Yet Another Solution to Galton’s Problem

Trevor Denton

For half a century, Galton’s problem has been a thorn in the side of cultural anthropologists, agitating them to print. Galton’s problem motivated the creators of all the major cross-cultural sampling frames to advocate selection of a single ethnographic description from each of up to hundreds of internally homogeneous culture areas. Why this happened is perplexing. Galton’s problem disappears when it is thought of as a theoretical issue of correct specification of a theoretical model. There are implications for the construction of new cross-cultural samples.

**Keywords:** Galton’s problem; theoretical model; cross-cultural sample

In the 1800s, E. B. Tyler proposed cross-cultural research as a method for discovering adhesions between behaviorally related variables y, x. The British scientist Sir Francis Galton countered that such a method could not distinguish adhesions of behaviorally related constructs (e.g., the independent family and neolocal postmarital residence) from adhesions of behaviorally unrelated constructs which diffuse together (e.g., hot dogs and catsup). Galton suggested that diffusion such as the latter would create many dependent, paired, replicate observations y, x of a single, original, independent observation.

The solution advocated here is to model all the behavioral relations connecting a construct y to its predictors in a correctly specified theoretical model, including the behavioral relations Galton had in mind. In such a correctly specified theoretical model, the problem of dependent observations disappears. Consideration is given to causal models, curve fitting, bivariate association and dimension reduction. Implications for the creation of cross-cultural samples are considered. Although earlier works on Galton’s problem

**Author’s Note:** The research reported in this article was begun when the writer was a faculty member at Brandon University. Data analyses were designed and executed by the writer using SAS V8.2 supplied by Brandon University to facilitate completion of the research. A Windows XP-Pro platform was used. The writer gratefully acknowledges referees’ suggestions for improvement.
(e.g., Levinson & Malone, 1980; Naroll, 1970) included discussion of diffusion as a behavioral phenomenon rather than simply a methodological hurdle, the solution suggested here breaks new ground.

**Galton’s Problem As Correct Specification of A Theoretical Model**

Suppose we find the Spearman correlation $r_{yx} \neq 0$ between a variable $y$ (e.g., the independent family) and $x$ (e.g., neolocal postmarital residence) to be significant. How do we account for the relation? We might conjecture at least three explanations, each of which is a variety of behavioral relation. Behavioral relation A might be posited to be a natural association that motivates independent invention of the independent family ($y$) in a society given that neolocal residence ($x$) is present in the society. Behavioral relation B might be posited to be present when behavioral relation A holds but the acquisition of $y|x$ (e.g., the independent family given that neolocal residence is present) is speeded up by the existence of an observable $y$ (i.e., the independent family) in neighboring societies. Behavioral relation C might be posited to be present when behavioral relation A does not hold but $y$ and $x$ diffuse together from neighbors for a reason such as conformity. Figure 1 diagrams these three behavioral relations. There may be more behavioral relations at work than these. Figure 1 consists of objective constructs without trying to get into the heads of the individuals involved. If the same subject matter were conceptualized as individual decision making, a different diagram might be appropriate.

How might we model the sorts of behavioral relations outlined in Figure 1? There appear to be at least three choices. A fully specified causal model (Asher, 1983) would be the most thorough description of what is going on. To date, the absence of adequate behavioral theory, accompanied by the absence of appropriate data, has produced few fully specified causal models in cultural anthropology.

Partial correlations might be considered as a second method for modeling the behavioral relations of Figure 1. The appendix does just that. In the appendix, binary random variables $Y = 1$ (independent family), $Y = 0$ (extended family) and $X = 1$ (neolocal residence), and $X = 0$ (postmarital residence with or near kin) are created from the Standard Cross-Cultural Sample (SCCS) coded data for V68 and V69 (Divale, 2004). The mean $\bar{x}$ and mean $\bar{y}$ in the region of each SCCS society is created from SCCS coded data for V68, V69, and V200 (a code for regions from Divale, 2004). In the appendix, a significant Spearman rho $r_{yx}$ detects the presence of an association...
The regional mean $\bar{y}$ is denoted $u$ and regional mean $\bar{x}$ is denoted $v$. Partial correlations $r_{yx,uv} > 0$ and $r_{yu,x} = 0$ give evidence that the association between $y$ and $x$ arises from Figure 1 behavioral relation A, or possibly behavioral relations A and B, but not behavioral relation C. Currently, there are no SCCS coded data (Divale, 2004) that might be used to distinguish between behavioral relations A and B in Figure 1.

The problem with a partial correlation method such as that used in the appendix is that it will not work in cases where many poorly understood behavioral relations are at work. The reason it appears to work in the appendix is that there may be few additional behavioral relations present in the subject matter.

Curve fitting is a third method that might be considered to model the behavioral relations of Figure 1. Linear models such as regression or logistic

---

**Figure 1**

Three Varieties of Behavioral Relations That Might Connect the Independent Family ($y$) to Neolocal Residence ($x$)

Note: Behavioral relation A posits the spontaneous, independent invention of the independent family $y$ in the presence of neolocal residence $x$. Behavioral relation B posits that the appearance of the independent family $y$ in societies where neolocal residence $x$ is present may be speeded up by the observable existence of the independent family $y$ in neighboring societies. Behavioral relation C posits the adoption of both the independent family and neolocal residence, $y$ and $x$ together, as a process of conformity to neighbors. The purpose of this figure is to illustrate patterned behavioral relations. Beyond A, B, and C, many other varieties of behavioral relations may be at work.

between $y$ and $x$. The regional mean $\bar{y}$ is denoted $u$ and regional mean $\bar{x}$ is denoted $v$. Partial correlations $r_{yx,uv} > 0$ and $r_{yu,x} = 0$ give evidence that the association between $y$ and $x$ arises from Figure 1 behavioral relation A, or possibly behavioral relations A and B, but not behavioral relation C. Currently, there are no SCCS coded data (Divale, 2004) that might be used to distinguish between behavioral relations A and B in Figure 1.

The problem with a partial correlation method such as that used in the appendix is that it will not work in cases where many poorly understood behavioral relations are at work. The reason it appears to work in the appendix is that there may be few additional behavioral relations present in the subject matter.

Curve fitting is a third method that might be considered to model the behavioral relations of Figure 1. Linear models such as regression or logistic

---

Note: Behavioral relation A posits the spontaneous, independent invention of the independent family $y$ in the presence of neolocal residence $x$. Behavioral relation B posits that the appearance of the independent family $y$ in societies where neolocal residence $x$ is present may be speeded up by the observable existence of the independent family $y$ in neighboring societies. Behavioral relation C posits the adoption of both the independent family and neolocal residence, $y$ and $x$ together, as a process of conformity to neighbors. The purpose of this figure is to illustrate patterned behavioral relations. Beyond A, B, and C, many other varieties of behavioral relations may be at work.

between $y$ and $x$. The regional mean $\bar{y}$ is denoted $u$ and regional mean $\bar{x}$ is denoted $v$. Partial correlations $r_{yx,uv} > 0$ and $r_{yu,x} = 0$ give evidence that the association between $y$ and $x$ arises from Figure 1 behavioral relation A, or possibly behavioral relations A and B, but not behavioral relation C. Currently, there are no SCCS coded data (Divale, 2004) that might be used to distinguish between behavioral relations A and B in Figure 1.

The problem with a partial correlation method such as that used in the appendix is that it will not work in cases where many poorly understood behavioral relations are at work. The reason it appears to work in the appendix is that there may be few additional behavioral relations present in the subject matter.

Curve fitting is a third method that might be considered to model the behavioral relations of Figure 1. Linear models such as regression or logistic
regression offer a method for constructing a theoretical model of a random variable Y (Kutner, Nachtscheim, & Neter, 2004, pp. 623-624). Although linear regression is the linear model most widely considered in statistics texts, logistic regression is the linear model most widely applicable to current cross-cultural coded data, most of which are categorical (Divale, 2004). Logistic regression models the probability \( P(Y = 1) \) that a binary random variable \( Y = 1, 0 \) is equal to 1. \( P(Y = 1) \) is modeled from predictors \( X_1, X_2, \ldots, X_{p-1} \). The predictors must be behaviorally related to \( Y \), but no causal relations such as those of Figure 1 need be specified. Simple association without causation may do. Provided the predictors adequately describe the behavior of \( Y \) and the data meet the premises of the linear model, the linear model is deemed a true statistical model of \( Y \) in regard to the data set to which it is fitted.

In regard to Galton’s problem, everything to be said about logistic regression will apply also to linear regression. Throughout this article, we will assume the existence of a well-defined population (Denton, 2007) and a sample from it.

For logistic regression (Kutner et al., 2004) the population (distribution) to be modeled consists of an infinite number of \( p \)-tupless \( Y, X_1, X_2, \ldots, X_{p-1} \). Such a population is conceptual. The criterion random variable \( Y_i \) is distributed Bernoulli, \( Y_i = 1, 0, i = 1, 2, \ldots \). The predictors \( X_1, X_2, \ldots, X_{p-1} \) may be categorical or continuous. We will assume they are continuous. Because each observation \( i = 1, 2, \ldots \) is a Bernoulli random variable \( Y_i = 1, 0 \), the population probability parameter of case \( i \) is \( P(Y_i = 1) = \pi_i \), which is the expectation \( E[Y_i] \) of observation \( i \). The population may be described as follows.

1. \( x_1, x_2, \ldots, x_{p-1} \) is a set of predictors such that \( x^T_i b = b_0 + b_1x_1 + b_2x_2 + b_{p-1}x_{p-1} \), \( b \) is the set of constant coefficients by which the conceptual population criterion random variable \( Y|x \) is distributed as in the logistic regression model.
2. \( P(Y_i = 1) \), the population expectation \( E[Y_i] \) of \( Y_i \), also denoted \( \pi_i \), may be stated in a variety of ways, all of which are equivalent:

\[
P(Y_i = 1) = \frac{\exp(x_i^T b)}{1 + \exp(x_i^T b)} \quad i = 1, 2, \ldots, n
\]

\[
= \frac{1 + \exp(-x_i^T b)}{1 + \exp(-x_i^T b)}^{-1}
\]

\[
= \ln \frac{\pi_i}{1 - \pi_i}
\]
A sample of \( n \) realizations from the conceptual population is \( n \) \( p \)-tuples \( y_i, x_{i1}, x_{i2}, \ldots, x_{ip} \), \( i = 1, 2, \ldots, n \) for which \( \hat{\pi}_i \) is the sample estimate of population \( \pi_i \). We will also use \( P(Y_i = 1) \) to denote \( \hat{\pi}_i \) provided the distinction between population parameter and sample estimate is clear.

Whether the predictors \( X_1, X_2, \ldots, X_{p-1} \) are random variables such as postmarital residence \( X = 1, 0 \) in the appendix or ordinary variables such as time and distance that are not to be considered random, \( Y \) is a random variable \( Y_i|x_i, i = 1, 2, \ldots, n \), realizations \( y_i \) of which are conditioned on the particular states of \( x_i \). Here, upper case and lower case distinguish between a variable before and after measurement of a realization.

The fitting of a logistic regression model may proceed in two stages—model building followed by model checking (Kutner et al., 2004). In model building, one starts with a set of candidate predictors \( X \) that theory conjectures are behaviorally related to the criterion variable \( Y \). Maximum likelihood estimation is applied to a log likelihood function (Kutner et al., 2004) to test whether a particular predictor or predictors may be discarded. The model-building stage is followed by a model-checking stage in which the model built is checked for goodness of fit.

Model checking proceeds as follows. In logistic regression, the residual (error) \( e_i = y_i - E[Y_i] \) is \( 1 - \pi_i \) if \( y_i = 1, -\pi_i \) if \( y_i = 0 \). Such a residual is not normally distributed. Using \( \hat{\pi}_i \) to denote the fitted estimate of the true population \( \pi_i \), the \( i \)th Pearson residual (Kutner et al., 2004) is calculated

\[
r_p = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}
\]

which, because it is distributed chi-square, provides a test of the goodness of fit of the fitted model. Randomly distributed \( r_p \) is evidence of good fit. The Pearson residual is not distributed chi-square unless there are sufficient numbers of realizations \( y \) at the same set of predictor states \( x \). The deviance residual is defined in a somewhat different way (Kutner et al., 2004). A likelihood ratio test based on the distribution of the deviance residual permits a second test of the goodness of fit of the fitted model. The Hosmer-Lemeshow statistic (Kutner et al., 2004) may also be used to test the goodness of fit of the fitted model.

The residual \( e_i = y_i - E[Y_i] \) and the Pearson residual \( r_p \) are, in fact, conditional random variables \( e_i|x_i, r_p|x_i \) conditioned on the predictor set measures \( x_i \). These residuals are usually scripted lower case but are random variables. Subject to the vagaries of random sampling, what determines whether the tests of \( e_i|x_i \) and \( r_p|x_i \) show goodness of fit is proper theoretical specification of the fitted model.
Resolution of Galton’s Problem by Curve Fitting: the Independent Family

In Table 1, a logistic regression model of the independent family is fitted to the SCCS coded data of the appendix. In Table 1, \( Y = 1, 0 \) is a binary random variable \( Y = 1 \) (the independent family is modal), 0 (the extended family is modal).

Behavioral theory suggests that Murdock and Provost’s (1973) modernization measure \( m \) may be a predictor of the independent family. Candidate predictor variables in powers from 1 to 3 of Murdock and Provost’s modernization scale \( m \) appear in Table 1 on the basis of known impact on a wide variety of subject matters (Denton, 2004). Murdock and Provost selected the 10 subscales that form modernization \( m \) because they best distinguish between cultures (and societies) along a continuum from the most ancient to the most recent form. Each of the 10 subscales is a measure of the amount of a separate subject matter present in a culture—the amount of writing and records, of fixity of residence, of food production, of community size, of technology, of land transport, of the presence of money, of density of population, of level of political integration, and of social stratification. The states 0, 1, 2, 3, 4 of each of the 10 subscales are known to have first appeared in the sequence 0, 1, 2, 3, 4. In the case of food production, for example, foraging preceded varieties of horticulture, which preceded varieties of intensive agriculture. The magnitude of modernization \( m \) present in a culture establishes the relative recency of the form of the culture, not its chronological recency. All 10 subscales from which modernization \( m \) is formed are pairwise monotonic increasing. That is to say, when any one subscale goes up, the conditional expectation of each of the rest goes up, conditioned on the first. Any single value of \( m \) implies a unique expected value of each of the 10 subscales from which it is formed. Given the nature of the 10 subject matters from which the modernization measure \( m \) is calculated, is it any wonder that many behaviors are correlated with \( m \)?

Behavioral impacts of modernization \( m \) are so well established (Denton, 2004; Levinson & Malone, 1980) that there is no need to assume that \( y \) and \( m \) might diffuse as a package with no behavioral relations connecting them beyond simple conformity to neighbors. However, our focus is on whether behavioral relation \( C \) in Figure 1 explains any association of \( y \) to \( x \).

As in the appendix, conformity to neighbors is measured in Table 1 using V200 of Divale (2004). SCCS V200 codes each SCCS society as one of six world regions—Sub-Sahara Africa, Circum-Mediterranean, East Eurasia, Insular Pacific, North America, or South America, each carefully defined.
The writer coded a diffusion measure for SCCS societies by calculating, for the binary random variable $Y = 1$ (independent family), 0 (extended family), the percentage $Y = 1$ of each region. Then, each society was assigned the percentage $Y = 1$ of the region in which the society is located. This is $\bar{y}$ in Table 1. Other measures of diffusion or conformity to neighbors might be considered. Also in Table 1 $v$ is percentage X = 1 ($\bar{x}$) of each region.

In case modernization m changes over time t of ethnographic observation, an interaction term mt in powers 1 to 3 is included in Table 1 as a candidate predictor. Time t, the year of ethnographic observation, is SCCS V838 of Divale (2004).

### Table 1

**Fitted Logistic Regression Model for the Independent Family**

<table>
<thead>
<tr>
<th>Criterion variable</th>
<th>The probability $P(Y = 1)$ that the independent family is the predominant form. $Y = 1$ if SCCS V68 = 1 or 2 or 3 or 4, otherwise $Y = 0$ (SCCS coded data from Divale, 2004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate predictor set</td>
<td>$&lt;m_y, m_y^2, m_y^3, m_x, m_x^2, m_x^3, \bar{u}, v, x&gt;$ where $u$ is mean regional $\bar{y}$, $v$ is mean regional $\bar{x}$ and $x$ is the realization of $X = 1$ (neolocal residence is the predominant form), 0 (kin-based residence is the predominant form) in each society</td>
</tr>
<tr>
<td>Fitted model</td>
<td>$P(X = 1) = \frac{1}{1 + \exp(-\mathbf{z}^T \mathbf{b})}$, $\mathbf{z} = \langle 1, m_y^2, m_x^3 \rangle$</td>
</tr>
<tr>
<td>Coefficients</td>
<td>$\mathbf{b}^T = \langle -0.0519, -215E-12 \rangle$; PR &gt; CHISQU for coefficients 0.8024, 0.0481</td>
</tr>
<tr>
<td>Goodness of fit Pr &gt; ChiSq:</td>
<td>Pearson residual 0.6081, Deviance residual 0.3763, Hosmer-Lemshow statistic 0.3249</td>
</tr>
<tr>
<td>Additional model statistics:</td>
<td>Receiver operating characteristic .570, $R^2$ .0223</td>
</tr>
</tbody>
</table>

Note: SCCS = Standard Cross-Cultural Sample. Candidate predictors $u$, $v$ and $x$ are defined in the appendix. Because $x$ is binary and $u$, $v$ have only six states, they are not regrouped. To test goodness of fit, modernization m and time t were regrouped as follows: Modernization m: $[0, 8] \rightarrow 4$, $[9, 16] \rightarrow 12.5$, $[17, 24] \rightarrow 20.5$, $[25, 32] \rightarrow 28.5$, $[33, 40] \rightarrow 36.5$. Time t: $[-1750, 1800] \rightarrow 1000$, $[1801, 1825] \rightarrow 1812.5$, $[1826, 1850] \rightarrow 1837$, $[1851, 1875] \rightarrow 1862$, $[1876, 1900] \rightarrow 1887$, $[1901, 1925] \rightarrow 1912$, $[1926, 1965] \rightarrow 1946$. The arrow (→) shows the group value given to a parenthetical interval of m or t. See text for model building and model checking.
The final candidate predictor set is \(<m^g, m^2g, m^3g, m^tg, m^{2tg}, m^{3tg}, u, v, x>\). The subscript \(g\) denotes that the modernization \(m\) and time \(t\) are grouped. Groupings are shown in Table 1. If predictor variables in a logistic regression model are continuous, model checking necessitates that they be grouped (Hosmer & Lemeshow, 2000). Candidate predictors \(m^g\) and \(m^{tg}\) in powers up to three allow for polynomial curves and for behavioral relations that change over time. Because there are only six states of each diffusion variable \(u\) and \(v\), and because \(x\) is binary, they are not grouped.

In the candidate predictor set \(<m^g, m^2g, m^3g, m^tg, m^{2tg}, m^{3tg}, u, v, x>\) the variables \(m^g, m^2g, m^3g, m^tg, m^{2tg}, m^{3tg}\) may be thought of as background variables. They are unlikely to be proximate causes (directly affecting individuals) of \(Y = 1, 0\) in a fully specified causal model. On the other hand, proximate causes of \(Y = 1, 0\) are likely themselves to be functions of modernization \(m\) and any interactions with time \(t\). The variable \(u\) denotes mean regional \(\bar{y}\), \(v\) denotes mean regional \(\bar{x}\), and \(x\) is the state of the neolocal residence random variable \(X = 1, 0\) in each society. Candidate predictors \(\bar{y}, \bar{x}\), and \(x\) might or might not be proximate causes of \(Y = 1, 0\) in a fully specified causal model. Here, \(x\) in each society is actually a realization \(x|m\) of random variable \(X|m\) conditioned on the magnitude of modernization present in the society. If \(x\) is thought of as a function of \(m\), \(P(X = 1) | m\) may be thought of as a function of \(m\) even if \(x\) is a predictor. In the case of random variable \(Y = 1, 0\), we know candidate background variables in \(m\) and time \(t\) and candidate, possibly proximate, causes \(x, \bar{y}, \bar{x}\), so we might as well test them all.

Model building in Table 1 started with best subsets variable selection (Kutner et al., 2004). A model with intercept was sought. The predictor set with the fewest predictors was sought subject to significant (\(\alpha = .05\)) \(p\) values for each predictor and no model with more predictors being significantly different from the more parsimonious model (likelihood ratio test). An intercept model was sought even if the \(p\) value of the intercept exceeded \(\alpha = .05\).

Initially, the predictor set \(<m^3g, x>\) or \(<m^{2tg}, x>\) appeared promising. Because \(x\) (the realization of the neolocal residence random variable \(X = 1, 0\)) is known to be behaviorally related to \(y\) (the realization of the independent family random variable \(Y = 1, 0\)), such predictor sets appear to be behaviorally meaningful. However, quasi-complete separation of data cases prevented maximum likelihood estimation of these predictor set models. Consequently, the latter models had to be abandoned as inconclusive. Because \(x\) is thought of as a realization of random variable \(X|m\), we will seek a solution in \(m\) and any interaction of \(m\) with \(t\).
Excluding the predictor sets $<m_g t_g^3, x>$ or $<m_g t_g^2, x>$ eliminated by quasi-complete separation of data cases, no predictor variable subset with more predictors, or even a different predictor, significantly improved on the predictor set $<m_g t_g^2>$ finally selected in Table 1. Because modernization $m$ is a behaviorally meaningful predictor, the fitted model is accepted. It should be noted that no model in $u, v$ ever warranted serious consideration. Even if $x$ is a predictor in the true statistical model of the infinite conceptual population $Y_1, Y_2, ...$ and the predictors $z$ of $Y = 0, 1$, the model shown in Table 1 gives a sufficient fit that they allow us to see what is going on from the perspective of diffusion. Because $u, v$ are excluded as predictors of $Y = 1, 0$, the issue of the diffusion behavioral relation $C$ in Figure 1 is resolved.

The sole predictors in the fitted model of Table 1 are the intercept and $m_g t_g^2$—the squared, grouped interaction of time $t$ multiplied by modernization $m$. Jackknife options in SAS PROC LOGISTIC were used to test for the effect of deletion of each case—one at a time—on overall CHISQU of the model with the case deleted. No model showing instability across samples would have been accepted. The absence of instability may be interpreted as evidence supporting validation of the fitted model.

Once a model was tentatively selected, model checking proceeded. Customary goodness of fit tests included the Hosmer-Lemeshow statistic, deviance residual, and Pearson residual tests all of which show good fit. The low $R^2$ in Table 1 is not a concern. The definition of $R^2$ in logistic regression differs from that in regression. Hosmer and Lemeshow (2000) caution that $R^2$ in logistic regression is generally much lower than in linear regression and might be omitted as a criterion for model selection. That the ROC (receiver operating characteristic) value in Table 1 is low is partly attributable to the fact that the fitted logistic regression curve is centered close to $P(Y = 1) = .5$. The curve is graphed in Figure 2. At $P(Y = 1) = .5$ residuals are approximately equally distributed above and below the curve which, for that reason, is a poor predictor of whether a data case is $y = 1$ or 0 at $P(Y = 1) = .5$.

The short of Table 1 is that we get an adequate model of $P(Y = 1)$ using only a single predictor—the impact of modernization $m$, which changes over time $t$. Model selection retains the interaction term $m_g t_g^2$ as a predictor but discards diffusion $u$ and $v$ as predictors. Diffusion $u$ tests behavioral relation $C$ in Figure 1. Galton’s problem is resolved during the model-building and model-fitting process. Simple diffusion $u$ is considered, but ultimately discarded, as a behavioral relation affecting $Y$.

One of Galton’s concerns was that where simple diffusion occurs in the absence of true behavioral relations, observations of $y$ are not independent.
If Figure 1 behavioral relation C is measured by predictors such as $\tilde{y}$, $\tilde{x}$, each pair of observations $y$, $x$ are independent, conditional realizations of random variables $(Y, X)|(\tilde{y}, \tilde{x})$. Any issue of dependent observations is resolved by the candidate regional predictors $\tilde{y}$, and $\tilde{x}$, which turn out to be eliminated during model building. The issue of randomness is resolved.
during curve fitting. Pearson and deviance residual tests in Table 1 establish that the residuals $e_i = y_i - E[Y_i | z_i]$, and therefore the conditional observations $Y_i | z_i, i = 1, 2, \ldots, n$, are randomly distributed. In an incorrectly specified theoretical model, randomness would be absent.

Linear regression models may be handled as above. When fitting a linear regression model to data, we include some measure $u$ of diffusion if behavioral theory gives us reason to so conjecture. We let fitting of the model decide whether diffusion $u$ is, or is not, a predictor in a true statistical model.

The success of the curve-fitting method of Table 1 depends on our ability to fit to data a curve that models $Y = 1, 0$. For subject matters such as hot dogs and ketchup, there would be no fit to a model in which modernization $m$ is the sole predictor. Predictors including modernization $m$, time $t$, and conformity might be needed. If so, this would still be a matter of correct specification of a theoretical model.

Even though curve fitting appears to resolve the issue of simple diffusion in Table 1, a fully specified causal model would be preferable. The model fitted in Table 1 appears to eliminate simple diffusion $u$ but does not otherwise illuminate the causal predictors of $P(Y = 1)$.

Galton’s problem will not necessarily arise in data analyses that use dimension reduction methods. Consider, for example, the case where one wishes to construct a latent variable measure of volume of exchange observed by indicators such as SCCS V149, V153, V154, V155 (Divale, 2004). What is at issue here is volume of exchange, not the origin of bivariate behavioral relations among indicators of the latent variable. If one had reason to do so, one might consider recalculating the bivariate correlations $r_{yx}$ by controlling ($r_{yx,d}$) for a variable $d$ of common origin or diffusion.

**Discussion**

Galton’s problem is not a methodological problem. It is a theoretical problem of correct specification of a theoretical model. If we include diffusion $d$ (Figure 1 behavioral relation C) as a possible behavioral relation in a theoretical model, we may leave it data to decide whether diffusion is, or is not, at work. When modeling a criterion random variable $Y$, the inclusion of a candidate predictor $u$ of Figure 1 behavioral relation C makes conditional observations $y/u$ independent. Model building decides whether candidate predictor $u$ may be discarded. Tests for randomness establish whether the fitted model residuals are random. Only if the theoretical model is inadequate, or suitable data measures are unavailable, will Galton’s problem become a methodological problem.
If we build diffusion or common origin into our models, we ought to be able (cost of coding permitting) to fit equations to large samples such as the entire Ethnographic Atlas (EA; Murdock, 1967) with many data cases from each internally homogeneous culture area. If a generalized linear model (Kutner et al., 2004, pp. 623-624)

\[ E[Y] = g(x^Tb), \quad x^Tb = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_{p-1} x_{p-1} \]

such as logistic regression or linear normal error regression is correctly specified, then observations become independent and the residual \( e_i = y_i - g(x_i^Tb) \) becomes randomly and independently distributed, as specified in the model.

A compelling case may be made that a large sample such as the EA (Murdock, 1967), of size exceeding 1,100, is much more representative of the real world than a smaller one such as the SCCS of size 186, pruned to eliminate (or reduce) data cases resulting from possible common origin or diffusion. Properly specified linear theoretical models may be fitted to the entire EA. They may also be fitted to a sample drawn from the EA. The SCCS is a set of 186 data cases such that each is drawn from a single culture area so as to mitigate Galton’s problem. Nevertheless, each of these 186 data cases will have been subject to the same processes of diffusion and pressures to conform as the data cases of the EA. In the latter regard, the SCCS remains a viable cross-cultural sample.

This article gives evidence, if any evidence were needed, that much remains to be accomplished in cross-cultural research. The key to Galton’s problem is correct specification of a theoretical model. With better behavioral theory will arrive the need to code new theoretical constructs far beyond the current SCCS inventory (Divale, 2004), perhaps for cross-cultural samples yet to be created.

**Appendix**

Independent Invention or Conformity to Neighbors? Spearman Rho Correlations and Partial Correlations for Family Form (Y) and Postmarital Residence (X)

**Variables**

1. \( Y = 1 \) if V68 = (1 or 2 or 3 or 4), \( Y = 0 \) if V68 = (5 or 6 or 7 or 8 or 9 or 10 or 11 or 12). V68 codes are from Divale (2004).
   \( Y = 1 \) denotes the independent family is modal. \( Y = 0 \) denotes the extended family is modal
Standard Cross-Cultural Sample (SCCS) proportions: p(Y = 1) = 77/185, p(Y = 0) = 108/185

2. $U_i$ denotes the mean $\bar{y}$ of the region to which data case $i$ belongs, $i = 1, 2, \ldots, 186$, using the six regions of SCCS V200. V200 codes are from Divale (2004).

3. $X = 1$ if $V69 = 5$, $X = 0$ if $V69 = (1$ or $2$ or $3$ or $4)$, $X = .$ denotes missing data. $X = 1$ denotes neolocal postmarital residence is modal. $X = 0$ denotes kin-based postmarital residence is modal. V69 codes are from Divale (2004).

SCCS proportions $p(X = 1) = 9/185$, $p(X = 0) = 176/185$

4. $V_i$ denotes the mean $\bar{x}$ of the region to which data case $i$ belongs, $i = 1, 2, \ldots, 186$, using the six regions of SCCS V200 from Divale (2004).

Simple correlation between $y$ and $x$

$$r_{yx} = 0.27, \ p = .0002$$

$H_1 \ r_{yx} \neq 0$ accepted

Partial correlation: Test for simple conformity to neighbors

$$r_{yu.x} = 0.11, \ p = .1398$$

$H_0 \ r_{yu.x} = 0$ accepted. Rejects behavioral relation C of Figure 1.

Partial correlation: Test for behavioral relation of $y$ to $x$ independent of conformity to neighbors

$$r_{yx.uv} = 0.28, \ p = .0001$$

$H_1 \ r_{yx.uv} \neq 0$ accepted. Supports behavioral relations A or A and B of Figure 1.

All $r_s$ are Spearman correlations ($n = 185$). In each case the hypotheses tested are $H_0 \ r_s = 0$, versus $H_1 \ r_s \neq 0$

The method of the appendix produces interpretable results because few extraneous variables appear to be at work. For most subject matters, Table 1 may provide a more interpretable method for testing for Figure 1 behavioral relations A and B versus C—simple conformity to neighbors.

References


**Trevor Denton** has a PhD in cultural anthropology from the University of Toronto. Before becoming a consultant based in Peterborough, Ontario, he pursued an academic career at several Canadian universities. He is a quantitative cultural anthropologist.