to as the general EIV model (Fuller, 1987). If the error variances are taken to be equal in the EIV model, it yields a line that minimizes "squared total Euclidean or orthogonal distances" (Carroll & Ruppert, 1996, p. 2) of the points from the line. A large number of OR models have been developed over the last century, and most of these models are reviewed by Madansky (1959) and by Riggs et al. (1978).

In the EIV model, the notion of measurement error is interpreted broadly to include any source of construct-irrelevant variance (e.g., sampling of observational methods, items or tasks, raters, occasions, contexts, time limits) that has an impact on one of the variables but not the other (Brennan, 2001; Carroll & Ruppert, 1996; Cronbach et al., 1972). One serious limitation in the use of OR models is the difficulty in estimating the total error for each of the variables; in the social sciences, researchers seldom have good estimates of more than one or two sources of error (Brennan, 2001). An even more serious limitation arises from the existence of sources or variability that are not uniquely attributable to error in either of the two variables. These extra errors, referred to as equation errors (Fuller, 1987), include nonlinearity in the relationship and natural variations (Ricker, 1973).

A principal-components analysis of data points defined by two standardized variables (i.e., z-scores) yields a line (the first principal component) for which the sum of squared perpendicular (or orthogonal) distances of the standardized data points from the line is minimized. The slope of this line is equal to the geometric mean of the slopes of the two lines generated by the OLS regressions of Y on X and X on Y (whether the equation is specified for the z-score scales, or is transformed to the observed-score scales for X and Y), and as a result, this model is generally referred to as the GM model (Draper, 1991; Madansky, 1959; Ricker, 1973; Riggs et al., 1978; Sprent, 1990). In deriving the GM model using principal components, it is not necessary to make assumptions about the errors in X or Y or about equation errors. The principal-components analysis with standardized scores adjusts for scale differences in X and Y, but it does not make any adjustments for the relative magnitudes of the errors in X and Y. It focuses on observed orthogonal deviations (on the z-score scales for X and Y) of the data points from the line. As shown later, the GM model can also be derived from the general EIV model by assuming that the two variables have the same reliability. The EIV model weights the X and Y components of orthogonal deviations of points from the line in terms of their relative error variances, and for the GM model to be special case of the EIV model, it is necessary that X and Y have the same reliability. Unlike the OLS models, the EIV and GM models are symmetric in the sense that the fitted line is independent of whether one takes Y as a function of X, or X as a function of Y.

OLS regression has become the standard approach to estimating linear relationships between two variables in education and the social sciences, in part because it is easy to estimate and in part because the asymmetric framework inherent in OLS regression is appropriate in many important contexts (e.g., any context in which the goal is to predict unknown future outcomes based on known test scores). However, there are contexts (e.g., in evaluating test bias, as distinct from predictive bias, by comparing linear relationships across groups with different mean scores) in which OR tends to be more appropriate.

The distinction between the OLS and orthogonal models can also be used to explicate Lord's (1967) paradox. Lord developed a hypothetical example in which an analysis of covariance based on OLS regression contradicted a simple and natural analysis of the data. We argue that if OR is substituted for OLS regression in Lord's example, the contradiction between the two analyses disappears.

As noted earlier, the EIV model assumes that all of the variability around the line representing the functional relationship can be accounted for by errors that can be assigned to X or Y (Carroll & Ruppert, 1996). Equation errors (Fuller, 1987) are not consistent with the assumptions built into the EIV model, which assumes that all of the variability around the best fitting line is due to errors of measurement in X and Y and therefore there is no equation error. In practice, most potential applications of OR involve some equation errors, and some potential applications contain substantial equation errors. The existence of equation errors limits the contexts in which the EIV model can be used; the potential impact of equation errors is discussed toward the end of this report.

In the next section, Specifying Best-Fitting Lines, we present the differences between the GM and OLS models, their underlying assumptions, and some of their properties within a common, traditional framework. This presentation allows us to highlight the differences in the assumptions made by the two models and the differences in the conclusions that can be drawn from applications of the models. In the third section, we present the more general EIV model (Fuller, 1987) and show how the OLS and GM models can be derived as special cases of this more general EIV model. In the fourth section, the properties of group-specific regressions used as criteria for test bias and predictive bias are explored for the OLS and GM models and simulated data is used to illustrate the behavior of OLS and ORs for this application. In the fifth section, we analyze Lord's (1967) paradox in terms of the differences between the OLS and GM models. In the last section, we provide some general conclusions.

As noted earlier, the statistical models in this paper (OLS, EIV, and GM) are well known, but EIV and GM are not used much in educational research. To help highlight situations where each model may be useful, we focus on the assumptions in these models, on differences in the implications of the models, and in particular, on the questions appropriately addressed by the different models. We provide two examples (evaluating predictive bias and test bias by comparing fitted lines across groups and Lord's paradox) that are intended to illustrate how the OLS model and the OR models address different questions about statistical relationships between two variables.

Specifying Best-Fitting Lines

Identifying the best fitting line for a set of data requires a criterion for evaluating the fit of any particular line. A commonly used approach minimizes some average squared deviation of the observed data points from the line. In developing such a least-squares analysis, it is necessary to specify the deviations that need to be minimized. Different least-squares criteria may be more or less reasonable in addressing different questions.

Note that in this paper, we will always take the vertical axis to be the Y axis and the horizontal axis to be the X axis. In some cases, this approach will be counter to convention (notably in considering the OLS regression of X on Y, where Y is the "predictor," or independent variable); however, the consistency will allow us to compare relationships within a single geometric framework, whether the lines are derived from an OR model, an OLS regression of Y on X, or an OLS regression of X on Y.

OLS Regression

The bivariate OLS regression model treats one variable as the independent variable (usually labeled X), which is assumed to be known (or manipulated) in advance without error, and treats the other variable as the dependent variable (usually labeled Y), which may contain measurement error as well as other sources of random variability, all of which are included in a single error term. That is, the dependent variable is taken to be contingent on observed values of the independent variable.

Y on *X*

The OLS regression of Y on X is chosen to minimize the sum of the squared vertical distances between the estimated and observed values of the dependent variable, Y. The resulting regression equation has the familiar form,

$$\hat{Y}_i = r_{YX}\frac{S_Y}{S_X}\left(X_i - \overline{X}\right) + \overline{Y}, \tag{1}$$

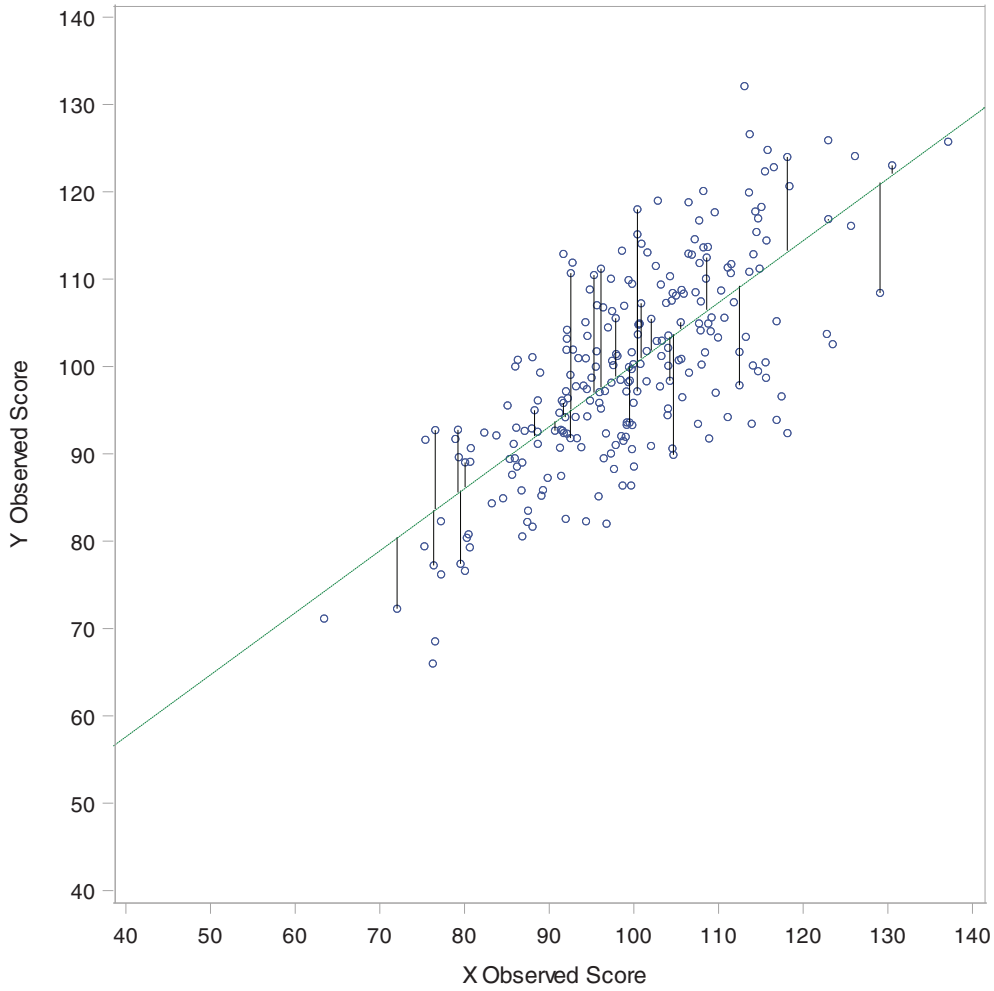

Figure 1 Scatterplot and ordinary least squares (OLS) regression of Y on X with vertical lines indicating deviations of observed values of Y from predicted values of Y based on the Y on X OLS line.

where $\widehat{Y}_i$ is the estimated (or predicted) value of Y_i given X_i, r_{YX} is the correlation between X and Y, s_X is the standard deviation of X, s_Y is the standard deviation of Y, and $\overline{X}$ and $\overline{Y}$ are the means of X and Y. The squared discrepancies to be minimized for the OLS regression of Y on X are illustrated by the vertical lines between observed values of Y and the Y on X line for a random subset of 30 data points from the scatterplot in Figure 1.

The Y on X line passes through the centroid, $(\overline{X},\overline{Y})$, of the joint distribution of X and Y, and the mean of $\widehat{Y}$ given by the regression equation is equal to $\overline{Y}$. However, the variance of $\widehat{Y}$ is smaller than the variance of Y by a factor of r^2_{XY}, indicating that $\widehat{Y}$ is generally closer than Y to $\overline{Y}$, and $\widehat{Y}$ is said to be "regressed" toward the mean (Campbell & Kenny, 1999; Galton, 1886).

Draper (1991) pointed out that one can think of the "fitted line" corresponding to Equation 1 as the specification of a line in the $X-Y$ plane (Draper, 1991, p. 7). Dropping the subscripts and the hat in Equation 1, we can represent this line as

$$Y = r_{YX}\frac{S_Y}{S_X}\left(X - \overline{X}\right) + \overline{Y}. \tag{2}$$

Equation 2 represents a line in the $X-Y$ plane, and as such does not give Y or X a special role as a dependent or independent variable.

This change in notation represents a shift in focus from the problem of estimating a dependent variable from an independent variable to the problem of representing the functional relationship between the two variables. This shift from prediction to drawing conclusions about a relationship corresponds to a change in the question being asked and

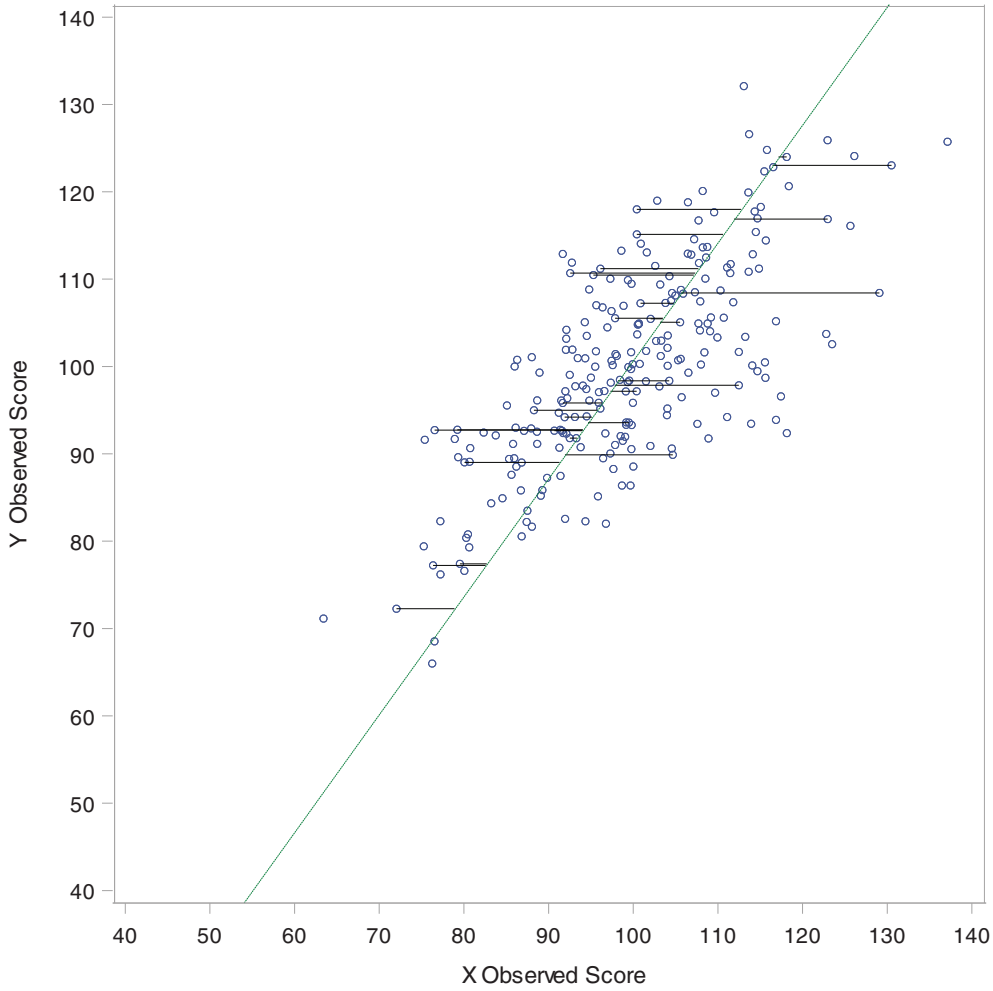

Figure 2 Scatterplot and ordinary least squares (OLS) regression of X on Y with horizontal lines indicating deviations of observed values of X from predicted values based on the X on Y OLS line.

a corresponding change in the interpretation of the "fitted line" (Draper, 1991). Note that we are not suggesting that Equation 2 is a good estimate of the true-score relationship between X and Y. In fact, it is well known that it is generally biased (often seriously biased) as an estimate of the true-score relationship (Carroll & Ruppert, 1996; Fuller, 1987; Linn & Werts, 1971), but it is one of a family of possible estimates of the true-score relationship, and it provides a convenient starting point.

X on *Y*

The OLS regression of X on Y identifies the line that minimizes the average squared difference between the observed values of X and the estimated values of X (or $\widehat{X}$), based on observed values of Y. In this reverse version of the OLS regression model, Y is taken to be the independent variable that is known without error, and X is taken to be the unknown, dependent variable. The OLS regression of X on Y minimizes the squared horizontal distances of the points from the line (assuming that X always represents the horizontal axis and Y represents the vertical axis), as in Figure 2, which uses the same scatterplot and illustrative data points as Figure 1.

The resulting OLS regression of X on Y has the following form:

$$\widehat{X}_i = r_{XY}\frac{S_X}{S_Y}\left(Y_i - \overline{Y}\right) + \overline{X} \tag{3}$$

where $\widehat{X}_i$ is the predicted value of X_i given Y_i, and the other terms in Equation 3 are the same as in Equation 1.

The "fitted line" corresponding to Equation 3 is

$$X = r_{XY}\frac{S_X}{S_Y}\left(Y - \overline{Y}\right) + \overline{X}. \tag{4}$$

Again, the fitted line is a line in the $X-Y$ plane, and it can be interpreted as representing a relationship between X and Y and not as a tool for predicting X from Y.

Transforming this equation, so that it expresses Y as a function of X, we get another formula for the same X on Y fitted line,

$$Y = \left(\frac{1}{r_{XY}}\right)\frac{S_Y}{S_X}\left(X - \overline{X}\right) + \overline{Y}. \tag{5}$$

Note that the fitted line represented by Equation 5 differs from the fitted line given by Equation 2 for the OLS Y on X line. In particular, assuming that the correlation between X and Y is less than 1.0, the X on Y line will have a larger slope than the Y on X line. For correlations that are close to 1.0, the two OLS lines will have similar slopes, but for correlations that are substantially less than 1.0, which is often the case, the two OLS lines will have markedly different slopes. The mean of the predicted values of X is equal to the mean of the observed values of X, and this X on Y line passes through the centroid $(\overline{X},\overline{Y})$, of the joint distribution of X and Y. So, the X on Y line and the Y on X line will intersect at the centroid $(\overline{X},\overline{Y})$.

Because of regression toward the mean, the X on Y line has a higher slope than it would have if the correlation between X and Y were 1.0, and it has a higher slope than the Y on X line (Campbell & Kenny, 1999; Galton, 1886). If we are interested in predicting one of the variables from known values of the other variable, the appropriate OLS regression line is optimal in the sense that it is expected to minimize the sum of squared errors of prediction. However, this lack of symmetry in the two OLS lines, that the Y on X line is different from the X on Y line, is a problem if one wants to estimate the true-score relationship between X and Y. How can researchers who are interested in the functional (or true-score) relationship, and not in predicting one variable from the other, choose between the two OLS-based lines?

Principal Components and GM Regression

OR models estimate the true-score relationship between X and Y when both variables are subject to error. These models minimize deviations in both the X and Y directions from the fitted line; for most OR models (e.g., the EIV model described later), the squared deviations in the two directions are weighted in some way.

In using the OR models, the goal is to identify a single line that best represents the functional (true-score) relationship between the two variables, rather than to predict one variable from the other. In introducing this kind of analysis, Kruskal (1953) made the following suggestion:

> In biological studies dealing with two or more quantitative characteristics of the individuals of a species or other grouping, the biologist may wish to represent the joint distribution of the characteristics by a single straight line. … Apparently there are two major motivations for working with this kind of representation: first, the desire for a concise (although, of course, incomplete) description of the distribution in order to indicate the general relative trends of the characteristics; and, second, the prospective use of such a line as a predicting mechanism. (p. 48)

Sprent (1969) characterized the relationships as being "law-like" (p. 29) or functional, and suggested that researchers tend to assume that:

> there is a basic mathematical relationship between variables with which their data would accord were it not for the fact that this relationship is obscured to some extent by "random" fluctuations, perhaps associated with both variables. (p. 29)

The focus is less on prediction, as such, and more on specifying the underlying relationship.

Principal-components analysis provides a symmetric approach to determining the line that best represents the functional (or true-score) relationship between X and Y by minimizing the sum of squared perpendicular (or orthogonal) distances of the data points from the line (Cliff, 1987, pp. 298–300). Given a set of data points in the $X-Y$ plane, principal-components analysis involves a simple rotation of coordinates to a new set of orthogonal axes, one of which (the first

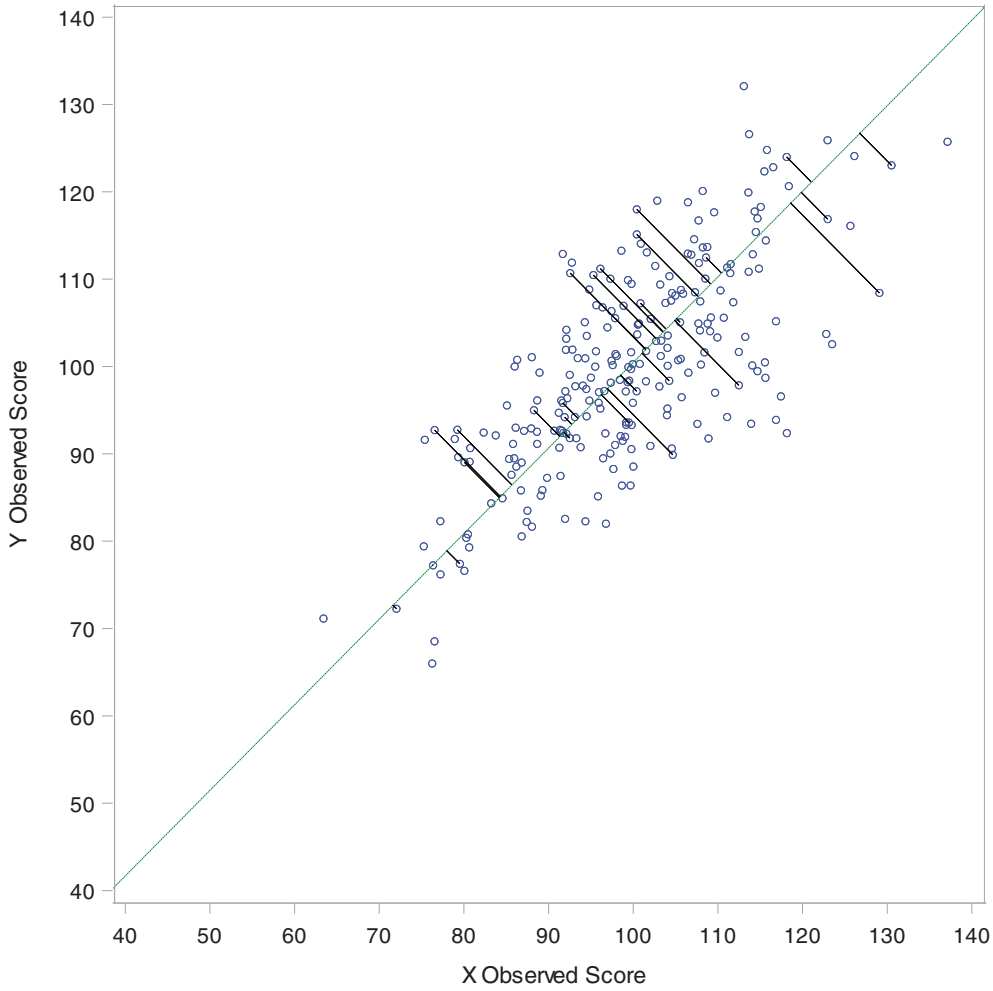

Figure 3 Scatterplot and geometric-mean (GM) regression with diagonal lines indicating deviations of observed values of X and Y from the GM line.

principal component) represents the direction of maximum variability and the other of which represents the residual variance perpendicular to the line defined by the first principal component. The direction of the axis with maximum variance (the first principal component) can be used to represent the linear relationship between X and Y. This line is illustrated in Figure 3, for the same scatterplot as Figures 1 and 2.

Although this approach identifies the line for which the sum of the squared orthogonal distances of the points from the line is minimized, it has the serious disadvantage of not being invariant under changes of scale in X or Y. Scale invariance can be ensured by transforming the two variables to the z-score scale before conducting the principal-components analysis.

Representing the variables as z scores (i.e., scores scaled to have a mean of 0 and a standard deviation of 1), the composite with maximum variance is represented by the following equation (Cliff, 1987):

$$Z_Y = sgn\left(r_{XY}\right) Z_X. \tag{6}$$

The *sgn* function indicates that the sign of the relationship (+ or -) between Z_X and Z_Y is given by the sign of the correlation between X and Y. To get Y as a function of X, we can convert Z_Y and Z_X to the original scale

$$\frac{Y-\overline{Y}}{S_Y} = sgn\left(r_{XY}\right)\frac{X-\overline{X}}{S_X}, \tag{7}$$

or

$$Y = sgn\left(r_{XY}\right)\frac{S_Y}{S_X}\left(X-\overline{X}\right)+\overline{Y}, \tag{8}$$

which, like the two OLS lines, passes through the centroid, $(\overline{X},\overline{Y})$. Note that the *sgn* function is necessary because the ratio of standard deviations is always positive, but for a decreasing function, r_{XY} would be negative, and the slopes of both OLS lines would be negative; under these circumstances, the slope of the GM line also needs to be negative.

Note that Equations 6–8 do not have hats on either variable; the result is not interpreted in terms of predicting Y from X or X from Y, but rather, as an estimate of the true-score relationship or underlying functional relationship between X and Y. As noted earlier, this line is generally referred to as the GM line because its slope is equal to the geometric mean of the OLS slopes in Equations 2 and 5.

In identifying a line that minimizes the total squared distances between data points and the line (horizontal and vertical), it is necessary that the scales for the two variables be comparable. By transforming the two variables to z-scores, the potential impact of scale differences is removed. Given the arbitrariness inherent in scales for most variables (e.g., Celsius vs. Fahrenheit), scale independence is a highly desirable characteristic.

The magnitude of the GM slope does not depend on the magnitude of the correlation between X and Y, but the sign (+ or −) is taken to be the sign of the correlation between the two variables. If the correlation is close to 1.0, the GM line incorporates most of the variability in the two variables and the data points cluster around this line. As the correlation decreases, the GM line incorporates less variability (Cliff, 1987). In general, the usefulness of the GM line in representing a linear trend in the data depends on the magnitude of the correlation, and this magnitude would need to be substantially different from zero for any line to be what Fuller (1987) referred to as "a reasonable approximation" (p. 35) of the relationship.

The variance in the values of Y computed using Equation 8, given the observed values of X, equals the variance in the observed values of Y. Because its slope does not include the correlation between X and Y, the GM line is not subject to regression toward the mean. The GM regression analysis is symmetric, in the sense that the GM regression of Y on X is algebraically equivalent to the GM regression of X on Y.

An EIV Model

The EIV model (e.g., Carroll & Ruppert, 1996; Deming, 1943; Fuller, 1987), allows both variables to contain substantial errors. What Fuller (1987) called the "classical errors-in-variables model" (p. 30) assumes that both the X and Y variables contain error,

$$Y = y + e, \tag{9}$$

and

$$X = x + u, \tag{10}$$

where y and x are the true scores for Y and X, respectively, and e and u are random errors that are uncorrelated with each other and with their respective true scores and that have means of 0. Because the errors are uncorrelated with the true scores, the variances in Y and X can be represented as

$$s_Y^2 = s_y^2 + s_e^2, \tag{11}$$

and

$$s_X^2 = s_x^2 + s_u^2. \tag{12}$$

The ratio of the error variance for Y to the error variance for X is given by

$$\delta = \frac{s_e^2}{s_u^2}, \tag{13}$$

and the value of this ratio is taken to be known, at least approximately (Carroll & Ruppert, 1996; Deming, 1943; Fuller, 1987). As discussed more fully later, estimating the ratio of the error variances, δ, is generally challenging, but in the examples discussed in this paper, δ can be unambiguously specified. By including only the measurement errors in X and Y in the EIV model, Fuller implicitly assumed that equation errors have no role in the analysis.

It is assumed (Fuller, 1987) that the true-score relationship between X and Y is linear:

$$y = \beta_0 + \beta_1 x. \tag{14}$$

Under these assumptions, a general maximum-likelihood solution for EIV regression yields the following estimates for the slope and intercept of the line (e.g., Carroll & Ruppert, 1996; Deming, 1943; Fuller, 1987; Madansky, 1959):

$$\widehat{\beta}_1 = \frac{s_Y^2 - \delta s_X^2 + \left[\left(s_Y^2 - \delta s_X^2\right)^2 + 4\delta s_{YX}^2\right]^{1/2}}{2s_{YX}}, \tag{15}$$

and

$$\widehat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}, \tag{16}$$

where s_{YX} is the covariance between Y and X. The likelihood is maximized by taking the positive square root in Equation 15 (Sprent, 1969), and it is clear that this ensures that the slope has the same sign as s_{YX}. This maximum-likelihood solution includes a large number of regression models as special cases. All of these EIV lines pass through the centroid.

The OLS lines and the GM line can be easily derived from the general EIV model. If we assume that the error variance for Y is 0, and the error variance for X is greater than 0, δ would be 0, and taking the positive root in Equation 15, the estimated slope would reduce to

$$\widehat{\beta}_1 = \frac{s_Y^2}{s_{YX}} = \frac{s_Y^2}{r_{YX}\, s_Y\, s_X} = \frac{s_Y}{r_{YX} s_X} = \frac{1}{r_{YX}} \frac{s_Y}{s_X}. \tag{17}$$

With this expression for β_1 and the expression for β_0 in Equation 16 as the slope and intercept, we have the fitted line for the OLS regression of X on Y in Equation 5. If we reverse the roles of X and Y in Equation 15 and assume that the error variance for X is 0, and take the positive root, we can derive the fitted line for the OLS regression of Y on X in Equation 2. In the intermediate cases where both error variances are greater than 0, the EIV line will be between the Y on X and the X on Y OLS lines.

We can derive Equation 8 for the GM line by assuming that the reliabilities (the ratios of true-score variance to observed-score variance) and therefore that the ratios of the error variance to the observed-score variance of the two variables are equal:

$$\frac{s_e^2}{s_Y^2} = \frac{s_u^2}{s_X^2}. \tag{18}$$

Under this assumption, δ is also equal to the ratio of the observed-score variances,

$$\delta = \frac{s_e^2}{s_u^2} = \frac{s_Y^2}{s_X^2}. \tag{19}$$

Substituting this expression for δ in Equation 13 and simplifying, we get

$$\widehat{\beta}_1 = sgn\left(r_{XY}\right) \frac{S_Y}{S_X}. \tag{20}$$

Using this expression for the slope and Equation 16 to define the intercept, we get the GM line in Equation 8.

Equation 8 plays a major role in linear equating and scaling (e.g., Holland & Dorans, 2006; Kolen & Brennan, 2014), where the goal is to establish or maintain a common scale for different forms of a test developed from common specifications. This application of the GM model is especially apt, because in the context of equating, where the two variables are scores on test forms designed to have the same reliability and to measure the same construct, the assumptions that the two variables have similar reliabilities and that the true-score relationship between X and Y is linear are especially plausible.

For the more general EIV model, the squared discrepancies to be minimized are the distances from the data points to the line in a direction that depends on the ratio, δ, of the error variances for the two variables (Carroll & Ruppert, 1996). If the errors in Y are much larger than the errors in X, and therefore δ is much larger than 1, the EIV line will be close to the OLS Y on X line. If the errors in X are much larger than the errors in Y, and therefore δ is much smaller than 1, the EIV line will be close to the OLS line for X on Y. If the error variances for Y and X are approximately equal relative to their observed-score variances, and therefore δ is approximately equal to $s_Y{}^2/s_X{}^2$, the EIV line will be close to the GM line. In cases where none of these three conditions holds, the general EIV model would still apply, assuming that the value of δ is known, with the EIV line between the two OLS lines.

Based on simulations, Riggs et al. (1978) suggested that the basic GM model described previously works about as well as the more general EIV model if the magnitudes of the error variances associated with the two variables relative to their respective observed-score variances do not differ by more than a factor of 2:

$$\frac{1}{2} < K^2 = \frac{s_e^2/s_Y^2}{s_u^2/s_X^2} = \delta\frac{s_X^2}{s_Y^2} < 2. \tag{21}$$

For higher or lower values of this ratio (greater than 2 or less than $1/2$), the GM model diverges substantially from the EIV model, and in cases where the equation errors are large relative to the measurement errors, neither model will work well.

Note that, like correlations, the fitting of a line to an $X-Y$ data set using either OLS or OR does not imply causation in either direction. The fact that there is a linear relationship between X and Y does not imply that X causes Y or that Y causes X. Making a case for a causal interpretation in either direction would require additional assumptions and additional evidence to support these assumptions (Holland & Rubin, 1983).

The application of the general EIV model requires a fairly accurate estimate of the ratio of the error variances, δ, (Fuller, 1987; Madansky, 1959; Wonnacott & Wonnacott, 1979), which usually requires accurate estimates of the two error variances (Carroll & Ruppert, 1996; Deming, 1943; Fuller, 1987). In many cases, good estimates of δ are hard, if not impossible, to obtain, and therefore, researchers need to forgo the use of the EIV model in some cases where they might like to use it. In other cases, they may choose to make do with fairly rough estimates (Carroll & Ruppert, 1996; Fuller, 1987; Sprent, 1990) with suitable cautions, as discussed more fully later. If it is reasonable to believe that the equation errors are modest and the reliabilities of the two variables are roughly equal, the GM line (Draper, 1991; Kruskal, 1953; Ricker, 1973), can be used as an approximate solution representing "a concise (although, of course, incomplete) description of the distribution in order to indicate the general relative trends" (Kruskal, 1953, p. 48).

Various sources of variability (e.g., nonlinearity, extraneous variables not included in the model, and context effects) in addition to measurement error, as such, can contribute to the scatter of data points around the line (Weisberg, 1985). These additional errors have been called *equation errors*, *errors in equations* (Carroll & Ruppert, 1996; Fuller, 1987; Sprent, 1990), and *natural errors* (Ricker, 1973). In estimating the relative magnitudes of the errors in X and Y, the variable to which the equation errors (or natural errors) should be assigned as error, is, by definition, not clear, and the EIV model does not provide guidance on what to do with errors that are not attributed to either X or Y.

Sprent (1990, p. 11) suggested that "it would seem that in the broadest context errors in equations may be either errors arising from *failure* to measure some relevant variable or may reflect inadequate (or inappropriate) model specification." Equation errors can be controlled to some extent by checking for nonlinearity and employing nonlinear models where appropriate and by including more relevant variables in the model (Carroll & Ruppert, 1996; Fuller, 1987; Sprent, 1990). In addition, it would be important to analyze the components contributing to the measurement errors for both variables as thoroughly as possible (Brennan, 2001; Carroll & Ruppert, 1996; Cronbach et al., 1972). To the extent that researchers do not estimate some sources of measurement error in X or Y, these errors will function as part of the equation error.

Note that the EIV model assumes that the true-score relationship is perfectly linear, as in Equation 14, and that all of the variability around this line is due to measurement errors, as in Equations 9 and 10. To the extent that equation errors are present and substantial, at least one of these assumptions is being violated. Substantial equation errors are problematic in EIV analyses, not only because these errors cast doubt on the linear model being assumed, but also because they suggest that the value of δ cannot be specified with confidence. Carroll and Ruppert (1996) suggested that equation errors be assigned to the Y variable, but this would destroy the symmetry of the analysis and shift the resulting line toward the OLS regression of Y on X. Draper (1991), under the heading of "practical advice" (p. 6) suggested that the maximum-likelihood solution in Equation 15 be used if δ "is known (or can reasonably be estimated)" (p. 6) and otherwise suggested that the GM model be used.

For the more extended examples discussed later in this paper, equation errors are not a problem. The simulated data examined in the next section do not include any equation errors, and the error variances and reliabilities are set to be equal. For the hypothetical case assumed in Lord's paradox, there is no indication of any equation errors.

Nevertheless, in general, equation errors constitute a serious limitation on the appropriate use of the EIV model because they are inconsistent with the basic assumptions of the model and because, by definition, they cannot be unambiguously assigned to X or Y and, as a result, interfere with the estimation of δ. If the equation errors are large compared to the errors of measurement in X and Y, the OR models should not generally be used to estimate functional relationships.

Group-Specific Lines

The relationships between test scores and other variables (e.g., other measures of the same attribute, variables that are theoretically related to the attribute, and criterion variables associated with the score interpretation) play a major role in the validation of test-score interpretations and the justification of score uses (Cronbach, 1971; Kane, 2006, 2013; Messick, 1989). In checking for systematic errors in assessment scores, these true-score relationships are generally expected to be invariant across subgroups (e.g., defined in terms of race, gender, disabilities), and differences in true-score relationships across subgroups can cast doubt on intended interpretations and uses of the scores (Camilli, 2006; Cleary, 1968; Cole & Moss, 1989; Dorans & Holland, 2000; Kane & Mroch, 2010).

Cleary's (1968) model takes a test to be biased if the estimated OLS regression of criterion scores on test scores yields substantially different lines for different groups. However, regression toward the mean can have a major impact on analyses that rely on OLS regression to make comparisons between the fitted lines for groups that differ in their mean scores on X and Y. As a result, the Cleary model can indicate that the test is biased across two groups even if the true-score relationships between test scores and criterion scores are exactly the same for the groups or that the test is not biased (or not substantially biased) even though there are systematic differences in the relationship. Cleary did not distinguish *predictive bias* and bias in the test as a measure of some construct. More recently, this distinction has been clearly recognized, and the 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) treat the comparison of group-specific OLS regressions as providing an empirical check on predictive bias as opposed to test bias or measurement bias. The *Standards* did not provide a comparable statistical check on test bias.

We will take the Cleary criterion (1968) as providing an indication of predictive bias, which involves systematic differences between groups in the optimal predictive relationship given by the OLS regressions of criterion scores on test scores. In contrast, we define test bias as a difference in the functional (or true-score) relationships between two variables for two groups.

Regression Toward the Mean

As indicated earlier, regression toward the mean is a statistical artifact in OLS regression, resulting from a less-than-perfect correlation between the two variables (Campbell & Kenny, 1999). One way to think of regression toward the mean is to start with the regression equation that would occur if the correlation were perfect (i.e., if true scores on X and Y are linearly related and both X and Y are observed without error). Under this assumption ($r_{YX} = 1.0$), the slope of the OLS regression of Y on X is given by s_Y/s_X. If we then assume that the correlation gradually decreases while the means and variances of X and Y stay the same, the slope of the OLS regression of Y on X would decrease in proportion to the correlation, eventually reaching 0.0 as the correlation decreased to 0.0. Because the OLS line passes through the centroid, it can be thought of as rotating about the centroid from its initial direction (with $r_{YX} = 1.0$ and a slope of s_Y/s_X) to a horizontal line represented by $Y = \overline{Y}$ with a slope of 0.0. This pattern is illustrated in Figure 4 for the regression of Y on X. The OLS regression of X on Y is essentially a mirror image of the OLS regression of Y on X, with the line rotating toward the vertical. The GM line is not subject to regression to the mean.

An Illustration with Simulated Data

In this section, we illustrate some properties of the OLS and GM lines with simulated data for two groups with different means on X and Y but otherwise identical distributions of true scores, errors, and observed scores. For both groups, the individual true scores on Y are equal to the corresponding true scores on X, and therefore, the true-score relationship is perfectly linear and the same in the two groups; so, there is no test bias in the data. We also assume in both groups that (a) each individual's observed score on X is equal to the individual's true score on X plus an independently sampled random error and (b) the individual's observed score on Y is equal to the individual's true score on Y plus an independently sampled random error.

In the high-scoring (H) group, the true scores on X are drawn from a normal distribution with a mean of 100 and a standard deviation of 10, and in the low-scoring (L) group, the true scores on X are drawn from a normal distribution with a mean of 70 and a standard deviation of 10. Because the true scores on Y are set equal to the true scores on X, the mean true score on Y will be equal to the mean true score on X for each group, and both groups will have the same true-score standard deviation for Y and X. In addition, the observed scores, X and Y, in the two groups contain independently

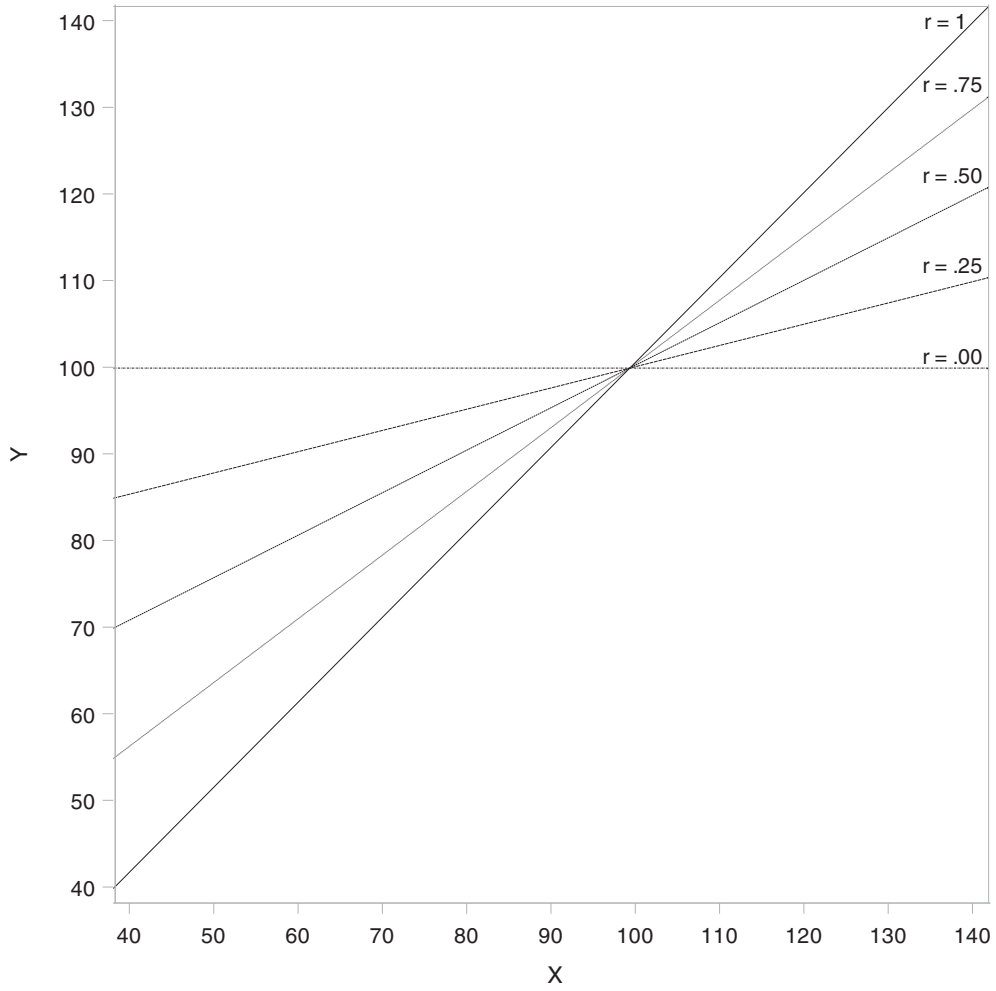

Figure 4 Illustration of the rotation of lines toward the horizontal, as the correlation between X and Y goes to 0.0.

Table 1 Simulated Data: Parameters

		True Score		
Group	Variable	Mean	*SD*	Error *SD*
H	*X*	100	10	6
	Y	100	10	6
L	*X*	70	10	6
	Y	70	10	6

sampled random errors drawn from a normal distribution with a mean of 0 and a standard deviation of 6. Table 1 contains a summary of these values.

Samples of 250 true scores for $y = x$ were randomly drawn for each group. Based on the parameters in Table 1, the means for X and Y for the two groups should be about 100 and 70, respectively, and the observed-score standard deviation should be about 11.66 (the square root of true score variance plus error variance, or the square root of $10^2 + 6^2$) for each variable in each group. The true scores on X and Y are perfectly correlated for both groups (i.e., $y = x$ for each individual), so the covariance between Y and X is equal to the true score variance for Y or X (10^2). The observed-score variances are equal ($10^2 + 6^2$), and therefore, the correlation between observed scores for Y and X for each group should be about

$$r_{YX} = \frac{(10)^2}{(10)^2 + (6)^2} = 0.735. \tag{22}$$

Table 2 Simulated Data: Observed Score Summary Statistics

		X Observed Score		*Y* Observed Score		
Group	*N*	Mean	*SD*	Mean	*SD*	*X*–*Y* Correlation
H	250	99.33	11.82	99.89	11.59	0.73
L	250	70.67	11.88	70.69	11.97	0.75

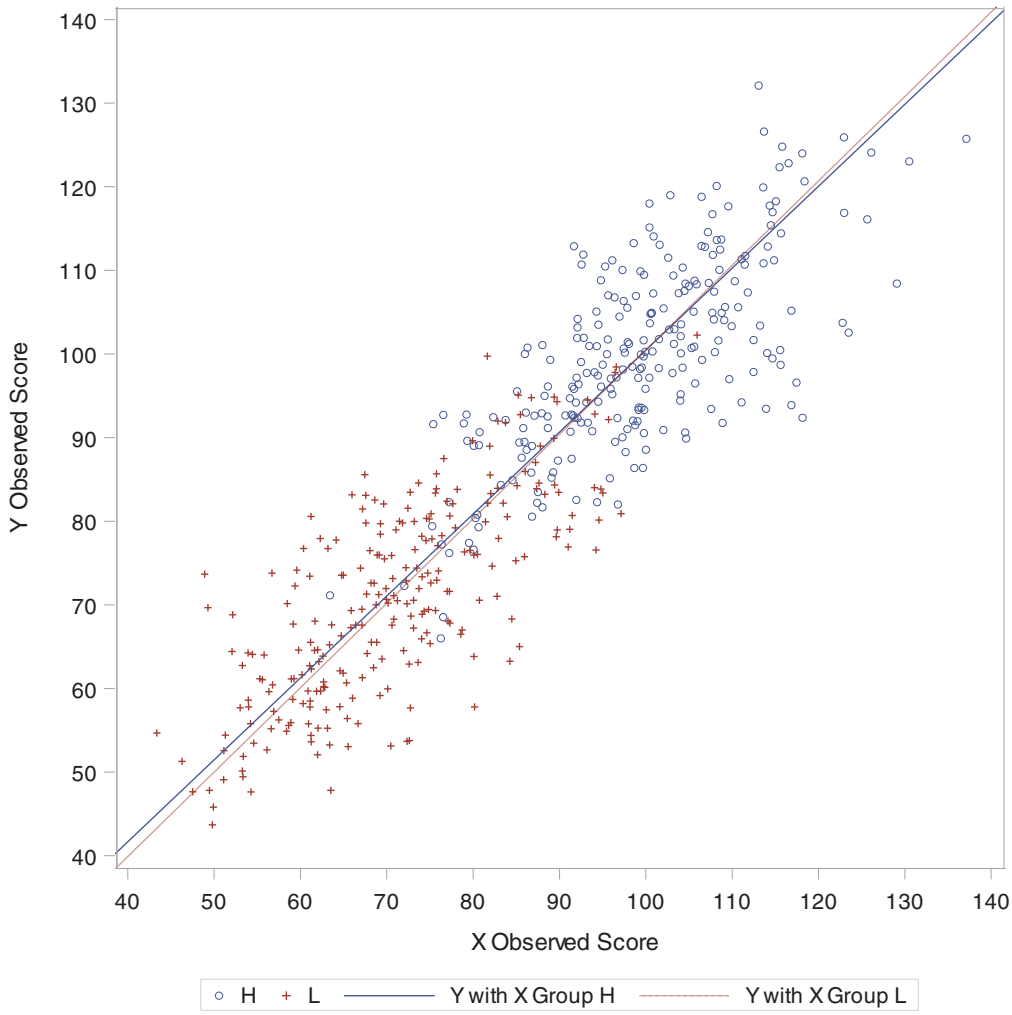

Figure 5 Scatterplot and group-specific geometric-mean (GM) regressions for simulated data with high-scoring (H) and low-scoring (L) groups.

In Table 2, we report summary statistics for the simulated data. The observed means, standard deviations, and correlations are close to the values expected, given the parameters used to generate the simulated data. The mean observed score for the high-scoring group is about 100, and the mean for the low-scoring group is about 70. The observed-score standard deviation for each group is within 0.3 of 11.66, and the correlation between *X* and *Y* for each group is close to 0.74.

Figure 5 provides a scatterplot, with the data points for the H group represented by "o," and the points for the L group represented by "+," along with the GM lines for the two groups. The two GM lines are very close, suggesting that the relationship between *X* and *Y* is essentially the same for the two groups and corresponds to the true-score relationship between *Y* and *X* (with the slight difference between the lines due to sampling variability). That is, in this case, the GM line recovers the underlying true-score relationship that was built into the data sets.

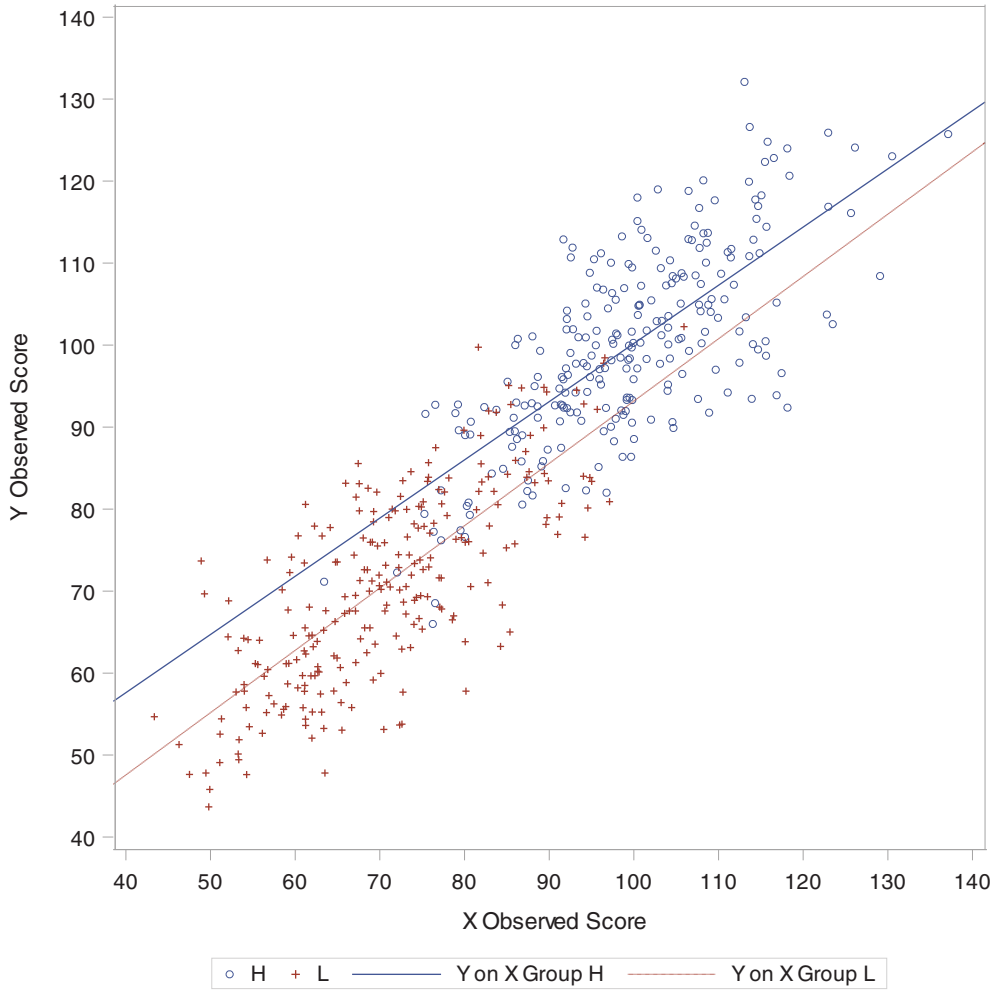

Figure 6 Scatterplot and group-specific ordinary least squares (OLS) regressions of Y on X for simulated data with high-scoring (H) and low-scoring (L) groups.

Figure 6 presents the OLS lines for Y on X for the two groups. The two OLS lines have slopes of $r_{YX}\frac{S_Y}{S_X}$, which are lower than the slope of 1.0 for the true score relationship built into the data, and because they pivot around different centroids, the two lines are separated, with the line for the H group above that for the L group.

Figure 7 presents the OLS regressions of X on Y for the two groups (with Y as the vertical axis and X as the horizontal axis). These lines have slopes that are greater than 1.0 by a factor of $\frac{1}{r_{YX}}\frac{S_Y}{S_X}$, causing the lines to be closer to vertical than the true-score relationship built into the data. These two lines also pivot around the different centroids, causing the lines to separate, with the OLS regression of X on Y for the higher scoring group being below that for the lower scoring group.

The OLS regressions yield different regressions for the two groups, and in each case, the regressions yield predictions with minimal squared errors of prediction. For the regression of Y on X, the line for the H group is above that for the L group, and for the regression of X on Y, the line for the H group is below that for the L group. Comparisons of the OLS regressions across groups provide a good empirical check for predictive bias (Cleary, 1968). However, the OLS-based comparisons do not provide an effective check for test bias; as noted previously, the simulated data for the two groups had identical true-score relationships between the two variables, but different means on the two variables and different OLS lines.

Note that if the true-score relationships are different across groups, say, in terms of an added or subtracted constant for X or Y (i.e., a group-specific systematic error), the regression-toward-the-mean effect inherent in OLS regression could eliminate, minimize, or even reverse the effect. The OLS model does not provide a dependable indication of differences in underlying true-score relationships across groups.

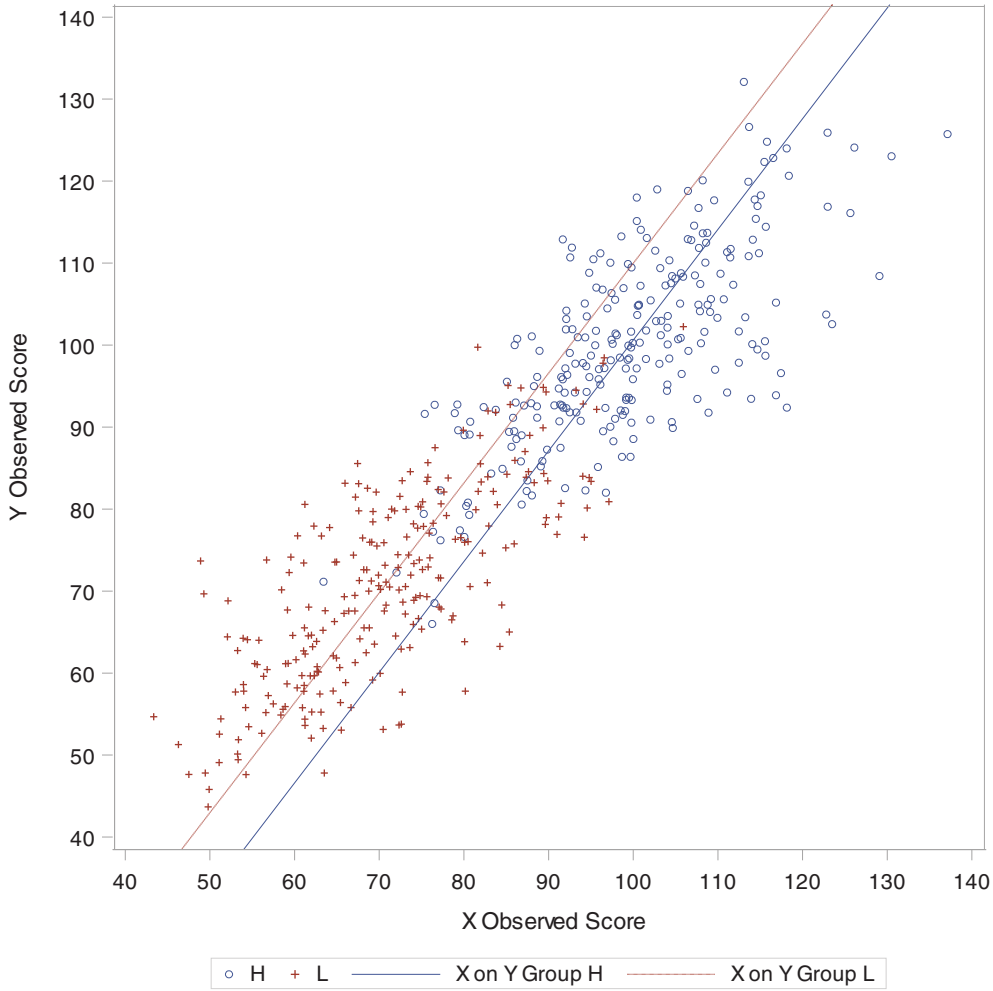

Figure 7 Scatterplot and group-specific ordinary least squares (OLS) regressions of X on Y for simulated data with high-scoring (H) and low-scoring (L) groups.

Y Intercepts for the Group-Specific Regressions of *Y* on *X*

By substituting 0.0 for X in group-specific versions of Equation 1, we can obtain the Y intercepts for the two groups. For the H group, the Y intercept is given by

$$Y_H int = \overline{Y}_H - r_{YX}\frac{S_Y}{S_X}\overline{X}_H. \tag{23}$$

For the L group, the Y intercept is given by

$$Y_L int = \overline{Y}_L - r_{YX}\frac{S_Y}{S_X}\overline{X}_L. \tag{24}$$

The difference between these two intercepts is given by

$$\Delta_{int} = Y_H int - Y_L int = \left(\overline{Y}_H - \overline{Y}_L\right) - r_{YX}\frac{S_Y}{S_X}\left(\overline{X}_H - \overline{X}_L\right). \tag{25}$$

The centroids for both groups are on the GM line, which has a slope of s_Y/s_X, and therefore

$$\left(\overline{Y}_H - \overline{Y}_L\right) = \frac{S_Y}{S_X}\left(\overline{X}_H - \overline{X}_L\right). \tag{26}$$

As a result, the difference between the intercepts given by Equation 25 can be written as

$$\Delta_{int} = \left(1 - r_{XY}\right) \frac{S_Y}{S_X} \left(\overline{X}_H - \overline{X}_L\right) = \left(1 - r_{XY}\right) \left(\overline{Y}_H - \overline{Y}_L\right). \tag{27}$$

So, if the correlation were to decrease from 1 to 0, the difference in Y intercepts would increase from 0.0 to $\left(\overline{Y}_H - \overline{Y}_L\right)$. In the absence of random error, $r_{YX} = 1.0$, the difference in Y intercepts is 0.0, and the two OLS lines coincide. As the correlation decreases, the OLS lines rotate around their centroids and the lines separate, even though the derivation assumes that the true-score relationship is the same in the two groups.

Lord's Paradox and Orthogonal Regression

In a deceptively simple, two-page paper, Lord (1967) called attention to some sources of confusion in drawing conclusions about the impact of an intervention using data from preexisting groups and the analysis of covariance (ANCOVA). "The situation is such that observed differences in the dependent variable might logically be caused by differences in the independent variable, and the research worker wishes to rule out this possibility" (Lord, 1967, p. 304).

Lord set up the paper as a thought experiment in which the data suggest a null result (no difference between groups), but ANCOVA suggests that there is a difference. In the 1967 paper, he developed the "paradox" in terms of an example, and in a subsequent publication, he extended his discussion of the paradox and introduced some additional examples.

We will focus on the example developed in the 1967 paper in which Lord assumed that a university is interested in the effects of the diet in its dining halls on student weight and in any differential impact between men and women and collects data on student weight on arrival in September and at the end of the academic year in June. These data, separated by gender, were analyzed by two statisticians.

The first statistician found that the mean weights of the girls at the beginning and end of the year are identical and, in fact, that the distributions of the girls' weights are also the same at the beginning and end of the year and the same is true for boys:

> Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year. (Lord, 1967, p. 304)

The average increase for both groups is 0.0, and in a bivariate plot of June weights against September weights, the centroids for both groups would be on the 45° line. Based on these results, the first statistician concluded the following:

> there is no evidence of any interesting effect of the school diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change. (Lord, 1967, p. 305)

This seems like a persuasive argument, as far as it goes.

The second statistician, working independently, conducted an ANCOVA, and finds that the slopes of the regressions of final weight on initial weight is the same for the two groups, but that there is a significant difference in the intercepts and concludes "the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes" (Lord, 1967, p. 305).

When pressed to explain this conclusion, the second statistician pointed out the following:

> If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distributions of initial weight, the relative positions of the regression lines shows that the subgroup of boys is going to gain substantially more during the year than the subgroup of girls. (Lord, 1967, p. 305)

Note that the second statistician seemed to be conceptualizing the comparison in terms of prediction, particularly in terms of differences between boys and girls in predicted June weights, given September weights.

Lord (1967) presents the question as one of possible causation, whether "observed differences in the dependent variable might logically be caused by differences in the independent variable" (p. 304). Several authors (Holland & Rubin, 1983;

Pearl, 2016; Wainer & Brown, 2007; Werts & Linn, 1969) have analyzed Lord's paradox in terms of causal models and in doing so have indicated how different conclusions can be generated by making different causal assumptions.

Our analysis of the paradox does not address the plausibility of claims that the dining room diet (or anything else) causes weight gain or differential weight gain; rather, we address the prior question of whether there is any differential weight gain. The paradox, as such, arises from the fact that the first statistician concluded that there is "no evidence of any differential effect" (Lord, 1967, p. 305) and the second statistician, using the same data, concluded that "boys showed significantly more gain in weight than the girls" (Lord, 1967, p. 305). Lord's scenario does not provide much, if any, evidence for any specific causal inference relating the dining room diet to changes in student weight between September and June. As the analyses by Holland and Rubin (1983), Pearl (2016), Wainer and Brown (2007), and Werts and Linn (1969) showed, even if one accepts the second statistician's conclusion, the data as described by Lord do not provide a good basis for concluding that the dining room diet is the cause of any such difference (rather than the vending machines in the dorms, the physical education curriculum, or natural age-related changes). The paradox results from the supposition that that two seemingly plausible analyses give different answers to the question of whether there was any differential change in the weights.

The questions posed by Lord's scenario are whether students tend to gain or lose weight between September and June and whether one group (boys or girls) tends to gain or lose more weight than the other. The goal is not to predict individual student weights in June from their weights in September; rather, it is to estimate the general relationship between weights at the beginning of the year (September) and at the end of the year (June). A positive answer to these questions could lead to analyses of possible causes, but as Holland and Rubin (1983) and others have pointed out, these causal investigations would require additional assumptions and additional data.

To the extent that regression analyses are to be used to address the prior question of whether there are changes to be explained, OR is more appropriate than OLS regression. In this context, the errors of measurement in student weights at the beginning and end of the year would include both the instrumental errors in the scales used and fluctuations in weight from hour to hour and day to day in September and June. In comparing weights between September and June, these fluctuations (person-occasion interactions) will function as sources of random error (Brennan, 2001; Cronbach et al., 1972). The second of these sources of error (variability over occasions) is likely to be much larger than the instrumental errors associated with the weight measurement, as such (assuming that the scales are decent), and both of these error sources are likely to be about the same in September and June. It is possible that some of the factors that lead to weight changes function differently in September and June, but in Lord's (1967) description of the thought experiment (with the weight distributions for both groups being the same in September and June), there is no indication that the error variances change from September to June. The error variances and the observed score variances, and hence the reliabilities, can be taken to be the same in September and June.

In agreement with Lord's (1967) tentative conclusion, an analysis in terms of OR (rather than OLS regression) indicates that the second statistician's analysis is far more suspect than that of the first statistician who concluded that because neither the boys nor the girls changed in weight on average, there is no evidence of any differential effect. The second statistician's argument based on ANCOVA is subject to serious criticism. As noted previously, both sets of weights, September and June, are subject to errors of various kinds. It seems reasonable to assume that the impact of these various sources of error will be similar in September and June, but in any case, it seems quite arbitrary to assign all of the error to the June weights, as is done in OLS regression. To the extent that the variability in the weights at the beginning of the year (i.e., in September) are comparable to the variability in the weights at the end of the year (i.e., in June), the OLS regressions employed in the ANCOVA analysis provide biased estimates of the true-score relationship between the September and June weights.

However, if the second statistician had used OR to investigate the relationships between September and June weights and, as suggested previously, had assumed that the errors of measurement were approximately the same in September and June, the GM lines would be appropriate (with equal error variances and equal observed-score variances) and would be identical for the two groups:

$$Y = \frac{S_Y}{S_X}\left(X - \overline{X}\right) + \overline{Y} = \left(X - \overline{X}\right) + \overline{Y}. \tag{28}$$

The GM lines provide more plausible estimates of the true-score relationships between the variables for each group than OLS regression (see Figure 5).

The OLS regressions developed by the second statistician would provide optimal predictions of individual student weights in June based on the students' weights in September. In this context, it is the student's observed weight in September that would be used to predict the student's weight in June, and this observed weight can be said to be known without error. Furthermore, the second statistician's analyses would indicate that such predictions could be improved by using separate OLS regressions for men and women. Using separate equations takes advantage of additional information about each student's gender, which in this case, is strongly related to weight.

At the end of his brief discussion, Lord (1967) concludes "with the data usually available for such studies, there is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups" (p. 305). It seems that Lord saw his analysis as suggesting that ANCOVA cannot be counted on to adequately account for differences between naturally occurring groups. We suggest that a basic problem with the ANCOVA analysis is the use of OLS regression to answer a question about group-specific, true-score relationships in a case in which both variables contain comparable levels of error and the two groups have substantially different means for X and Y.

Note that there are a number of potential sources of bias in drawing inferences based on comparisons across naturally occurring, preexisting groups (Werts & Linn, 1969), and the use of OR in place of OLS addresses only one of these threats to the accuracy of the conclusions. However, the threat addressed, regression toward the mean, tends to be a particularly serious threat when the groups differ substantially in their mean scores and both variables contain substantial error.

Limitations of the Orthogonal Regression Models

As mentioned earlier, there are a number of difficulties associated with the application of the OR models. Although we will suggest some possible ways of addressing these issues, we recognize that we have not solved the problems, and some of them may not have general solutions.

Probably the most serious limitations are the difficulty in estimating the total errors (due to all potential sources of measurement error) in the two variables and the question of how to deal with equation errors. The EIV model only requires the ratio of these error variances, but except in some special cases (e.g., Lord's thought experiment, equating), estimating the ratio generally requires estimates of the two error variances. In cases where a substantial part of the overall variability around the fitted line is not accounted for by the errors of measurement in the two variables (i.e., where the equation error is substantial), the EIV model cannot be unambiguously applied.

If a thorough analysis of potential sources of error (Brennan, 2001; Carroll & Ruppert, 1996; Cronbach et al., 1972) accounts for most of the variability around the line, the equation error could be assigned to the two error variances in some way (e.g., half to each) and the resulting line might be acceptable as a useful approximation. If the question being asked is one that fits the OR models, it may be more reasonable to use the EIV model with the best available estimate of δ and to explicitly recognize this source of uncertainty in the analysis (e.g., by providing upper and lower bounds) than to use a less appropriate model. One way to implement this strategy would be to also estimate separately the EIV line that results from assigning all of the equation-error variance to the error variance for X and the EIV line that results from assigning all of the equation-error variance to the error variance for Y; these two lines would pass through the centroid and would provide an indication of the uncertainty in the slope of the line due to the equation errors. If the equation errors are large, this uncertainty will be large, and at some point, it would be prudent not to rely on the EIV model (Draper, 1991; Ricker, 1973). More work is need in this area (see Carroll & Ruppert, 1996).

As suggested earlier, if the variables and context are such that we can reasonably assume that the two variables have approximately equal reliabilities, as in the two illustrations in this paper and in equating, the GM model can be used, and it does not require estimates of the error variances or their ratio. In cases where measurement errors are small and most of the variability is what Ricker (1973) calls natural variability rather than measurement error (e.g., in comparing the heights of brothers and sisters), the GM model can provide a natural approach to estimating a functional relationship that provides a concise summary of the trend in the data (Kruskal, 1953).

Another issue concerns applications in which there are more than two variables. All of the OR models are stated in terms of the relationship between two variables, X and Y, and do not allow for multiple variables on either side of the equation. This is a limitation of these models, but it is not clear that it is a serious limitation in practice. OLS regression analyses

often involve multiple independent variables and one dependent variable because the goal is to get good predictions of the dependent variable and prediction tends to get better if additional, relevant independent variables are added to the analysis. If one is interested in predicting a dependent variable from several independent variables, one should ordinarily use OLS regression. In cases where one is interested in the functional relationship between X and Y, additional variables that have impacts on X or Y tend to cloud the issue and are likely to be treated as sources of error in X or Y or as equation error.

In cases (e.g., in bias studies) where researchers are interested in the true-score relationships between a criterion variable and several predictor variables (e.g., in order to identify any predictors that exhibit measurement bias), it would probably be best do the analyses for each potential predictor variable separately. If one of the predictor variables exhibits more test bias than the others, one might want to consider dropping that variable.

If researchers are interested in the relationship between a composite variable (e.g., including test scores and high school grade-point average) and some criterion variable (college grade-point average), the composite can be taken as a single variable, and the OR analysis can be conducted using this composite variable and the criterion. Optimal weightings for the composite can be estimated using OLS regression. This potential application of OR will be difficult (if not impossible) in many cases because of large equation errors.

If a researcher is interested in examining true-score relationships among three or more variables, the pair-wise relationships between variables can be examined, but a full, integrated analysis would generally require a more complex model. So, given the main use that we have proposed for the OR models discussed here—investigating true-score relationships between variables—the fact that the models involve only two variables is not necessarily a serious problem in many applications.

Culpepper et al. (2019) provided a sophisticated joint analysis of predictive bias and measurement bias, based on a factor model for a criterion variable and multiple independent variables. Their analyses yielded a number of interesting results but made a number of strong assumptions about the factor structure; they did not consider the issue of equation errors. These different models (OLS, EIV, GM, and the Culpepper et al. model) make different assumptions and draw different conclusions, and the appropriate model to use in a specific case will depend on the assumptions that are considered plausible and the questions being asked.

Conclusion

OLS and ORs address different questions and seek to minimize different mean squared deviations. OLS regression provides an optimal least-squares prediction of a dependent variable contingent on one or more independent variables. OR estimates a line that represents the functional (or true-score) relationship between two variables.

OLS regression provides a natural approach when the goal is to use a known observed score to predict an unknown outcome variable (e.g., in selection). It is relatively easy to implement, and predictive score interpretations have always been important in educational measurement (Gulliksen, 1950); so, it is not surprising that OLS regression has gotten a lot of attention in educational research in general, and in educational measurement in particular. If one's goal is to predict one variable (the dependent variable) from one or more other variables (the independent variables), the assumptions in the OLS model are highly plausible and the resulting line provides optimal predictions.

OLS regression assumes that the dependent variable is contingent on a known independent variable, and all of the error (or uncertainty) is assigned to the dependent variable, which is to be predicted. That the independent variable is measured without error is not generally a plausible assumption, but in the context of prediction, it is the observed value of X that is used as the basis for the prediction.

However, OLS regression has limited utility if the goal is to examine the relationship between true scores on two measures where both measures contain substantial error because regression toward the mean reduces the slope of the resulting line (e.g., Berry, 1993; Berry & Feldman, 1985; Draper & Smith, 1998; Lewis-Beck, 1980). When used to compare true-score relationships for groups with different means on the two variables, OLS regression introduces bias in the estimates of the slopes of the lines, which causes the lines to separate even if the true-score relationship is exactly the same for the two groups.

OR has been around for more than a hundred years (Adcock, 1877; Pearson, 1901) but has gotten less attention than OLS regression because it is more complicated to implement (Carroll & Ruppert, 1996; Deming, 1943; Mandel, 1964) and does not provide optimal predictions. However, in many contexts, reearchers are interested in the functional relationships

between constructs, or hypothesized true values of the variables, and not so much in relationships between observed scores, containing various amounts of error (Cronbach, 1971; Cronbach & Meehl, 1955; Dorans & Holland, 2000; Kane, 2006, 2013; Loevinger, 1957; Messick, 1989).

Differences between the group-specific OLS regressions provide an empirical check on predictive bias (Cleary, 1968). OR provides more accurate estimates of true-score relationships, and therefore, in addressing questions of test bias, OR is generally more appropriate than OLS regression. Differences in true-score relationships across groups would indicate that one or both of the variables involved somewhat different meanings for the groups and, therefore, would indicate that one or the other or both variables are biased in some way (Kane & Mroch, 2010).

In Lord's (1967) hypothetical scenario, the question that gives rise to the paradox is whether there were any systematic differences between boys and girls in weight changes between September and June. Lord's exposition did not indicate any interest in predicting individual June weights from September weights, and the role of the dining room diet as a possible cause of any such difference was not seriously considered. If OR had been used by the second statistician in Lord's scenario instead of OLS regression, the two statisticians would have arrived at the same conclusion (that there was no difference in overall weight changes between boys and girls), thus resolving their disagreement.

In general, it is important to choose a model that reflects the question being asked and one for which its assumptions are satisfied. If the question focuses on estimating the true-score relationship between two variables, some form of OR may be appropriate. If the goal is to predict some variable, Y, based on known values of another variable, X, the OLS regression of Y on X would provide an optimal least-squares line.

When the goal is to identify the line that provides the best overall fit to the data (for both the X and Y deviations) for each group, rather than predicting one variable from the other, OR is preferred, if it can be applied. In cases where the reliabilities are approximately equal for X and Y, the GM model can be used. If the reliabilities are not equal but the ratio of the error variances can be estimated, the EIV model would be the preferred approach. However, in many educational contexts, there are extraneous factors that contribute substantial variability that goes beyond measurement error as such and therefore the equation errors are likely to be large, making it difficult to implement the OR models.

However, in contexts where most of the variability tends to be due to measurement errors that can be assigned to one or the other of the two measures and, therefore, the equation errors are relatively modest (Kane & Mroch, 2010), OR can be applied and can provide useful insights about some issues (e.g., test bias, as distinct from predictive bias).

References

Adcock, R. (1877). Note on the method of least squares. *The Analyst*, *4*(6), 183 – 184.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, *5*(3), 207 – 212. https://doi.org/10.2307/3001936

Berry, W. (1993). *Understanding regression assumptions.* Sage.

Berry, W., & Feldman, S. (1985). *Multiple regression in practice*. Sage.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221 – 256). American Council on Education and Praeger.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. Guilford Press.

Carroll, R. J., & Ruppert, D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician*, *50*(11), 1 – 6. https://doi.org/10.2307/2685035

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115 – 124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Cliff, N. (1987). *Analyzing multivariate data*. Harcourt Brace Jovanovivch.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201 – 219). American Council on Education and Macmillan.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443 – 507). American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, *84*(1), 285–309. https://doi.org/10.1007/s11336-018-9649-2

Deming, W. E. (1943). *Statistical adjustment of data*. Dover.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Draper, N. R. (1991, April 28–30). *Straight line regression when both variables are subject to error*. Paper presented at the 3rd Annual Conference on Applied Statistics in Agriculture, Manhattan, KS. https://doi.org/10.4148/2475-7772.1414

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). Wiley.

Fuller, W. A. (1987). *Measurement error models*. Wiley.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Holland, P. W., & Dorans, N. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). American Council on Education and Praeger.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement *(pp. 3–35)*. Erlbaum.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M., & Mroch, A. (2010). Modeling group differences in OLS and orthogonal regression: Implications for differential validity studies. *Applied Measurement in Education*, *23*(3), 215–241. https://doi.org/10.1080/08957347.2010.485990

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.

Kruskal, W. (1953). On the uniqueness of the line of organic correlation. *Biometrics*, *9*(1), 47–58. https://doi.org/10.2307/3001632

Lewis-Beck, M. S. (1980). *Applied regression: An introduction*. Sage.

Linn, R., & Werts, C. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*(1), 1–4.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph Supplement 9]. *Psychological Reports*, *3*, 635–694. https://doi.org/10.2466%2Fpr0.1957.3.3.635

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304–305. https://doi.org/10.1037/h0025105

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, *54*(285), 173–205.

Mandel, J. (1964). *The statistical analysis of experimental data*. Dover.

Messick (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Nievergelt, Y. (1994). Total least squares: State-of-the-art regression in numerical analysis. *SIAM Review*, *36*(2), 258–264. https://doi.org/10.1137/1036055

Pearl, J. (2016). Lord's paradox revisited–(oh Lord! kumbaya!). *Journal of Causal Inference*, *4*(2), 20160021. https://doi.org/10.1515/jci-2016-0021

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *6*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pedhazur, E. J., & Schmelkin, L. (1991). *Measurement, design, and analysis*. Erlbaum.

Ricker, W. (1973). Linear regressions in fishery research. *Journal of Fishery Research Board Canada*, *30*(3), 409–434. https://doi.org/10.1139/f73-072

Riggs, D., Guarnieri, J., & Addelman, S. (1978). Fitting straight lines when both variables are subject to error. *Life Sciences*, *22*(13–15), 1305–1360. https://doi.org/10.1016/0024-3205(78)90098-X

Sprent, P. (1969). *Models in regression and related topics*. Methuen.

Sprent, P. (1990). Some history of functional and structural relationships. In P. J. Brown & W. A. Fuller (Eds.), *Statistical analysis of measurement error models and applications*. American Mathematical Society.

Wainer, H., & Brown, L. M. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26 Psychometrics* (pp. 893–918). Elsevier.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, *11*(3), 284–300.

Weisberg, S. (1985). *Applied linear regression*. Wiley.

Werts, C., & Linn, R. L. (1969). *Donald Campbell's advice applied to Lord's paradox* (Research Bulletin No. RB-69-08). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00168.x

Wonnacott, R. J., & Wonnacott, T. H. (1979). *Econometrics* (2nd ed.). Wiley.

Suggested citation:

Kane, M. T., & Mroch, A. A. (2020). *Orthogonal regression, the Cleary criterion, and Lord's paradox: Asking the right questions* (Research Report No. RR-20-14). Educational Testing Service. https://doi.org/10.1002/ets2.12298

Action Editor: John Mazzeo

Reviewers: Rebecca Zwick