J. Coll. admve Sci., Univ. of Riyadh, Vol. 7, *pp. 29—43 (1979-1980)*

Detecting Sources of Hospital Data Error; Application of Causal Model Approach

Dr. Jahangir Khan
College of Administrative Sciences, University of Riyadh, Riyadh, Saudi Arabia.

Data collected through survey usually contains problems of interaction and multicollinearity. Such is the case with a set of data collected from 480 U.S. hospitals. Responses from the Directors of the hospitals collected by two independent sources on a number of similar questions about the hospital programmes were found to be discrepant. A step-wise regression analysis and, later a Pearson's Product moment correlation identified four variables that correlated positively with the dependent variable.

Assuming that interaction between and within independent and dependent variables exists, this paper uses Blalock's causal model approach to identify the most critical variables influencing data discrepancy. For this purpose, four Models were developed and tested in relation to "Goodness of Fit" criterion for predicted and actual correlations. The best predictive model was found to be one which assumes an independent influence of both organizational variables and personal characteristics of the hospital Directors; *i.e.*, equations 13 to 18 give the best possible results in explaining why sets of responses are discrepant.

Introduction

It is axiomatic, that correlational analysis gives one certain inter correlations which merely indicate as to how a set of independent variables behave

relative to the dependent variable. This, however, does not directly furnish the answer to the question of causality[1]. Primarily this is a problem of control, i.e. we have problems of multi collinearity and interactions of variables even where we have, at least theoretically, some control, like in regression analysis[2]. This can be clarified if one considers the linear regression equation:

$$Y = \alpha + b_1X_1 + E$$

The main assumption in this equation is that E (the expression for error term) is not correlated with either X_1 (and another set of X_s if they were placed in the equation) or Y. A concommitant assumption is that all the X_s are also un-correlated in relation to their "independent" influence on Y. As may be seen these assumptions are untenable as far as reality is concerned. In such instance, it is necessary to use some other form of analysis. The current paper offers suggestions as to how these problems could be overcome. Using empirically generated correlations, this paper examines the causal influence of a set of predictors on a dependent variable; in so doing, the paper introduces the reader to Blalock's Causal Model Procedure.

Methods and Materials

Blalock's Causal model approach may be seen as a way to overcome the complications enunciated above in connection with the regression and correlational statistics[3]. The method makes explicit assumptions about the system of variables one is looking at. Briefly these assumptions include: (1) giving on explicit defition to only a finite set of variables, (2) representations of these variables in causal terms, (3) clarifying the confounding influence of exogeneous variables. The data for the paper is drawn from two sources; one, a study of Hospitals (N=480) in U.S. reporting data on

---

1. Vide Hubert Blalock, *Social Statistics,* New York, McGraw Hill, 1960, pp. 337—343.

2. Multi collinearity, briefly, refers to the confounding influence of high correlations among predictors. Interaction, on the other hand, refers to the situation in which the independent and dependent variables are so correlated that they are not only mutually causative, but might be related through other exogenous variables that are conventionally subsummed under the error term. For more technical and detailed discussion, see John Johntson, *Economic Methods,* New York: McGraw Hill, 1963. See also, Hubert Blalock, *Theory Construction: From Verbal to Mathematical Formulation,* New Jersey: Prentice Hall, 1969.

3. Hubert Blalockm *Gausal Inferences in Non-Experimental Research,* Chapel Hill: University of North Carolina Press, 1964.

various programmes carried out by the hospitals and two, American Hospitals Association reporting on the same data. The two data were found to be discrepant. Estimations of these errors were conducted by performing step vise regression and are reported elsewhere[4]. These estimations gave an idea of the kinds of factors that tend to influence the dependent variable; i.e. data discrepancy in this case. However, the question of causality remained unresolved. In order to resolve this dilema, Blalock's causal inference procedures have been used here to answer the causality question.[5] For this purpose, certain empirically obtained inter-correlations between a set of independent and dependent variables were placed in a general model i.e. Figure 2. An attempt was then made to construct several alternative models and to causally evaluate these using causal inference procedures.[6] Blalock's procedure is essentially in line with Sewell Wright's Path Analysis Method.[7]

Prior to the construction of these models, two necessary conditions were kept in view — one, that the variables included in the system should be those that the original analysis indicated as the most important ; and two, the specification of interrelationship of variables in a manner that certain variables, on a *priori* basis, are considered logically prior, in a causal chain, to certain others.

General Model & Correlation Matrix

By way of meeting the first condition, eight most important variables, i.e., those that accounted for most of the variance according to original step-wise regression analysis, were picked up. A Pearson's Product moment correlation was then computed, running each of these eight against the (dependent variable) discrepancy of data. In addition, intercorrelations

---

4. Jahangir Khan and James Veney, "Sources of Data Discrepency in a selected sample of Hospitals" a paper read at the 99th Annual meeting of the American Health Association and Related Organization, Minneapolis, Minn. October 10-15, 1971.

5. Sewell Wright. "The Method of Path Coefficients" *Annals of Mathematical Statistics*, 5 (September, 1934). pp. 167-215. See also John Juckey. "Causation. Regression and Path Analysis", in *Statistics and Mathematics in Biology*, Oscar Kempthone, et al. (eds). Iowa: American Iowa State College Press, 1954 pp. 35-66.

6. Jahangir Khan and James Veney, Op. Cit. 1971.

7. A certain amount of simplication is introduced as a result of using only a sub-set of variables, as in Figure 2; however, both the criteria of specifications of variables and the high correlations for the variables included in the system, make them the more critical for present purposes.

using the same procedure were computed for the eight predictors. The four, out of these eight, which correlated most highly with the dependent scores were picked up as the candidate variables for later analysis.

Relative to the second condition, i.e., the specification of linkages and causal priority of certain variables, a simple but logically valid conceptual schema was constructed (See Figure 1).

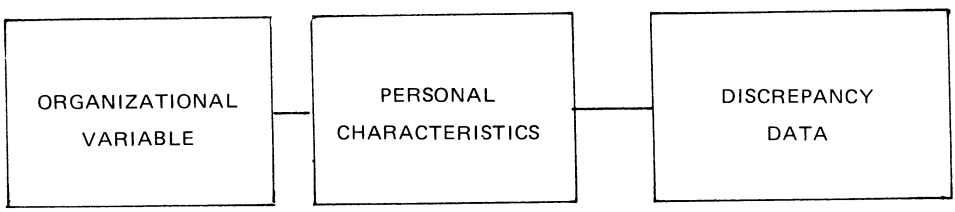


Figure 1: General Scheme of Causal Relationships of Variables

The schema suggests, that the organizational variables, in a causal chain, come before the variables titled as personal characteristics, which, in turn, are prior to the data discrepancy.

Evaluation of Models

Figure 2 shows the set of intercorrelations of the four variables as well as their correlation with the dependent variable.[7] There are four causal models that will be tested here. While a large number of variations of these models are possible, for purposes of this paper, these four seem to include the most critical paths that the predictors can possibly take to influence the dependent variable.

Blalock's procedure further involves generating correlation estimates of variables in different critical paths by comparing the actual (empirically derived) values of correlations with those that are generated by predictive equations as in Tables 1-VI.

Evaluating Critical Paths

Model 1 (Figure 3) assumes a fairly simple, but reasonable sequence of influence of the variables. It suggests that organizational factors influence the characteristics of the administrators which, in turn, affect the data

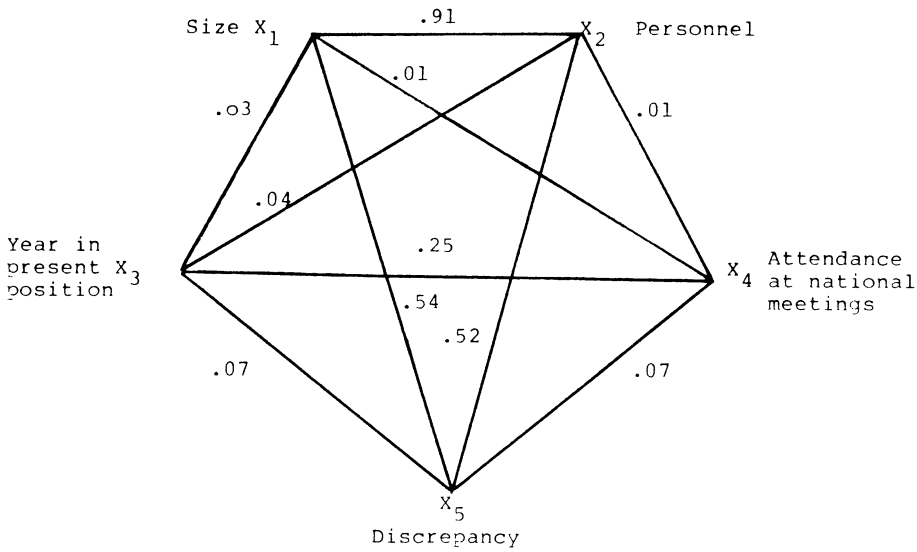


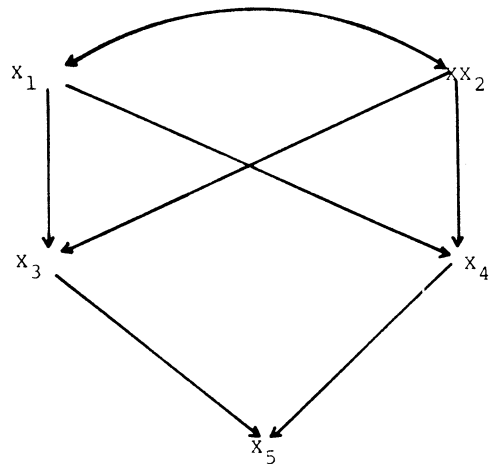
Figure 2: General Picture of Correlations

discrepancy. More specifically, the small size hospitals tend to support a relatively less conforming (professionally speaking) administrator who has been in the job for a minimum number of years. The small size hospital, its concomitant small number of personnel as well as the minimum professional conformity of the administrator and the fewer years of his incumbency in the present job tend to produce an effect on data discrepancy.

The model furnished three predictive equations, i.e. equations 1-3 in Table I.[8] The predicted values for the three sets of equations provide a very poor fit with the actual scores given in the Table. Considerable gap exists between the predicted and the actual values. While the actual relationship between x_1 and x_5 is considerably higher, i.e., .54, the relationship between these through x_2 and x_3 is fairly poor. Similar argument can be given for r25 and r35. This gap suggests, therefore, that the sequential causal influence of organizational variables via personal characteristics of the administrator to the dependent variable, as posited in Model 1, is not in that direction. It also suggests that the explanation of discrepancy of data

8. Actually, in terms of the number of unknowns that the model should predict, we should have had four equations; the fourth,i.e., r^{12} is, however, represented by a curved arrow to distinguished it from the rest of predictive equation. The curved arrow, as a convention, represents either the lack of knowledge of why a correlation occurred if it occurred or the case of an obvious correlation because of the very nature of correlation variables as in the present case. (The correlation r^{12} is so universalistic that it has been left out of the predictive set of equations.)

in terms of an interaction of these two sets of variables is empirically ill-founded.



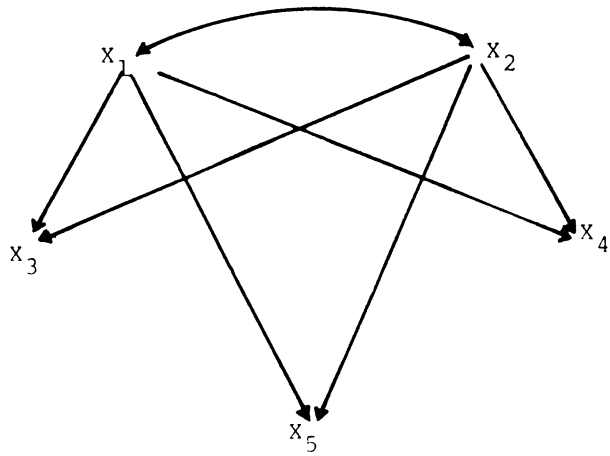
Model 1:

Figure 3: Causal Influence of Organizational Variables on Dependent variable through Personal Characteristics of the Administrator.

Table I. Actual vs predicted correlation values for model 1.

Equations		Predicted	Vs	Actual
1)	r25 = r23.r35 + r24.r45 = (.04)(.07) + (.01)(.07)	.004		.52
2)	r15 = r13.r35 + r14.r45 = (.03)(.07) + (.01)(.07)	.003		.54
3)	r34 = r23.r24 + r14.r13 = (.04)(.01) + (.01)(.03)	.001		.25

In order to provide a better explanation of the discrepancy scores, Model 2 (Figure 4) was stipulated. What this model suggests is again a fairly simple theoretical proposition — that the organizational variables causally influence both the dependent variable, i.e., discrepancy of data, and the personal characteristics of the administrator. In concrete terms, x_1 and x_2 are independent causes of x_3, x_4 and x_5. This model further posits that any connection between x_3 and x_4 and x_5 is only spurious, i.e., as a result of the organizational variables x_1 and x_2. More specifically, it is suggested that



Model 2

Figure 4: Causal Influence of Organization Variables Separately on Personal and Dependent Variables

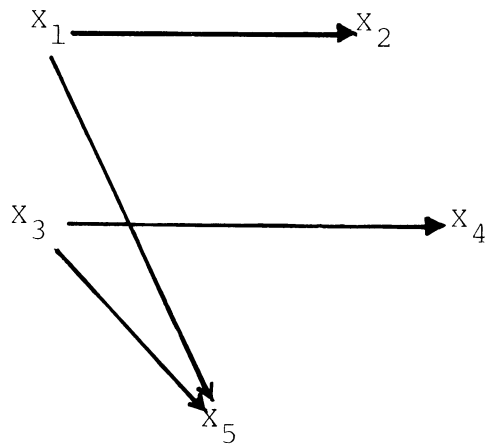
Table II. Actual vs predicted correlation values for model 2 .

Equations		Predicted	Vs	Actual
4)	r45 = r25.r25 + r14.r15 = (.01)(.52) + (.01)(.54)	.01		.07
5)	r35 = r15.r13 + r23.r25 = (.52)(.03) + (.04)(.52)	.04		.07
6)	r34 = r24.r23 + r14.r13 = (.01)(.04) + (.01)(.03)	.001		.25

while the small hospital size may influence the acquisition of minimally conforming administrators who are in their jobs for a certain period of time, these organizational variables, however, operate independently to generate discrepancy in the data. Furthermore, whether or not an administrator is minimally conforming to professional culture does, in no way, influence the discrepancy of data; it may well be an independent function of the structure of the hospital.

The model gives us three predictive equations for three paths. Looking at the predicted and the actual values, one can see fairly large gaps. The only equation that is a little better than the others is equation five. On the whole, the model furnishes a fairly poor picture of reality scores. Therefore, the second model is also somewhat inadequate.

A third model (Figure 5) proposes basically that the data discrepancy is a function of both the organizational variables and the personal characteristics of the administrator. In other words, both sets of factors influence the data discrepancy independently of each other, i.e., the hospital structure and the individual administrator independently influence the data discrepancy. In conceptual terms, two changes have been made in the model. First, an arrow has now been drawn from x_1 to x_2, on a reasonably valid assumption that the organizational size influences the number of



Model 3

Figure 5: Model showing Independent Causal Influence of Organizational and Personal Variables on the Dependent Variable.

Table III. Actual vs predicted correlation values for model 3.

Equations		Predicted	Vs	Actual
7)	r13	0		.03
8)	r14	0		.01
9)	r23	0		.04
10)	r24	0		.01
11)	r25 = r12.r15 = (.91)(.54)	.49		.52
12)	r45 = r34.r35 = (.25)(.07)	.02		.07

personnel, and not vice versa; second, an arrow leads from x_3 to x_4 on the assumption that the years an an administrator spends in the job influences to some degree the level of his attendance at the national meetings. Therefore, although this model assumes independence between sets of predictors, it assumes interdependence among the predictors belonging to the different set.

This model results in six predictive equations (equations 7-12). The values for predicted equations and the actual values are fairly close to the actual values. Of particular interest are equations 11 and 12, both of which determine the independent and separate influence of two sets of factors. Both of them are fairly close, although equation 11 shows better fitting results (i.e. .49 vs .52). The rest of the equations are also quite accurate in relation to the actual data. This model appears to be preferable in comparison to Model 1 and 2 and it furnishes a good fit between the predicted and the actual values.

Testing Predictive Equations

A question of considerable importance can be raised at this juncture; that is, if one changes the directions of the arrows while, at the same time, retains the theoretical context of Model 3 (Figure 5), would that introduce any change in the predictive equations or not; more specifically, suppose one reversed the directions of the arrows in Figure 5, e.g., instead of an arrow coming from X_1 to X_2, one changed it to X_2 to X_1, in a way that Figure 5 may assume either of the shapes in Figure 6 or Figure 7, one could then legitimately ask about the possible changes that the reversal may tend to produce in the original predictive equations for Model 3. In a sense, we are testing here whether or not the orginal predictions resulted due to mere chance or whether there is indeed some substance to the theoretical position assumed in that model.

Variation A (Figure 6) is a slight modification of the original model and is based on the premise that independent predictors within their own class can assume any direction such that X_2 can be assumed to cause independently X_1 and X_5; a rather weak assumption in as much as it assumes that the number of personnel can determine the size of a hospital, has been made. Nevertheless, in real world, there is reason to believe that the number of personnel may influence the size of the hospital, either by their professional policies established over the years, or the relationships

the personnel has been able to cultivate with the hospital management, etc. A general point is that whenever such a high correlation exists as between X_1 and X_2, it is indeed valid to assume mutual causation. Under these circumstances, reversing the direction of arrows for determining the real impact does make a considerable sense. By the same token, variation A also assumes a reversal of arrows from X_4 to X_3, specifying a direct cause of X_3 and X_5 to X_4. Table IV gives the predictive equations for this variation of the model. The predicted and the actual values show consider-

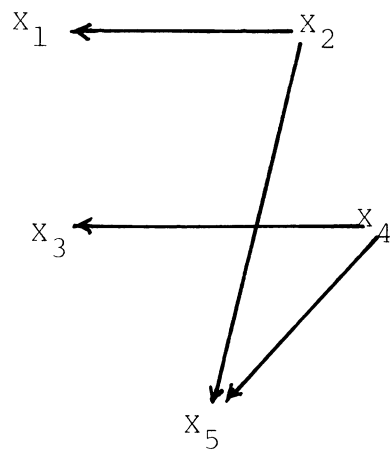


Figure 6: Variation A — For Model 3.

Table IV. Actual vs predicted correlation values for variation A of model 3.

	Equations	Predicted	Vs	Actual
13)	r13	0		.03
14)	r14	0		.01
15)	r23	0		.04
16)	r24	0		.01
17)	r15 = r21.r25 = (.91)(.52)	.47		.52
18)	r35 = r43.r45 = (.25)(.07)	.02		.07

able correspondence. More important, there is almost no difference in the predicted scores in this set up from the ones calculated for the original model. Except for equation 17 which shows a slight drop of only .02 from that of the original predicted correlation, all other equations are exactly of the same magnitude.

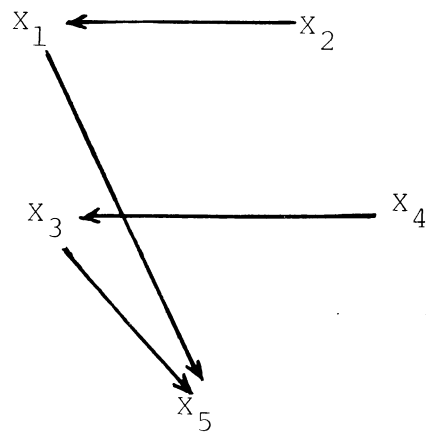


Figure 7: Variation B — For Model 3.

Table V. Actual vs predicted correlation values for variation B of model 3 .

	Equations	Predicted	Vs	Actual
19)	r13	0		.03
20)	r14	0		.01
21)	r23	0		.04
22)	r24	0		.01
23)	r25 = r12.r15 = (.91)(.54)	.49		.52
24)	r45 = r43.r35 = (.25)(.07)	.02		.07

Variation B for the Model 3 (Figure 7) is another possibility indicating a logical chain of influence running in one instance from X_2 through X_1 to X_5 and, in another instance, running from X_4 through X_3 to X_5. The

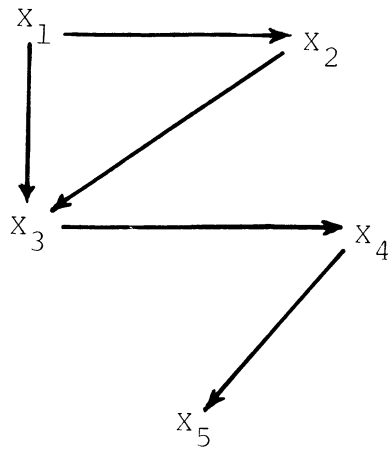
variation is built on the premise, that, within the separate sets of predictions, there could possibly be a pretty straight forward chain of influence relative to the dependent variable. The results, in Table V, indicate, that all the equations, under this variety, are exactly the same as in the original model.

It is, therefore, safer to suggest, that the original results obtained are reasonably dependable, to the extent, that any variations, in the assumed paths, do not alter the original predictions. A point of further interest, however, is whether, within separate classes of variables, one should assume the independent influence of one single factor to dependent variables, e.g., X_3 in Figure 5; a concomitant question of interest is, should one assume a sequence of influence running, for example, from X_4 through X_3 to X_5? It is not possible to answer these questions here. The more important result of the present evaluation is that, regardless of the flow of influence within separate classes of variables, the two classes of variables do, in fact, influence the dependent variable separately and independently.

Re-examining Variable Interaction

The final model (Figure 8) to be examined is basically a modification of Model 1. It only proposes a different set of paths to examine the independent influence of organizational variables on the dependent variable through the personal characteristics of the administrator. While conceptually similar, it differs from Model 1 in that a path has been established from X_1 to X_2, thus linking X_1 to X_5 both through X_2 and X_3. The arrow linking X_3 and X_5 has been eliminated, thus making it possible to predict the influence of critical variables through X_4 all the time. In conceptual terms, it makes sense to determine whether or not any difference in predictive power will occur, if certain paths are eliminated (e.g., paths 24 and 14, etc.) and certain new ones are introduced (e.g. path 12). This is to allow a maximum likelihood for the organizational variables to exert their influence through intervening variables, i.e., characteristics of the administrators.

The model provides five predictive equations. On examining Table VI, one may notice that while two of the equations (i.e., number 13 and 14) are fairly good, the rest are extremely poor. Equations 16 and 17 are especially very poor fit, with actual scores (i.e., .0001 vs. .54 and .007 vs. .52). What this lack of correspondence means is, that, even when providing



Model 4
Figure 8: Variation A for Model 1

Table VI. Actual vs predicted correlation values for model 4

	Equations	Predicted	Vs	Actual
25)	r14 = r13.r34 + r12.r23.r34 = (.03)(.25) + (.91)(.04)(.25)	.02		.01
26)	r24 = r23.r34 = (.04)(.25)	.01		.01
27)	r35 = r34.r45 = (.25)(.07)	.02		.07
28)	r15 = r13.r34.r45 + r12.r23.r34.r45 = (.03)(.25)(.07) + (.91)(.04)(.25)(.07)	.0001		.54
29)	r25 = r23.r34.r45 = (.04)(.25)(.07)	.007		.52

maximum alternative paths in a manner that allows the organizational variables to exercise maximum possible influence via the personal characteristics of the administrators, few gains in terms of accuracy of prediction will result. Essentially, therefore, the organizational variables do not seem to affect data discrepancy via the personal characteristics of the administrator.

Conclusions

On the basis of these four models and their causal assessment, one can say, fairly accurately, that data discrepancy is independently influenced by both organizational variables as well as the personal characteristics of the administrator. There does not seem to be anything in the prediction equations that will allow one to establish any causal influence running from organizational variables through personal characteristics to the dependent variables. In a sense, it fits reality better and provides somewhat better chances to planners to manipulate at least some of the variables outside of the context of the hospital, thus making it possible to control some portion of data discrepancy.

Cited References

1. **Blalock, Hubert** (1960) *Social Statistics*, New York, McGraw Hill, pp. 337-343.

2. **Johnston, John,** (1963) *Econometric Methods*, New York: McGraw Hill.

3. **Blalock, Hubert** (1969) *Theory Construction: From Verbal to Mathematical Formulation.* New Jersey: Prentice Hall.

4. **Khan, Jahangir and James Veney,** (1971) "Sources of Data Discrepancy in a selected sample of Hospitals", a paper read at the *99th Annual meeting of the American Public Health Association* Minneapolis, Minn. October 10-15.

5. **Blalock, Hubert** (1964) *Causal Inferences in Non-Experimental Research*, Chapel Hill: University of North Carolina Press.

6. **Tuckey, John** (1954) Causation, Regression and Path Analysis, in "*the Statistics and Mathematics in Biology*", **Kempthone, Oscar** *et al.* (eds). Iowa: American Iowa State College Press, pp.35-66.

7. **Wright, Swell** (1934) "The Method of Path Coefficients", *Annals of Mathematical Statistics*, 5; September, pp. 167-215.

تعيين مصادر خطأ البيانات الميدانية : تطبيق مدخل النموذج السببي

الدكتور جهانجير خان

أستاذ مشارك بقسم الإدارة العامة، كلية العلوم الإدارية، جامعة الرياض.

تشترك معظم الدراسات المسحية المعتمدة على بيانات ميدانية في احتوائها على مشكلتين أساسيتين هما التفاعل (Interaction) والازدواج الخطي (Multicollinearity) ، ولقد أكد ظهور هذه الحالة دراسة حديثة بنيت على بيانات مجمعة من ٤٨٠ مستشفى في الولايات المتحدة، حيث وجد تباين واضح في البيانات التي جمعت من مديري هذه المستشفيات وذلك باستخدام أسلوبين مستقلين لتجميع نفس البيانات من معامل هؤلاء المديرين. وباستخدام طريقة تحليل الانحدار المرحلي المتعددة، وكذلك معامل (بيرسون) للارتباط اتضح أن هناك أربعة عوامل ترتبط طرديا مع العوامل التابعة.

يستخدم هذا البحث طريقة النموذج السببي لبلالوك (Blalock) لتحديد أهم المتغيرات المسؤولة عن تباين البيانات وذلك بافتراض وجود التفاعل بين وداخل مجموعتي المتغيرات المستقلة والتابعة، ولهذا الغرض فقد تم فحص سبع نماذج باستخدام (قوة التطابق) لمعاملات الارتباط الحقيقية والتنبؤية. أوضح التحليل أن أفضل النماذج التنبؤية هو الذي يفترض استقلالية التأثير لكل من المتغيرات التنظيمية والخصائص الشخصية لمديري المستشفيات أي أن المعادلات رقم ١٣ الى ١٨ أعطت أفضل النتائج لشرح التباين في اجابات المستجوبين.