Spatial Autocorrelation Models for Galton's Problem[1]

Colin Loftin and Sally K. Ward°

The effects of Galton's problem are discussed within a framework provided by the linear regression model. We examine five illustrative diffusion models and evaluate alternative estimation procedures (especially Naroll's linked pairs test and Wirsing's second order partial correlation). While no one procedure is adequate for all models, the specification of a diffusion model provides guidance in the selection of an appropriate estimation procedure.

[Accepted for publication: November, 1980.]

° *Colin Loftin is an assistant professor in the Department of Sociology and the Center for Research on Social Organization at the University of Michigan. His research interests are in the area of social control and comparative social organization.*

Sally K. Ward is an assistant professor in the Department of Sociology and Anthropology at the University of New Hampshire. Her current research interests involve an analysis of income inequality in U.S. communities and a comparison of the concept of dominance in world system and urban political economy work.

Introduction

Many disussions of the effects of Galton's problem on estimates of relationships between variables in cross-cultural studies have presented adjustment procedures that might be used to deal with the problem. In this paper, we develop some theoretical models of diffusion processes; we analyze several sets of simulated data to demonstrate some of the effects of Galton's problem on the estimates of relationships; and we evaluate the adjustment strategies that have been presented in the literature. In particular, we evaluate Naroll's (1961, 1964, 1970) linked pairs test and Wirsing's (1975) use of second order partial correlation. Finally, we suggest alternatives for dealing with Galton's problem.

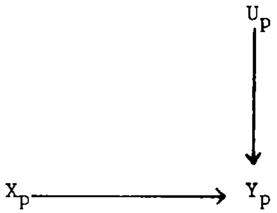
The Linear Regression Model

We begin with a discussion of the linear regression model which provides a theoretical basis for the discussion of Galton's problem. The basic elements in the model are three variables linked by a linear equation:[2]

$$Y_p = \alpha + \beta X_p + U_p \qquad (1)$$

The model represents a dependent variable (Y_p) which is assumed to be produced by an explanatory or independent variable (X_p), and one or more disturbance variables (U_p) which represent all factors that influence the dependent variable other than the independent variable.[3] For example, one might be investigating the influence of technological complexity of societies (X_p) on their organizational complexity (Y_p), but many other variables (e.g., environmental factors, energy resources, historical factors) also influence organizational complexity. This residual category of factors other than technological complexity is represented in the model by the disturbance variable (U_p). It is, in fact, an aggregate of all

Figure 1. Causal Model Described by Linear Regression Equation



variables that "disturb" the relationship being investigated. The two Greek letters in the equation (α and β) are constant terms representing the intercept and the slope of the linear relationship. Figure 1 provides a diagram of the causal model that would be described by the regression equation.

The use of the regression model for statistical inferences is facilitated if we can make the following assumptions about the distribution of the disturbance variables:

1) The disturbances have a mean of zero. Some factors increase Y, but others decrease it, so that in the long run we expect that the effect on the mean value of Y will be zero.
2) The disturbances are normally distributed and have a constant variance.
3) The disturbances are independent of each other, i.e., knowledge of the disturbances at point p does not allow one to predict the value of the disturbances at any other point.
4) The disturbances are independent of the explanatory variables.[4]

Given this model, statisticians have developed a set of procedures which allow one to estimate the unknown constants of the equation (the slope, β and the intercept α) and to make probability statements about the distribution of these estimates. Some of the statistical properties of the model are very helpful in understanding Galton's problem, and it is therefore useful to describe them in detail. We make use of a data simulation technique to illustrate some of the statistical properties of the linear regression model.

The strategy of the simulation is very simple; we generated twenty hypothetical studies that conform to the assumptions of the linear regression model. We then applied the statistical procedures to those data to see to what degree we would have arrived at the correct conclusions about the mechanism that generated the data. By averaging our results over twenty studies we can illustrate the general behavior of the statistics. Later we modify some of the assumptions of the model, repeat the simulation and the analysis, and examine the effects of the modification.

The specific characteristics of the initial simulation were as follows: (1) sixty random X_p values were drawn from a uniform distribution with a range of zero to ninety-nine; (2) sixty U_p values were drawn from a normal distribution with a mean of zero and a standard deviation of ten, (3) the slope (β) was assigned the value of one and the intercept (α) the value of zero, and (4) Y_p was calculated according to formula (1). We repeated the procedure twenty times, generating twenty hypothetical studies of sixty cases each. We then analyzed the data using linear regression techniques.

The results of the analysis are presented in Table 1. We focus attention on estimation of the slope coefficient because it is the value of primary theoretical importance in most studies; however, our general conclusions would hold equally well for estimation of the intercept.

There are three important properties of the ordinary least squares (OLS) regression estimates[5] that we want to emphasize:

1) The regression estimate of the slope is an unbiased estimate;
2) the regression estimate of the slope is the most efficient estimate of the slope;
3) the regression estimate of the standard error of the slope is an unbiased estimate. These three properties can be derived deductively from the definition of the normal linear regression model, and proofs are available in the statistical literature.[6] Our purpose is simply to explain their meaning; later we will show how Galton's problem influences them.

Column two of Table 1 illustrates the first property of the OLS estimates of the slope. The twenty slope estimates vary around the true value--some are larger and some are smaller--but the average of the twenty estimates (.996) is almost exactly equal to the true value (1.0). The distribution of the estimates of the slope around the true slope is an empirical analog of the theoretical concept of the sampling distribution of the slope. The sampling distribution of a statistical estimator is the distribution that would be obtained if an infinitely large number of estimates were made and the results tabulated. Our sample of twenty estimates can be thought of as an estimate of the sampling distribution of the particular statistic under consideration. In fact, it is a sample size of twenty from that sampling distribution and, like all samples, we do not expect it to correspond exactly to the characteristics of the population. It will, however, provide a useful tool for estimating the characteristics of the population.

The statement that the regression estimate of the slope is an unbiased estimate refers to the fact that the mean of the sampling distribution of the slope estimate will be exactly equal to the true value of the slope. Thus, our simulated data illustrate this principle to the extent that they have a mean that is very close to the true value of the slope. They do not, of course, have a mean of exactly 1.0

Table 1. Ordinary Least Squares Estimates of Linear Regression Model

(1) Sample Number	(2) Slope Estimate	(3) Estimated Standard Error of Slope	(4) Coefficient of Determination
1	.951	.052	.851
2	1.064	.047	.898
3	1.106	.048	.903
4	1.047	.058	.848
5	1.033	.047	.894
6	1.007	.044	.902
7	.989	.039	.917
8	.996	.048	.882
9	.947	.032	.937
10	.919	.045	.876
11	1.037	.046	.899
12	.981	.046	.889
13	.925	.040	.901
14	1.005	.056	.846
15	1.013	.042	.909
16	.972	.034	.935
17	.993	.044	.898
18	.936	.041	.901
19	.975	.047	.883
20	1.018	.036	.931
Mean	.996	.045	.895
Standard Deviation	.048	.007	.026

Twenty Samples of Sixty Observations Each because they represent only a sample from the population of all possible estimates of the slope.[7]

The second and third characteristics of the OLS estimates refer to the amount of variation of the slope estimates around the true values--that is, to the dispersion of the sampling distribution of the slope estimates. The statement that the regression estimates are the most efficient estimates of the slope means that, compared to alternative estimation procedures, the OLS estimates have the smallest variance around the true value.[8] It is important to have an estimation procedure with a small variance because, in the long run, the estimates would be closer to the true value than they would be were we using a less efficient estimation procedure. The standard deviation of the slope estimates provides us with a measure of their efficiency. Since the estimates are unbiased and thus their expected value is the true value, their standard deviation provides us with an approximation of the standard deviation of the true sample distribution (i.e., the standard error). We emphasize that these values are only approximations because of the relatively small number of replications that we have conducted.

The third important property of the OLS estimates is that the <u>estimate of the standard error of the slope</u> is an unbiased estimate. The standard error of any statistic is the standard deviation of its sampling distribution. An estimate of this value gives the researcher an idea of how much one can expect the estimates to differ from the true value in the long run, and provides the basis for tests of significance. It can be demonstrated that the OLS procedures provide estimates of the standard error of the slope that have an expected value identical to the true standard error of the slope (Beals 1972: 237). This property is illustrated in Table 1, which shows the mean of the twenty estimates of the standard error (.045) as almost exactly equal to the standard deviation of the slope estimates (.048). In other words, the mean of column 3 is approximately equal to the standard deviation of column 2; this is to be expected, since both of these values represent estimates of the standard error of the slope estimates.

The fourth column of Table 1 provides the coefficient of determination (the square of the product moment correlation coefficient) for the twenty sets of data. Previous discussions of Galton's problem have focused almost exclusively on this value; some of the misunderstanding of the nature of the problem is a result of the failure to distinguish between the correlation coefficient and such other statistics as the slope estimates and the estimates of the standard error of the slope--which estimate completely different aspects of the regression model. The coefficient of determination estimates the proportion of the variation in the dependent variable that can be attributed to (explained by) the variation in the independent variable. The mean of the values in Table 1 (.895) corresponds closely to the true value for the model (.90). In subsequent discussion we do not emphasize the coefficient of determination or the correlation coefficient, because the most important effects of Galton's problem on statistical estimators can be seen more clearly with respect to the estimates of the slope and the standard error of the slope.

The Nature of Galton's Problem

Given the linear regression model, we argue that Galton's problem is not a single phenomenon or issue. Rather it is a special problem that frequently arises when dealing with historically interdependent social units influenced by many common geographically distributed factors. In what way are the assumptions of the model violated by interdependence of units and what are the consequences for the statistical estimators of that model? We discuss five specific models which represent different types of violations (or apparent violations) of the assumptions; then we demonstrate with simulated data how the usual OLS estimates are influenced by the modifications of the assumptions; finally, we apply two adjustment procedures to see whether they provide better estimates.

Five Diffusion Models

The diffusion models used in this study are similar to those used in earlier studies by Loftin (1975) and by Wirsing (1975); the models assume that diffusion moves in only one dimension along an array which is analogous to Naroll's (1961: 26) diffusion arcs.[9] Diffusion is simulated in the models by constructing the values of a variable at point p such that they are partially determined by the values of the same variable at an adjacent point (p-1), but not by other points in the array. The assumptions built into the models, especially that diffusion moves in only one dimension, are significant simplifications and should not be ignored in subsequent work. For present purposes, however, they provide a way to investigate the properties of estimators such as those suggested by Wirsing (1975) systematically and to clarify the theoretical nature of Galton's problem.

Figures 2-6 summarize the assumptions of the five diffusion models used in our study. Note that there are two versions of each model; version A is distinguished from version B only by the value assigned to the slope (1.0 in version A and 0 in version B). In other respects, including the actual numbers used in the simulations, the two versions of the models are identical. Tables 2-6 summarize the most relevant results of the data analysis.

Model 1.

Model 1 (see Figure 2) is constructed so that only the independent variable is diffusing: X_p is a function of X_{p-1} and the disturbances represented by V_p; the dependent variable is a function of the independent variable and the disturbances represented by U_p, but the disturbances of X_p and the disturbances of Y_p are different and independent of each other. Naroll (1964: 866) has argued that in cases such as this, where only one variable is diffusing, Galton's problem does not exist. Since none of the assumptions of the linear regression model is violated, Naroll's conclusion is supported by the logic

Figure 2. Model 1: Only the Independent Variable Diffuses

$$\begin{tikzcd}
V_{p-1} \arrow[dr] & & & U_{p-1} \arrow[dl] \\
& X_{p-1} \arrow[r] \arrow[dd] & Y_{p-1} & \\
V_p \arrow[dr] & & & U_p \arrow[dl] \\
& X_p \arrow[r] & Y_p &
\end{tikzcd}$$

$$Y_p = \alpha + \beta X_p + U_p$$

$$X_p = \phi X_{p-1} + V_p$$

Where:

V_p and U_p are normally and independently distributed random variables with mean zero and standard deviation ten (μ=0, σ=10).

X_0 is a uniformly distributed random variable with range 0 to 99.

$\alpha = 0$

β = 1.0 (for version A), 0 (for version B)

$\phi = 0.9$

of the model. It can be demonstrated that all of the desirable properties of the regression estimators will be retained when only the independent variable diffuses.[10] The results in Table 2 are consistent with these expectations; the mean of the OLS estimates of the slope is very close to the true value (1.014 is the estimate of the true value of 1 for version A and .014 is the estimate of the true value of 0 for version B); the mean of the estimated standard errors of the slope (.064) is very close to the standard deviation of the twenty slope estimates (.06) and no errors would have been made in either version of the model had the slopes been tested for statistical significance.[11]

While it is true that the OLS procedures provide valid estimates for Model 1, the model points up some serious errors in the research strategy that Naroll has suggested

Table 2. Analysis for Model 1

Estimation Method	*Mean of Slope Estimates*	*Mean of Estimates of Standard Error of the Slope*	*Standard Deviation of Slope Estimates*	*Number of Errors in Tests of Significance**
		Version A (B = 1.0)		
OLS Regression	1.014	.064	.060	0
Wirsing's Method	1.040	.118	.122	0
Durbin's Method	1.015	.063	.058	0
		Version B (B = 0)		
OLS Regression	.014	.064	.060	1
Wirsing's Method	.040	.118	.122	0
Durbin's Method	.015	.063	.058	3

*Two tail test, .05 level of significance
(Twenty Samples of Sixty Observations Each)

for analyzing data such as those generated according to Model 1. In version B of Model 1 the application of a test such as Naroll's (1964, 1970) linked pair test would lead one to the appropriate conclusion for this model. That is, X would be significantly autocorrelated, but Y would not be, so one would ignore Galton's problem and draw inferences on the basis of the OLS estimates of the model. In contrast, note what would happen in a case like version A of Model 1--where both X and Y are autocorrelated because X is autocorrelated and it has a direct effect on Y. The mean autocorrelated coefficients for the twenty replications of version A are: r_{XX} = .831 and r_{YY} = .662. Following Naroll's logic, one would conclude that Galton's problem exists and that some precautions should be taken in drawing inferences from the OLS estimates. In fact, there are no problems with inferences from the OLS estimates; the slope estimate is unbiased and efficient, and the standard error estimate is unbiased. A judgment based on the magnitude of the autocorrelation coefficients would be misleading.

It is interesting to see what would happen were we to apply an adjustment procedure such as Wirsing's second order partial regression to the data generated according to Model 1.[12]

Table 2 shows that Wirsing's estimates of the slope and the standard error of the slope are very close to their true values (i.e., they are unbiased), but they are less efficient than the OLS estimates. This is reflected by the larger mean of the estimates of the standard error of the slope (.118 for Wirsing's method as opposed to .064 for the OLS estimates), and the increase in the number of errors that one would make in testing the significance of the version B slopes (one with the Wirsing method and none with the OLS estimates). The application of Wirsing's method thus leads, in the case of Model 1, to estimates that are slightly less desirable than those which would have been obtained by applying the usual regression procedures.

Table 2 and subsequent tables also present the results of another estimation procedure closely related to Wirsing's but which requires a few additional calculations. This procedure has been suggested by Durbin (1960) for use in time-series analysis.[13] It requires two steps: The first is identical to Wirsing's method; the dependent variable is regressed on the independent variable at point p, the p-1 value of the independent variable (X_{p-1}), and the p-1 value of the dependent variable (Y_{p-1}). The model is as follows:

$$Y_p = \alpha* + \beta X_p + \gamma X_{p-1} + \Theta Y_{p-1} + U_p$$

In the second stage the estimated coefficient of the Y_{p-1} variable, which we will call Θ, is used to construct two new variables: (Y_p - Θ Y_{p-1}) and (X_p - Θ X_{p-1}). Then a new regression analysis is conducted with the model:

$$(Y_p - \Theta Y_{p-1}) = \alpha* + \beta(X_p - \Theta X_{p-1}) + U_p$$

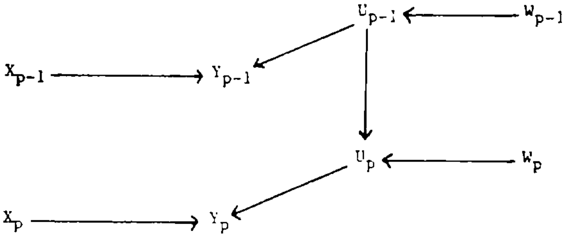
where $\alpha* = \alpha(1 - \Theta)$.

Durbin has shown that the OLS estimates of the $\alpha*$ and β coefficients will be efficient estimates for large samples. Note that in Table 2 the Durbin method provides estimates that are virtually identical to the OLS estimates and better estimates than those derived from Wirsing's procedure.

Model 2.

Model 2 (see Figure 3) provides an interesting contrast to Model 1; in this case only the disturbance variables are diffusing. This is a violation of the assumption of the linear regression model that disturbances are independent of each other (see assumption number three above), and will lead to problems with the OLS estimators. Note, however, that since the independent variable is not diffusing (the mean of the autocorrelation coefficients for X_p in Model 2 is .00001) Naroll's linked pair test would lead one to the conclusion that Galton's problem is not an issue in this analysis, and one would proceed as though there were no problems. Such a conclusion would be somewhat misleading. It can be shown (Johnston 1972: 247-249) that where the disturbances are interdependent, as they are in Model 2, the OLS estimates of the slope will remain unbiased, but the

Figure 3. Model 2: Only the Disturbance Variable Diffuses



$$Y_p = \alpha + \beta X_p + U_p$$

$$U_p = \theta U_{p-1} + W_p$$

Where:

W_p is a normally and independently distributed random variable with mean zero and standard deviation ten ($\mu=0$, $\sigma=10$).

X_p is a uniformly distributed random variable with range 0 to 99.

U_0 is a normally distributed random variable with mean zero and standard deviation 22.94 $\left(\mu=0,\ \sigma=10\sqrt{\frac{1}{1-\theta^2}}\right)$.

$\alpha = 0$

$\beta = 1.0$ (for version A), 0 (for version B)

$\theta = 0.9$

estimate of the standard error of the slope will be biased. However, the extent of the bias will be negligible where the disturbances are the only variables that are autocorrelated. Only when the disturbance and the independent variable are both autocorrelated will the bias be serious; we deal with this case in Model 3. We believe that in most empirical situations, it would be quite unusual to find a case like Model 2. These qualities of the OLS estimates under the assumption of Model 2 are reflected in the simulated data (Table 3) by the fact that the mean of the OLS estimate of the standard error of the slope (.099) is very close to the standard deviation of the twenty slope estimates derived from the data analysis (.092).

Table 3. Analysis for Model 2

Estimation Method	Mean of Slope Estimates	Mean of Estimates of Standard Error of the Slope	Standard Deviation of Slope Estimates	Number of Errors in Tests of Significance*
		Version A (B = 1.0)		
OLS Regression	.979	.099	.092	0
Wirsing's Method	1.010	.047	.038	0
Durbin's Method	1.009	.035	.033	0
		Version B (B = 0)		
OLS Regression	-.021	.099	.092	1
Wirsing's Method	.010	.047	.038	1
Durbin's Method	.009	.035	.033	1

*Two tail test, .05 level of significance
(Twenty Samples of Sixty Observations Each)

Note that the slope estimates remain unbiased even though the disturbance variables are interdependent. This is an important fact because previous discussions of Galton's problem have suggested that interdependence of units of analysis tends to produce spuriously high estimates of relationships between variables.

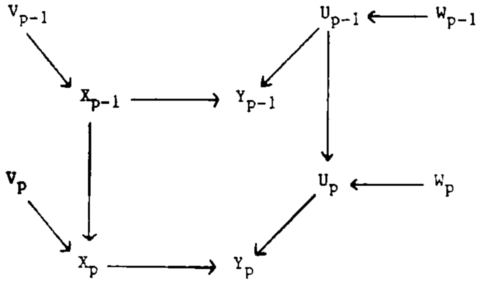
The Wirsing method provides better estimates for Model 2 than OLS regression, in that the estimates of the slope are more efficient. However, the Durbin method estimates are even more efficient than the Wirsing estimates and they, too, are unbiased.

Model 3.

Model 3 combines Models 1 and 2 to produce a case where both the independent variable and the disturbances are diffusing (see Figure 4). This is the only one of the three models discussed where Naroll's linked pair test would consistently lead to the appropriate conclusion about the existence of Galton's problem; but note that it would not allow one to distinguish between version A of Model 1, where there is no problem with the OLS estimates, and version A and B of Model 3, where there will be serious problems with the OLS estimates. Our primary concern in this paper is not with tests for the existence of Galton's problem, but with those instances where the dependent variable is continuously distributed and regression analysis is appropriate. A better test than any of those suggested by Naroll would be the Durbin-Watson statistic which has been used extensively in time series analysis. The Durbin-Watson statistic uses the estimated disturbances of the OLS regression analysis to test for the existence of autocorrelation in the disturbances--see Beals (1972: 348-352); Wonnacott and Wonnacott (1970: 52-53).[14] Where it is not possible to estimate the disturbance variables, other techniques will have to be used, but our analysis points out an important class of errors that may be made by relying on the linked pair test.

Like Model 2, the interdependence of the disturbances in Model 3 is an explicit violation of an assumption of the

Figure 4. Model 3: The Independent and Disturbance Variables Diffuse Independently



$$Y_p = \alpha + \beta X_p + U_p$$

$$X_p = \phi X_{p-1} + V_p$$

$$U_p = \theta U_{p-1} + W_p$$

Where:

V_p and W_p are normally and independently distributed random variables with mean zero and standard deviation ten ($\mu=0$, $\sigma=10$) and $V_p \neq W_p$

X_0 is a uniformly distributed random variable with range 0 to 99.

U_0 is a normally distributed random variable with mean zero and standard deviation 22.94 $\left(\mu=0,\ \sigma= 10\sqrt{\frac{1}{1-\theta^2}}\right)$.

$\alpha = 0$

$\beta = 1.0$ (for version A), 0 (for version B)

$\phi = \theta = 0.9$

linear regression model, and the consequence for the OLS estimates are similar (see Table 4); the estimates of the slope remain unbiased, but they are no longer as efficient. Moreover, the estimates of the standard error of the slope are rather severely biased; the mean of the estimates of the standard error of the slope is only about one fourth of the standard deviation of the twenty estimates of the slope

Table 4. Analysis for *Model 3*

Estimation Method	*Mean of Slope Estimates*	*Mean of Estimates of Standard Error of the Slope*	*Standard Deviation of Slope Estimates*	*Number of Errors in Tests of Significance**
		Version A (B = 1.0)		
OLS Regression	1.066	.123	.460	0
Wirsing's Method	.959	.117	.196	0
Durbin's Method	.940	.112	.211	0
		Version B (B = 0)		
OLS Regression	.066	.123	.460	15
Wirsing's Method	-.041	.117	.196	4
Durbin's Method	-.060	.112	.211	5

*Two tail test, .05 level of significance
(Twenty Samples of Sixty Observations Each)

derived from the simulated data. These results are consistent with proofs that are readily available in the statistical literature (Johnston 1972: 247-249). In Model 3, the estimate of the standard error is more severely biased than in Model 2 because both the independent variable and the disturbances are autocorrelated.[15]

In view of previous discussions of Galton's problem, the fact that the slope estimates are unbiased is an especially important feature of our argument. The simulation illustrates that in a situation such as Model 3, where the independent variable and the disturbances are autocorrelated, slope estimates are not spuriously high. Problems arise with the variability of the slope estimates, not with their mean.

One might suspect that the lack of spuriously high estimates of the relationship is somehow a feature of our choice of the slope coefficient to measure the relationship. Thus if we had selected other measures of association such as Kendall's Tau or Goodman and Kruskal's Gamma we would have found that the relationships were spuriously high. This, however, is not the case. The major features of our argument are not affected by the choice of the measure of association. For example, had we selected the product moment correlation coefficient as our measure of association in version B of Model 3 we would have made exactly the same number of errors as we did using the slope estimate and ordinary least squares regression. In fifteen out of the twenty samples we would have concluded that there was a statistically significant relationship, when in fact the true relationship was zero. In six of the fifteen erroneous inferences we would have concluded that the relationship was negative and significant and in nine of them we would have concluded that it was positive and significant.

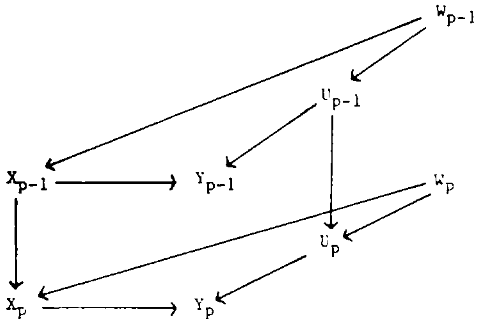
There is one important difference between correlation and regression slope coefficients in this context. Because the autocorrelation of the disturbances leads one to underestimate the unexplained variance, the absolute value of the correlation coefficient, but not the regression coefficient, tends to be overestimated. This does not influence the mean correlation, since some will be positive and some will be negative. However, were one to look only at the absolute magnitude of the correlation or only at the coefficient of determination (r^2), one would conclude that the associations were spuriously high.

The Wirsing and the Durbin estimates of the slope are also unbiased and are more efficient than the OLS estimates by about the same amount. However, estimates of the standard error of the slope derived from the Wirsing and the Durbin procedures also appear to be biased downward. The magnitude of the bias is less than with the OLS estimates (only about forty to fifty percent rather than seventy-five percent for the OLS estimates) and thus the number of errors that would be made in tests of significance for version B of the model is reduced (from fifteen to four or five), but the bias will still be a source of errors in inferences about the true model.

Model 4.

Model 4 is like Model 3 in that both the independent variable and the disturbance variables are diffusing (see Figure 5). There is a difference, and a very important one; the model is constructed so that the variables diffuse according to a similar pattern. In terms of the model, both X_p and U_p (and therefore Y_p) share the term W_p which represents a common diffusion process. Model 4 thus violates two of the assumptions of the linear regression model; like Models 2 and 3, the disturbance variables are not independent of each other, and in addition, the disturbances are not independent of the explanatory variable (see assumption number four above). The effects of this modification of the model on the statistical estimates are dramatic (see Table 5). The most important effect is that all of the estimates of the slope (OLS, Wirsing's and Durbin's) are severely biased upward. The reason for the bias is that the covariation between X_p and Y_p that is due to the common diffusion factor (the W_p disturbance factors in Figure 5) is attributed to the direct relationship between X_p and Y_p. The impact of the bias is particularly serious for version B, where there is no true

Figure 5. Model 4: The Independent and Disturbance Variable Diffuse Together



$$Y_p = \alpha + \beta X_p + U_p$$

$$X_p = \phi X_{p-1} + W_p$$

$$U_p = \theta U_{p-1} + W_p$$

Where:

W_p is a normally and independently distributed random variable with mean zero and standard deviation ten ($\mu=0$, $\sigma=10$).

X_0 is a uniformly distributed random variable with range 0 to 99.

U_0 is a normally distributed random variable with mean zero and standard deviation 22.94 $\left(\mu=0,\ \sigma= 10\sqrt{\frac{1}{1-\theta^2}}\right)$.

$\alpha = 0$

$\beta = 1.0$ (for version A), 0 (for version B)

$\phi = \theta = 0.9$

relationship between X_p and Y_p. Here all of the slope estimates are significantly greater than zero, and one would make an error in all twenty replications if one followed the usual procedures and concluded that there is a true non-zero relationship between X_p and Y_p. All of the three desirable properties of the OLS estimates are lost in a case like Model 4: the slope estimates are biased (column 1), the estimates of the standard errors of the slope are biased

Table 5. Analysis for Model 4

Estimation Method	Mean of Slope Estimates	Mean of Estimates of Standard Error of the Slope	Standard Deviation of Slope Estimates	Number of Errors in Tests of Significance*
		Version A (B = 1.0)		
OLS Regression	1.779	.060	.259	0
Wirsing's Method	1.723	.057	.222	0
Durbin's Method	1.717	.055	.222	0
		Version B (B = 0)		
OLS Regression	.779	.060	.259	20
Wirsing's Method	.723	.057	.222	20
Durbin's Method	.717	.055	.222	20

*Two tail test, .05 level of significance
(Twenty Samples of Sixty Observations Each)

downward (compare column 2 with column 3), and the estimates are inefficient (column 3). Of course, the relatively small estimates of the standard error of the slope are not completely erroneous. The slope estimates are not widely dispersed around their mean. The problem, however, is that since they are biased estimates, they are dispersed around the wrong value. Nevertheless, this bias in the slope estimates is confounded by an additional bias in the estimates of the standard error of the slope. This can be seen in Table 5, where the mean of the OLS estimates of the standard error of the slope (.060) is only twenty-three percent of the standard deviation of the twenty slope estimates derived from the simulations (.259).

It is important to note that Model 4 is the only one of the models that produces the kind of spurious relationships generally thought to characterize Galton's problem. This illustrates the importance of distinguishing among different configurations of the problem and the utility of the linear regression model as a way of thinking about the problem.

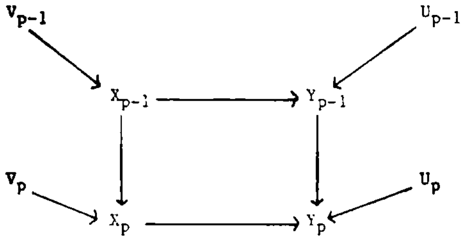
Returning to Table 5, note that the Wirsing method and the Durbin method are no better than the OLS estimates of the model. They provide no "solution" to Galton's problem when the disturbance variables are confounded with the independent variable. The only way to derive better estimates would be to obtain direct measures of the common disturbance factors and bring them explicitly into the equation, and then estimate the effect of X on Y controlling for the effects of the common variables. Therefore, the solution requires an expansion of the theoretical model to include additional independent variables. There are no estimation procedures which will improve the situation, because the source of the problem is a theoretical misspecification.

Model 5.

Model 5, the final model to be considered, is one in which both the independent and the dependent variables are diffusing but the disturbance variables are not (see Figure

6 and Table 6). No apparent violations of the OLS regression model are built into the specification of Model 5, yet there are serious problems with the OLS regression results presented in Table 6.[16] For Version A of the model, the OLS regression produces slope estimates which are biased upward and standard error estimates which are biased downward and are not efficient. For version B of model, the slope estimates are <u>not</u> biased, but the standard error estimates are--leading to a large number of errors in the tests of significance.

Figure 6. Model 5: The Independent and Dependent Variables Diffuse Independently



$$Y_p = \alpha + \beta X_p + \gamma Y_{p-1} + U_p$$

$$X_p = \phi X_{p-1} + V_p$$

Where:

V_p and U_p are normally and independently distributed random variables with mean zero and standard deviation ten ($\mu=0$, $\sigma=10$).

X_0 is a uniformly distributed random variable with range 0 to 99.

$\alpha = 0$

$\beta = 1.0$ (for version A), o (for version B)

$\phi = 0.9$

$\gamma = 0.9$

Table 6. Analysis for Model 5

Estimation Method	Mean of Slope Estimates	Mean of Estimates of Standard Error of the Slope	Standard Deviation of Slope Estimates	Number of Errors in Tests of Significance*
		Version A (B = 1.0)		
OLS Regression	3.359	.544	1.151	0
Wirsing's Method	.983	.134	.137	0
Durbin's Method	.936	.312	.281	0
		Version B (B = 0)		
OLS Regression	.030	.116	.306	10
Wirsing's Method	-.010	.134	.129	0
Durbin's Method	.006	.127	.126	2

*Two tail test, .05 level of significance
(Twenty Samples of Sixty Observations Each)

The explanation for these results is similar to the explanation raised in the discussion of Model 4; that is, the OLS regression with X_p as the sole independent variable is a misspecification of the theoretical model which generated the data. The dependent variable is caused by both X and by the dependent variable at the previous point in the array. By omitting this second cause of Y, we are allowing the X variable alone to account for variations in the dependent variable, and we are forcing the error term of the regression equation to be autocorrelated because of the omitted lagged value of the dependent variable. Clearly, the situation is undesirable. This is a case where Naroll's linked pair test would lead to the correct conclusion that Galton's problem exists and that caution should be exercised in interpreting the OLS estimates. The mean autocorrelation coefficients for the model are: $\bar{r}_{XX}$ = .884; $\bar{r}_{YY}$ = .980 (version A); and $\bar{r}_{YY}$ = .841 (version B).

There are, however, adjustments which can improve the estimation of the model. The best solution to the problems raised by this model is to identify correctly the theoretical model that generated the data. When we analyzed this model, using the correct specification of two independent variables (X_p and Y_{p-1}), the regression estimates were unbiased and efficient.[17] In practice, however, it may be difficult to determine, a priori, what the appropriate model is. In a situation such as this one, the Wirsing procedure produces good results because it employs a model that is very close to the true model used to generate the data. The Wirsing estimate of the slope (.983) is unbiased; the estimate of the standard error of the slope (.134) is unbiased (compare columns 2 and 3), and the slope estimate is far more efficient than the OLS regression estimate (.137 vs. 1.151). The Durbin procedure is also superior to OLS since it yields unbiased slope and standard error estimates (.936 and .312 respectively), yet the estimates are not as efficient as the Wirsing estimates (.281 vs. .137). In this case, the Durbin procedure is less desirable because it is generally used to adjust for autocorrelation of the disturbances and in this model, the disturbances are not autocorrelated once the model has been specified correctly.

The results for this model also illustrate the advantage of using the Durbin-Watson statistic to test for autocorrelation over a technique such as Naroll's linked pair test. The Durbin-Watson statistic for the OLS regression for version A of the model is significant in each of the twenty samples we analyzed. This is an indication that the disturbances are autocorrelated and that there are potential problems with OLS procedures to estimate the model. The same conclusion follows from the linked pair test. However, the Durbin-Watson statistic is a more sensitive indicator of the nature of the problem. When the model is specified correctly, as with the Wirsing procedure, the corresponding Durbin-Watson statistics are not significant. Thus the statistic can be used to identify what configuration of Galton's problem is at hand, while the linked pair test has a much more restricted use. In short, if the Durbin-Watson statistic is significant, the first step is to rethink the model under examination to guard against omitting causal variables. If the Durbin-Watson statistic is still significant for respecified models, the problem lies in the autocorrelated disturbances, and the strategies suggested under our discussion of Models 2 and 3 are appropriate.

Comparison of OLS, Wirsing and Durbin Methods

It is apparent from our empirical analysis that Wirsing's procedure provides more efficient estimates of the slope coefficients than does OLS regression for Models 2, 3, and 5; it is less efficient in the case of Model 1, and provides little, if any, improvement over OLS estimates in the case of Model 4. On the other hand, the Durbin procedure provides estimates that are as good as or better than OLS and Wirsing estimates for Models 2 and 3.

For Model 1, no adjustment procedure is necessary, because none of the assumptions of the linear regression model is violated. OLS regression is appropriate; the Durbin-Watson statistic for the OLS regression indicates that the disturbances are not autocorrelated, therefore one can proceed with OLS estimation.

None of the three estimation procedures discussed here (OLS, Wirsing, and Durbin) provides unbiased, efficient estimates for Model 4, because this model not only violates the assumption that the disturbances are independent of each other, but also represents a severe violation of the assumption that the disturbances are independent of the explanatory variables.

Model 5 is the most complicated model we have analyzed, and strategies for adjustment procedures are different than for Models 2 and 3. The problem that occurs with OLS estimates for Model 5 is that the model has not been correctly specified. The Wirsing procedure is a good approximation of the correct model, so estimates with this procedure are better than OLS estimates. The Durbin procedure is unnecessary and undesirable because once the model has been correctly specified, the disturbance variables are not autocorrelated.

For models with autocorrelated disturbances (e.g., Models 2 and 3), there is empirical and theoretical justification for choosing the Durbin procedure over the Wirsing procedure. The problem with those models is the violation of the assumption that the disturbance variables are independent of each other (assumption number three above). It is important to realize that so long as the disturbance variables are independent of each other, it makes no difference whether the dependent variable and the independent variables are autocorrelated.[18] Interdependence of variables other than the disturbances does not violate any of the assumptions of the linear regression model, and thus the OLS estimates of the slope and its standard error will be unbiased and efficient if the model is correctly specified. Wirsing's justification for his procedure is misleading in that he suggests that it "is able to control for the diffusional effects of both sociocultural traits" (Wirsing 1975: 150).[19] The problem with Models 2 and 3 is not that both sociocultural traits diffuse; rather it is that the disturbances diffuse. Wirsing's estimation procedures are sometimes more efficient than OLS regression estimates, but not for the reasons that he suggests. The estimates are better because Wirsing's procedure includes a control for the effects of the diffusion in some of the models. For example, in Model 3 we can derive Wirsing's estimation equation from the definitions that are provided in Figure 4. The equations for Y_p and U_p are as follows:

$$Y_p = \alpha + \beta X_p + U_p$$

$$U_p = \Theta U_{p-1} + W_p$$

Substituting the second equation into the first gives:

$$Y_p = \alpha + \beta X_p + \Theta U_{p-1} + W_p$$

The Y_{p-1} values are generated by the same process and can be written as:

$$Y_{p-1} = \alpha + \beta X_{p-1} + U_{p-1}$$

This can be rearranged and substituted into the Y_p equation as follow:

$$U_{p-1} = Y_{p-1} - (\alpha + \beta X_{p-1})$$

$$\begin{aligned} Y_p &= \alpha + \beta X_p + \Theta U_{p-1} + W_p \\ &= \alpha + \beta X_p + \Theta(Y_{p-1} - \alpha - \beta X_{p-1}) + W_p \\ &= \alpha + \beta X_p + \Theta Y_{p-1} - \Theta\alpha - \Theta\beta X_{p-1} + W_p \\ &= \alpha\ (1 - \Theta) + \beta X_p + \Theta Y_{p-1} - \Theta\beta X_{p-1} + W_p \end{aligned}$$

The last expression is the model that Wirsing uses for his estimation procedure. While it may seem that this equation will meet all of the assumptions of the linear regression model, it does not. Because the equation contains Y_{p-1}, there is a dependence between W_p and the values of Y subsequent to Y_p, and therefore assumption number four is violated. It can be demonstrated that because of this dependence the estimates of the coefficients in the equation will be biased and often relatively poor in the sense that they will have a high variance. However, as the sample size increases, they will tend to converge on the true value and thus provide estimates that are close to the true value (Wonnacott and Wonnacott 1970: 146-147; Beals 1972: 367-368). In addition, it can be shown that Durbin's procedure, which uses the estimate of Θ from a model which is identical to Wirsing's estimation equation to transform the X_p and Y_p values, will generally provide better estimates of β than the estimates that are derived from the first stage (i.e., Wirsing's estimates of β).[20]

Our simulation fails to illustrate clearly this property of Durbin's estimates because the number of cases in each set of data is relatively large (N = 60); therefore, the estimates of β derived from Wirsing's procedure are good estimates and are not improved by the use of Durbin's transformation. However, it is important to emphasize that this is because of the large number of cases in the samples, and we would expect that the Durbin procedure would have provided an improvement over the Wirsing estimates had the number of observations in each sample been smaller.

Conclusions

In summary, Galton's problem is actually a series of potential problems rather than one unique problem. We have presented five models which are different configurations of the diffusion process, and the nature of the problem and the appropriate correction procedures vary across the models. There is no unique "solution" to Galton's problem, just as there is no unique configuration of the problem. There are, however, analysis strategies that are useful for specific research problems.

For any particular research problem, it is necessary to test first for the presence of diffusion, and second, to identify the configuration of diffusion that is operating.

The Durbin-Watson statistic can be used to test for diffusion of the disturbance variables, while the autocorrelation coefficients for the independent and dependent variables can help in identifying the nature of the diffusion process. If the Durbin-Watson statistic does not show that the disturbances are diffusing, then no adjustment procedures are necessary; the regression estimates are unbiased and efficient even in those cases where the independent variable is diffusing (see Model 1 results). If the Durbin-Watson statistic does show that the disturbances are diffusing, adjustment procedures may be necessary. In this case, the autocorrelation coefficients should be examined in conjunction wih the Durbin-Watson statistic. If the autocorrelation coefficients do not indicate diffusion, then the effect of diffusion of the disturbances on the OLS estimates will not be severe, and the Durbin procedure we have reviewed will be a sufficient correction technique (see Model 2). If, on the other hand, the independent variable is also diffusing, as indicated by a significant autocorrelat- ion coefficient, then the biasing effects of diffusion will be more pronounced. In this case, however, the bias affects the variance of the estimates rather than the mean; the Durbin procedure is again appropriate, although Wirsing's procedure will also provide better estimates than OLS (see Model 3).

The configurations of Galton's problem which involve theoretical misspecification of the model are the most complex and the most difficult to deal with. The statistical procedures we have discussed will not help in those cases where the model has not been correctly specified, nor will they provide guidelines for indicating whether such misspecification is relevant. Models 4 and 5 illustrate the point. Both are misspecified by the OLS model, and the undesirable properties of both are clear in the results. However, partialling "works" as a solution to Model 5 because the inclusion of lagged values for the independent and dependent variables models the true mechanism which generated the data. Partialling does not work for Model 4 because of the omission of the variable which is causing the disturbances to diffuse. In practice, of course, it would not be possible to distinguish between Model 4 and Model 5, so the best strategy here, as with all models, is a careful examination of the theoretical specification of the model.

In short, Galton's problem does not necessarily lead to bias in the estimates of relationships, but it may, in some cases. There is no one solution to the problem, short of very careful theoretical specification and rigorous checks on the estimates obtained. We have relied on regression analysis to illustrate the nature of the problem(s) and various adjustment procedures. However, the problems we have discussed are not unique to the statistics presented here. Any statistical procedure will be problematic; the nature of the effects of diffusion is perhaps more clear with regression estimates, but alternative estimation procedures are certainly not a guarantee against faulty inferences where Galton's problem is involved.

Finally, we have relied on the statistical literature on time series to build our models and to examine the effects of diffusion. Throughout the discussion, we have assumed that our cases are "aligned" in a meaningful fashion, just as cases in a time series are, by definition, aligned in a meaningful way. This is a crucial simplifying assumption, and the application of our discussion depends on the ability to model spatial diffusion to produce the time series analogy. This is by no means a straightforward task.

NOTES

[1] We thank Purushottam M. Kulkarni, Larry Hamilton, Michael Timberlake, Rolf Wirsing, and Raoul Naroll for helpful comments on earlier drafts of this paper, and Steve Schmaltz for invaluable computer assistance. The research was supported in part by grant Soc. 75-23382 from the National Science Foundation.

[2] Additional discussion of the linear regression model may be found in Blalock (1979: 382-396) and Beals (1972: Chap. 1). More technical treatments may be found in Goldberger (1964: 151-212, especially pp. 161-162) and Johnston (1972: Chap. 8).

[3] The subscript "p" is used throughout this paper to stand for a particular point in an ordered array, such as one of Naroll's (1961: 24-29) diffusion arcs.

[4] In many specifications of the model, X_p is assumed to be a nonstochastic variable with values fixed in repeated samples. This assumption makes the demonstration of the properties of the regression estimates of the model easier, but it is not a realistic assumption for non-experimental studies. The weaker assumption, that the disturbances are independent of the explanatory variable, does not modify the major derivations that can be made from the model and is more consistent with cross-cultural research. Also, strictly speaking, two other assumptions are necessary: that the values of X_p in the sample must not all be equal to the same number; and that if there is more than one explanatory variable, no exact linear relationships exist among them (see Goldberger 1964: 161-162).

[5] In our discusion we use the term OLS regression estimates to refer to estimates derived from ordinary least squares procedures. The special designation is useful because we discuss other estimation procedures.

[6] See, for example, Beals (1972: 235-240) and Johnston (1972: 123-127).

[7] If the number of samples were infinitely large, the mean of the distribution would equal the true parameter.

[8] More precisely, this statement is restricted to linear and unbiased estimates and should be stated as follows: "Within the class of linear unbiased estimators

of α (or β), the least square estimator has minimum variance" (Wonnacott and Wonnacott 1970: 21).

[9] Our models draw on the analogy between Galton's problem and the first-order autoregressive model, the properties of which have been analyzed extensively in the context of time series analysis. See, for example, Kmenta (1971: 269-297). Blalock (1968: 175) first pointed out the importance of the time series analogy.

[10] Formulas are available which allow one to derive the true standard error of the slope of our models (see Johnston 1972: 146-249). For Model 1 the OLS estimates of the standard error of the slope are unbiased and therefore are expected to be equal, in the long run, to the true standard error of the slope.

[11] The test is of the model that $\beta = 0$, with .05 level of significance. The critical value of t for a two-tailed test with approximately sixty degrees of freedom is 2.0.

[12] Wirsing used partial correlation, while we use partial regression. The issues are clearer in the case of regression, and therefore we rely on these statistics. Our conclusions are general and apply to both correlation and regression.

[13] A discussion of Durbin's procedure can be found in Johnston (1972: 263-264).

[14] The Durbin-Watson statistic is closely related to the autocorrelation coefficient for the disturbances. It is preferable to the autocorrelation coefficient because it has a known distribution which can be used for tests of significance, while the distribution of the autocorrelation coefficient is difficult to determine (Beals 1972: 348-349).

[15] There are two sources of error in the estimate of the standard error; one affects the estimate of the variance of the regression coefficient, and the other affects the

estimate of the variance of the disturbance. Both are more severe when the disturbance and the explanatory variable are diffusing (see Johnston 1972: 247-249).

[16] In fact, there is a violation of the OLS regression model. Since Y_{p-1} is an explanatory variable, the disturbance variables are not independent of all of the explanatory variables (see assumption four above). But, so long as the disturbances are independent (assumption three) this will not be a serious problem. The OLS estimates can be expected to perform well. In technical terms they will be "consistent." See Johnston (1972: 305-306) for additional discussion.

[17] For version A of the model, the mean of the slope estimates is 1.007; the mean of the estimates of the standard error of the slope is .069; and the standard deviation of the slope estimates is .059. For version B, the comparable results are .025, .063, and .062, respectively.

[18] Of course, autocorrelation of the explanatory variable contributes to the bias in the standard error estimates, but only where the disturbances are also autocorrelated. This is demonstrated by the contrast between models 1, 2, and 3. There is no bias in the standard error estimates in Model 1, where only the explanatory variable is autocorrelated.

[19] It is clear from the context that the two sociocultural traits referred to are the independent and dependent variables.

[20] Durbin has shown that the second stage estimates not only converge on the true value as the sample size increases, but also that the variance of the estimates will be smaller than any other "consistent" estimator (i.e., it is an asymptotically efficient estimator). For a discussion see Rao and Griliches (1969).

REFERENCES

Beals, Ralph E.

1972 Statistics for Economists. Chicago: Rand McNally.

Blalock, Hubert M.

1968 Theory Bulding and Causal Inferences. In Methodology in Social Research. Hubert M. Blalock and Ann B. Blalock, eds. pp. 155-198. New York: McGraw-Hill.

1979 Social Statistics. Revised second edition. New York: McGraw-Hill.

Durbin, J.

1960 Estimation of Parameters in Time-Series Regression Models. Journal of the Royal Statistical Society, Series B, 22 (January): 139-153.

Goldberger, Arthur S.

1964 Econometric Theory. New York: Wiley.

Johnston, J.

1972 Econometric Methods. New York: McGraw-Hill.

Kmenta, Jan

1971 Elements of Econometrics. New York: Macmillan.

Loftin, Colin

1975 Partial Correlation as an Adjustment Procedure for Galton's Problem. Behavior Science Research 10: 131-141.

Naroll, Raoul

1961 Two Solutions to Galton's Problem. Philosophy of Science 28: 15-39. Reprinted In Readings in Cross-Cultural Methodology. Frank Moore, ed. pp. 221-245. New Haven: HRAF Press, 1961.

1964 A Fifth Solution to Galton's Problem. American Anthropologist 66: 863-867.

1970 Galton's Problem. In A Handbook of Method in Cultural Anthropology. Raoul Naroll and Ronald Cohen eds. pp. 974-989. Garden City, N.Y.: Natural History Press. Reprinted 1973. New York: Columbia University Press.

Rao, Potluri and Zvi Griliches.

1969 Small-Sample Properties of Several Two-Stage Regression Methods in the Context of Auto-Correlated Errors. Journal of the American Statistical Association 64: 253-272.

Wirsing, Rolf

1975 Second Order Partials as a Means to Control Diffusion. Behavior Science Research 10: 143-159.

Wonnacott, Thomas H. and Ronald J. Wonnacott

1970 Econometrics. New York: Wiley.