A Regression Paradox for Linear Models: Sufficient Conditions and Relation to Simpson’s Paradox

Aiyou CHEN, Thomas BENGTSSON, and Tin Kam HO

An analysis of customer survey data using direct and reverse linear regression leads to inconsistent conclusions with respect to the effect of a group variable. This counterintuitive phenomenon, called the “regression paradox,” causes seemingly contradictory group effects when the predictor and regres-
sand are interchanged. Using analytical developments as well as geometric arguments, we describe sufficient conditions under which the regression paradox will appear in linear Gaussian models. The results show that the phenomenon depends on a distribution shift between the groups relative to the predictabil-
ity of the model. As a consequence, the paradox can appear naturally in certain distributions, and may not be caused by sam-
pling error or incorrectly specified models. Simulations verify that the paradox may appear in more general, non-Gaussian set-
ings. An interesting, geometric connection to Simpson’s para-
dox is provided.

KEY WORDS: Group effect; Monotone regression; Regression paradox; Reverse regression; Simpson’s paradox.

1. INTRODUCTION

One important use of statistical methodology is in the as-
sestment of group differences. When the data are collected by
means of a controlled experiment, the designation of response
and explanatory variables is unambiguous, and the data col-
lection procedure typically outlines the subsequent statistical
analysis. For experimental data, the group effect then may be
simply defined as the mean difference in the response vari-
able between groups. In contrast, when the data stem from
an observational study, such methodological neatness is atyp-
ical, and depending on how the data are approached, statisti-
cal analyses may even produce seemingly contradictory con-
clusions. Probably the best-known example of such complexity
is Simpson’s paradox (Simpson 1951), with the University of
California Berkeley graduate data set as a well-known example
(Bickel, Hammel, and O’Connell 1975).

In the linear regression setting with observational data, a simi-
lar but less well-known phenomenon may occur when group
effects are compared from direct and reverse regressions. In this
case, exchanging the roles of the regressand and the predictor
variable produces conflicting results with respect to the group
effect. This phenomenon has been reported in studies of salary
discrimination (Conway and Roberts 1983), where in the pres-
ence of continuous covariates, linear regression with dummy
variables is used to test for gender differences. In Conway
and Roberts (1983) and Conway’s and Robert’s (1983) analysis,
when the roles of the continuous predictor (i.e. qualification)
and regessand (i.e. salary) variables are switched, the authors
reported contradictory results with respect to the gender effect.
Specifically, when controlling (statistically) for differing levels
of qualification, women appear to have a lower expected salary.
Consequently, to obtain the same expected earnings, women re-
quire higher qualification levels. Paradoxically, however, when
controlling for salary differences, men appear to have higher
qualification levels (Conway and Roberts 1983). A somewhat
heated debate on the merits of reverse regression in salary dis-
rimination studies ensued (e.g., Conway and Roberts 1984;
Goldberger 1984; and Green and Ferber 1984), but it did not
resolve the phenomenon technically. Other work has attributed
the paradox to collinearity among predictor and group variables
(Whiteside and Narayanan 1989), but, as we show later, ex-
cessive association between the regressors actually removes the
paradox and thus cannot fully explain the phenomenon. Some
authors have considered the phenomenon in terms of incor-
rectly specified measurement processes (Samsa 1992) or mod-
els (Racine and Rilstone 1995). A brief introduction to statisti-
cal paradoxes has been given by Geng (2006).

To shed light on the regression phenomenon, our work pro-
vides sufficient conditions for the paradox to appear in linear,
Gaussian settings. As we show later, the paradox can arise natu-
raly in some scenarios and is not necessarily the result of sam-
ping error, collinearity, or misspecified models, as has been
suggested previously. Simulations further show that the phe-
nomenon is possible in more general, non-Gaussian settings.
We also provide an interesting geometric connection between
the regression and Simpson’s paradox.

This article is organized as follows. The next section de-
cribes a recent study of customer satisfaction data, in which we
rediscovered the regression paradox. Section 3 then gives suf-
cient conditions for the paradox to occur in a linear, Gaussian
system and provides a geometric interpretation of the phenom-
emon. Section 4 discusses generalizations to non-Gaussian set-
tings, and Section 5 connects the phenomenon to Simpson’s
paradox. Section 6 concludes the article with a brief discussion.

2. THE REGRESSION PARADOX

To provide context for the technical conditions of the regres-

sion paradox, in this section we describe the data analysis prob-
lem in which we encountered the phenomenon. The overall purpose of this analysis is to gain insight into which factors are the most important to customer satisfaction. Sections 2.1 and 2.2 describe the data and the variable selection step. Section 2.3 describes the regression paradox in the framework of a linear model.

2.1 Data Description
The available data consist of two databases pertaining to the year 2006. The first database, referred to as AR (assisted requests), details customer-reported network issues and resolutions. The AR database contains about 100 fields describing the customer and the product, along with problem solution descriptions; example fields include the type of network problem reported (e.g., hardware or software), to which product family and business unit the problem belongs, and how the issue was resolved. The database also contains time-tracking information (e.g., report, target, and actual resolve dates). The second database, termed CS (customer satisfaction), contains responses to a customer satisfaction survey. The CS data are collected after the resolution of an AR problem; thus each CS record can be matched to at least one entry in AR. The survey contains approximately 20 multiple-choice questions assessing overall satisfaction as well as satisfaction with specific aspects of the process (e.g., whether a timely response is provided). The survey scores take values ranging from 1 (least satisfied) to 10 (most satisfied), with a median score of around 8.

2.2 Data Analysis
Because the information in AR precedes (in time) that of CS, we initially conceptualize and model customer satisfaction scores as a function \( f(\cdot) \) of the variables contained in AR; schematically, we let

\[
CS = f(AR) + \epsilon. \tag{1}
\]

Here \( \epsilon \) is a noise term representing unexplained variability. The functional relationship in (1) was modeled using classical methods (e.g. stepwise linear and logistic regression procedures), and classification and regression trees (CART) were used for variable selection. As a result of these analyses, Resolve Duration (i.e. time in days between the opening and closing of a ticket) and Region (regions A and B) stand out as the most significant predictor variables of Satisfaction Scores.

We further note that because the data are observational, we could attempt to reverse the roles of the CS and AR variables in (1). For instance, assume that while conditioning on satisfaction scores, we are interested in identifying whether the customer receives the same level of service. To exemplify this, Figure 1 shows boxplots of resolve duration (given in log base 2) conditional on satisfaction score and grouped by region (region A, top panel; region B, bottom panel). Although some sample variability is present, the figure indicates that the median values of resolve duration may be monotonically decreasing with satisfaction score. This is not surprising; high satisfaction scores are associated with shorter resolve durations, and, as can be seen, the result appears to hold true for both regions.

Next we describe the paradox generated when the regional effect is modeled using regression techniques.

![Figure 1. Boxplots of resolve duration (log base 2) conditional on region and satisfaction score. The data for region A are plotted in the top panel; that for region B, in the bottom panel. Both predictor variables are represented along the y-axis.](image)

2.3 Regression Paradox
As suggested by the boxplots in Figure 1, satisfaction score and resolve duration appear to be related through a monotone function. Thus, to approximate the regional effect for both satisfaction score and resolve duration, we may fit linear models with the group effect modeled by a dummy variable. The regression paradox then can be described by comparing the fitted values for each region. Toward this end, using data from both regions, we propose a linear model of the form

\[
(Satisfaction Score)_{ij} \sim \alpha + \beta(Resolve Duration)_{ij} + \gamma R_i + \xi_{ij}. \tag{2}
\]

In (2), \( R = 1 \) if the data are for region B and \( R = 0 \) otherwise, and \( \xi_{ij} \) is independent, mean-0 noise with variance \( \sigma^2_\xi \). For this model, the additive group effect is estimated as \( \hat{\gamma} = 0.73 \), with an approximate standard error of 0.10. Thus, given the same resolve duration, the expected satisfaction score is 0.73 points higher for region B customers compared with region A customers. We also fit a linear model for the reverse regression and obtain an estimated regional effect of \(-1.49\) (with standard error 0.19), implying that region B customers have shorter expected resolve duration given the same satisfaction score. The fitted lines from both regressions are shown in Figure 2. The estimated regression coefficients from the direct and reverse models are \(-0.086\) and \(-0.32\), with standard errors of 0.012 and 0.047, respectively. To facilitate visualization, in this plot the raw data have been jittered.

For these data, the regression paradox now can be realized as follows. The fitted lines in Figure 2(a) indicate that for a given resolve duration, the expected satisfaction score is uniformly higher for region B than for region A. Thus, controlling for the negative effects of resolve duration, the result seems to suggest...
that region B customers are happier with our service than region A customers, and to obtain the same level of expected satisfaction, region B customers wait longer and thus appear to be more patient than region A customers. The results pertaining to the reverse regression seems to contradict this conclusion, however. In particular, Figure 2(b) shows that given satisfaction score, region A customers wait longer on average, and thus appear to be more patient than region B customers.

Although the model specified in (2) is intuitive, it is possible that fitting a more flexible statistical model could remove the contradiction. To account for this possibility, we fit four separate monotone regressions (one for each region in both the forward and reverse cases). This analysis imposes fewer restrictions on the model, but the results from this analysis (not shown) lead to the same contradictory conclusion regarding the group effect as was obtained from the linear regression fits.

For the data analysis presented here, it may be argued that choosing, fitting, and validating a more appropriate statistical model may remove the contradiction. Certainly, the error processes are hardly Gaussian, and thus alternative models might affect the results. As mentioned earlier, previous work has considered the paradox in the context of incorrect measurement error processes (Samsa 1992), model misspecification (Racine and Rilstone 1995), and multicollinearity (Whiteside and Narayanan 1989). Other work has interpreted the phenomena in the framework of causal inference (Holland 1986; Wainer and Brown 2004; Arah 2008). This work attributes the phenomenon to the nature of observational and descriptive studies, with their associated lack of control and measurement of crucial variables. Our present work addresses the problem from a different vantage point. In particular, we show that the regression paradox is consistent with some joint distributions of the regressand and predictor variables, and is not necessarily the consequence of inadequately specified statistical models or poor data analysis.

In the next section, we give sufficient analytic conditions for the paradox in a simple Gaussian setting.

3. INVESTIGATION OF THE REGRESSION PARADOX UNDER A LINEAR GAUSSIAN MODEL

As shown by the figures, neither resolve duration or satisfaction score is Gaussian-distributed, and their expected values do not appear to be linearly related. But to establish the conditions under which the paradox occurs, as well as for analytic convenience, here we assume a Gaussian linear framework. Before proceeding, we explicitly state the paradox.

3.1 Contradictory Group Effects

Suppose that there are two groups, A and B. For each group, we assume that the expected value of $Y$ conditional on $X$ is given by a straight line with the same slope coefficient but different intercepts. Figure 3(a) shows an example with negative slope, where the two groups are offset in the $y$- and $x$-directions by constant terms $df$ and $d'$. In this case, for the direct regression, $df > 0$ represents the group effect, and $d' > 0$ is the horizontal offset required to obtain the same predicted value for $Y$ in the two groups. We now reverse the regression and consider the specific (and surprising) scenario depicted in Figure 3(b). As can be seen, in this plot the (vertical) offset between groups A and B is negative (i.e. $d_r < 0$), implying that given any $Y = y^*$, the prediction for $X_B$ is less than that for $X_A$. But this group effect contradicts the forward regression, which requires $x_B > x_A$ to obtain the same predicted value $y^*$ for $Y$. For linear models, inconsistent conclusions of this kind appear only when the $x$-displacement required in the forward regression to obtain the same expected $y$-value in both groups differ in sign from the group effect in the reverse regression, and we take this to be...
Figure 3. Illustration of the regression paradox. The lighter lines represent group A; the thicker lines, group B. (a) The forward regression with $d_f, d' > 0$. (b) The reverse regression with $d_r < 0$.

the meaning of the paradox. Equivalently, when the slope coefficient is negative, the paradox occurs when the group effects from the direct- and reverse regression (i.e., $d_f$ and $d_r$) differ in sign. (This argument assumes that the slopes do not change sign when we move between forward and reverse regression, a fact that is trivially verified.)

As noted earlier, the paradox cannot occur in the absence of noise, because, in this case, the vertical and horizontal offsets are not altered between the forward and reverse models; that is, for this scenario, $d' \equiv d_r$ and sign$(d_f) = $ sign$(d_r)$. To clarify precisely when the phenomenon occurs, we next delineate sufficient conditions for the paradox.

### 3.2 Sufficient Conditions for Paradox

Suppose that we have two random variables, $X$ and $Y$, and two groups, A and B. Within each group, we assume that the pair $X, Y$ follows the bivariate Gaussian distribution. Given these assumptions, there exist slope and intercept parameters, as well as mean-0, independent noise terms $\epsilon_A$ and $\epsilon_B$, such that the linear models

\[
Y_A = \beta_A X_A + \gamma_A + \epsilon_A \quad \text{and} \quad Y_B = \beta_B X_B + \gamma_B + \epsilon_B.
\]

describe the joint distribution of $Y$ and $X$ for each group. In what follows, without loss of generality, we parameterize the group effect by setting $\gamma_A \equiv -\gamma$ and $\gamma_B \equiv \gamma$. In this section we also take $\beta_A = \beta_B \equiv \beta$ and further assume that the correlation between $X$ and $Y$ is the same in both groups. A discussion of the impact of different correlations in each group is deferred to Section 3.3. Moreover, to parallel the data example of Section 2, we take $\beta < 0$. Thus we can write

\[
Y_i = \beta X_i + \gamma I_i + \epsilon_i, \quad i \in \{A, B\},
\]

where $I_A = -1$ and $I_B = 1$ are (nonrandom) indicator variables denoting group membership.

Now we look at the reverse regression. Again, given our assumptions, there exist parameters and random variables such that

\[
X_i = \frac{1}{\beta_r} Y_i - \frac{\gamma_r}{\beta_r} I_i + \eta_i, \quad i \in \{A, B\},
\]

where the noise processes $\eta_A$ and $\eta_B$ have mean 0 and are independent of $Y_A$ and $Y_B$. The parameterization in (5) may seem awkward, but it can be justified by solving for $X_i$ in (4). In particular, as we show later, this form yields a straightforward and intuitive condition for the paradox that can be easily illustrated geometrically (cf. Figure 4). In passing, we note that $\text{sign}(\beta) = \text{sign}(\beta_r) = \text{sign}(\text{cov}(Y, X))$, and because $\beta < 0$ and $\beta/\beta_r = (\text{corr}(X, Y))^2$, we have $\beta_r < \beta$.

As noted earlier, the regression paradox appears whenever the intercepts $\text{sign}(\gamma) \not= \text{sign}(\gamma_r)$. Now, using (4) and (5), we can write

\[
2\gamma = E(Y_B) - E(Y_A) - \beta(E(X_B) - E(X_A)) \quad \text{and} \quad 2\gamma_r = E(Y_B) - E(Y_A) - \beta_r(E(X_B) - E(X_A)),
\]

and, by taking the ratio of the foregoing two equations on both sides, we obtain

\[
\frac{\gamma}{\gamma_r} = \frac{s - \beta}{s - \beta_r},
\]

where $s = (E(Y_B) - E(Y_A))/(E(X_B) - E(X_A))$ measures the slope of the mean difference between the two groups. We can now state the following proposition.

**Proposition 3.1.** For the linear model considered here, with $\beta < 0$, if $E(X_B) \not= E(X_A)$, the regression paradox appears if and only if $\beta_r < s < \beta$.

The result follows by noting that the inequality condition produces a negative ratio in (6). Then, as defined in Section 3.1, the paradox appears. Using the same developments as before, the proposition can be easily extended to the case $\beta > 0$. 

The American Statistician, August 2009, Vol. 63, No. 3 221
The proposition makes clear that the paradox cannot occur when the slope for the means differs greatly from those of the regression lines, and, in particular, $s$ must have the same sign as the slope coefficients. Moreover, in agreement with intuition (e.g. Sec. 3.1), the phenomenon is unlikely to occur when the noise level is low. To see this algebraically, write $t_u$ tuition (e.g. Sec. 3.1), the phenomenon is unlikely to occur sign as the slope coefficients. Moreover, in agreement with in-position 3.1. Moreover, the paradox cannot occur if $\gamma$ lies in the shadowed area determined by the forward and reverse regression lines. The thicker lines represent group A, and the thinner lines represent group B. Let $\beta$ represent the reverse regression; the lighter lines represent the forward regression of $\gamma$. Thus Figure 4 shows that the paradox occurs if and only $\beta_r < s < \beta$. (b) The paradox condition is not satisfied with $s > \beta$.

3.3 The Regression Paradox With Group-Specific Models

Here we consider the scenario with group-specific slopes and intercepts. For this case, a proof similar to that of the previous section based on first- and second-moment properties can be constructed to show when the paradox occurs. To keep the argument transparent, here we give only a geometric justification. As we show, when the slopes and intercepts are unconstrained, the paradox always appears in some “conflict zone” in sample space.

Consistent with the description of the paradox given in Sections 3.1 and 3.2, we construct the geometric argument by delineating sets of $x$, $y$-values where the paradox occurs. Toward this end, we specify regions in $x$-space needed to obtain the same $y$-prediction in the forward model. These regions, given by the inequalities in (i) and (ii) in the next paragraph, describe where the $x$-displacement is positive or negative (cf. $d'$, Sec. 3.1). The inequalities given in (iii) and (iv) specify regions in which $E(X_A|Y = y) > E(X_B|Y = y)$ (and vice versa). An intersection of these regions then yields the paradox, as is illustrated with a specific case.

We follow the parameterization in (3) and write the forward models in terms of expected values, that is, $E(Y_A|X = x) = \beta_A x + \gamma_A$ and $E(Y_B|X = x) = \beta_B x + \gamma_B$. With different slopes, the lines (expected values) for the two groups A and B intersect at $(R_x, R_y)$. Below and above $R_y$, we have either

- (i) $x_A(\tilde{y}) < x_B(\tilde{y}) \quad \forall \tilde{y} < R_y$ and
- $x_A(\tilde{y}) > x_B(\tilde{y}) \quad \forall \tilde{y} > R_y$ or
- (ii) $x_A(\tilde{y}) > x_B(\tilde{y}) \quad \forall \tilde{y} < R_y$ and
- $x_A(\tilde{y}) < x_B(\tilde{y}) \quad \forall \tilde{y} > R_y$,

where $x_A(\tilde{y})$ and $x_B(\tilde{y})$ denote the values of $x$ required to predict a fixed $\tilde{y}$ for group A and group B, respectively.
The reverse models assume a similar form and intersect at \((S_x, S_y)\). That is, with \(g_A(y) \equiv E(X_A|Y = y)\) and \(g_B(y) \equiv E(X_B|Y = y)\), we have either

\[
\begin{align*}
(iii) & \quad g_A(y) < g_B(y) \quad \forall y < S_y \\
& \quad g_A(y) > g_B(y) \quad \forall y > S_y
\end{align*}
\]

or

\[
\begin{align*}
(iv) & \quad g_A(y) > g_B(y) \quad \forall y < S_y \\
& \quad g_A(y) < g_B(y) \quad \forall y > S_y
\end{align*}
\]

Now the locations of \(R_y\) and \(S_y\) divide the range of \(y\) into three intervals, and it can be seen that regardless of which of the forward and reverse models are in effect, in at least one of these intervals the two sets of models yield contradictory conclusions in terms of the group effect. To illustrate, we assume that \(\beta_A < \beta_B < 0\), \(R_y < S_y\), and \(R_y < S_y\), and that (i) and (iv) are true (see Figure 5). Now, as mentioned earlier, the half-space above \(R_y\) will overlap with the half-space below \(S_y\), and the paradox occurs when we compare the forward and reverse models in the interval \(y < R_y\) or \(y > S_y\), where the offsets in the \(x\)-direction experience the same kind of sign change as described in Section 3.1 (cf. Figure 5). Concretely, for the forward model, when \(y > R_y\), to obtain \(E(Y_A|X = x_A) = E(Y_B|X = x_B)\) we require \(d' = x_B - x_A < 0\); however, for the reverse model, for any value of \(y\) in the interval \(y > S_y\), we have \(E(X_B|Y = y) - E(X_A|Y = y) > 0\), producing seemingly contradictory results. Furthermore, suppose we were to move from (iv) to (iii) for the reverse model; then the group effect would agree only outside the interval \(R_y < y < S_y\), again producing the paradox.

It can be shown that, depending on the relative magnitudes and signs of \(\beta_A\) and \(\beta_B\) and which scenarios of the forward and reverse models hold, we have the paradox occurring either everywhere in the interval between \(R_y\) and \(S_y\), or everywhere outside of this interval. Either situation would be possible if the slopes \(\beta_A\) and \(\beta_B\) were of the same sign; if they were of opposite signs, then the paradox would occur only between \(R_y\) and \(S_y\). Some distributions would give rise to a very narrow interval between \(R_y\) and \(S_y\), or an interval located far away from the centers of the data distributions. In any case, the relevancy of the paradox in the group-specific case depends greatly on the amount of distributional mass residing in any such conflict zones. It is interesting to note that the scenario \(\beta_A = \beta_B\) can be viewed as a special case in which the intersection points \(R_y\) and \(S_y\) approach \(+\infty\). In this case the paradox may or may not be present (the conflict zone is either \((−\infty, +\infty)\) or the empty set), and it occurs under the conditions given in Section 3.

We next consider under what conditions the paradox can appear more generally.

**4. REGRESSION PARADOX IN NONLINEAR, NON–GAUSSIAN MODELS**

Although the sufficient conditions for the paradox are quite simple, the assumptions that lead to Proposition 3.1 are limiting. In particular, joint Gaussianity of \(X\) and \(Y\) ensures existence of the direct and reverse regression models (4) and (5), but such linear decompositions cannot be expected to hold for general bivariate distributions. It may be possible to extend our result to distributions with spherical kernels, but we do not pursue this avenue here. Rather, to show that the paradox may indeed appear in more general settings, we use the bootstrap to simulate a large sample from the joint non-Gaussian empirical distribution of \(\{Y = \text{Satisfaction Score}, X = \text{Resolve Duration}\}\) conditional on \(R = \text{Region}\). Thus, given a large (non-Gaussian) sample, we verify that nonparametric estimates of \(E(Y|X, R)\) and \(E(X|Y, R)\) have the necessary properties for the paradox. (It might seem that we already addressed this scenario in Sec. 2.3, but the smoothing results exemplified therein are subject to estimation error and thus cannot confirm the existence of the paradox in a non-Gaussian population.)

We generate a sample from the joint, non-Gaussian (empirical) distribution \(\hat{p}_{Y,X|R} \equiv \hat{p}(Y, X|R)\) as follows. First, we draw a sample of size \(N = 10^6\) from \(\hat{p}(Y|R)\). Roughly 60% of the data pertain to region A, and we sample according to this proportion. Conditional on these draws, we sample from \(\hat{p}(X|Y, R)\). This step is based on Figure 1, and the resulting joint draws will inherit the non-Gaussian characteristics of \(\hat{p}_{Y,X|R}\).

To estimate the conditional expected values \(E_{\hat{p}}(Y|X, R)\) and \(E_{\hat{p}}(X|Y, R)\), we fit nonparametric local regression models using LOESS (Cleveland 1979). Not surprisingly, the results of Figure 6 demonstrate model fits similar to those obtained by fitting monotone regression models. (The figure shows only a subsample of points.) As can be seen, the estimated conditional expectations produce regional effects consistent with the paradox. This verifies the existence of a non-Gaussian population, \(\hat{p}_{Y,X|R}\), for which the paradox is present.

Next, we describe connections with Simpson’s and Lord’s paradoxes.

**5. RELATIONSHIP TO THE PARADOXES OF SIMPSON AND LORD**

Simpson’s paradox also can be related to our problem. Figure 7 shows a situation with negative slopes where the regression paradox is present (cf. Figure 4(a)), and where the fourth
Figure 6. Regression paradox in a non-Gaussian setting. (a) Given resolve duration, the estimated expected satisfaction score is greater for region B (represented by ‘+’) than for region A (‘◦’). (b) Given satisfaction score, the estimated expected resolve duration is greater for region A than for region B.

quadrant is divided into three disjoint areas. (Here the point P serves as origin.) Now, using arguments similar to those presented in Section 3.2, we can easily show that if Q is moved into region I, with cov(X, Y) kept constant in both groups, then we have $E(Y_A|X) > E(Y_B|X)$ but $E(Y_A) < E(Y_B)$, producing relationships between the conditional and marginal means consistent with Simpson’s paradox. Similarly, placing Q in region III and reversing the prediction, we obtain $E(X_A|Y) < E(X_B|Y)$ but $E(X_A) > E(X_B)$, again yielding Simpson’s paradox. We further note that when Q is located in region I (or region III), the association between X (or Y) and the group variable is too high for the regression paradox to occur.

Interestingly, then, with $\text{cov}(X, Y) < 0$, we see that when Q is located in the fourth (or second) quadrant, either the regression paradox or Simpson’s paradox will be present. (When Q is located in first or third quadrant, neither paradox will be present.) Thus in the case where X, Y are jointly random, there is always the potential for confusion regarding the group effect, from either the regression paradox or Simpson’s paradox. This confusion follows from the fact that when Q lies in the fourth (or second) quadrant and $\text{cov}(X, Y) < 0$, the effects due to a difference in marginal means and the regression effect push the conditional group difference in opposite directions, and the relative weight of these two contributions determines which type of paradox will be present, with geometry as given in Figure 7.

We note that this problem is not likely to appear in a controlled experiment, where randomization is used to ensure that preexisting groups are closely matched on variables of potential importance to the outcome.

Lord (1967) described a similar paradox within a linear model framework such that $E(Y_A) = E(Y_B)$ and $E(X_A) = E(X_B)$, but $E(Y_B | X_B = x) - E(Y_A | X_A = x) > 0$, uniformly in x (i.e., $E(Y_B - X_B | X_B = x) - E(Y_A - X_A | X_A = x) > 0$). According to Lord, the marginal group means seem to indicate no group effect (i.e., $E(Y_B - X_B) = E(Y_A - X_A)$), yet the comparison of conditional expectations appears to contradict the lack of a group differential effect. It is interesting to note that Lord’s paradox can be illustrated geometrically in the same way that we have depicted the regression phenomenon. To show this, we can add reverse regression lines to figure 1 of Lord (1967) and verify that the slope of the line connecting the group means falls between those of the direct and reverse regression lines. Lord (1967) did not provide a definite explanation of the phenomenon but did note that group comparisons are troublesome when there are preexisting differences between groups. In a brief discussion of the problem, Geng (2006) similarly noted that group comparisons are mostly meaningful when

Figure 7. Relating the regression paradox and Simpson’s paradox. The fourth quadrant is split into three disjoint regions. The regression paradox is present in region II, and Simpson’s paradox occurs in regions I and III.
the $X$-distribution within each group are the same. With additional, casual model assumptions, Holland and Rubin (1983) provided an insightful description of Lord’s paradox in the context of causal inference.

It would be nice to be able to provide with a simple explanation for the regression paradox. But, as discussed earlier, because the paradox depends on a combination of the mean shift between groups and the signal-to-noise level, no simple explanation appears to be available. Nonetheless, we observe that as any prediction in a noisy system regresses toward the mean, when the correlation between $X$ and $Y$ goes to zero, the differences in prediction between groups approach their marginal mean shift, and if the ratio of mean shifts in the $X,Y$-plane (i.e. $s$) has the same sign as $\text{cov}(X,Y)$, then the paradox will occur. Viewed in this way, the paradox may be considered a regression toward the difference in group marginal means, with its occurrence determined by the strength and direction of linear association between $X$ and $Y$. Lord’s paradox also may be understood in this manner. Moreover, as the within-group noise level increases, the group effect modeled by a linear regression will approach the mean shift between groups, producing a situation without paradox. Put more straightforwardly, as $X$ and $Y$ approach independence, it should not be surprising that different conclusions may be reached with respect to the (statistical) effect of the group variable.

6. CONCLUSION

Our analysis suggests that some caution should be maintained when applying a well-established statistical methodology to open-ended contexts, such as mining observational data without a predefined hypothesis. Specifically, in a context where the distinction between cause and response is not obvious, one may wish to take care when considering the interpretation of estimated group effects. Our work also underscores the fact that empirical laws derived from fitting simple models to experimental data (e.g. regression models) are not necessarily usable as algebraic identities.

[Received October 2008. Revised April 2009.]

REFERENCES


